RiCL

Research in Corpus Linguistics

RiCL 12/2 (2024)

Special Issue

"Innovations in the compilation and analysis of spoken corpora"

edited by Robbie Love

aelinco
Asociación Española de Lingüística de Corpus

RiCL 12/2 (2024)

Editors

Paula Rodríguez-Puente and Carlos Prado-Alonso

ISSN 2243-4712

https://ricl.aelinco.es/

RiCL Research in Corpus Linguistics

Official journal of aelinco Asociación Española de Lingüística de Corpus

RiCL Research in Corpus Linguistics

12/2 (2024)

Articles

Book Reviews

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

RiCL Research in Corpus Linguistics

"We've lost you Ian": Multi-modal corpus innovations in capturing, processing and analysing professional online spoken interactions

Anne O'Keeffe[a] – Dawn Knight[b] – Geraldine Mark[b] – Christopher Fitzgerald[a] – Justin McNamara[a]
Svenja Adolphs[c] – Benjamin Cowan[d] – Tania Fahey Palma[e] – Fiona Farr[f] – Sandrine Peraldi[d]

Mary Immaculate College, University of Limerick[a] / Ireland
Cardiff University[b] / United Kingdom
University of Nottingham[c] / United Kingdom
University College Dublin[d] / Ireland
University of Aberdeen[e] / United Kingdom
University of Limerick[f] / Ireland

Abstract – Online communication via video platforms has become a standard component of workplace interaction for many businesses and employees. The rapid uptake in the use of virtual meeting platforms due to COVID-19 restrictions meant that many people had to quickly adjust to communication via this medium without much (if any) training as to how workplace communication is successfully facilitated on these platforms. The *Interactional Variation Online* project aims to analyse a corpus of virtual meetings to gain a multi-modal understanding of this context of language use. This paper describes one component of the project, namely guidelines that can be replicated when constructing a corpus of multi-modal data derived from recordings of online meetings. A further aim is to determine typical features of virtual meetings in comparison to face-to-face meetings so as to inform good practice in virtual workplace interactions. By looking at how non-verbal behaviour, such as head movements, gaze, posture, and spoken discourse interact in this medium, we both undertake a holistic analysis of interaction in virtual meetings and produce a template for the development of multi-modal corpora for future analysis.

Keywords – Online workplace communication; corpus pragmatics; multi-modal corpus linguistics; corpus construction; transcription

1. Introduction[1]

The pandemic has acted as a catalyst for change and has impacted on the behaviours of producers and consumers of digital interactional content. Businesses have changed their

[1] This research/project was funded by the *Arts and Humanities Research Council* (UKRI-AHRC) and the *Irish Research Council* (IRC) under the *UK-Ireland Collaboration in the Digital Humanities Research Grants Call* (grant numbers AH/W001608/1 and IRC/W001608/1).

interaction with customers, relying more on social media to reinforce their brand image (Ibrahim and Aljarah 2023); cultural organisations have embraced different forms of digital delivery of content, often co-produced by their audiences, such as book readings broadcast online as well as in-person; education has seen the large-scale adoption of online modes of instruction and interaction through both synchronous and asynchronous means of tuition. The list goes on.

There is a need now to examine whether existing paradigms for analysing verbal and non-verbal discourse in both face-to-face and virtual contexts are fit-for-purpose and develop technical protocols for capturing and analysing online multi-modal interaction. The *Interactional Variation Online* project (IVO)[2] draws on the expertise of researchers and their collaborators in the UK and Ireland to evolve standardised ways of approaching questions about multi-modal communication that are accessible and (re)producible by other researchers and non-technical experts. These will inform research practice relating to the gathering, storage, processing and analysis of multi-modal data through community-building aspects of the project for multi-modal corpus linguistic research. This paper outlines the phases of corpus design and construction undertaken by the IVO project, including:

1. Surveying and partner engagement
2. Establishing a design frame
3. Data collection
4. Transcription
5. Coding and mark-up
6. Establishing an analytical framework.

Innovation in corpus linguistics is inextricably linked to technological developments that drive changes in corpus construction and analysis. Concurrently, analysis of human interaction in and with digital technologies is an ongoing concern of researchers in the digital humanities (Mackenzie 2020). Until recently, capturing video recordings of business meetings with each participant individually framed, for the specific purpose of looking at verbal and non-verbal behaviour, would have required vast amounts of hardware and intrusion on the meeting space (see Knight and Adolphs 2020). Now that virtual meetings have proliferated and become normalised, particularly during and since COVID-19, this

[2] https://ivohub.com/

type of data is easily captured using the users' own hardware. This, in turn, has provided a capacity to analyse non-verbal communication in meetings without the need for a laboratory-type set-up with multiple cameras to capture audio and video of all participants. The past two decades have seen early developments of multi-modal spoken corpora and analysis software, and many areas in applied linguistics have begun to investigate modes other than speech, e.g., in social semiotics (Harrison 2003), multi-modal (inter) action analysis (Cohn 2016), conversation analysis (Mondada 2019) and gesture studies (Cienki 2016). These studies are making advances in establishing a greater understanding of the interconnected network of modes that construct meaning (Levinson and Holler 2014; Holler and Levinson 2019).

Within Multi-Modal Corpus Linguistics (MMCL), a multi-modal corpus aligns multiple discursive modes (e.g., textual transcription, video and/or audio data), and provides the tools to examine interaction within and between different modalities in the generation of meaning. Allwood (2008: 210) provides the following rationale for collecting and analysing multi-modal corpora: "they provide material for more complete studies of 'interactive face-to-face sharing and construction of meaning and understanding' which is what language and communication are all about." Most current multi-modal corpora are 'specialised', so built to examine a particular discursive context, such as meeting rooms (Friedland *et al.* 2009), academic supervisions (Knight and Adolphs 2008), political interviews (Trotta *et al*. 2020), and/or to meet the requirements of a particular research area/project. There currently exist no 'general' large-scale multi-modal corpora, with data from a range of discursive contexts and/or socio-demographic groups, and few of the corpora that do exist are freely available to the research community. The CLARIN website[3] provides links to some of those multi-modal corpora that are accessible. Over a decade ago, when reflecting on the future for multi-modal corpora, Knight (2011) outlined a range of methodological and technical issues and challenges faced by researchers in MMCL, the majority of which remain pertinent today. Annotating and analysing multi-modal corpora remains an expensive, time-consuming and technically complex process. However, the proliferation of workplace communication via virtual post COVID-19 pandemic meetings gives rise to an opportunity for multi-modal corpus construction that we aim to illustrate in this paper.

[3] https://www.clarin.eu/resource-families/multimodal-corpora

2. User-driven design

For the IVO project, the team adopted a user-driven approach to the research and corpus design (i.e., one in which practitioners and end-users co-construct the design from the start to ensure that it has "relevance and application to real-world problems and uses beyond the academic context" (Knight *et al.* 2021: 44)). To achieve this, we looked beyond the immediate research team to gain a baseline understanding of the general population's working behaviours, and their perceptions of working online during the COVID-19 pandemic. This was undertaken via an online survey[4] which was circulated to the project partners and their networks and publicised via social media platforms. The survey attracted 371 responses from individuals working in a range of vocations including academic, pharmaceuticals, finance, real estate, IT, media, the creative arts, medicine and for charitable organisations, of whom 54 per cent were from Ireland, 20 per cent from the UK, 18 per cent from Malta and 8 per cent from other locations. Likewise, 54 per cent defined themselves as female and 46 per cent male. Age ranges of respondents are shown in Table 1.

Age range	**Percentage**
18–24	11
25–34	26
35–44	20
45–54	20
55–64	21
65+	2

Table 1: Age ranges of survey respondents

Results showed a substantial increase in online meetings during the pandemic, with 41 per cent of respondents saying that they never had online meetings prior to the pandemic. Just 3 per cent of respondents said that they never had online meetings at the time of the survey (January 2022, emerging from pandemic restrictions). When asked whether they preferred face-to-face or online meetings for specific types of meetings (e.g., whole organisation meetings, social events), most respondents (76%) were in favour of face-to-face social events but were happy to have other meetings virtually. Connected to this, there was a strong sense of loss of social interaction in online environments, something that is seen as more pervasive in face-to-face interaction. This is also noted in Milz *et al.*'s (2023) study of online public planning meetings during COVID-19.

[4] https://ivohub.com/wp-content/uploads/2022/10/IVO_Baseline_Infographic.pdf

Free-text responses to the question 'what doesn't work well in virtual meeting environments?' included: 1) 'less informal interaction'; 2) 'you can't pick up the mood music of the room'; 3) 'difficult to build team spirit'; 4) 'no face-to-face, presenter cannot see facial expressions'.

Whilst the survey gained a relatively small number of respondents, with most from just three English-speaking countries (Ireland, United Kingdom and Malta), some useful insights into the broad preferences towards certain platforms for different types of work interactions were gained. It also helped to highlight the perception that specific platforms are chosen depending on the relative formality of the event, for example, respondents showed a preference for *Zoom*[5] in social meetings and for *Microsoft Teams*[6] in team meetings. These results also show both positive and negative sentiment towards virtual meetings. They underscore the desire for an increase in the social interaction which is lost in this environment and a maintenance of the convenience that is gained through online meetings. Our results tally with early studies on the efficacy of virtual meetings (e.g., Panteli and Dawson 2001) and overlap with some of the findings in Milz *et al.* (2023), such as a preference for holding large team meetings online rather than face-to-face.

3. Challenges and considerations

3.1. Developing multi-modal corpora

Research in this space faces on-going challenges relating to the forms of data to be included, namely:

1. The modalities to be captured and represented, i.e., what hardware to use to track gaze direction.
2. Where to source the data (and how).
3. The format for storage, i.e., which encrypted shared platforms to use.
4. The method of transcription and coding and deciding on whether speech-to-text tools are preferable to manual transcription.

[5] https://zoom.us/
[6] https://www.microsoft.com/en-us/microsoft-teams/log-in

5. The best way to align annotations and different modalities in a meaningful way to map temporal and/or semiotic relationships between these, e.g., deciding on a tool like *ELAN*[7] to create a system of co-occurring tiers.

Some reflections on these elements are discussed below, with further guidance provided on our website.[8]

3.2. Data collection

3.2.1. Establishing a sampling frame

The starting point for any corpus project is to define 'what' data is to be included/recorded and, important in the case of multi-modal corpora, 'how' it is to be recorded. This scoping process is scaffolded using a corpus design frame, also known as a 'sampling frame', and defined by Knight *et al.* (2021) as a rubric that specifies which texts, from which genres, and in what proportions are to be sampled for use in a corpus. Design frames ensure that the data collection is principled so that the resultant corpus provides, as far as possible, an accurate representation of the communicative contexts it seeks to capture (and represent).

To construct a design frame for a corpus of online meetings, we needed to define both 'meeting' and 'agenda' in terms of how we intend to use these terms for our purposes. Schwartzman (1989: 7) defines a meeting as "a communicative event involving three or more people who agree to assemble for a purpose ostensibly related to the functioning of an organization or a group." We extend this definition to include meetings of two people, to account for dyadic interaction, those which are agenda-driven, and which take place in virtual environments. The purpose of an agenda, according to Svennevig (2012: 54), is to provide "the participants with a 'template' for the topics to be addressed and the activities to engage in during the meeting." For the purposes of this project, the design frame we adopted involves data from online meetings which we define as communicative events, involving two or more people who agree to assemble online for a purpose, with a predetermined formal or informal agenda, related to the functioning of an organisation or a group.

Decisions regarding what should be included in a design frame, and its associated design taxonomy (i.e., its explicit categorisation framework), are typically driven by the

[7] https://archive.mpi.nl/tla/elan
[8] https://ivohub.com/resources/

specific aims of the corpus/associated research project. Efforts have been made to establish frameworks for representativeness and corpus design (e.g., Egbert *et al.* 2022) and there is a general understanding that any taxonomy used "must be consistent and transparent so that corpus users can navigate the corpus with ease" (Knight *et al.* 2021: 28). The development of a design frame is often iterative and dynamic, undergoing changes as the context is understood more while data is being collected. While this leads to a design frame that is more tailored to the dataset as it is collected, this requires detailed documentation and justification throughout the process. The initial design frame provided criteria for data sources that were seen as essential variables in the construction of a corpus of virtual meetings for the IVO project. These encompassed three broad factors: 1) the meeting type (e.g., team meetings, one-to-one meetings), 2) the sector (public/private) and 3) the meeting context or goal (e.g., transactional, pedagogic, team updates).

Subsequently, as we collected recordings to be included in the IVO corpus, it became apparent that a focus on the sector categories was needed, as it would provide a design frame that would encompass a spread of team meetings of various sizes and configurations with enough variables in terms of goals (desired outcomes of a meeting) and context (organisational setting) to create a corpus that would suit the aims of the research project. From pursuing early versions of sampling frames that were focused on meeting type and context, we had inherently acquired data that fell into multiple categories but were biased towards the public sector. The adoption of a private/public sector-focused framework (as these categories are defined by Esteve and Ysa 2011) led us to pursue access to data aligning with these categories (see Table 2). To this end, following a review of industry categories based on data collected, we designed a sampling framework that would take, as its principle, the categorisation set out in Table 2.

Sector	**Organisation types**
Private	Designated activity companies, sole proprietorships, partnerships, limited liability companies and considerations.
Public	Educational institutions, NGOs, government and health.

Table 2: Private/public sector-focused framework for data collection

To include a spread of organisational types, we aimed to collect data from each of the categories (in Table 2), although no set wordcount or prescriptions for balance were defined at the start, as these were likely to be somewhat driven by opportunism. We were essentially open to receiving whatever data was offered by those we contacted, and no prerequisites

were provided by the research team aside from the meetings taking place over video conferencing software. Thus, no prescriptions were established regarding the overall size of the corpus, the topics discussed within given meetings, nor the optimal lengths of recordings. Establishing targets is certainly advisable for corpus projects that intend to be large in scale, however. When building the *Spoken British National Corpus* 2014 (BNC2014; Love 2020), for example, it was vital that participants were asked "to make recordings of no less than 20 minutes in length" (Love 2020: 45) as it would have been near impossible to reach the *Spoken BNC2014*'s >11-million-word target in a timely way with shorter excerpts. By not imposing a time limit or restraint on the meetings recorded, the data is also more natural and authentic in style and that collection is therefore driven by participants, reflecting the actual process of online meetings.

3.2.2. Recording practicalities

When building spoken and multi-modal datasets in non-virtual environments, decisions regarding 'how' data is to be recorded also need close consideration. Access to equipment for recording and data storage can often be a challenge, and practical aspects such as 'where' to position the equipment, where participants will be in relation to this, and so on, need consideration. We posit that the digital pivot has certainly afforded a more streamlined approach here, whereby the decision making is somewhat not by the 'researcher' but the 'researched'.

In terms of camera settings, we were not prescriptive and essentially accepted all variants from those including recordings where some participants had their cameras off or had their audio muted. We also collected video-only and hybrid options as this reflects the reality of participant behaviour in virtual meetings. This variability, and the inclusion of multiple parties in the talk, increased the complexity and richness of the data, which needed to be factored into the time dedicated for annotation and analysis. This is because, as Goodwin (1994: 607) states

> like transcription, any camera position constitutes a theory about what is relevant within a scene - one that will have enormous consequences for what can be seen in it later - and what forms of subsequent analysis are possible.

The only requirement we did have was for data to be recorded by a representative from each of the meetings, using the built-in recording functionality of the given videoconferencing

software used (i.e., *Teams* or *Zoom*). Data could then be easily shared with members of the project team and was stored securely for subsequent analysis (guidelines on how to do this are available on the project website).[9] Such functionalities, and ever-increasing access to extensive cloud and desktop storage solutions, again, makes this stage of the corpus development process far quicker and easier than in face-to-face recording contexts which have resulted in other multimodal corpora such as the dyadic and triadic conversations that are the components of the *Freiburg Multimodal Interaction Corpus* (FreMIC; Rühlemann and Ptak 2023). However, a reliance on third-party software and the internet connectivity of participants can lead to technical issues, such as video and audio drop-out, resulting in participants relying on phrases such as that in the title of this paper to highlight such deficits to other participants.

3.2.3. Ethical considerations

Formal written consent was received from all participating organisations and, where possible, individuals as *a priori* for the development of the IVO corpus, and permission to re-use images/screenshots used in this paper were acquired from those participants who feature here. In cases where data were already in the public domain (for example, on company *YouTube* channels), explicit permission was granted from the organisations who made the data public and, where possible, the individuals participating in the recordings were contacted to request their consent. This is in line with best practice and in accordance with guidelines for best practice, such as that produced by the *British Association for Applied Linguistics* (BAAL 2021). Unfortunately, due to restrictions in copyright, publication and distribution, only excerpts of the dataset of the IVO corpus are publicly available for other users. The lack of availability and reusability of multi-modal corpora is an on-going issue within the field (see Knight and Adolphs (2020) and Knight (2011) for further discussions on this).

3.3. Orthographic transcription

Denham and Onwuegbuzie (2013) list four elements of spoken language as likely lost in transcriptions: 1) proxemics (the interpersonal space in the communication), 2) chronemics (the speed of the delivery and the length of silences), 3) kinesics (body language

[9] https://ivohub.com/resources/

and posture) and 4) paralinguistics (including volume, pitch and voice quality). Though the IVO corpus preserves much of the visual and auditory content of the original event, orthographic transcription is still required to enable searchability of spoken items after the corpus is constructed. While no agreed standard for transcription necessarily exists (i.e., the 'what' of transcription), shared practices are common across general spoken corpora as, for instance, the *Cambridge and Nottingham Corpus of Discourse in English* (CANCODE: Carter and McCarthy 2004) and/or national corpora with spoken components, like the *Spoken BNC2014*). In these cases, the "value" of spoken corpora is partly in revealing the "normal dysfluency" of speech (Biber *et al.* 1999: 1048), so there is an emphasis on transcribing *verbatim*, i.e., without standardising the content. This approach was also taken by the IVO team, using an adapted version of the CANCODE conventions (see website resource on transcribing multi-modal data).[10]

The actual process of transcription (i.e., the 'how' of transcription) has been noted as being a particularly time-consuming and arduous task (Knight and Adolphs 2022). As Lin and Chen (2020: 72) note, it can take "an hour to annotate the intonation and rhythm patterns in a single minute of speech," and "a further hour of video to conduct a detailed annotation for one minute of video." This is on top of the time taken to transcribe speech orthographically, whereby an hour of speech is estimated to take a trained researcher up to 14 hours (two working days) to transcribe (O'Keeffe *et al.* 2007). To speed up the process of transcription, the affordances of using speech-to-text and automatic captioning technologies have been explored by developers of spoken corpora. Love (2020: 104–107), for example, experimented with the use of a beta version of *Trint*,[11] an automatic speech-to-text transcription and editing tool, when constructing the *Spoken BNC2014*. He discovered that whilst the "time alignment and editing functionalities of the tool were very good, the accuracy of transcription appeared to be very low" (Love 2020: 107), with a "poor ability to separate turns according to the speakers who produced them" (*ibid*: 107). Love (2020) also tested other similar tools but concluded that, at the time of developing the *Spoken BNC2014*, they were all unfit for purpose as they did not produce fine-grained accurate outputs that are required for linguistic analysis.

In light of the 'digital pivot' and the increasing number and ubiquity of speech-to-text tools, and the fact they are now integrated directly into the main online meeting software

[10] https://ivohub.com/transcribing-mm-data/
[11] https://trint.com/

Microsoft Teams and *Zoom*, for example, it seemed appropriate to revisit the potential for using speech-to-text tools here. To this end, *Otter.ai*[12] was used in the first instance to generate a 'first-pass' of the collected data. This was then carefully checked and edited through close viewing and listening. *Otter.ai* is oriented towards creating transcriptions that prioritise legibility and coherence rather than preserving all elements of the original speech. In the process of automated transcription, items such as backchannels, repetitions and hesitations (for example, *uh, um, ah*) are omitted and so require manual addition by the analyst during the checking phase. In addition, for some recordings featuring strong regional or national accents, the accuracy of the transcription was low and, as a result, required a lot more manual input/editing. Despite these shortcomings, *Otter.ai* proved effective for turn separation and time alignment, offered ease of editing in its user interface, and generally increased the speed of transcription, so it was deemed more of benefit than *cost* to use.

The speed of transcription checking per minute is strongly determined by the transcript that is being reviewed and the elements of the transcript that are determined as necessary elements of the review process. For the IVO project, these elements of review were necessary but time-consuming components of the process, and they include 1) checking for accuracy and editing accordingly; 2) inserting fillers (such as *uh* and *um*) which *Otter.ai* is programmed to ignore; 3) inserting symbols and codes for items like interruptions, coughs and non-verbal sounds; and 4) anonymising any content that might reveal the identity of the participants or organisations involved.

3.4. Coding and mark-up

As non-verbal behaviours (such as gesture, gaze, posture, head nodding) are not readily analysable units, annotation is required as a precursor to the analysis phase. As noted by Allwood *et al.* (2007b: 274), "annotation schemes often reflect the specific requirements that drive the creation of such a [multi-modal] corpus" and these different needs and requirements often result in the use of bespoke coding schemes for marking-up non-verbal behaviours. Despite this lack of universality, there are broadly two types of schemes: 1) those which focus on form and 2) those which focus primarily on communicative function.

Form-based schemes typically concentrate on marking up non-verbal behaviour purely in kinesic terms, capturing, for example, the size, shape and relative position of sequenc-

[12] https://otter.ai/

es of movements that form non-verbal behaviours. Examples of these include McNeill's (2000) *Gesture Phase Coding Scheme*, which allows the modelling of a range of bodily movements, but predominantly concentrates on defining sequences of hand movement, and Ekman and Friesen's (1978) *Facial Action Coding Scheme* (FACS), which focuses on classifying facial expressions through the movement of specific facial muscles, known as *Action Units* (AUs). Other schemes include the *Bielefeld Speech and Gesture Alignment Corpus* (SaGA; Lücking *et al.* 2010), the *REmote COL-laborative and Affective Interactions Corpus* (RECOLA; Ringeval *et al.* 2013) and *Video-mediated English as a Lingua Franca Conversations Corpus* (ViMELF; Brunner and Diemer 2021), which are designed to code gestures and signs which co-occur with talk.

Function-based schemes, which are more relevant to this current research, enable the mark-up of the semiotic and/or pragmatic relationship between verbalisations and non-verbal behaviour, that is, the communicative function of multi-modal interaction. These are schemes which annotate, for example, where non-verbal behaviours co-occur (or not) with speech, and the basic discursive function of the non-verbal behaviour and speech within and across such patterning. Examples of these types of coding schemes include Holler and Beattie's (2002) *binary coding scheme for iconic gestures*, and Allwood *et al.*'s (2007a) *MUMIN coding scheme*. To account for the simultaneous annotation of the form, pragmatic meaning and prosodic profile of gestures, the *M3D System* (Rohrer *et al.* 2020), instead, proposes a multidimensional approach to labelling gestures that goes beyond traditional systems, such as McNeill's (2000), which are solely descriptive. In addition, the *Database of Speech and Gesture* corpus (DoSaGE; Pak-Hin Kong *et al.* 2015), annotated via three independent tiers: a tier for linguistic information of the transcript, a tier for forms of gestures, and a tier with functions for each gesture used.

The IVO project, as described in Knight *et al.* (in press), used an annotation scheme that considered both form and function in an approach to analysing head nods in virtual meetings. This entailed creating two tiers for each speaker and annotating form on one tier and function on the other. The form categories were frequency, speed and range, while the functions were the categories of backchannels described by O'Keeffe and Adolphs (2008).

3.5. Establishing an analytical framework

Close qualitative viewing of the data collected led us to construct a suitable descriptive framework for online meeting stages and practices as outlined in Figure 1, below, and Table 2 (cf. Section 3.2.1). This is loosely based on Handford's (2010) business meeting stages and discursive practices. Handford's model consists of six meeting stages, including three pre- and post-meeting stages, representing access to participants before, during and after the data collection, and which accommodate the intertextual nature of the meeting (i.e., references made to emails, agendas and other communications outside of the meeting). The data collected for the IVO project contains only data recorded and collected during a scheduled meeting time, with no access to participants, before and after the meeting recording span. The simplified structural model we designed for our corpus (informed by Handford's model) is composed of four main stages (1–4) identified in our data, illustrated in Figure 1 and exemplified in Table 2 (Section 3.2.1), respectively. Figure 1 represents the broader context within which the meetings occur. It shows additional exchanges and activities (e.g., email communications), both virtual and face-to-face, which typically take place between participants whilst preparing for meetings (e.g., agenda creation, slides for item presentation) and responding to activities after the event (completing action points, writing up meeting notes) in spatio-temporal contexts, other than the meeting itself. These are labelled 'meeting preparation' and 'post meeting follow-up' and are illustrated in grey before and after the four main stages of the meeting. While we do not have corpus data from these two peripheral stages, we recognise their relevance to the recorded meeting data; for example, we have evidence, either through direct reference to an agenda or from the way meetings are progressed, that the preparation of an agenda before the meeting is central to the management of each of the meetings analysed. In summary, in our model, we include the pre-meeting and post-meeting phases as a component to understand inter-textuality, e.g., participants may reference emails sent or discussions between individuals before meetings.

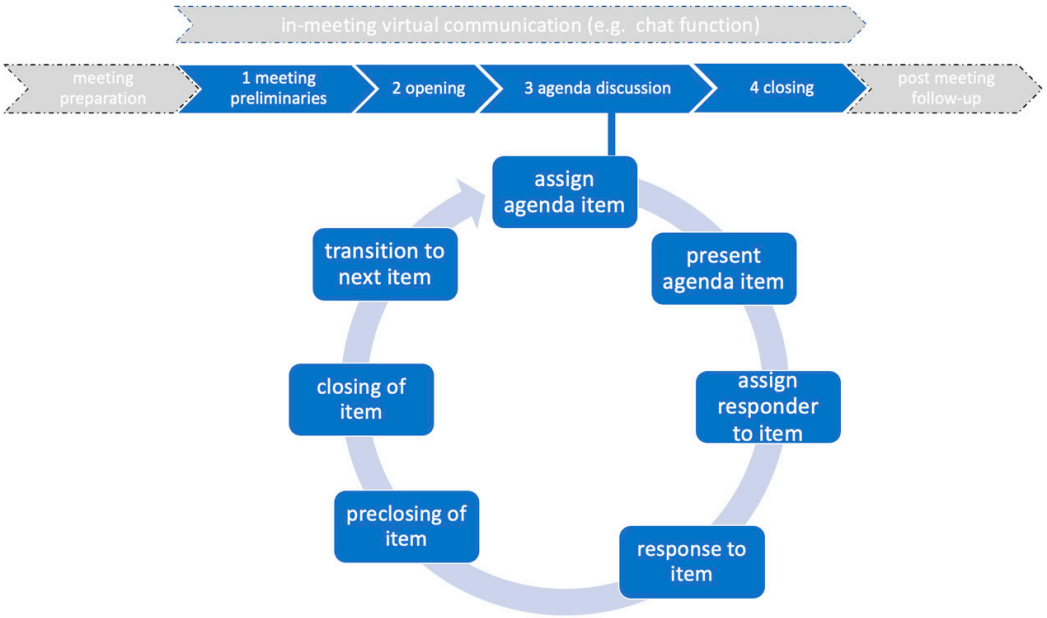


Figure 1: Linear and cyclical meeting stages

As Figure 1 illustrates, there are four core meeting stages which are fixed in order, and clearly identifiable in the data. Stage 1 represents a preamble to the meeting and may include work-related talk and small talk (Mirivel and Tracey 2005) as well as, crucial to this context, technology-related content specifically related to the setting up of the virtual meeting or the visibility of participants (particularly in meetings of large attendance, i.e., 20 participants or more). Stage 2 represents the start of the meeting proper. Stage 3 is typically a cyclical iterative stage, driven by multiple agenda items. Stage 4 can also contain more phatic communication once the 'business' of the meeting is over. In our corpus, stages 1 and 4 are regulated by the data contributors and the intactness of what takes place in these stages depends on when the recording is started and stopped.

As Handford (2010) notes, participants do more than simply progress through each of the structural stages of the meeting; they engage in a variety of goal-oriented dynamic discursive practices (e.g., setting the agenda, bringing the discussion back on track, checking shared understanding, moving from one agenda item to the next, bringing the meeting to a close). To scaffold and investigate these practices, Handford adopts McCarthy's (1998) four strands of linguistic behaviour found in spoken discourse (expectations, formulations, recollections and instantiations) and exemplifies them in terms of discursive practices, e.g., setting the agenda, bringing discussion 'on track' and bringing meeting to a close (Handford 2010: 77). Through our structural framework, we adopt a similar ap-

proach, isolating sub-sections within each of these stages, to examine specific discursive practices, at specific meeting stages, (e.g., managing turn-taking in the agenda discussion stage, closing of agenda items, timekeeping in the opening stage and agenda discussion stage). This gives us a means to systematically explore how these stages and practices are managed and executed both verbally and non-verbally, investigating whether certain verbal and non-verbal routines are realised at varying levels of granularity, at a stage, practice or linguistic feature level of analysis. It also allows for further cross-categorisation from different perspectives (e.g., identifying whether certain practices are characteristic of certain linguistic strands, examining the relational and transactional verbal and non-verbal behaviours and their co-occurrence at particular stages or with specific practices, or within the different discourse communities within our sampling frame). The stages and examples of practices and the verbal and non-verbal linguistic items used to enact them are set out in Table 3. We note that, while the meeting stages are relatively fixed, the practices listed and exemplified in Table 3 are not confined to these stages of the meeting. They provide us with targeted text external and internal to meetings (i.e., via structural/contextual factors or by linguistic components of the texts). Though not annotated as a component of the corpus or part of the corpus construction process, this framework was established from engagement with corpus after construction and provides us with a means of targeting language via meeting stages. These stages and discursive practices are essentially ways into our data for analytical purposes. In addition to the discursive practices enacted verbally and nonverbally, online meetings facilitate virtual means of enacting these via the use of the chat box, virtual reactions and emojis such as hand raises. Though our recordings do not include these, we have evidence that they are being used in verbal responses, e.g., “did you have your hand up?”

Stages and Discursive Practices	Exemplars
1. Premeeting (participants are present/joining)	
Setting up technology and hosting administration	*so yeah sorry we were just getting the live stream sorted there*
Introducing members	*we welcome X who is my director shadow*
Greetings	*hi everyone;* waves (physical/virtual); *hello, hi* comments in chat box
Engaging in small talk	*it's a beautiful sunny day here*
Transition move to opening	*okay, good; right, this meeting won't take too long*
2. Opening	
Reference to previous meeting	*as we said in the previous meeting*
Time keeping	*we're going to keep the presentations to 10 minutes*
Housekeeping	*please keep yourself on mute if you're not talking*
Previewing meeting	*this meeting is really going to be just about*
Acknowledging absentees and late arrivals	*X can't make it today*
3. Agenda discussion	
Assigning agenda item with nomination	*first up we have X over to you*
Contributing agenda item	*thanks everyone I'll just give you an update on ...*
Assigning responders to agenda item	X did you want to come in there; go ahead X
Request to contribute	hand up (physical/virtual); *can I just pop/jump in here*
Responding to agenda item (e.g. expressing gratitude, praise, encouragement; requesting more information/clarification; adding commentary; summarising; acknowledging contribution / endorsement of update or work done; displaying support)	*thank you and all your staff for the hard work you do; good; great; fantastic;* thumbs up; hand clap (physical/virtual); *we've covered a lot there*
Preclosing of agenda item	*so a massive thanks for the presentation um I think the questions have shown there's lots of interest in all the work you're doing...we'll move on to our next presentation; can I take that motion as adopted?*
Closing with upshot/gist	*we'll follow up on that again*
Transition move to next agenda item	*up next is item eight*
4. Closing	
Opening up closings	*any other business; does anyone have anything else to say?*
Concluding meeting	*alright. lovely to see you all and hopefully see many of you in person next week and we'll be in touch over over email over the coming days*
Goodbyes	*see you next time*; *bye*; waves, comments in chat box

Table 3: Meeting stages and discursive practices

3.6. (Re)presentation for analysis

While there is a dearth in the existence of freely available and widely used multi-modal corpora (Knight and Adolphs 2020), a range of digital tools and resources that exist to support the analysis of bespoke multi-modal datasets do exist. Examples of such software include *Transana*,[13] which enables users to integrate, transcribe, categorise and code their data then search and explore it in more detail (Halverson *et al.* 2012) and *ExMARaLDA* (Schmidt and Wörner 2014), which facilitates time-aligned transcription and annotation of multi-modal data. Perhaps the most popular, free and frequently used tool is Max Planck Institute's *ELAN* (Wittenburg *et al.* 2006). As seen in the screenshot in Figure 2, *ELAN* enables the annotation and analysis of data across multiple 'tiers' of information.

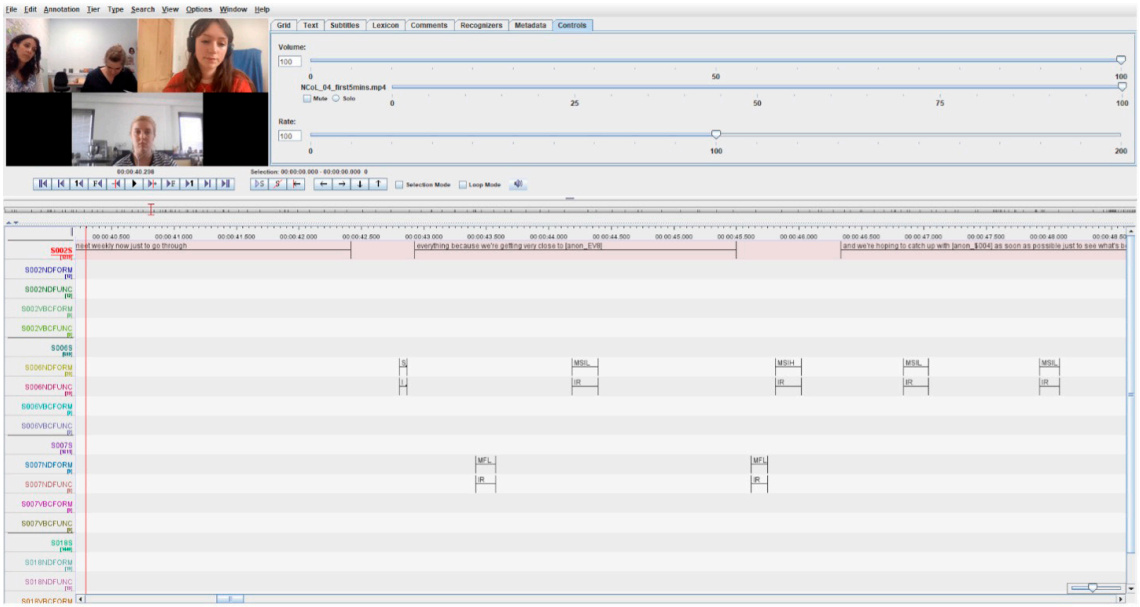


Figure 2: Screenshot of transcribed and coded data in *ELAN*

ELAN's tiers support frame-based analyses of multiple modes of time series data, from audio and video data to sensor outputs, which allows for synchronisation of tiers of data through the media timeline. Another feature of *ELAN* is its interoperability and flexibility in accommodating a range of file types, both for importing and exporting purposes. For example, different types of files can be imported into different tiers, e.g., .txt files or segmented .srt and Audacity files can be imported for transcription of speech, *Praat* files for analysis of sound, .mp4 files for video. Due to the ease of accessibility and use of *ELAN*, this software was the preferred option for the present study. Other studies which have used *ELAN* include the *Human-Computer Interaction Technologies* corpus (HuComTech; Pápay *et al.* 2011), DoSaGE (Pak-Hin Kong *et al.* 2015) and the *NEUROpsychological GESture*

[13] https://www.transana.com

corpus (NEUROGES; Lausberg 2019). Following the process of cleaning-up transcripts in otter, to enable the transcripts obtained from *Otter.ai* to be used in *ELAN*, two steps needed to be taken. These can be usefully repeated in other studies/projects of this nature:

1. Exporting transcription as SubRip (.srt) file. This file format is predominantly used for the creation of subtitles for integration into video files. To be used as subtitles, these files are timestamped to align with video to allow for future re-alignment in video annotation and analysis software. *Otter* has various options for line and character breaks, which result in different segmentation parameters in the subsequent transcription. We set the max number of lines to 1 and max characters per line to 2, which segments transcripts to approximate inter-pausal units when imported subsequently into tiers in *ELAN*.
2. Isolating individual speakers for use as single-speaker tiers in *ELAN*. To separate individual speakers to be treated as single-speaker tiers in *ELAN*, the .srt files are processed by a *python* code which isolates individual speakers. These files are then ready to be imported into *ELAN* as individual speaker tiers.

The above process results in an *ELAN* project that has the speech of individual speakers separated onto individual tiers. Additional tiers are then added for the annotation of non-verbal behaviour. Once these tiers have been defined, an *ELAN* template with tiers and a controlled vocabulary with set descriptions for annotations on tiers is created that can be used in other projects with the same analytical focus. For example, for an analysis of head nods as backchannels (see Knight *et al.* in press), we set a controlled vocabulary that was used on tiers for both form and function of backchannels. Having annotated data in *ELAN*, projects can be exported in a range of formats such as tab-delimited text, inter-linear text and subtitles text file. For our purposes, tab-delimited text files were used to observe incidence, frequency and co-occurrence of backchannels in excel spreadsheets.

4. Conclusion

Innovation is often based on how previous approaches are integrated in new ways. In this paper, we have outlined how, through the design, construction and analytical framing of the IVO corpus (see Knight *et al.* in press), we have engaged in practices that are innovative in how they integrate approaches in the following four areas:

1. Design frame: taking an approach to a design frame that is both user-based and focused on sectors.
2. Data collection: integrating data sourced from project partners that was recorded for this project with pre-existing recordings in the public domain.
3. Corpus construction: integrating speech to text transcription with a tiered system of multi-modal corpus analysis.
4. Analytical frame: integrating and adapting frameworks from previous workplace discourse to establish a framing that facilitates approaches to the data based on meeting stages, discursive practices and discourse features.

The challenges outlined in this paper regarding components of the corpus construction process are preceded by obstacles faced when attempting to acquire data. According to a survey carried by *KPMG International Limited* in 2022,[14] both businesses and consumers are growing evermore concerned about privacy and data security. In this environment, it is challenging for organisations to submit recordings of meetings for research purposes. To ensure the acquisition of data from sectors that fit into the IVO sampling frame, we have drawn upon a network of partners and investigators who have taken interest in and trusted the project from the outset. This trust-building has been essential to both the acquisition and sharing of data that constitutes the IVO corpus. Thus, having several project members with connections to various industries has been integral to the IVO corpus construction.

The construction of multi-modal corpora is still a relatively new endeavour. While recordings of virtual meetings promise a representation of an event that is close to what the participants in that event experienced, it remains challenging to ensure a process of corpus construction that is both efficient and reusable. The temptation to annotate everything in fine-grained detail is superseded by the understanding that this is extremely time-consuming, laborious and challenging in a context where you may be presented with thirty panels of speakers on a screen at one time. As with any project of this nature, we have found that clearly defined analytical goals (with a framework such as that outlined in this paper), and research questions aid the process of determining what to annotate and how. The decisions we have made in approaching data collection and analysis have ensured that we can both collect data that represents online meetings in a reasonably representative way and gain insights into this data in a manner that is achievable within the limited scope of a project such as this.

[14] https://advisory.kpmg.us/articles/2021/bridging-the-trust-chasm.html

References

Allwood, Jens. 2008. Multimodal corpora. In Anke Lüdeling and Merja Kytö eds. *Corpus Linguistics. An International Handbook*. Berlin: Mouton de Gruyter, 207–225.

Allwood, Jens, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta and Patrizia Paggio. 2007a. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation* 41/3: 273–287.

Allwood, Jens, Stefan Kopp, Karl Grammer, Elisabeth Ahlsén, Elisabeth Oberzaucher and Markus Koppensteiner. 2007b. The analysis of embodied communicative feedback in multimodal corpora: A prerequisite for behavior simulation. *Language Resources and Evaluation* 4/3: 255–272.

BAAL = The British Association for Applied Linguistics. 2021. *Recommendations on Good Practice in Applied Linguistics* (fourth edition). Available at www.baal.org.uk

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Brunner, Marie-Louise and Stefan Diemer. 2021. Multimodal meaning making: The annotation of nonverbal elements in multimodal corpus transcription. *Research in Corpus Linguistics* 9/1: 63–88.

Carter, Ronald and Michael McCarthy. 2004. Talking, creating: Interactional language, creativity and context. *Applied Linguistics* 25/1: 62–88.

Cienki, Alan. 2016. Cognitive Linguistics, gesture studies, and multimodal communication. *Cognitive Linguistics* 27/4: 603–618.

Cohn, Neil. 2016. A multimodal parallel architecture: A cognitive framework for multimodal interactions. *Cognition* 146: 304–323.

Denham, Magdalena A. and Anthony John Onwuegbuzie. 2013. Beyond words: Using nonverbal communication data in research to enhance thick description and interpretation. *International Journal of Qualitative Methods* 12/1: 670–696.

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

Ekman, Paul and Wallace V. Friesen. 1968. Nonverbal behavior in psychotherapy research. In John M. Shlien ed. *Research in Psychotherapy Volume III*. Massachusetts: American Psychological Association, 179–206.

Esteve, Marc and Tamyko Ysa. 2011. Differences between the public and the private sectors? Reviewing the myth. *ESADEgov e-bulletin* https://esadepublic.esade.edu/posts/post/differences-between-the-public-and-the-private-sectors-reviewing-the-myth

Friedland, Gerald, Hayley Hung and Chuohao Yeo. 2009. Multi-modal speaker diarization of real-world meetings using compressed-domain video features. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*. Tapei: IEE, 4069–4072.

Goodwin, Charles. 1994. Professional Vision. *American Anthropologist* 96/3: 606–633.

Halverson, Erica Rosenfeld, Michelle Bass and David Woods. 2012. The process of creation: A novel methodology for analysing multimodal data. *The Qualitative Report* 17/11: 1–27.

Handford, Michael. 2010. *The Language of Business Meetings.* Cambridge: Cambridge University Press.

Harrison, Claire. 2003. Visual social semiotics: Understanding how still images make meaning. *Technical Communication* 50/1: 46–60.

Holler, Judith and Geoffrey Beattie. 2002. A micro-analytic investigation of how iconic gestures and speech represent core semantic features in talk. *Semiotica* 142/1: 31–69.
Holler, Judith and Stephen C. Levinson. 2019. Multimodal language processing in human communication. *Trends in Cognitive Sciences* 23/8: 639–652.
Ibrahim, Blend and Ahmad Aljarah. 2023. The era of Instagram expansion: Matching social media marketing activities and brand loyalty through customer relationship quality. *Journal of Marketing Communications* 29/1: 1–25.
Knight, Dawn. 2011. The future of multimodal corpora. *Brazilian Journal of Applied Linguistics* 11/2: 391–416.
Knight, Dawn and Svenja Adolphs. 2008. Multi-modal corpus pragmatics: The case of active listenership. In Jesús Romero-Trillo ed. *Pragmatics and Corpus Linguistics: A Mutualistic Entente*. New York: Mouton De Gruyter,175–190.
Knight, Dawn and Svenja Adolphs. 2020. Multimodal corpora. In Stefan Gries and Magali Paquot eds. *A Practical Handbook of Corpus Linguistics*. Paris: Springer, 353–371.
Knight, Dawn and Svenja Adolphs. 2022. Building a spoken corpus. In Anne O'Keeffe and Michael McCarthy eds. *The Routledge Handbook of Corpus Linguistics.* London: Routledge, 21–34.
Knight, Dawn, Steve Morris, Laura Arman, Jennifer Needs and Mair Rees. 2021. *Building a National Corpus: A Welsh Language Case Study.* London: Palgrave.
Knight, Dawn, Anne O'Keeffe, Geraldine Mark, Chris Fitzgerald, Justin McNamara, Svenja Adolphs, Benjamin Cowan, Tania Fahey-Palma, Fiona Farr and Sandrine Peraldi. In press. *Interactional Variation Online* (IVO): Corpus approaches to analysing multi-modality in virtual meetings. *International Journal of Corpus Linguistics*.
Lausberg, Hedda. 2019. *The NEUROGES® Analysis System for Nonverbal Behavior and Gesture*. *The Complete Research Coding Manual including an Interactive Video Learning Tool and Coding Template*. Berlin: Peter Lang.
Levinson, Stephen C. and Judith Holler. 2014. The origin of human multi-modal communication. *Philosophical Transactions of the Royal Society B: Biological Sciences*. https://doi.org/10.1098/rstb.2013.0302
Lin, Phoebe and Yaoyao Chen. 2020. Multimodality I: Speech, prosody and gestures. In Svenja Adolphs and Dawn Knight eds. *The Routledge Handbook of English Language and Digital Humanities*. London: Routledge, 66–84.
Love, Robbie. 2020. *Overcoming Challenges in Corpus Construction.* London: Routledge.
Lücking, Andy, Kirsten Bergmann, Florian Hahn, Stefan Kopp and Hannes Rieser. 2010. The *Bielefeld Speech and Gesture Alignment* corpus (SaGA). In Nicoletta Calzolari, Khalid Choukri, Bente Maegaard, Joseph Mariani, Jan Odijk, Stelios Piperidis, Mike Rosner and Daniel Tapias eds. *Proceedings of the 7th International Conference on Language Resources and Evaluation.* Valletta: English Language Resource Association, 92–98.
Mackenzie, Jai. 2020. Digital interaction. In Svenja Adolphs and Dawn Knight eds. *The Routledge Handbook of English Language and Digital Humanities*. London: Routledge, 49–65.
McCarthy, Michael J. 1998. *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.
McNeill, David. 2000. Action and Thought. In David McNeill ed. *Language and Gesture*. Cambridge: Cambridge University Press, 139–140.

Milz, Dan, Atul Pokharel and Curt D. Gervich. 2023. Facilitating online participatory planning during the COVID-19 pandemic. *Journal of the American Planning Association*: 1–14.

Mirivel, Julien C. and Karen Tracy. 2005. Premeeting talk: An organizationally crucial form of talk. *Research on Language and Social Interaction* 38/1: 1–34.

Mondada, Lorenza. 2019. Contemporary issues in conversation analysis: Embodiment and materiality, multimodality and multisensoriality in social interaction. *Journal of Pragmatics* 145: 47–62.

O'Keeffe, Anne and Svenja Adolphs. 2008. Using a corpus to look at variational pragmatics: Response tokens in British and Irish discourse. In Anne Barron and Klaus P. Schneider eds. *Variational Pragmatics*. Amsterdam: John Benjamins, 69–98.

O'Keeffe, Anne, Michael J. McCarthy and Ron Carter. 2007. *From Corpus to Classroom – Language Use and Language Teaching*. Cambridge: Cambridge University Press.

Pak-Hin Kong, Anthony, Law Sam-Po, Connie Ching-Yin Kwan, Cristy Lai and Vivian Lam. 2015. A coding system with independent annotations of gesture forms and functions during verbal communication: Development of a *Database of Speech and GEsture* (DoSaGE). *Journal of Nonverbal Behavior* 39/1: 93–111.

Panteli, Niki and Patrick Dawson. 2001. Video conferencing meetings: Changing patterns of business communication. *New Technology, Work and Employment* 16/2: 88–99.

Pápay, Kinga, Szilvia Szeghalmy and István Szekrényes. 2011. HuComTech Multimodal Corpus Annotation. *Argumentum* 7: 330–347.

Ringeval, Fabien, Andreas Sonderegger, Juergen Sauer and Denis Lalanne. 2013. Introducing the RECOLA multimodal corpus of remote collaborative and affective interactions. In Rama Chellappa, Xilin Chen, Qiang Ji, Maja Pantic, Stan Sclaroff and Lijun Yin eds. *Proceedings of the 10th IEEE International Conference on Automatic Face and Gesture Recognition*. Shanghai: Curran Associates, 1–8.

Rohrer, Patrick Louis, Ingrid Vilà-Giménez, Júlia Florit-Pons, Núria Esteve-Gibert, Ada Ren, Stefanie Shattuck-Hufnagel and Pilar Prieto. 2020. *The Multimodal Multidimensional (M3D) Labelling Scheme for the Annotation of Audiovisual Corpora*. *Gesture and Speech in Interaction Conference.* Stockholm: University of Stockholm.

Rühlemann, Christoph and Alexander Ptak. 2023. Reaching beneath the tip of the iceberg: A guide to the Freiburg Multimodal Interaction Corpus. *Open Linguistics* 9/1: 20220245. https://doi.org/10.1515/opli-2022-0245

Schwartzman, Helen B. 1989. *The Meeting: Gatherings in Organizations and Communities*. New York: Plenum Press.

Schmidt, Thomas and Kai Wörner. 2014. EXMARaLDA. In Jacques Durand, Ulrike Gut and Gjert Kristoffersen eds. *Handbook on Corpus Phonology*. Oxford: Oxford University Press, 402–419.

Svennevig, Jan. 2012. The agenda as resource for topic introduction in workplace meetings. *Discourse Studies* 14/1: 53–66.

Trotta, Daniela, Alessio Palmero Aprosio, Sara Tonelli and Elia Annibale. 2020. Adding gesture, posture and facial displays to the polimodal corpus of political interviews. In Nicoletta Calzolari (Conference chair), Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk and Stelios Piperidiseds eds. *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille: European Language Resources Association, 4320–4326.

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
Anne O'Keeffe
Department of English Language and Literature
Mary Immaculate College
University of Limerick
South Circular Rd
V94 VN26
Limerick
Ireland
Email: anne.okeeffe@mic.ul.ie

received: June 2023
accepted: January 2024

RiCL Research in Corpus Linguistics

Building LANA-CASE, a spoken corpus of American English conversation: Challenges and innovations in corpus compilation

Elizabeth Hanks[a] – Tony McEnery[b] – Jesse Egbert[a] – Tove Larsson[a] – Douglas Biber[a] – Randi Reppen[a]
Paul Baker[b] – Vaclav Brezina[b] – Gavin Brookes[b] – Isobelle Clarke[b] – Raffaella Bottini[b]

Northern Arizona University[a] / United States
Lancaster University[b] / United Kingdom

Abstract – The *Lancaster-Northern Arizona Corpus of Spoken American English* (LANA-CASE) is a collaborative project between Lancaster University and Northern Arizona University to create a publicly available, large-scale corpus of American English conversation. In this article, we describe the design of LANA-CASE in terms of the challenges that have arisen and how these have been addressed – including decisions related to operationalizing the domain, sampling the data, recruiting participants, and selecting instruments for data collection. In addressing these challenges, we were able to draw on and further develop strategies established in the creation of other spoken corpora (including the British English counterpart to LANA-CASE, the *Spoken British National Corpus 2014*) as well as to implement recent theoretical and technical innovations related to each step. We hope that this discussion can inform future projects focused on the design and construction of spoken corpora.

Keywords – spoken corpora; conversation; corpus compilation; LANA-CASE

1. Introduction[1]

Corpora can provide meaningful insights into language, and they have a wide range of applications in research, teaching, and beyond (McEnery and Wilson 2001). While a great number of written corpora exist, the number of corpora that contain spoken language is more limited, which is likely due to the additional demands on time, resources, and ethical considerations that compiling a spoken corpus entails (McEnery and Brookes 2022). How-

[1] We gratefully acknowledge *Lancaster University's Global Advancement Fund*, *Northern Arizona University's Faculty Course-based Undergraduate Research Experience Development Grant*, the *Northern Arizona University Corpus Lab*, and *Northern Arizona University's SGS Award*, whose support has made this project possible. We are also very grateful to our collaborators who have helped with recruitment and data management, contributors who have supported data collection, and participants who have submitted conversations.

ever, recent innovations[2] can help researchers overcome such challenges to a degree. For example, the compilers of the *Spoken British National Corpus 2014* (Spoken BNC2014; Love *et al.* 2017), a large-scale corpus containing 11.5 million words of British English conversation, implemented some innovations in response to the challenges they confronted during the corpus compilation process. The work discussed in the present paper builds on such innovations as well as innovations from other recently compiled spoken corpora.

The present article reports on the early compilation phases of the *Lancaster-Northern Arizona Corpus of American Spoken English* (LANA-CASE), a large-scale corpus of American English conversation, which we are currently compiling to be made freely available for linguists and language teachers upon completion. At the time of writing, about 600 hours of conversation recordings have been submitted and over two million words have been transcribed, with the goal of collecting and transcribing eight to ten million words in total.

This project incorporates several innovations in operationalizing the domain (conducting a domain analysis following the recommendations of Egbert *et al.* 2022), sampling (adopting an iterative sampling process which captures gender, race/ethnicity, communicative purpose, and other demographic and situational variables), recruiting participants (through piloting various recruitment methods and selecting the most effective ones —such as social media— to invest in), and instruments used in data collection (utilizing data collection software such as Phonic that is adopted in a series of discrete steps). We describe how these innovations can help compilers of spoken corpora overcome practical challenges, using the first year of conceptualization and data collection for LANA-CASE as a case study. These challenges and innovations will be addressed in turn in Section 2 (domain analysis), Section 3 (planning the sample), Section 4 (recruitment), and Section 5 (instruments).

2. Domain analysis

Corpus design should ideally encompass what Egbert, Biber, and Gray *et al*. (2022) refer to as the "domain considerations" by describing the domain, operationalizing the domain, and planning the sample (see Egbert *et al*. 2022, Chapter 4). Within this framework, the first phase in building a corpus involves conducting a domain analysis. A key consideration within this phase entails first learning as much as possible about the target domain, or the

[2] For the purposes of this paper, we define 'innovation' as any methodological decision made that we have not seen implemented in the compilation of previous spoken corpora.

real-world language domain that the corpus aims to represent. This step is followed by operationalizing the domain, which is done by identifying the set of texts from which the corpus can realistically be collected. The third step involves choosing a sampling method and collecting the sample of texts. Establishing the target domain and operational domain allows the researcher to evaluate the degree to which the operational domain represents the real-world domain, and the degree to which the corpus sample represents the operational domain. We followed these guidelines to describe the domain of conversational American English (i.e., the population of texts that the corpus will ideally represent) and the operationalized domain (i.e., the subset of texts we could feasibly collect for inclusion in the corpus). We use this framework to guide the design and compilation of LANA-CASE. In this paper we focus primarily on the second step of a domain analysis: operationalizing the domain.[3]

2.1. Definition of the target domain

The target domain for LANA-CASE is spoken American English conversation. We define 'conversation' as an interactive spoken exchange of any length which is co-constructed by interlocutors (Hanks in preparation). Conversation can refer broadly to a wide range of communicative exchanges. Examples include an interaction that serves purely social functions, such as much of the conversation captured in the Cambridge and Nottingham Corpus of Discourse in English (CANCODE; McCarthy 1998) as well as an interaction that helps accomplish a task, such as much of the conversation captured in the Michigan Corpus of Academic Spoken English (MICASE; Simpson-Vlach and Leicher 2006). American English conversation specifically takes place between interlocutors who speak a variety of English that is typical within the United States (U.S.). Conversation may take place between interlocutors of diverse individual identities or characteristics — including such factors as age, race/ethnicity, and gender.

2.2. Description of the operational domain

The operational domain for LANA-CASE reflects the domain of spoken American English conversation in the following ways: it contains unplanned and unedited interactive

[3] A full description of our target domain is beyond the scope of the present article and will be documented in forthcoming publications.

spoken discourse that takes place in both face-to-face and remote modes between speakers of a variety of English that is typical within the U.S., regardless of individual identities or characteristics. However, it is restricted in that it includes only data from consenting participants who are 18 years or older, from participants who have lived in the U.S. prior to attending elementary school, and of conversations that take place between only two or three interlocutors. We discuss these decisions below.

We determined that segments of conversation must be recorded to be included in the corpus, and ethically, conversation should only be recorded with all interlocutors' knowledge and prior consent. It is possible that the observer effect may result in some differences between the conversations included in LANA-CASE and unrecorded conversations that will not be represented in the corpus (e.g., Saha *et al.* 2023). Building upon the findings of Love (2020), we strived to increase the reliability of speaker identification when transcribing (i.e., the ability for transcribers to attribute speech to the correct speaker) by operationalizing the domain as conversation between only two or three interlocutors (see Love 2020).

As a way of operationalizing what it entails to speak a variety of English that is typical within the U.S., we decided that all conversations must be between participants who have lived in the U.S. since before elementary school. The reason for this decision is that self-identification of language background is an inconsistent measure, especially when considering the complex nature of language input, output, community, and identity (Davies 1991). We therefore opted to operationalize the domain in practicable terms by collecting data from only one (quite large) population of American English speakers, using criteria that are objective and can result in more consistent data. Specifically, we chose to collect data only from speakers who have lived in the U.S. since before elementary school and speak English as a primary language. Additionally, while participants can be interlocutors of diverse ages, race/ethnicities, genders, and regions within the U.S., only interlocutors who are at least 18 years old are eligible to participate to simplify the process of ensuring informed consent.[4]

Because conversation may refer to a variety of communicative exchanges, we have provided participants little guidance in terms of what types of conversation they may submit. The instructions we provide are limited on our website to the following: "record

[4] We have taken great care to ensure all participants in the corpus have provided informed consent. The Terms and Conditions each participant signs are available in Appendix A.

your group talking about any topic(s) while completing your day-to-day tasks (e.g., during drinks with friends, a work meeting, getting ready for the day, etc.) and communicate as you normally would."

3. Planning the sample

Planning the sample required us to consider what and how much to sample. We discuss these points in this section by describing how sampling issues have been addressed in the compilation of LANA-CASE.

3.1. What to sample

We have planned the sample based on a) participant demographics such as age and b) situational characteristics such as conversation setting. Because sampling equally across all possible strata would not be logistically feasible, we streamlined the sampling process by defining selection and descriptive criteria for participant demographics, following the approach used in the *British National Corpus 1994* (BNC1994; Aston and Burnard 1998).

The planned sample covers data from specific demographic groups that represent four key individual variables ('selection criteria'); we also collect metadata that do not specifically guide our sampling but will be useful for corpus users ('descriptive criteria'). The LANA-CASE selection criteria include: 1) age, 2) race/ethnicity, 3) gender, 4) geographic region, and 5) residential setting (urban/suburban or rural). These selection criteria were set in part to ensure adequate representation from demographic variables that could influence language (e.g., Labov 1997). We have built upon the demographic data collected in the Spoken BNC2014 by collecting information about participants' race or ethnicity while also sampling based on gender, allowing participants to identify as any gender rather than restricting participants to a binary selection. The descriptive criteria include information about participant demographics: additional languages, educational background, and occupation. To avoid excess influence of linguistic features by a single contributor, the number of conversations that any individual participant can submit is limited to a maximum of four hours of conversation. The decision to limit each speaker's contribution was taken to maximize diversity in the sample.

Although we have prioritized planning the sample based on participant demographics, we also collect metadata about situational characteristics of conversations: interlocutors' relationship, setting (home, restaurant, etc.), and communicative purposes (using a list developed by Biber *et al.* 2021). We have adopted an iterative sampling process following Biber (1993) in which we consistently monitor the sample structure to detect imbalances in the submissions (e.g., to ensure balance across gender), which we have been able to address in recruitment efforts (see section 4).

3.2. How to sample

We aim to make LANA-CASE suitable for a wide range of research strands (e.g., quantitative analyses of lexicogrammatical features, qualitative analyses of pragmatics, analyses of sociolinguistic and register variation, discourse analysis, data-driven learning, lexicography, etc.). As such, we seek to build a corpus that is as large as possible, given inevitable constraints on time and funding. We expect that the completed corpus will be between eight and ten million words. The estimated size of each demographic sub-stratum has been established based on population data from the most recent U.S. Census (U.S. Census Bureau n.d.) with the goal of proportionally representing different ages, race/ethnicities, genders, demographic regions, and settings (urban/suburban or rural) within the U.S. These proportions provide rough guidelines as to a) the ideal proportion of our sample that should fall into each category (in the case of region and residential setting) and b) a minimum benchmark in terms of the representation from minority groups (in the case of participants' age, race/ethnicity, and gender). The estimated proportions we have used to guide our sample are shown in Table 1. While it is unlikely that the data in the final corpus will fall into these estimates perfectly, they are guidelines which we have strived for in terms of recruitment.

Selection criteria	Population	Estimated proportion of selection criterion
Age	18–25 years old	25%
	26–39 years old	22%
	40–65 years old	33%
	66 years old and over	20%
Race/Ethnicity (percent estimates account for intersectionality)	White	60%
	Hispanic or Latinx	18%
	Black or African American	13%
	Asian	5%
	American Indian or Alaska Native	2%
	Native Hawaiian or Pacific Islander	2%
Gender	Male	47%
	Female	47%
	Other (e.g., nonbinary)	6%
Geographical region	South	28%
	West	24%
	Midwest	21%
	Northeast	17%
Residential setting	Urban/Suburban	80%
	Rural	20%

Table 1: Proportion guidelines for data sampling

4. Recruitment

In order to approximate the sampling distribution described above, careful recruitment is necessary. Recruitment is a challenge in many studies (e.g., Farrokhi and Mahmoudi-Hamidabad; 2012 Dworkin *et al.* 2016), and it is further complicated in a project such as this where participant activities are relatively demanding (as this can mean fewer potential participants are willing to sign up), participants cannot participate in a single sitting (as this can lead to high attrition rates), and the researchers do not have easy access to the population of interest. We sought to preempt some of these issues by piloting different recruitment strategies (including some that had not been utilized in previous corpus compilation projects), aiming to build rapport with participants, and offering incentives. These are discussed in the following sub-sections.

4.1. Piloting recruitment strategies

We have explored which recruitment strategies are most effective at: a) recruiting many participants, b) recruiting participants from hard-to-reach populations, and c) recruiting participants with lower rates of attrition (i.e., participants who follow-through by submitting all materials over the course of several days or weeks). The list below contains the recruitment strategies we have piloted, with asterisks marking those that have been particularly effective and therefore warrant continued use.

1. Cold-calling senior centers and asking them to post flyers and/or host conversation events.
2. Cold-calling scout councils and inviting scouts to earn badges or awards by helping with recruitment.
3. Posting flyers at the cash register at gas stations in several U.S. states.
4. Passing out flyers in person.
5. Posting recruitment videos on social media, including *TikTok*,[5] *Instagram*,[6] *Facebook*,[7] *YouTube*,[8] and *Twitter*.[9]*
6. Offering extra credit (i.e., bonus points that supplement a students' overall grade in a course) to students for participating*.
7. Sending emails to trade schools and community colleges, requesting a recruitment message be sent to students.
8. Sending recruitment emails to alumni listservs at Lancaster University and Northern Arizona University.
9. Contacting local news stations and inviting them to cover the project.
10. Holding a booth at farmer's markets to advertise the project.
11. Advertising on *Hulu*, a U.S. television and movie streaming service.
12. Contracting a market research panel to gather participants from minority populations.
13. Hosting conversation activities at assisted living centers*.
14. Inviting personal contacts such as friends and family to participate (word of mouth)*.

---

[5] https://www.tiktok.com/@lana_linguistics?is_from_webapp=1&sender_device=pc
[6] https://www.instagram.com/lana_linguistics/
[7] https://www.facebook.com/profile.php?id=100088160239514
[8] https://www.youtube.com/channel/UCf8g41kI3d5QOov5RgxT9uQ
[9] https://twitter.com/LANA_corpus?ref_src=twsrc%5Egoogle%7Ctwcamp%5Eserp%7Ctwgr%5Eauthor

15. Inviting participants to recruit their personal contacts in order to receive additional monetary incentives.

In addition to these recruitment strategies, we are currently piloting several others, such as hosting recruitment events at restaurants, which we will report on upon completion of recruitment efforts. As can be seen above, the most effective strategies thus far have been posting recruitment videos on social media, offering extra credit to students, hosting conversation activities at assisted living centers, and recruiting personal contacts. Further information about how we have implemented each of these strategies, their effectiveness, and their strengths and limitations is provided in Table 2.

Strategy	Explanation	Effectiveness	Strengths	Limitations
Posting recruitment videos on social media, including *TikTok*, *Instagram*, *Facebook*, *YouTube*, and *Twitter*	A social media presence was initially established by posting five days a week (videos were created on *TikTok* and then shared to all other platforms). Many videos are specifically related to the project, inviting viewers to participate, while others are related to linguistics more generally to create more engagement with the channel and therefore boost rapport and overall viewership.	45 percent of recorders were recruited through this method. 17 percent of these recorders then followed through with submitting at least one conversation.	• Posting videos is free • Participants come from diverse backgrounds in terms of region, setting, and race/ethnicity. • Relatively large viewership (most videos receive several thousand views, up to over a million views). • Interest in LANA-CASE and in linguistics more generally is generated.	• Attrition rates are high. • Most participants are under 35 years old. • Participants tend to ask questions in the comments rather than reading about them on the website. • Posting regularly is time-consuming and can be intellectually and emotionally taxing.
Offering extra credit to students for participating	We have offered extra credit (i.e., bonus points that supplement a students' overall grade in a course) to our university students for submitting 45–60 minutes of conversation. All participants receive credit for participating, whether they and/or their interlocutor(s) are eligible to contribute data to LANA-CASE or not, and whether they choose for their conversation to be used in the corpus or not. Linguistics faculties at other universities across the U.S. have implemented this strategy in their classes as well.	11 percent of recorders were recruited through this method. 76 percent of these recorders then followed through with submitting at least one conversation.	• Attrition rates are low. • Data is free. • Participants follow instructions carefully. • Several participants from each class.	• Participant diversity is limited. • Some data must be discarded because participants are not eligible and/or choose not to contribute to the corpus. • Encouraging participants to provide any missing data after the end of a semester is challenging.

Table 2: Recruitment strategies

Strategy	Explanation	Effectiveness	Strengths	Limitations
Hosting conversation activities at assisted living centers.	We have hosted conversation activities at a local assisted living center during which time residents conversed with volunteer university students. The university students recorded the conversation and ensured all materials, including informed consent, were submitted.	0.5 percent of recorders were recruited through this method. 100 percent of these recorders then followed through with submitting at least one conversation.	• Participants come from older age ranges, a population that is particularly challenging to reach. • Attrition rates are low. • Community involvement in corpus creation is encouraged.	• Relatively few recorders can be recruited at a single event. • Several volunteers to record and submit residents' conversation may be necessary. • Some data must be discarded because residents at the assisted living center are not eligible and/or not legally able to provide informed consent.
Inviting personal contacts such as friends and family to participate (word of mouth)	We have invited our own friends and family to participate. Recorders who have been remunerated also share this opportunity with their own personal contacts.	12 percent of recorders were recruited through this method. 17 percent of these recorders then followed through with submitting at least one conversation.	• Relatively easy to send participants reminders and/or ask questions.	• Attrition rates are high. • Participant diversity is limited. • Participants may disregard formal submission procedures (e.g., sending recordings through personal email rather than completing the surveys provided to them).

Table 2: (Continuation)

As shown, one of the most common limitations we have confronted in recruitment is high attrition rates, in which participants do not follow through by submitting all required materials after signing up. We have attempted to address this issue by sending participants (bi)monthly email reminders about the deadline for their submissions. Another challenge we have faced is limited diversity, as most participants are white and under 35 years of age. Social media has helped to alleviate this problem to an extent by reaching a broader audience of various races/ethnicities, and hosting conversation activities at assisted living centers has helped reach participants from older age ranges.

An additional step we have taken to recruit more participants who are over 35 years old and come from diverse racial/ethnic backgrounds is to collaborate with a market research panel who specifically recruits participants from underrepresented demographic categories, a recruitment strategy employed in the creation of the original BNC1994. While this endeavor brought in 364 recorders from hard-to-reach populations (17% of our total pool of recorders), only two participants followed through with submitting at least one recording. It is possible this method did not prove effective because market research panels have access to millions of people who are primarily motivated by monetary incentives. The incentive we offered may not have aligned with these individuals' expectations in light of the activities we asked that they complete.

4.2. Building rapport

We have sought to build rapport with participants by sending them monthly emails with reminders about the status of their submissions, including how many minutes of conversation they have submitted and how many more are necessary for them to receive remuneration. We also keep an active presence on social media and send remuneration in a timely manner.

4.3. Offering incentives

In addition to building rapport, we have incentivized participants through monetary remuneration. Once their submitted conversations add up to two hours, participants receive an *Amazon* e-gift card. Recorders have the option to choose to receive the gift card or to donate it back to the project. While some have chosen to donate their gift card, many have elected to be paid. We also use monetary incentives to encourage participants to submit

several conversations of various lengths. Each conversation submission enters recorders into a monthly raffle to win an additional gift card.

In addition, we work to incentivize participants by sharing ideas about possible applications of the corpus. For example, LANA-CASE could be used to create more equitable learning environments by comparing the language in textbooks to the language of Latinx and other racial/ethnic minority groups in the U.S. Recruits have responded to such ideas on social media with enthusiasm (e.g., "I LOVE LINGUISTICS. THIS IS SO COOL" and "Absolutely fascinating. I now love linguistics").

Finally, since this is a long-term data collection process that is expected to last at least two years, we have encountered the need to incentivize participants to submit their conversation recordings promptly. Through monthly deadlines, we ask participants to submit their conversations by the 15th of the month for them to receive remuneration on the 16th. This has allowed for a steady flow of submissions that is necessary to adopt an iterative sampling process (see, e.g., Biber 1993).

5. Instruments

In order to collect recorded conversations from a large number of diverse participants across the U.S., we determined that creating instruments to allow for Public Participation in Scientific Research (PPSR; Shirk *et al.* 2012), following methods used in the Spoken BNC2014 (Love *et al.* 2017) and the *National Corpus of Contemporary Welsh* (CorCenCC; Knight *et al.* 2021), would be most effective. With the help of these instruments, participants followed instructions to record their own conversations and submit them along with all necessary metadata. When implementing PPSR in this way, clear yet simple instructions as well as easily navigable data collection instruments are necessary. This section describes steps we have taken to develop relatively user-friendly data collection processes and instruments.

We adopted a data collection process similar to both the BNC1994 (Leech 1993) and the Spoken BNC2014 (Love *et al.* 2017) wherein one participant signs up as a recorder to submit conversation recordings. Placing the responsibility of submitting recordings along with all required information on a single participant rather than a group enabled us to more easily contact the individual concerned and distribute remuneration to them. We believe that recruiting recorders —as opposed to groups— also allowed us to establish a workflow which encouraged submissions of more naturalistic conversations.

Recorders are directed to our website[10] to learn more about the project and get involved. Once they decide to contribute data, they are asked to fill out three short electronic surveys. First, they take a two- to three-minute survey in *Qualtrics*,[11] where they sign up, provide informed consent, and answer demographic questions about themselves. They are then encouraged to record their everyday conversations, specifically conversations that would have happened regardless of whether they were recording them (e.g., eating lunch with friends, cleaning the kitchen with a roommate, driving across town with a partner, etc.). The conversation recording(s) should be submitted as part of the second survey, hosted on the platform *Phonic*[12] (phonic.ai) at their convenience. The *Phonic* survey asks that all participants introduce themselves vocally in a brief 15–second recording, to facilitate speaker identification in transcription. It is not until after the conversation is submitted that the recorder is asked to complete the third (and final) step: a demographic survey for the other participant(s) in the conversation. The full process is depicted in Figure 1 below.

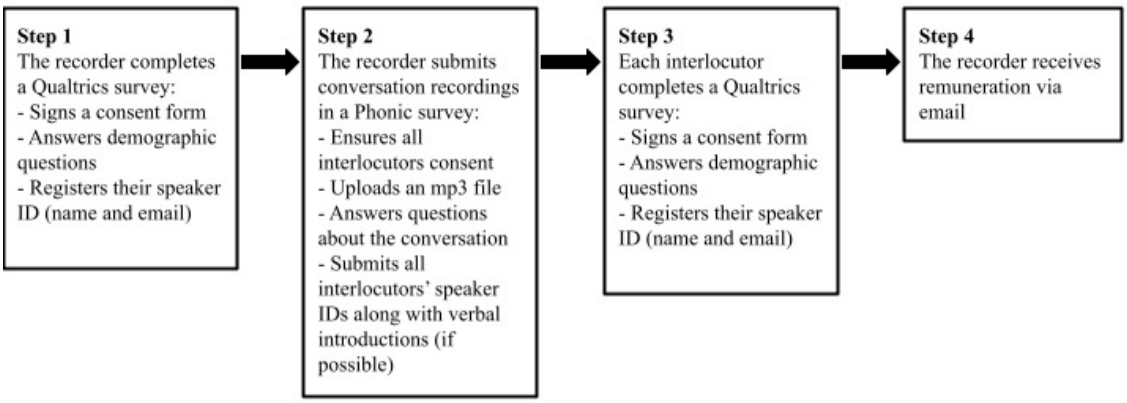


Figure 1: Steps for recorders to participate in data collection

Before arriving at this data collection process, we had considered several other possible workflow models, such as requiring all demographic information to be submitted along with the conversation recording itself in a single survey as well as requiring all demographic surveys to be submitted before the conversation is uploaded. Although the final process we arrived at is more time-consuming in the post-processing stage than other possible workflow models (because it requires matching the appropriate demographic surveys to each conversation), we believe it encourages participants to submit conversations that occur

[10] http://tinyurl.com/yc4su4z5
[11] https://www.qualtrics.com/
[12] https://www.phonic.com/

more naturally because recorders can begin recording conversations spontaneously with minimal intrusion while still submitting all required documentation.

We had also considered utilizing a crowdsourcing app for data collection. Yet, while applications have been shown to be effective tools for corpus creation (e.g., Knight *et al.* 2021), we determined a series of questionnaires to be better suited to our needs because downloading an application may a) require more commitment on the participants' part, thus reducing the number of people who register and b) make it more challenging for participants not familiar with using such apps to participate.

Because this data collection process is demanding of participants' time and energy, we sought to streamline the process as much as possible, which required balancing our desire for extensive metadata with participants' possible aversion to lengthy surveys. Thus, the demographic survey contains minimal questions so that it should take participants up to only three minutes to complete (the full list of questions and answer options can be found in Appendix B).

6. Conclusion

There are many challenges associated with compiling spoken corpora, including those discussed in this paper as well as those that fall beyond its scope, such as transcription, part-of-speech tagging, and preparing data for public release. The challenges we have faced in the LANA-CASE project to date include: 1) planning the sample (sampling largely based on participants' demographic variables), 2) recruiting participants (building rapport, providing incentives, and recruiting diverse and reliable participants), and 3) designing instruments (encouraging submissions of naturalistic conversations and using simple yet descriptive surveys). Each of these challenges has required creative problem solving. This has resulted in innovative approaches to corpus building, including carrying out a domain analysis (following Egbert *et al.*'s 2022 recommendation), sampling iteratively based on demographic and situational variables, recruiting participants by piloting several recruitment methods and investing in the most effective ones (e.g., social media such as *TikTok*), and adopting a new software called *Phonic* as part of a series of discrete data collection steps. Yet, as the process is still ongoing, we have yet to evaluate the success rate of our efforts. We also do not expect our decisions to be the only solutions to such issues; however, we do

hope that they may stimulate further discussion and spark new ideas for future compilers of spoken corpora to build on.

References

Aston, Guy and Lou Burnard. 1998. *The BNC Handbook: Exploring the British National Corpus with SARA*. Edinburgh: Edinburgh University Press.

Biber, Douglas. 1993. Representativeness in corpus design. *Literary and Linguistic Computing* 8/4: 243–257.

Biber, Douglas, Jesse Egbert, Daniel Keller and Stacey Wizner. 2021. Towards a taxonomy of conversational discourse types: An empirical corpus-based analysis. *Journal of Pragmatics* 171: 20–35.

Davies, Alan. 1991. *The Native Speaker in Applied Linguistics*. Edinburgh: Edinburgh University Press.

Dworkin, Jodi, Heather Hessel, Kate Gliske and Jessie H. Rudi. 2016. A comparison of three online recruitment strategies for engaging parents. *Family Relations* 65/4: 550–561.

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

Farrokhi, Farahman and Asgar Mahmoudi-Hamidabad. 2012. Rethinking convenience sampling: Defining quality criteria. *Theory & Practice in Language Studies* 2/4: 784–792.

Hanks, Elizabeth. (In preparation). Exploring the register of conversation: Uncovering linguists' insights about its situational characteristics.

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 and Evaluation* 55: 789–816.

Labov, William. 1997. Linguistics and sociolinguistics. In Nikolas Coupland and Adam Jaworski eds. *Sociolinguistics: A Reader*. London: Palgrave Macmillan, 23–24.

Leech, Geoffrey. 1993. 100 million words of English. *English Today* 9/1: 9–15.

Love, Robbie. 2020. *Overcoming Challenges in Corpus Construction: The Spoken British National Corpus 2014*. New York: Routledge.

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.

McCarthy, Michael J. 1998. *Spoken Language and Applied Linguistics*. Cambridge: Cambridge University Press.

McEnery, Tony and Andrew Wilson. 2001. *Corpus Linguistics: An Introduction.* Edinburgh: Edinburgh University Press.

McEnery, Tony and Gavin Brookes. 2022. Building a written corpus: What are the basics? In Anne O'Keeffe and Michael McCarthy eds. *The Routledge Handbook of Corpus Linguistics.* London: Routledge, 35–47.

Saha, Koustuv, Pranshu Gupta, Gloria Mark, Emre Kıcıman and Munmun De Choudhury. 2023. Observer effect in social media use. https://doi.org/10.21203/rs.3.rs-2492994/v1

Shirk, Jennifer, Heidi Ballard, Candie Wilderman, Tina Phillips, Andrea Wiggins, Rebecca Jordan, Ellan McCallie, Matthew Minarchek, Bruce Lewenstein, Marianne Krasny

and Rick Bonney. 2012. Public participation in scientific research: A framework for deliberate design. *Ecology and Society* 17/2: 1–20.
Simpson-Vlach, Rita C. and Sheryl Leicher. 2006. *The MICASE Handbook: A Resource for Users of the Michigan Corpus of Academic Spoken English*. Ann Arbor: University of Michigan Press.
U.S. Census Bureau. n.d. *Explore census data*. https://data.census.gov/(June 2022).

Corresponding author
Elizabeth Hanks
Northern Arizona University
College of Arts and Letters
English Department
S San Francisco St.
Flagstaff
AZ 86001
United States
Email: eah472@nau.edu

received: May 2023
accepted: February 2024

Appendix A: Terms and conditions

Project information

You are being invited to participate in a project titled *The Lancaster-Northern Arizona Corpus of American Spoken English* (LANA-CASE). Data collection for this project is being conducted by Jesse Egbert, Tove Larsson, Elizabeth Hanks, Doug Biber, and Randi Reppen from Northern Arizona University in Flagstaff, Arizona and Tony McEnery, Vaclav Brezina, Paul Baker, Gavin Brookes, Isobelle Clarke, and Raffaella Bottini from Lancaster University in the United Kingdom.

The purpose of this project is to create a resource for linguistic research. We are collecting samples of spoken American English that will be used to inform research into the English language as well as the development of teaching materials for language learners. The recordings will be transcribed and then made into a publicly available resource in both audio and transcribed (written) form.

Participant eligibility criteria

You are eligible to participate in this project if you speak English as one of your primary languages, have lived in the United States since before elementary school, and are at least 18 years old.

Participant activities

If you agree to take part in this project, you will be asked to complete several steps:

1. The first step is for recorders only and comprises a questionnaire that asks you to agree to the terms and conditions, answer demographic questions about yourself, and register your speaker ID. This questionnaire will take about three minutes to complete and only has to be completed once per recorder. The data will be anonymized.
2. The second step requires participants to record a conversation and answer a brief questionnaire about the conversation. You will meet with a group of two–three people who are all eligible to participate and agree to be recorded. Please record the audio of your conversation(s) using a personal device. Conversations of any

length will be accepted. The subsequent questionnaire will ask you to upload your conversation and answer a few questions. This will take about two minutes to complete. The recorder may submit up to four hours of conversation recordings (usually broken down into multiple submissions of shorter conversations). We recommend connecting to Wi-Fi so that the audio recording uploads quickly. The conversation you record in this step can be about any topic(s) you would like to discuss. To protect the speakers, however, we recommend avoiding discussions of illegal activity.

3. In the third step, recorders ask their conversation partners to complete a brief questionnaire. The questionnaire asks speakers to agree to the terms and conditions, answer demographic questions, and register a speaker ID. This questionnaire will take about three minutes to complete and only has to be completed once per speaker. It may be completed by the speakers themselves or by the recorder on behalf of the speaker (with their express permission). The data will be anonymized.

By participating, you agree that the research team has permission to store indefinitely, transcribe, and otherwise use recordings of your speech, and you agree that such data may be stored and used in perpetuity. You also agree that other researchers throughout the world have permission to use recordings and transcriptions of your speech for research and/or language teaching indefinitely.

Participant compensation

The recorder is eligible to receive a $25 *Amazon* e-gift card for every two hours of conversation that they submit as part of Step 2. The Amazon e-gift card will be sent to the email you provide. Payment will be sent on the 16th of each month until all necessary data has been collected. Only one recorder may receive remuneration for each two-hour block.

Each recording submitted will enter recorders into a drawing to win an additional $50 *Amazon* e-gift card. Results from the drawing will be publicized on the 16th of each month until all necessary data has been collected.

The following requirements must be met in order to be remunerated:

1. All questions in the questionnaires are answered in full.

2. The recording includes a conversation between 2–3 speakers who have registered speaker IDs.
3. The audio in the recording is clear and of good enough quality that 90% of the conversation can be understood (e.g., record in a relatively quiet location without much background noise and keep the recording device in a central location so it captures audio from all speakers)

Protection of risks

As with any online-related activity, the risk of a breach of confidentiality is possible. We will minimize this risk by saving all data on an encrypted, password-protected server. Additionally, the research team will protect your privacy by removing personal information (such as references to people, places, and institutions) from the transcription. Your recording and transcription will be available only to researchers who have completed a data use agreement and are accessing the data strictly for research and/or language teaching purposes.

Your participation in this project is completely voluntary and you can withdraw at any time. If you choose not to participate, it will not result in any penalty or loss of benefits to which you are otherwise entitled.

Contact information

If you have questions about this project, you may contact the research team at ShareYourVoiceEnglish@gmail.com

* If recorders participate as a school assignment, they are eligible to receive class credit as determined by their instructor rather than monetary compensation.

Appendix B: Demography Survey

Question	Possible answers
Do you agree to the Terms and Conditions?	Yes No
Have you lived in the United States since before elementary school?	Yes No
Are you 18 years old or over?	Yes No
What is your birth year?	[open-ended]
What is your gender?	Male Female Other (please specify): [open-ended]
What is your race/ethnicity? (check all that apply)	White Hispanic or Latino Black or African American Asian American Indian or Alaska Native Native Hawaiian or Pacific Islander Other (please specify): [open-ended]
What language(s) do you speak at home?	English Spanish Other (please list the language(s))
What language(s) do you speak outside of the home?	English Spanish Other (please list the language(s))
What is the highest level of education you have completed?	Less than high school High school graduate Trade school certificate (e.g., electrician, commercial driver, cosmetology, etc.) Undergraduate degree Graduate degree
What is/are your occupation(s) (e.g., nurse, student, construction worker, etc.)?	[open-ended]
What best describes your living situation? Feel free to add more details, if necessary.	I live in an urban or suburban area I live in a rural area
In what state do you currently live?	[drop-down of all 50 states and Washington D.C.]
Have you lived in one state for more than half your life? If yes, which state?	Yes No [drop-down of all 50 states and Washington D.C.]
Where did you find out about this project? (For example, *Facebook*, a flyer at a coffee shop, a friend, etc.)	[open-ended]

RiCL Research in Corpus Linguistics

Compiling a corpus of African American Language from oral histories

Sarah Moeller[a] – Alexis Davis[a] – Wilermine Previlon[a] – Michael Bottini[a] – Kevin Tang[b/a]

University of Florida[a] / United States
Heinrich-Heine University Düsseldorf[b] / Germany

Abstract – African American Language (AAL) is a marginalized variety of American English that has been understudied due to a lack of accessible data. This lack of data has made it difficult to research language in African American communities and has been shown to cause emerging technologies such as Automatic Speech Recognition (ASR) to perform worse for African American speakers. To address this gap, the *Joel Buchanan Archive of African American Oral History* (JBA) at the University of Florida is being compiled into a time-aligned and linguistically annotated corpus. Through Natural Language Processing (NLP) techniques, this project will automatically time-align spoken data with transcripts and automatically tag AAL features. Transcription and time-alignment challenges have arisen as we ensure accuracy in depicting AAL morphosyntactic and phonetic structure. Two linguistic studies illustrate how the *African American Corpus from Oral Histories* betters our understanding of this lesser-studied variety.

Keywords – African American English; oral history; Automatic Speech Recognition; natural language processing; corpus linguistics; morphosyntax

1. Introduction[1]

African American Language[2] (henceforth AAL) is one of the most studied marginalized varieties of American English, yet AAL linguistic data that is annotated, compiled and searchable in the form of a corpus is mostly inaccessible for linguists. Only in 2018 was the first corpus of African American speech, namely the *Corpus of Regional African American Language* (CORAAL; Kendall and Farrington 2021), made available. For the first time, the general population and linguists alike could access interviews of varying African American regional accents. However, with CORAAL being the only such

[1] This research is part of the project *Reanimating African American Histories of the Gulf South* which was supported by the National Endowment for the Humanities (PW–277433–21). We thank the editors (Robbie Love and Carlos Prado-Alonso) and the anonymous reviewers for their valuable feedback.
[2] The term 'African American Language' is also often referred to as 'African American English' (AAE) and 'African American Vernacular English' (AAVE).

resource of its kind, lack of data creates an obstacle for much needed research on language and race in diverse African American communities. Further, the effect of this gap extends to emerging technologies such as automatic speech recognition (henceforth ASR) that perform more poorly for African American speakers (Blackley *et al.* 2019; Koenecke *et al.* 2020; Martin and Tang 2020). As technology expands into everyday life (Lee *et al.* 2022; Yoon *et al.* 2023; Davis *et al*. 2024), healthcare, and hiring (Martin and Wright 2022, and references therein), racial disparity in technology could have dire consequences for African Americans.

A time-aligned corpus of linguistically annotated AAL audio and transcripts will provide a much-needed increase of speech data for African American corpus linguistics. The *Joel Buchanan Archive of African American Oral History* (JBA) at the University of Florida contains a growing collection of over 600 oral history interviews.[3] This is larger than many publicly available spoken corpora of any language or dialect and could have a significant impact on AAL studies, as well as provide a tool to hold major technology developers like *Amazon* and *Apple* accountable to address their racially biased systems, (see Blodgett *et al.* 2020 for a survey of biases in natural language processing, henceforth NLP).

This paper describes the compilation and initial analysis of AAL linguistic features of a time-aligned and linguistically annotated AAL corpus from JBA. Our work aims to produce two major deliverables: 1) a large African American speech corpus enriched for the first time with linguistic annotations and 2) a time-aligned representation of the audio files and transcribed texts. We want to identify and analyze the distinctive features of AAL in hopes of advancing language science, improving NLP, increasing awareness of the richness of language in the USA, and supporting educational endeavors related to African American culture and history. We are also building NLP systems specific to AAL and adapting a forced alignment model from General American English (henceforth GAE) to the recordings and transcripts of AAL.

We provide a background by describing the nature of oral history collections, existing African American linguistic corpora, including the *Joel Buchanan Archive* that we work with, and the AAL (cf. Section 2). We provide a general outline of the compilation process from an oral history collection to corpus for linguistic study (Section

[3] https://ufdc.ufl.edu/collections/ohfb

3). We then describe the specific challenges that have arisen while compiling our corpus, detailing issues related to transcription and time-alignment (Section 4). To illustrate how such a corpus can better our understanding of a lesser-studied language variety, we include two linguistic case studies of the distribution of AAL features (Section 5.2) and the syntactic structures that signal the presence of the AAL feature habitual *be* (Section 5.3) based on an initial sample of the oral history collection described in Section 5.1. Finally, we summarize a computational method that we have developed for the tagging of linguistic features that are distinctive to AAL (Section 5.4).

2. Background

With oral history collections containing a wealth of untapped linguistic data, it is important to understand their construction and the stories they house. Additionally, our methods are also inspired by previous corpus work (cf. Kendall and Farrington 2021 or Fitzgerald 2022, among others). In this section, we describe endeavors to catalog naturalistic AAL data via corpora, as well as oral history collections. Finally, we introduce the oral history collection that we are harnessing for corpus compilation, and briefly describe the specific challenges we face with AAL.

2.1. What are oral history collections?

Oral history interviews are a method of documenting history through audio/video recorded stories of individuals. These projects are centered around the experiences of narrowly defined groups, such as first-generation college students (e.g., the *Machen Florida Opportunity Scholars Oral History Program*)[4] or people who knew or worked with important persons, (e.g., the *John F. Kennedy and Robert F. Kennedy Oral History Collection*)[5] or specific racial groups, such as African Americans. A common method of collecting is using community networks to find volunteers who are willing to share their personal stories. They may also focus on certain regions, or topics such as historical events. However, most oral history collections remain largely inaccessible to corpus linguistics, their power to enlighten through linguistic analysis lying untapped. There are however a few exceptions. Schiffrin (2002) uses an oral history transcription of one

[4] https://oral.history.ufl.edu/projects/machen-florida-opportunity-scholars-program-mfos/
[5] https://www.jfklibrary.org/archives/about-archival-collections/oral-histories

Holocaust survivor to investigate her relationship with her mother and friends. The analysis is based on linguistic construction, such as the variation in the use of referring terms and reported speech. Similarly, Fitzgerald (2022) reports on the compilation and usage of the *Corpus of Irish Historical Narratives*[6] using an archive of Irish oral history documents, the *Irish Bureau of Military History*,[7] which consists of 238 oral testimonies. By applying corpus linguistic methods, Fitzgerald investigates the commitment to truth and what is meant by truth by examining the use of a set of mental process verbs, such as *think*, *remember*, *suppose* and *believe*, and expectation markers such as *actually*, *in fact* and *of course*. Finally, the most note-worthy use of oral history collection is the *Freiburg English Dialect Corpus* (Kortmann and Wagner 2005), which contains transcriptions and audio recordings, totaling 372 interviews which comprise 2.5 million words of text and 300 hours of speech.

Compiling a corpus of linguistically annotated audio and transcripts from existing oral history collections can provide a much-needed increase of speech data for AAL. This increase of data can support more equitable language technology, specifically for ASR. African Americans are a population largely absent from focused corpus linguistics studies as well as NLP (Dacon 2022; Martin 2022; Martin and Wright 2022). African Americans have been the focus of several oral history projects.[8] Generally speaking, these projects serve as chronicles of African Americans who lived through the transatlantic slave trade, the Jim Crow era, the Civil Rights Movement, the wars of the twentieth century, and the first Black presidency. Linguists will find a wealth of conversational speech data, sociolinguistic dynamics, and phonological and morphosyntactic structures.

2.2. African American speech corpora

With AAL's marginalized status being a consistent obstacle for linguistic data collection, linguists have begun to create their own repositories of AAL data. CORAAL (Kendall and Farrington 2021) has been publishing collections of sociolinguistic interviews since 2018 (Kendall and Farrington 2022), and these projects focus on different regional varieties of AAL. CORAAL works in tandem with researchers (often African American)

[6] http://corpas.ria.ie/
[7] https://www.militaryarchives.ie/collections/online-collections/bureau-of-military-history-1913-1921/
[8] For a compiled list of oral history projects with a focus on African Americans, see https://guides.library.duke.edu/africanamericanoralhistories/collections

to make this data accessible to both linguists and the general public. Kendall and Farrington (2022: 191) have also reiterated the importance of their work and argued that their point is "that new advances and better science can be done if there are *more* public and larger data sets." As much AAL data is not readily available for linguists, let alone published in an open space for hobbyists or educators to interact with, CORAAL is pioneering a more accessible approach to disseminate linguistic research. This work has inspired one of our main objectives with this project as well.

As already stated in Section 1, the current paper deals with the compilation of a new speech corpus from an oral history collection called JBA, rather than compiling a corpus using sociolinguistic interviews like CORAAL. JBA is a large and growing collection, containing more than 600 oral history interviews with African Americans that were recorded in the state of Florida and across the Southeastern United States. The archive houses a corpus of approximately 6.5 million words and 1,100 hours of audio. It combines several different regional projects under the *Samuel Proctor Oral History Program* (SPOHP),[9] with the earliest interviews from the 1970s, and is continually updated. Interviewees consist of student participants, community elders, prominent citizens (e.g., pastors, politicians, union workers), ranging from teenage to elderly in age. Through snowball sampling, a non-probability sampling technique taken from sociological methods, where participants recommend other individuals in the community, this collection records individual life experiences to shed light on the complex histories of communities. This collection is maintained by the *University of Florida Digital Collection*[10] through George A. Smathers Libraries.[11]

2.3. AAL

AAL has a rich set of distinctive phonological and morphosyntactic features. Unique syntactic features of AAL can pose challenges for language technology, as many products are built using GAE resources, which often do not account for the differing grammatical structures that are foundational for AAL. Because AAL and GAE do have these differences, it is necessary to build language technology from AAL data. Given the limited number of repositories to make this endeavor easier, we are tasked with adding to

[9] https://oral.history.ufl.edu/
[10] https://ufdc.ufl.edu/
[11] https://ufdc.ufl.edu/collections/flaac

the data that already exists. Most of our effort to date has focused on six morphosyntactic features that are significant and frequent characteristics of AAL and differ from GAE (Green 2002) and are illustrated below. These are: a) Person/number disagreement (absence of third person singular -*s*), as in example (1); b) habitual *be* (cf. 2); c) multiple negation (cf. 3); d) remote past *bin* (cf. 4); e) existential *it/dey* (cf. 5); and f) null copula, as in (6).

(1) Saying he just want to remember.
Saying he just wants to remember.

(2) I be in my office by 7:30.
I am usually in my office by 7:30.

(3) I ain't step on no dog.
I didn't step on a dog.

(4) We been adding cinnamon to the cookies.
We've added cinnamon to the cookies for a long time now.

(5) Dey some coffee in the kitchen.
There is some coffee in the kitchen.

(6) We the county champs.
We are the county champs.

3. Compiling Linguistic Corpus from Oral History Collection

This section describes general steps to compile a linguistic corpus from a collection of oral histories. It presents issues that corpus linguists may expect to encounter. The first step is to edit the oral history transcriptions. Second, the transcriptions need to be annotated with linguistic features. Third, compiling a spoken corpus involves aligning the transcribed utterances to the timestamps of corresponding speech in audio files (Harrington 2010).

3.1. Oral histories transcriptions

Oral historians who focus on the experiences of minority communities are likely to encounter language varieties that may have distinctive features which are not accurately represented in the standard orthography. Anyone may struggle to understand another person's accent, recognize regional linguistic features, or correctly interpret

colloquialisms. How oral historians handle these features during transcription is informed by their needs. Therefore, the first step for compiling a linguistically useful corpus is to decide whether to edit the collection's transcriptions to suit the linguists' needs.

A general lack of orthographic norms for regional language varieties complicates transcription (Ghyselen *et al.* 2020). Encountering a less common language variety reduces transcription accuracy by humans. The *National Court Reporters Association*[12] (NCRA) sets a standard at 95 percent accuracy to be certified as a court reporter, yet professional and certified transcribers with thorough training and examination measured as low as 59.5 percent when transcribing African American speakers. In oral history projects, transcribers are often student volunteers and are unlikely to have the same level of training or linguistic awareness to help them hear and transcribe regional linguistic variations (see Appendix B: Data Availability).

Linguists wishing to compile a corpus from an oral history collection should carefully review the oral history program's transcription guidelines. The field of oral history has no universal transcription standards and individual programs post their own guidelines which range in expectations and details. The examination of a handful of these guidelines (Samuel Proctor Oral History Program 2007, 2016; Oregon Department of Transportation Research Section 2010; Strong *et al.* 2018; Samuel Proctor Oral History Project 2020, 2023) reveals a tendency to place priority on informational content, which is a historian's primary interest. Guidelines seem to reflect a philosophy that transcribed dialogue does not need to be completely reflective of a person's speech. For example, the *Columbia Center for Oral History Research's* (CCOHR) transcription style guide (Columbia University Center for Oral History Research 2022: 2, 15) states the following:

> The characteristics of *how* individual speakers communicate —in terms of syntax, grammar, and word usage— are welcome in the transcript so long as they do not interfere with the written clarity of *what* speakers meant to communicate. Fidelity to key characteristics of each individual's speech holds a lower priority.

While oral history guidelines do tend to emphasize authenticity to the spoken word, they often encourage, and sometimes require, adherence to standard orthography. Some programs allow only 'dictionary spelling' and specifically prohibit non-standard representations, including requirements that contracted forms be spelled out (e.g., *won't*

[12] https://www.ncra.org/

rather than *will not*). Some guidelines recommend inserting dropped elements, such as pronouns or the third person singular verb agreement suffix *-s* which may be attributed to 'hurried speech' rather than potential regional linguistic markers. The few oral history guidelines that expect some faithfulness to the spoken sounds find compromises to handle regional linguistic variation. For example, the Samuel Proctor Oral History Project's (2020) transcription guide states that it is important to be faithful to the uniqueness of each person's voice and emphasizes preservation of an individual's personal manner of speaking, allowing reduced forms such as *kinda*, *gonna*, and *wanna* as well as non-standard grammar. The Samuel Proctor Oral History Project's (2016: 8) guidelines also permit transcribers to leave in double negations, explicitly stating "do not change improper grammar said by the speaker." At the same time, the guidelines prohibit transcription of conversational fillers (*um*, *er*, etc.) that might be expected for a linguistic analysis of conversation.

The guidelines vary in how to represent a variety's well-known features. Some recommend that the distinctive features of non-standard varieties be 'corrected' with the explicit purpose of not embarrassing the speaker, as if desiring to present every interviewee as polished, formal, and educated. For example, the oral history transcription guide by Oregon Department of Transportation Research Section (2010: 6) states that "Slang such as 'y'all' is acceptable —very occasionally— if that's what was spoken, although it should not be used extensively for regional approximations *á la* Mark Twain." Notably, no guidelines we examined contain sections dedicated to the transcription of non-standard or regional vocabulary, pronunciation, or grammar. Our corpus adheres to the SPOHP transcription guidelines except for the additional guideline we created to account for African American morphosyntactic features (see section 4).

3.2. Time alignment

Time alignment increases the accessibility and usability of oral histories. Time alignment is the matching of a transcription excerpt to an excerpt of audio. Oral history audio files are often archived separately from transcription files, only sharing a file name (with a different file extension). This lack of alignment between the text and audio makes it difficult to investigate the spoken aspects such as phonetic features tied to the interviewee's social background or emotional state, and non-literal meaning such as irony and humor. Without an aligned connection between the transcription and recordings, it is

more difficult to observe linguistic features (e.g., pronunciation variations) in the acoustics and annotate them in the text. Likewise, addressing transcription errors benefits from immediate access to the spoken utterance.

Fortunately, time alignment can be automated, via a procedure referred to as 'forced alignment'. Forced alignment can be performed at the utterance-level and at the word/phone-level using existing toolkits, such as *Aeneas* (Pettarin 2017). *Aeneas* generates utterance-level alignment by first synthesizing speech from the orthographic representations and then comparing the synthesized speech with the actual speech. This comparison obtains an approximation of each utterance' timestamps. Word/phone-level time alignment can be obtained with another tool, namely the *Montreal Forced Aligner* (MFA; McAuliffe *et al.* 2017), which relies on acoustic models of phonetic units. Both *Aeneas* and MFA were used to align our corpus. Our approach to time alignment is discussed in Section 4.

Unsurprisingly, the quality of alignments depends on the quality of the recordings. Oral history interviews take place in a variety of settings, such as a home, restaurant, school, and even outdoors. This mirrors the environmental settings of linguistic or anthropological fieldwork recordings (Whalen and McDonough 2015). When time-aligning the *Spoken British National Corpus*,[13] Coleman *et al.* (2011) reported several challenges. Everyday conversations were particularly difficult to align due to factors such as overlapping speakers, background noise, variable signal loudness, reverberation, distortion, and poor speaker vocal health. The alignment was suboptimal for phonetic acoustic research (only 24% of the phoneme boundaries were within 20 milliseconds of expert human labels) although it achieved sufficient accuracy for users to navigate to the desired audio portion of the transcription (83% of the phoneme boundaries were within two seconds of their correct positions). For this reason, time-alignment remains an important step when compiling a linguistic corpus from an oral history collection.

Forced aligners suffer from limited or poor-quality data and improve with increased training data. Forced aligners for majority languages have usually been trained on hundreds or even thousands of hours (DiCanio *et al.* 2013). This amount of data is not available for AAL. However, when data is not available, data from related languages or from languages with a similar phone inventory has been shown to improve forced

[13] http://www.phon.ox.ac.uk/AudioBNC

alignment (see Tang and Bennett 2019; Pandey *et al.* 2020, and references therein). So, GAE models can be fine-tuned for AAL.[14] Alternatively, a model trained on data with significant background noise removed can improve the alignment quality (Johnson *et al.* 2018). While such preprocessing of the audio data can improve the alignment, it is labor intensive. Johnson *et al.* (2018) report that to clean one hour and 42 minutes of recordings required an undergraduate research assistant to complete 120 to 150 hours of manual editing.

3.3. Annotating linguistic features of interest

Linguistic annotation is an important tool for linguistic study. Annotation consists of adding information to texts about features such as anaphora, parts-of-speech (POS), phonetics, semantic roles, and syntactic structure (see Schiel *et al.* 2012: Chapter 8). Oral history collections are not usually enriched with linguistic information. Those wishing to undertake in-depth linguistic investigation via oral history must undertake linguistic annotation.

In our experience, annotation of oral history transcriptions does not present special considerations or require a particular tool or method. Linguistic annotation can be undertaken with various tools and guidelines that can be designed to a project's goals to fit the team's workflow. However, if one wants to apply automatic annotation, it should be noted that most state-of-the-art NLP tools are not trained or optimized for spontaneous speech (Moore *et al.* 2015; Dinkar *et al.* 2023) and may need to be fine-tuned (Rohanian and Hough 2021) or require a final step of manual corrections.

4. CHALLENGES DURING COMPILATION

Compiling a linguistic corpus from an oral history collection gives rise to challenges, particularly with a non-standard language variety. This section details challenges we encountered during transcription and time-alignment and summarizes how we handled the challenges.

[14] See Magnotta (2022) for a comparison of forced aligners that were trained on either AAL speech or GAE speech data.

4.1. Transcription challenges and errors

Transcription of spontaneous speech is a challenging task. Transcribers misperceive what they hear due to background noise, poor audio quality, lackluster listening equipment, limited transcriber training, or the transcriber's fatigue (Meyer *et al.* 2013; Tang 2015). Transcription errors are unavoidable but mistranscriptions may have downstream negative effects. Critically, some mistranscriptions may not be considered errors by oral historians (see Section 3.1) but may nevertheless hinder identification of linguistic information. Additionally, if a corpus is used to train ASR systems, the systems may reflect bias against certain speakers due to the transcription choice (Blackley *et al.* 2019; Koenecke *et al.* 2020; Martin and Tang 2020).

Most oral history projects detail a pipeline for transcribing audio files. SPOHP employs its own group of in-house human transcribers who have received training based on the SPOHP's own transcription guidelines (Section 3.1). At SPOHP, each transcription goes through three passes. In the first pass, a first draft transcription is created. This is followed by an audit pass to correct errors and clarify areas marked as unsure. A third pass finalizes the transcription. To understand the transcription of JBA oral histories better, we took a detailed look at 14 transcriptions. We examine first draft transcripts (rather than final drafts) because oral history collections are constantly growing and at any given time, the bulk of the interviews are in first-draft status. First drafts provide an opportunity to examine transcribers' first impressions of AAL. Also, we found that not all mistranscriptions that could be attributed to misperceiving dialectal variation are corrected in the final drafts.

We identified mismatches between audio and transcripts and grouped them into four categories: 'omission', 'insertion', 'substitution', and 'unsure' (Hennink and Weber 2013; Stolcke and Droppo 2017; Zayats *et al.* 2019). Omission takes place when a word or phrase is present in the audio but not in the transcript, as illustrated in (7). Insertion occurs when a word or phrase is not present in the audio but is in the transcript, as in (8). Unsure involves cases in which the transcribers indicated they were unsure of their work, e.g., [*inaudible at 6:04*].

(7) Audio: A lot of my friends *had* got drafted and had already got killed.
Transcript: A lot of my friends got drafted and had already got killed.

(8) Audio: I went to a lady was teaching school in a wooden building.
Transcript: I went to a lady *who* was teaching school in a wooden building.

Substitution is found when the transcript misrepresents a word or phrase in the audio. Word-level substitutions differ from the audio by an entire word, as shown in (9). Character-level substitutions differ in no more than two letters or else a single inflectional morpheme differs, as illustrated in (10).

(9) Audio: In middle school I didn't go to a *white* school…
Transcript: In middle school I didn't go to a *black* school…

(10) Audio: I work for him until I went in the service.
Transcript: I work**ed** for him until I went in the service.

Across the 14 first-draft transcripts we found 1,041 mistranscriptions. The distribution is shown in Table 1. 82 percent were unconscious errors (omission, substitution, insertion). The most common type is that of substitutions (47%). Omissions (27%) and insertions (7%) are less common.

Total Errors	**Substitution**	**Omission**	**Insertion**
1,041	493	283	76

Table 1: Distribution of transcription errors in sample texts

We found notable trends that relate to misrepresentation of AAL. For example, character-level substitutions frequently occurred with verb tense markers which misrepresents AAL characteristic person/number disagreement, as shown in example (11). Also, word-level substitution could result in sparse representation of possible AAL signifiers, as in (12).

(11) Audio: I can't believe that name escape me.
Transcript: I can't believe that name escape**d** me.

(12) Audio: And my *father* came back from the war, and they fell in love…
Transcript: And my *pop* came back from the war, and they fell in love…

There were 38 occurrences of five of the six AAL features (remote past *bin* did not occur) and 31 were transcribed correctly. Occurrences of perfect *done* and multiple negations were always transcribed correctly. Interestingly, AAL features were artificially inserted two times, as in example (13).

(13) Audio: No, **I'd be** passing by…
Transcript: No, **I be** passing by…

Existential *it/dey* proved the most difficult for transcribers (and annotators) to identify. Occurrences were usually substituted for the GAE existential construction, as in example (14).

(14) Audio: …**it was** so many of us today that they had three.
Transcript: …**there were** so many of us today that they had three.

The null copula was mistranscribed three times. In example (15), we see an instance of null copula being incorrectly omitted by the insertion of *was*.

(15) Audio: …the same street the Duncan Brothers funeral home on.
Transcript: …the same street the Duncan Brothers funeral home **was** on.

When compiling an oral history collection into a corpus for linguistic study, transcription issues should be addressed in a way that best serves the compiler's goals. Our project intends to serve oral historians as well as linguists so, when making decisions about handling transcription issues, we sought a balance between accurate linguistic representation and easy-to-read transcripts. We decided to change the transcription guidelines only where they misrepresent the AAL grammatical features we are currently interested in (see Section 2.3). We correct other general issues such as larger missing or incorrect transcriptions to accomplish forced time alignment of the transcriptions and audio files. Such issues that related to several seconds of audio hinder time alignment. Our more accurate representation of the AAE features also seems to improve forced alignment and increase utility for training NLP models for AAE. We work with final draft transcriptions where possible. In the future, we hope to address how best to represent phonological features while still maintaining readability.

In our project, revising transcriptions has been integrated into the pipeline of annotation and time-alignment. We developed simple short recorded video training with exercises about the AAL features of interest. The training makes transcribers aware of these features and AAL in general, guides them to preserve standardized spelling wherever possible, but to refrain from 'correcting' potential dialectal markers. Readers who wish to inform their own corpus compilation work may find our training materials in the publicly-available repository mentioned in Appendix B (Data Availability).

We currently focus annotation efforts on the six distinctive AAL morphosyntactic features listed in Section 2.3. Annotation is performed by students who are native or near native English speakers of any English variety (including AAL), having varying

familiarity with AAL and varying backgrounds in linguistics. Annotators are trained and tested on their ability to identify the six AAL morphosyntactic features before they are allowed to work independently. They listen to the audio recordings and annotate the location and type of feature in the transcriptions using a software tool named 'Rezonator' (DuBois *et al.* 2020). In this tool, annotators highlight a word or phrase with a mouse and select from a list of feature labels. Rezonator text annotations can be exported as CSV files with units in rows and their labels in cells on the same row. In Figure 1, the linguistic feature of person/number disagreement is shown as annotated within an interview. We found it efficient to have multiple rounds of annotation. In each round, annotators focus on one or two features rather than all six at once, while also checking for features they had previously annotated but may have missed.

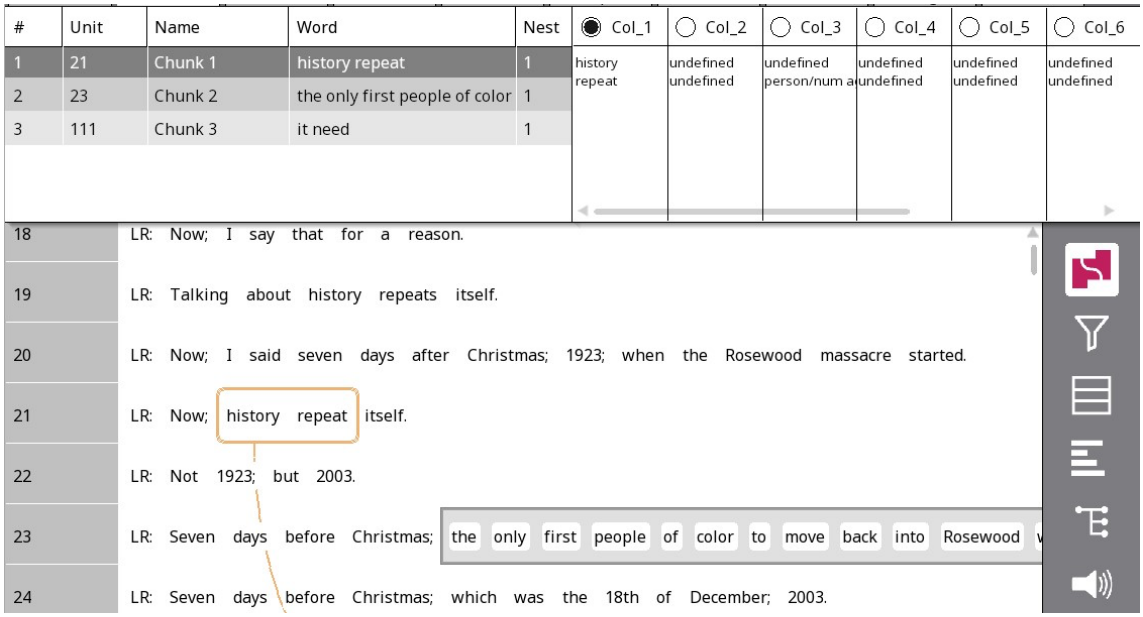


Figure 1: An annotation within an interview

4.2. Time-alignment: Challenges and errors

Time-alignment of audio and text in non-GAE varieties presents a set of challenges that stems from various sources such as mismatches in transcriptions, cross-talk, and lack of customized phonetic resources. In some instances, we have been able to address these challenges. For others, we merely highlight here obstacles that are likely to be encountered by anyone compiling a corpus from oral histories.

4.2.1. Challenges due to mistranscription

One challenge stems from the mistranscriptions of AAL (see Sections 3.1 and 4.1), which can result in misalignment, or even complete omission of distinctive features during forced alignment. As an example, the clause *...when Live Oak got **they** team...* in AAL was mistranscribed as *...when Live Oak got **their** team...* (*they* has the equivalent of *their* in GAE). If time-aligned files output from *Aeneas* and from MFA are constructed from such a mistranscription, the mistranscription would lead MFA to attempt aligning the vowel in *they* [eɪ] using an acoustic model of the vowel in their [ɛ] and it would also try to align some audio signals to an acoustic model of a rhotic phone [ɹ] in *their*, which does not exist in the audio. The misalignment of the phones results in an overall poorer word-level alignment of this and surrounding words, as illustrated in Figure 2.

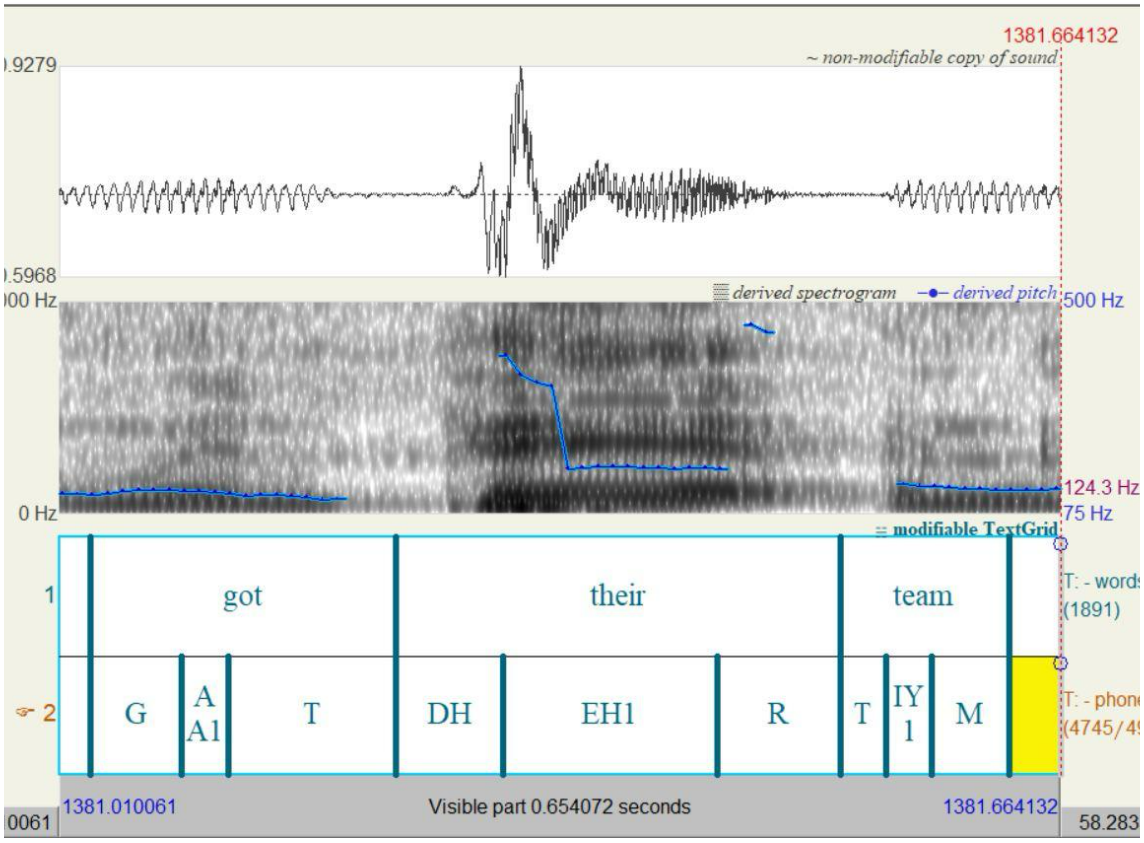


Figure 2: Time-aligned files that, due to mistranscription, do not contain the AAL feature *they*

To mitigate this issue, we are correcting the transcriptions prior to time-alignment (see Section 4.1). While automatic transcription error detection exists, such as Kisler and Schiel (2018), they require a pre-existing set of corrected transcriptions before a language-specific model can be trained. We decided to leave automatic transcription error detection for future work as we must first create a large enough set of corrected

transcriptions. We decided on steps that involve focusing first on interview transcriptions in the final draft version and then manually correcting any transcripts that contain AAL features in the audio but not in the transcripts (see Section 4.1). Once the transcript correctly matches the features in the audio, MFA is able to time-align the audio signals with the transcriptions more accurately. Remaining alignment errors can be quickly corrected by hand.

4.2.2. Cross-talk

Cross-talk occurs when two or more speakers produce an utterance simultaneously. It often results in one speaker cutting off their utterances upon hearing the other. The transcription in Figure 3 illustrates two instances of the interviewer being interrupted by the interviewee (fourth and third-to-last lines). Not all cases of cross-talks are clearly indicated in the transcripts as an interruption (e.g., a dash line at the end of a line) because the speaker being interrupted may finish their utterances as if there is no interruption. Crucially, the onset and offset time of cross-talks (i.e., the exact portion where both speakers overlap) may not be indicated in oral history transcripts.

```
T:      My momma's Naomi Bryan. Father, unknown.

W:      And when was your mother born?

T:      Let me see now; she was born in December the 12th, 1910.

W:      How about her mother and father? What do you know about her—

T:      I don't know nothing, I don't know about them.

W:      Do you have brothers and sisters?

T:      One sister.

W:      Can you state her name?

T:      Sarah Cross.

W:      And is she still—

T:      Deceased.

W:      Deceased.
```

Figure 3: Original transcription of two cross-talk examples

The identification of words and phones can be hampered due to the conflicting voice activity. One solution is to split speakers' audio into separate channels or even entire files if the audio was recorded with a dual channel setup. Another solution is to label an utterance as belonging to a certain speaker (speaker diarization). This can be done automatically using NLP models trained specifically for this task and it is supported by MFA. We are currently testing this method but, as with most language technologies, speaker diarization is known to perform worse for non-standard varieties (Tevissen *et al.* 2023). Currently, we are performing the speaker diarization task manually by listening for cross-talk in the audio and using indicators of interruption in the oral history transcription. We then separate the time-aligned words into speaker tiers in textgrid format, thus allowing cross-talk to be represented in a way that reduces misalignment.

4.2.3. Pronunciation dictionaries missing AAL variants

Pronunciation dictionaries for forced alignment exist for different languages, but their diversity as it relates to English varieties is limited. AAL does not have a pronunciation dictionary in the MFA database. Thus, phonetic representations may not be comprehensive for AAL. For instance, the MFA ARPABET dictionary for English does not include the AAL *th*-stopping pronunciation, as can be seen in Figure 4, which contains only the *DH* mapping to the IPA [ð], but not the *D* mapping to the IPA [d] AAL variants for the words *that* and *they*. This is a challenge we have not yet been able to address adequately.

```
that    DH AE1 T
that    DH AH0 T
that'   DH AE1 T AH0
that'd  DH AE1 T IH0 D
that'll DH AE1 T AH0 L
that'n  DH AE1 T AH0 AH0 N
that're DH AE1 T AH0 R
that's  DH AE1 T S
that'sh DH AE1 T AH0 SH
that'th DH AE1 T IH0 TH
that'ud DH AE1 T IH0 AH0 D
that've DH AE1 T AH0 V
```

```
they    DH EY1
they'd  DH EY1 D
they'l  DH EY1 AH0 L
they'ld DH EY1 D
they'll DH EY1 L
they'm  DH EY1 AH0 M
they'n  DH EY1 AH0 N
they're DH EH1 R
they's  DH EY1 Z
they'se DH EY1 S
they've DH EY1 V
```

Figure 4: Dictionary entries for variations of *that* and *they*, which omit the *th*-stopping feature found in AAL

AAL phonetic/phonological features can be modeled using a phone-level forced-alignment model which can capture pronunciation variation due to accent and regional differences (Yuan and Liberman 2011; McLarty *et al.* 2019; Kendall *et al.* 2021). This is achieved by allowing the model to evaluate multiple pronunciations of the same word. For example, given the acoustic signal of the word *running*, which can be produced with a velar nasal sound or with an alveolar nasal sound (velar nasal fronting), the model will assign the most likely pronunciation. Some well reported phonological features include: *th*-stopping (e.g., *that* [dæt]; Thomas and Bailey 2015), velar nasal fronting (e.g., *running* [ɹʌnɪn]; Tagliamonte 2004), final consonant cluster reduction (e.g., *test* [tes], *hand* [han]; Green 2002: 107), monophthongization of /ai/ (e.g., *buy* [ba:]; Rahman 2008: 147) and /ɹ/ vocalization (e.g., *court* [koət], *bear* [beə]; Green 2002: 120). Some grammatical features that resemble a pronunciation variation can also be automatically annotated as such, for instance, the absence of *-s* verb tense inflection (e.g., *He goes* [goʊ]; Rahman 2008: 147).

The latest version of MFA has an advanced feature that enables the selection of the correct pronunciations given the acoustic likelihood with a set of prior pronunciation probabilities. However, we cannot easily estimate the pronunciation probabilities of each variant without a large and accurate phonetically transcribed spoken AAL corpus.[15] Therefore, we cannot make use of an advanced feature of MFA. For this reason, we decided to generate the possible pronunciation variations of each word type with equal probabilities. We have extracted sentences from JBA interview transcripts to identify any words that had pronunciations missing from the MFA dictionary, then generated more comprehensive dictionaries that account for AAL pronunciations of these words. We are currently testing this method, as we continue to address time-alignment challenges.

5. Discovering Linguistic Information in Oral Histories

The process of compiling a linguistically enriched corpus from oral histories provides insights into the language variety spoken by the interviewees in natural settings. This section describes insights we have gleaned so far about AAL. The insights pertain to the

[15] CORAAL can potentially be used for this purpose as it has been fully forced aligned with phonetic transcriptions (Farrington and Kendall 2019). However, we have not investigated to what extent CORAAL handled pronunciation variations and whether they were manually verified.

representativeness of an oral history collection (Section 5.1), the distribution of AAL features in speech (Section 5.2) and modeling syntactic structures that signal the presence of the AAL feature habitual *be* (Section 5.3). Finally, we discuss the potential for modeling AAL linguistic information preserved in oral history collections with NLP systems, which in turn can provide automated assistance to annotators (Section 5.4).

5.1. Preliminary description of the UF AAL Spoken Corpus

The JBA (see Section 2.2) was chosen due to its abundance of AAL speakers and its conversational nature that is conducive to naturally occurring instances of AAL features (Roller 2015). Our goal is to compile a corpus with no less than 500 JBA interviews that are completely time-aligned and featurally annotated. The compiled corpus will aid the linguistic investigation of AAL and the development of NLP tools that reduce bias against AAL.

The notion of representativeness is an important consideration in corpus compilation for linguistic investigation. However, oral historians' sampling technique is not governed by linguistic representativeness (as mentioned in Section 2.2, snow-ball sampling was used in JBA). Egbert *et al.* (2022: 28–51) report a survey of conceptualizations of representativeness in corpus linguistics, but all these conceptualizations that tie to random or stratified sampling are not applicable to a corpus compiled from an oral history collection. However, one conceptualization, which Egbert *et al.* (2022) argue against, is that a very large corpus is a *de facto* representative corpus. In this view, corpus size is the primary consideration of corpus design and, as Sinclair (1991: 8) states, "... a corpus should be as large as possible and should keep on growing." A spoken corpus with 500 JBA interviews is arguably a very large corpus, indeed larger than the other significant AAL corpus (CORAAL). When completed, the full corpus should serve as a corpus of the AAL spoken in the state of Florida and across the Southeastern United States. What specific linguistic features and socio-demographic dimensions it will be representative of remain to be determined upon completion.

For the purpose of NLP tool development, we chose a subset of 58 interviews. It is important to note that not all African Americans speak AAL. Therefore, our initial corpus subset was compiled from a preliminary inspection of the JBA transcribed materials conducted to identify speakers who consistently used AAL, while giving strong weight

to audio quality (see Section 3.2). We also made practical considerations, such as if the transcription was completed and if the recording had no more than three AAL interviewees. Our current compiled corpus comprises speech from 18 interviewees across 16 interviews that have been fully annotated. In the 18 interviewees, gender is evenly distributed (nine male speakers and nine female speakers). 15 speakers are adults ranging from 26–80 years while the remaining three are teenagers. The interviews were recorded over a ten-year span from 2008 to 2018 and were located through Mississippi and northern Florida. The corpus contains 59,388 tokens, averaging 4,568 per document (see Appendix A for more details).

5.2. Distribution of AAL features

Early in the project, we performed a pilot annotation round on 16 interviews that were annotated fully in one round for all six features. These interviews were annotated by our trained annotators, but they were also annotated independently by a class of graduate and undergraduate students who were given an abbreviated version of the training. Although the more thoroughly trained annotators were better at hearing some features, the two teams of annotators confirmed each other's findings regarding the distribution of features.

The pilot study allowed us to investigate the distribution of AAL features and understand the prevalence of each feature better. The results are displayed in Table 2. Null copula (34%) and person/number disagreement (30%) together comprise over half the AAL features in the data. Multiple negation (16%) and existential *it/dey* are less frequent but still prominent and together comprise one third of the AAL features. The other three features are much less frequent. Remote past *bin* and habitual *be* are the rarest, while perfect *done* occurs only slightly more frequently. Our observations on remote past *bin* match other work where it is noted as a rare feature (Green *et al.* 2022).

Text	Null copula	Person/ Number	Multiple Negation	Existential *it/dey*	Perfect *done*	Remote past *bin*	Habitual *be*	Total
1	25	9	1	6	0	0	1	42
2	0	1	0	3	0	0	0	4
3	3	15	4	3	2	0	0	27
4	2	0	0	3	0	1	0	6
5	10	7	5	4	3	1	0	30
6	2	7	4	5	0	0	0	18
7	1	1	0	0	0	0	0	2
8	5	7	8	3	0	0	0	23
9	7	3	4	0	0	0	0	14
10	2	3	3	2	0	0	0	10
11	15	13	7	1	1	0	1	38
12	6	2	1	0	0	0	1	10
13	0	0	0	0	0	0	0	0
14	4	0	1	0	0	0	0	5
15	5	7	5	6	1	0	0	24
16	0	1	0	4	0	0	0	5
Total	87	76	43	40	7	2	3	258

Table 2: AAL features found by annotation team in a pilot study of 16 oral histories[16]

We found that the features are not equally recognizable to annotators. Existential *it/dey* is difficult to identify despite its frequency and requires second passes. This feature may be difficult to identify because it requires attention to the larger context, whereas the other features can mostly be identified by examining the sentence or phrasal context. Another feature that is problematic for annotators is remote past *bin*. We suspect the motivation for this are the feature's rarity, semantic load, and contextual constraints which are more complex than what is described in the literature. On the other hand, multiple negation, which is found in other varieties of English, is one of the most frequent features and seemingly the easiest to recognize.

5.3. Syntactic environments of habitual be

It is known that certain syntactic environments correlate with AAL features like habitual *be* (Fasold 1972; Green 2002). We used the oral history data to investigate whether these described environments hold in naturalistic data. We investigated POS and syntactic dependency relations in the environment of habitual and non-habitual occurrences of *be*. To analyze the structures, we first ran the *NLTK POS* tagger (Bird 2009) and the *spaCy*

[16] The list of texts by interviewee is shown in Appendix A.

Universal Dependency syntactic parser[17] on all sentences containing *be*. Both models are originally trained on GAE. Then, we analyzed the syntactic patterns and, from this analysis, built a rule-based machine learning classifier to identify *be* as either habitual or non-habitual.[18] In the process, we confirmed POS and syntactic dependencies commonly found in the environment of habitual *be* and uncovered new ones. Further information about the NLP results of the efforts described here are available in Previlon *et al.* (2024).

5.3.1. POS environments

We first leveraged POS patterns distinguishing habitual and non-habitual usage of *be* as described by Green (2002). For example, patterns indicative of the habitual meaning include a pronoun immediately preceding *be*, as in *...they be like, what you finna do?* or a verb ending in *-ing* immediately following it, as in *But LeBron be passing though*. We coded these POS environments described in Green (2002) as Boolean (True/False) Python rules to filter out many non-habitual *be* instances. These filtering rules do not capture all instances of non-habitual *be*, but they also do not flag any false positives.

Examining non-habitual instances that were not filtered, we uncovered POS patterns not described in the literature. Because our data is limited, we labeled these new patterns *ad-hoc*. Future study may determine if they are generally applicable. *Ad-hoc* POS rule 1 states that *be* is non-habitual if it is immediately followed by a deverbal noun and immediately preceded by neither a personal pronoun nor a noun, as in (16) below. *Ad-hoc* POS rule 2 states that *be* is non-habitual if it is immediately preceded by an adverb and immediately preceding that adverb is either a verb or modal verb, as in (17).

(16) I will **never** be **going** there again.

(17) You **should regularly** be trimming the dog's nails.

The power of the known and the *ad-hoc* POS patterns to disambiguate habitual/non-habitual meanings was tested by applying a machine learning classifier trained on the output of the filtering rules to 5,133 instances of *be* in the CORAAL corpus (Santiago *et al.* 2022). The results were compared to the manually annotations and are displayed in

[17] https://spacy.io/

[18] It would be more accurate to say that we disambiguate standard uses of *be* from non-standard, because we do not distinguish between habitual *be* and other non-standard invariant forms, such as emphatic *be*, which occur in nearly identical syntactic environments (Harris 2019).

Table 3. The POS-based classifier correctly tags 79 percent of non-habitual instances, and only 13 percent of instances of habitual *be* were false positives.

The *ad-hoc* rules increased non-habitual filtering accuracy over just the known environments, but unlike the known POS patterns, they also incorrectly filtered some habitual instances. Future analysis may show whether these false positives can be reduced by refining the *ad-hoc* analysis or whether these environments are indeed ambiguous.

	Tagged as 'non-habitual'	**Not tagged**	**Total**
Non-habitual	3,662	994	4,656
Habitual	61	416	477

Table 3: Number of habitual and non-habitual usages of *be* and how they are tagged by rules based on POS environments[19]

5.3.2. Dependency syntax environments

As part of the process of building NLP tools for AAL, we explored how GAE-trained NLP tools model the syntactic environments that signal the presence of habitual *be*. We contrasted those environments with the modeling of non-habitual *be* syntax. We parsed the dependency trees of 250 sentences (132 habitual and 118 non-habitual) containing *be* in the JBA data. Nine significant patterns were identified. Of note for developers of NLP tools to disambiguate habitual and non-habitual *be* is that the relevant dependency relations are primarily constrained by immediate parent, child, and sibling relations between *be* and other words. Zoomable figures illustrating these dependency structures are available on the Open Science Framework (see Appendix A: Data availability).

The nine patterns can be stated as rules. Rules 1–3 identify habitual patterns while rules 4–9 identify non-habitual patterns. In large part, our analysis of computational modeling confirms Green's (2002) description of habitual *be*. Rule 1, 'main verb', is found when *be* is POS-tagged as a verb and has a child that is an adjective, adverb, or preposition (48 instances). Rule 2, 'aux to main verb', is when *be* is both POS-tagged and has a dependency relation of auxiliary while its parent is labeled as a verb (70 instances). Rule 3, 'subordinate clause', occurs when *be* is POS-tagged as an auxiliary with the

[19] It should be borne in mind that the POS rule-based tagger identifies non-habitual usages, so 'not tagged' does not necessarily imply habitual meaning.

dependency relation of either a relative clause modifier or a clausal complement (6 instances).

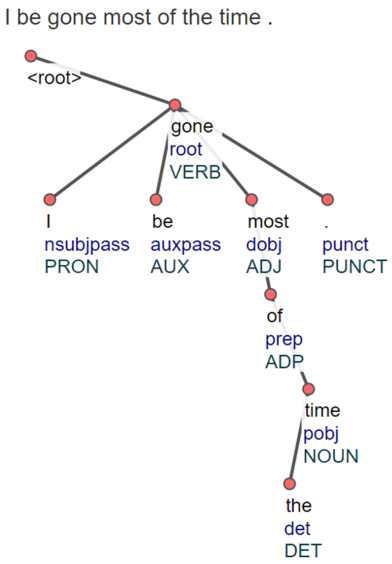


Figure 5: Illustration of habitual dependency rule 2

Under the pattern of rule 2, an interesting phenomenon is found when *be* is followed by *gone* as in *I be gone most of the time*. This is illustrated in Figure 5, which shows that the parser, although useful for capturing the habitual pattern described in rule 2, nevertheless mischaracterizes the syntactic and semantic structure of AAL. It labels *be* as an auxiliary and *gone* as a verb. In GAE, *gone* functions as a verb when used as the past participle of *go* (*She had gone to the store*), but takes on an adjectival role when describing a state of existence, as in *The bad days are gone*. In this example, it can be argued that, with the habitual meaning, *gone* leans more towards an adjectival interpretation than a verbal one, since the speaker is not actively engaging in the act of 'gon-ing', but rather *gone* describes a repeated state of being. That is, the subject is often not here, but it is not the case that it has often left for a place.

Five patterns can be stated as dependency rules of non-habitual meanings. Rule 4, 'auxiliary child', takes place when *be* has a child that is both POS-tagged as auxiliary and has a dependency relation of auxiliary (36 instances). Rule 5, 'auxiliary sibling', is when *be* has a sibling that is both POS-tagged as auxiliary and has a dependency relation of auxiliary (22 instances). Rule 6, 'governed by conjunction', is when *be* has *and* as a sibling and *be* is POS-tagged as an auxiliary with a dependency relation of conjunction. Also, *be* has a child that is an adjective, adverb, or noun (four instances). Rule 7, 'noun child', is when *be* is immediately preceded by a pronoun and has a child that is a noun

(one instance). Rule 8, 'infinitival child', is when *be* has an immediate child that is POS-tagged as a particle and has a dependency relation of auxiliary (51 instances). Typically, the child is infinitival *to* (46 instances), or its phonetic variations *ta* (two instances), or *na* as in *gonna* and *wanna* (three instances). Finally, rule 9, 'particle child', is when *be* has a child that is POS-tagged as a particle with the dependency relation of auxiliary (one instance).

The nine rules discussed above capture all but eight of the 250 sentences. Examining these eight sentences, we learned that the parser also mischaracterizes AAL because of non-standard orthography, such as *g*-dropping, when the velar nasal fronting or variable (ING) is represented orthographically as *-in*,[20] presumably as a reflection of the pronunciation (Tagliamonte 2004; Hazen 2008). Similarly, the non-standard spelling of *Imma*, as in *Imma be talking in a minute*, is incorrectly labeled as a proper noun and *wanna*, as in *...and you never wanna be here tomorrow*, is sometimes labeled as a verb. Similar cases of tagging errors by GAE-trained models on common AAL lexical items have previously been reported (Dacon 2022).

Our analysis of the oral history data reveals that both POS patterns and dependency structures are needed to describe and identify the habitual *be* construction. We tested the expediency of including the syntactic environments in an automatic habitual *be* tagger and this reveals that POS-based descriptions do not sufficiently disambiguate habitual and non-habitual meanings. The addition of dependency-based rules allows the tagger to correctly identify habitual *be* when the POS-based rules do not. For example, the use of syntactic dependencies allows the identification of a *be* that is POS-tagged as a verb (matching the first dependency parsing rule) as habitual, but a POS-only model would flag it as non-habitual. Similarly, in another example, the second dependency parsing rule ('aux to main verb') correctly flags habitual *be* where the POS-only approach does not.

5.4. Automatic annotation of syntactic features

Annotation of distinctive AAL morphosyntactic features is necessary for the study of the language variety. Unfortunately, manual annotation is prohibitive in terms of time and cost, particularly because oral history programs continually add interviews and transcriptions. To assist annotation, innovative methods with NLP can be applied.

[20] The six instances were *tryin*, *walkin*, *willin*, *throwin*, *hurtin*, and *laughin*.

Numerous NLP tools exist that work well for GAE and the same cannot be said for AAL with some exceptions (Jørgensen *et al.* 2016; Blodgett *et al.* 2018). As seen in Section 5.3.2, pre-trained models for GAE cannot however be readily transferred to AAL (see Ziems *et al.* 2022). Therefore, we are designing our own AAL feature taggers.

State-of-the-art NLP models that will identify the features automatically are dependent on massive amounts of annotated data. Manual annotation is necessary to create training examples. When data is limited, then linguistic analysis can compensate, to a certain extent. We decided to leverage syntactic patterns that statistically correlate with AAL features as input data to machine learning models.

Our general steps for developing an AAL feature tagger are 1) identifying linguistic contexts of the feature in the data, 2) analyzing the contextual patterns, 3) coding those patterns as Boolean (True/False) rules that filter non-occurrences and flag likely occurrences of the feature, and 4) using the rules to train a classifier to identify whether a string of text contains the feature of interest. The results can then be integrated into the annotation process by 5) presenting the computer's 'annotations' to human annotators for checking, and 6) using the human corrections to retrain and improve the model. This builds a machine-in-the-loop cycle of annotation that should be faster and more accurate than either purely manual or purely automatic work.

After developing the habitual *be* tagger, we tested it on 5,133 manually annotated sentences of CORAAL (Kendall and Farrington 2021), which were kept separate from the data used to analyze the syntactic environments. Four machine learning models were implemented with *scikit-learn* (version 1.2.2; Pedregosa *et al.* 2011) and a transformer was implemented with *fairseq* (Ott *et al.* 2019). The best model achieved a 0.96 F1 score and beat a baseline that does not leverage syntactic patterns. This shows robust results even in the face of data sparsity. The detail of these model developments and experiments can be found in Previlon *et al.* (2024).

However, the results included many false negatives, indicating additional analysis is needed. In the meantime, an effective machine-in-the-loop annotation cycle only requires annotators to verify true positives and false positives of habitual *be*. We were able to reduce the tagger's false tagging of the habitual *be* as non-habitual by increasing the model's recall. We also applied data augmentation with synthesized habitual *be* sentences *á la* Santiago *et al.* (2022). This created a more balanced training corpus because habitual *be* is rare. A corpus with nearly equal examples of the two classes

positively impacts results by reducing statistical bias towards the more frequent non-habitual *be*.

6. CONCLUSION

This study presented initial efforts in compiling a spoken corpus of AAL using recordings and transcriptions from the oral history discipline (Section 2). This corpus will both address the AAL data gap and allow technology developers to correct racially biased systems. We acknowledge that our project is not the first to compile a set of 'legacy' audio recordings into a linguistic corpus. For instance, Olsen *et al.* (2017) reported on a pipeline to deal with transcription and time alignment issues with a legacy speech corpus consisting of sociolinguistic interviews (Pederson *et al.* 1986) which contain African American speakers. Nonetheless, our project involves the compilation of data that is not only historical but that was originally collected for purposes other than linguistic research, which brings a unique set of challenges (Sections 3 and 4). Furthermore, additional challenges arise due to a lack of reliable computational tools and established transcription standards for AAL, which we are addressing through the process of compiling this corpus (Sections 4 and 5).

In Section 5, we demonstrate the wealth of linguistic information that can be extracted from oral histories, such as a frequency distribution of AAL linguistic features and models of the features' syntactic patterns. As our annotation continues, these analyses will be updated. Discovering distributional information from just a handful of interviews reveals the possibilities for oral history work to enhance corpus linguistics research. No longer limited to self-collected data in small amounts, or publicly accessible data that contains unverifiable sources (e.g., Blodgett *et al.* 2018), linguists may expand their research to address concerns in specific populations. It has enabled us to examine speech features of AAL speakers across the Gulf South on data compiled for non-linguistic purposes. Following the guidelines of Kendall and Farrington (2022) about managing the African American sociolinguistic data, in the future, we may find connections between regionality and the use of various features by annotating what sociolinguistic metadata the oral historians collected or that are found within the oral histories themselves.

Finally, our work contributes to the development of NLP for AAL. Modern speech technology is not possible without time-aligned language data. We are currently testing

the ability of forced alignment to perform on multiple pronunciation features of AAL. Since AAL cannot be fully represented within the limits of GAE orthography, additional linguistic annotation is needed. This annotation, in turn, provides training data for NLP models that will be effective on AAL. The last part of Section 5 demonstrates development of NLP tools designed for annotating AAL linguistic features. We developed an automatic classifier for tagging the AAL feature called habitual *be*. Our work improves an NLP model using insights from syntax. We are extending this method to automatically tag other features such as the neutralized person/number agreement characteristic of AAL. Crucially, our spoken corpus of oral histories can contribute both in providing authentic spoken data for AAL as well as the cultural references mentioned in the topic-guided oral history interviews.

References

Bird, Steven, Edward Loper and Ewan Klein. 2009. *Natural Language Processing with Python*. California: O'Reilly Media Inc.

Blackley, Suzanne V., Jessica Huynh, Liqin Wang, Zfania Korach and Li Zhou. 2019. Speech recognition for clinical documentation from 1990 to 2018: A systematic review. *Journal of the American Medical Informatics Association* 26/4: 324–338.

Blodgett, Su Lin, Johnny Wei and Brendan O'Connor. 2018. Twitter universal dependency parsing for African-American and mainstream American English. In Iryna Gurevych and Yusuke Miayo eds. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics* (Volume 1: Long Papers). Melbourne: Association for Computational Linguistics, 1415–1425.

Blodgett, Su Lin, Solon Barocas, Hal Daumé III and Hanna Wallach. 2020. Language (technology) is power: A critical survey of "bias" in NLP. In Dan Jurafsky, Joyce Chai, Natalie Schluter and Joel Tetreault eds. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online publication: Association for Computational Linguistics, 5454–5476.

Coleman, John, Mark Liberman, Greg Kochanski, Jiahong Yuan, Sergio Grau, Chris Cieri and Lou Burnard. 2011. Mining years and years of speech. *Phonetics Laboratory of the University of Oxford*: 1–23. https://diggingintodata.org/sites/diggingintodata.org/files/miningayearofspeechwhitepaper.pdf.

Columbia University Center for Oral History Research. 2022. *Columbia University Oral History Transcription Style Guide*. https://www.ccohr.incite.columbia.edu/s/CCOHR-Transcript-Style-Guide-2022-htpm.pdf (accessed 31 January 2024.)

Dacon, Jamell. 2022. Towards a deep multi-layered dialectal language analysis: A case study of African-American English. In Su Lin Blodgett, Hal Daumé III, Michael Madaio, Anika Nenkova, Brendan O'Connor, Hanna Wallach and Qian Yang eds. *Proceedings of the 2nd Workshop on Bridging Human–Computer Interaction and*

Natural Language Processing. Seattle: Association for Computational Linguistics, 55–63.

Davis, Alexis, Joshua L. Martin, Eric Cooks, Melissa J. Vilaro, Danyell Wilson-Howard, Kevin Tang and Janice Krieger. 2024. From English to "Englishes": A process perspective on enhancing the linguistic responsiveness of culturally tailored cancer prevention interventions. *Journal of Medical Internet Research* preprint: 57528. https://preprints.jmir.org/preprint/57528

DiCanio, Christian, Hosung Nam, Douglas H. Whalen, H. Timothy Bunnell, Jonathan D. Amith and Rey Castillo García. 2013. Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment. *The Journal of the Acoustical Society of America* 134/3: 2235–2246.

Dinkar, Tanvi, Chleé Clavel and Ioana Vasilescu. 2023. Fillers in spoken language understanding: Computational and psycholinguistic perspectives. *arXiv* preprint arXiv:2301.10761: 1–20. https://arxiv.org/pdf/2301.10761.pdf

DuBois, John W., Terry DuBois, Georgio Klironomos and Brady Moore. 2020. From answer to question: Coherence analysis with rezonator. In Sophia Malamud, James Pustejovsky and Jonathan Ginzburg eds. *Proceedings of the 24th Workshop on the Semantics and Pragmatics of Dialogue - Short Papers*. Waltham, New Jersey: SEMDIAL, 1–4. http://semdial.org/anthology/Z20-Bois_semdial_0031.pdf

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

Farrington, Charlie and Tyler Kendall. 2019. *The Corpus of Regional African American Language: MFA-Aligned*. Version 2019.06. http://lingtools.uoregon.edu/coraal/aligned/.

Fasold, Ralph. 1972. *Tense Marking in Black English: A Linguistic and Social Analysis*. Washington: Center for Applied Linguistics.

Fitzgerald, Chris. 2022. *Investigating a Corpus of Historical Oral Testimonies: The Linguistic Construction of Certainty*. London: Routledge

Ghyselen, Anne-Sophie, Anne Breitbarth, Melissa Farasyn, Jacques Van Keymeulen and Arjan van Hessen. 2020. Clearing the transcription hurdle in dialect corpus building: The Corpus of Southern Dutch Dialects as case study. *Frontiers in artificial intelligence* 3/10. https://doi.org/10.3389/frai.2020.00010

Green, Lisa J. 2002. *African American English: A Linguistic Introduction*. Cambridge: Cambridge University Press.

Green, Lisa, Kristine M. Yu, Anissa Neal, Ayana Whitmal, Tamira Powe and Deniz Özyıldız. 2022. Range in the use and realization of BIN in African American English. *Language and Speech* 65/4: 958–1006.

Harris, A. Nicole. 2019. *The Non-Aspectual Meaning of African American English Aspect Markers*. New Haven: Yale University ProQuest Dissertations Publishing.

Harrington, Jonathan. 2010. *Phonetic Analysis of Speech Corpora*. Hoboken: John Wiley & Sons.

Hazen, Kirk. 2008. A vernacular baseline for English in Appalachia. *American Speech* 83/2: 116–140.

Hennink, Monique and Mary B. Weber. 2013. Quality issues of court reporters and transcriptionists for qualitative research. *Qualitative Health Research* 23/5: 700–710.

Johnson, Lisa M., Marianna Di Paolo and Adrian Bell. 2018. Forced alignment for understudied language varieties: Testing prosodylab-aligner with tongan data. *Language Documentation & Conservation* 12: 80–123.

Jørgensen, Anna, Dirk Hovy and Anders Søgaard. 2016. Learning a POS tagger for AAVE-like language. In Kevin Knight, Ani Nenkova and Owen Rambow eds. *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego: Association for Computational Linguistics, 1115–1120

Kendall, Tyler and Charlie Farrington. 2021. *The Corpus of Regional African American Language.* Version 2020.05. http://oraal.uoregon.edu/coraal (accessed 25 June 2023.)

Kendall, Tyler and Charlie Farrington. 2022. Managing sociolinguistic data with the Corpus of Regional African American Language (CORAAL). In Andrea L. Berez-Kroeker, Bradley McDonnell, Eve Koller and Lauren B. Collister eds. *The Open Handbook of Linguistic Data Management.* Massachusetts: The MIT Press, 185–94.

Kendall, Tyler, Charlotte Vaughn, Charlie Farrington, Kaylynn Gunter, Jaidan McLean, Chloe Tacata and Shelby Arnson. 2021. Considering performance in the automated and manual coding of sociolinguistic variables: Lessons from variable (ING). *Frontiers in Artificial Intelligence* 4. https://doi.org/10.3389/frai.2021.648543

Kisler, Thomas and Florian Schiel. 2018. MOCCA: Measure of confidence for corpus analysis: Automatic reliability check of transcript and automatic segmentation. In Nicoletta Calzolari ed. *Proceedings of the 11th International Conference on Language Resources and Evaluation*. Miyazaki: European Language Resources Association, 1781–1786.

Koenecke, Allison, Andrew Nam, Emily Lake, Joe Nudell, Minnie Quartey, Zion Mengesha, Connor Toups, John R. Rickford, Dan Jurafsky and Sharad Goel. 2020. Racial disparities in automated speech recognition. In Judith T. Irvine and Ann Arbor eds. *Proceedings of the National Academy of Sciences* 117/14: 7684–7689. https://doi.org/10.1073/pnas.1915768117

Kortmann, Bernd and Wagner, Susanne. 2005. The Freiburg English Dialect Project and Corpus (FRED). In Bernd Kortmann, Tanja Herrmann, Lukas Pietsch and Susanne Wagner eds. *Volume 1 Agreement, Gender, Relative Clauses.* Berlin: De Gruyter Mouton, 1–20.

Lee, Donghee N., Myiah J. Hutchens, Thomas J. George, Danyell Wilson-Howard, Eric J. Cooks and Janice L. Krieger. 2022. Do they speak like me? Exploring how perceptions of linguistic difference may influence patient perceptions of healthcare providers. *Medical Education Online:* 27/1: 2107470. https://doi.org/10.1080/10872981.2022.2107470

Magnotta, Sierra. 2022. *Analysis of Two Acoustic Models on Forced Alignment of African American English*. Georgia, U.S.: University of Georgia dissertation.

Martin, Joshua L. 2022. *Automatic Speech Recognition Systems, Spoken Corpora, and African American Language: An Examination of Linguistic Bias and Morphosyntactic Features*. Gainesville, Florida: University of Florida dissertation.

Martin, Joshua L. and Kevin Tang. 2020. Understanding racial disparities in automatic speech recognition: The case of habitual 'be'. *Interspeech*: 626–630.

Martin, Joshua L. and Kelly E. Wright. 2022. Bias in automatic speech recognition: The case of African American language. *Applied Linguistics* 44/4: 613–630.

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

McLarty, Jason, Taylor Jones and Christopher Hall. 2019. Corpus-based sociophonetic approaches to postvocalic R-lessness in African American language. *American Speech* 94/1: 91–109.

Meyer, Julien, Laure Dentel and Fanny Meunier. 2013. Speech recognition in natural background noise. *PloS one* 8/11. https://doi.org/10.1371/journal.pone.0079279

Moore, Russell, Andrew Caine, Calbert Graham and Paula Buttery. 2015. Incremental dependency parsing and disfluency detection in spoken learner English. In Pavel Král and Václav Matoušek eds. *Text, Speech, and Dialogue.* New York: Springer International Publishing, 470–479.

Olsen, Rachel M., Michael L. Olsen, Joseph A. Stanley, Margaret E. L. Renwick and William Kretzschmar. 2017. Methods for transcription and forced alignment of a legacy speech corpus. *Proceedings of Meetings on Acoustics,* 1–13. https://doi.org/10.1121/2.0000559

Oregon Department of Transportation Research Section. 2010. *Guide to Transcribing and Summarizing Oral Histories.* https://www.oregon.gov/odot/Programs/ResearchDocuments/guide_to_transcribing_and_summarizing_oral_histories.pdf (accessed 25 June 2023.)

Ott, Myle, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier and Michael Auli. 2019. Fairseq: A fast, extensible toolkit for sequence modelin. In Ammar Waleed, Annie Louis and Nasrin Mostafazadeh eds. *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*. Minneapolis: Association for Computational Linguistics, 48–53.

Pandey, Ayushi, Pamir Gogoi and Kevin Tang. Understanding forced alignment errors in Hindi-English code-mixed speech–a feature analysis. 2020. In *Proceedings of First Workshop on Speech Technologies for Codeswitching in Multilingual Communities,* 13–17. http://festvox.org/cedar/WSTCSMC2020.pdf

Pederson, Lee, Susan Leas McDaniel and Carol M. Adams eds. 1986. *Linguistic Atlas of the Gulf States*. Georgia: University of Georgia Press.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, Jake Vanderplas, Alexandre Passos, David Cournapeau, Matthieu Brucher, Matthieu Perrot and Édouard Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research* 12: 2825–2830.

Pettarin, Alberto. 2017. *Aeneas: Automagically Synchronize Audio and Text*. https://www.readbeyond.it/aeneas/ (accessed 29 June 2023.)

Previlon, Wilermine, Alice Rozet, Jotsna Gowda, Bill Dyer, Kevin Tang and Sarah Moeller. 2024. Leveraging syntactic dependencies in disambiguation: the case of African American English. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*. (Preprint available at https://doi.org/10.31234/osf.io/ph7q8).

Rahman, Jacquelyn. 2008. Middle-class African Americans: Reactions and attitudes toward African American English. *American Speech* 83/ 2: 141–76.

Rohanian, Morteza and Julian Hough. 2021. Best of both worlds: Making high accuracy non-incremental transformer-based disfluency detection incremental. In Chengqing Zong, Fei Xia, Wenjie Li and Roberto Navigli eds. *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics.* Online publication: Association for Computational Linguistics, 3693–3703.

Roller, K. 2015. Towards the 'oral' in oral history: Using historical narratives in linguistics. *Oral History*: 73–84.

Samuel Proctor Oral History Program. 2007. *Style Guide: Guidelines for Transcribing and Editing Oral Histories*. https://ufdc.ufl.edu/IR00002513/00001 (accessed 25 June 2023.)

Samuel Proctor Oral History Program. 2016. *Style Guide: Guidelines for Transcribing and Editing Oral Histories.* https://oral.history.ufl.edu/wp-content/uploads/sites/15/SPOHP-Style-Guide-2016.pdf (accessed 25 June 2023.)

Samuel Proctor Oral History Project. 2020. *Learn to Transcribe Oral History the SPOHP Way*. https://www.youtube.com/watch?v=_aKXmOLQINw (accessed 23 June 2023.)

Samuel Proctor Oral History Project. 2023. *Machen Florida Opportunity Scholars Program (MFOS).* https://oral.history.ufl.edu/projects/machen-florida-opportunity-scholars-program-mfos/ (accessed 27 June 2023.)

Santiago, Harrison, Joshua Martin, Sarah Moeller and Kevin Tang. 2022. Disambiguation of morpho-syntactic features of African American English: The case of habitual be. In Bharathi Raja Chakravarthi, B Bharathi, John P McCrae, Manel Zarrouk, Kalika Bali, Paul Buitelaar eds. *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*. Dublin: Association for Computational Linguistics, 70–75.

Schiel, Florian, Christoph Draxler, Angela Baumann, Tania Ellbogen and Alexander Steffen. 2012. *The Production of Speech Corpora*. München: Open Access Ludwig-Maximilians-Universität München. https://doi.org/10.5282/ubm/epub.13693.

Schiffrin, Deborah. 2002. Mother and friends in a holocaust life story. *Language in Society* 31/3: 309–353.

Sinclair, John. 1991. *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Stolcke, Andreas and Jasha Droppo. 2017. Comparing human and machine errors in conversational speech transcription. *Interspeech*: 137–141.

Strong, Liz, Mary Marshall Clark and Caitlin Bertin-Mahieux. 2018. *Columbia University Oral History Transcription Style Guide*. Columbia: Columbia University Center for Oral History Research. https://incite.columbia.edu/publications-old/2019/3/13/oral-history-transcription-style-guide (accessed 25 June 2023.)

Tagliamonte, Sali A. 2004. Someth[in]'s go[ing] on!: Variable *ing* at ground zero. In Britt-Louise Gunnarsson, Lena Bergström, Gerd Eklund, Staffan Fidell, Lise H. Hansen, Angela Karstadt, Bengt Nordberg, Eva Sundergren and Mats Thelander eds. *Language Variation in Europe: Papers from the Second International Conference on Language Variation in Europe.* Uppsala: Uppsala Universitet, 390–403.

Tang, Kevin. 2015. *Naturalistic Speech Misperception*. London: University College London dissertation.

Tang, Kevin and Ryan Bennett. 2019. Unite and conquer: Bootstrapping forced alignment tools for closely-related minority languages (mayan). In Sasha Calhoun, Paola Escudero, Marija Tabain and Paul Warren eds. *Proceedings of the 19th International Congress of Phonetic Sciences*. Canberra: Australasian Speech Science and Technology Association Inc, 1719−1723-

Tevissen, Yannis, Jérôme Boudy, Gérard Chollet and Frédéric Petitpont. 2023. Towards measuring and scoring speaker diarization fairness. *CoRR* abs/2302.09991. https://doi.org/10.48550/arXiv.2302.09991.

Thomas, Erik R. and Guy Bailey. 2015. Segmental phonology of African American English. In Jennifer Bloomquist, Lisa J. Green and Sonja L. Lanehart eds. *The*

Oxford Handbook of African American Language. Oxford: Oxford University Press, 403–419.

Whalen, Douglas H. and Joyce McDonough. 2015. Taking the laboratory into the field. *Annual Review of Linguistic*s 1/1: 395–415.

Yoon, Sunmoo, Peter Broadwell, Frederick F. Sun, Maria De Planell-Saguer and Nicole Davis. 2023. Application of topic modeling on artificial intelligence studies as a foundation to develop ethical guidelines in African American dementia caregiving. *Studies in Health Technology and Informatics* 305, 541–544.

Yuan, Jiahong and Mark Liberman. Automatic detection of "g-dropping" in American English using forced alignment. 2011. In the *2011 IEEE Workshop on Automatic Speech Recognition and Understanding*. Hawaii: Curran Associates Inc, 490–493. https://doi.org/10.1109/ASRU.2011.6163980

Zayats, Vicky, Trang Tran, Richard Wright, Courtney Mansfield and Mari Ostendorf. 2019. Disfluencies and human speech transcription errors. *Interspeech:* 3088–3092.

Ziems, Caleb, William Held, Jingfeng Yang and Diyi Yang. 2022. Multi-VALUE: A framework for cross-dialectal English NLP. *CoRR* abs/2212.08011. https://doi.org/10.48550/arXiv.2212.08011.

Corresponding author
Kevin Tang
Heinrich-Heine-University Düsseldorf
Department of English and American Studies
Sekretariat Ulrike Kayser
Geb. 23.21.02.102
Universitätsstraße 1
40225 Düsseldorf
Germany
Email: kevin.tang@hhu.de

received: July 2023
accepted: February 2024

APPENDIX A: LIST OF TEXTS BY INTERVIEWEE

	Interviewee	Gender	Age at recording	Date of interview	Place of interview	Tokens
(1)	Alexis Cooper	Female	Teen	09/21/2013	Moorhead, MS	5,231
(2)	Breyanna Hooper	Female	Teen	09/21/2013	Sunflower, MS	4,137
(3)	Cornelius Towns	Male	83	11/25/2012	Bland, FL	3,181
(4)	Darron L Edwards	Male	42	09/23/2011	Ruleville, MS	3,181
(5)	David Faison	Male	84	03/11/2016	Silver Springs, FL	5,184
(6)	Deloris Johnson	Female	65	05/25/2010	Gainesville, FL	8,316
(7)	Diana Bell	Female	60s	06/12/2009	Alachua County	4,711
(8)	Ernest Sneed	Male	Elder	09/06/2016	Alachua County	2,548
(9)	Eugene Martin	Male	70s	06/24/2016	High Springs, FL	5,323
(10)	Jeanette G. Jackson	Female	80s	02/26/2008	Alachua County	3,148
(10)	Doris Marie Perry Ryan	Female	70s	02/26/2008	Alachua County, FL	3,148
(10)	Orien A. Hills	Male	Elder	02/26/2008	Alachua County, FL	3,148
(11)	John Booth	Male	62	01/13/2009	Gainesville, FL	4,969
(12)	John Due	Male	83	06/18/2017	Unknown, MS	3,212
(13)	Shirely Felton	Female	63	06/21/2018	Fort Myers, FL	4,977
(14)	Vendarae Lewis	Female	56	07/27/2012	Bartow, FL	3,225
(15)	Yolanda Veal	Female	26	09/24/2010	Indianola, MS	4,862
(16)	Zacchaeus McEwen	Male	17	09/18/2013	McComb, MS	4,266

APPENDIX B: DATA AVAILABILITY AND AUTHORSHIP CONTRIBUTION STATEMENT

All figures are available on Open Science: http://doi.org/10.17605/OSF.IO/4HGBW. Our training materials for transcribers to become aware of AAL features and AAL in general may be found at https://doi.org/10.17605/OSF.IO/X9WHN.

Sarah Moeller and Kevin Tang contributed equally to this paper, and they served as the senior and corresponding authors. We follow the CRediT taxonomy (https://credit.niso.org/).

Conceptualization: Sarah Moeller and Kevin Tang; **Formal Analysis:** Sarah Moeller, Kevin Tang, Alexis Davis, Wilermine Previlon and Michael Bottini; **Funding acquisition:** Sarah Moeller and Kevin Tang; **Investigation:** Sarah Moeller, Kevin Tang, Alexis Davis, Wilermine Previlon and Michael Bottini; **Methodology:** Sarah Moeller, Kevin Tang, Alexis Davis, Wilermine Previlon and Michael Bottini; **Resources:** Sarah Moeller and Kevin Tang; **Software:** Wilermine Previlon; **Supervision:** Sarah Moeller, Kevin Tang and Alexis Davis; **Visualization:** Alexis Davis and Wilermine Previlon; **Writing original draft:** Sarah Moeller, Kevin Tang, Wilermine Previlon and Alexis Davis; **Writing review and editing:** Sarah Moeller, Kevin Tang, Wilermine Previlon and Alexis Davis.

RiCL Research in Corpus Linguistics

Addressing comparability and retrieval issues in conversation corpora: A case study on the *Spoken British National Corpora* (1994 and 2014), using the past perfect

Nicholas Smith[a] – Cristiano Broccias[b] – Cathleen Waters[c]
University of Leicester[a] / United Kingdom
University of Genoa[b] / Italy
Independent researcher[c]

Abstract – This paper addresses issues in comparison and analysis of conversation corpora. We focus on the demographically-sampled spoken portions of the *British National Corpora* (BNC), representing British English in 1994 and 2014, for the purposes of studying recent language change and sociolinguistic variation. Issues of comparability and representativeness of the two BNCs have been raised before (see Love 2020), with several measures taken to ensure backwards compatibility of the Spoken BNC2014 with its 1994 counterpart. However, we believe further considerations and solutions merit attention, relating to sampling, transcription, annotation, and corpus querying. The BNClab subcorpus (Brezina *et al.* 2018a), a sociolinguistic judgment sample derived from the parent BNCs, provides a very promising basis for analysis, although arguably its mixed geographical representativeness affects cross-time comparability. To address this, we make some proposals for modifying the BNClab subcorpus to improve comparability. Then, we use the modified sample to address issues in retrieval and quantification of grammatical constructions in the spoken BNCs, namely a) determining an appropriate frequency metric, b) retrieving a comprehensive but manageable set of examples from 'messy' spoken data, and c) handling transcription inaccuracies. Finally, we discuss the case study findings and wider methodological implications for users of these corpora.

Keywords – spoken BNCs; corpus comparability and representativeness; grammatical retrieval; precision and recall; past perfect

1. INTRODUCTION[1]

This paper addresses two related kinds of challenge: a) comparability issues in spoken language corpora, and b) issues in retrieval and analysis of grammatical constructions in such corpora. While these issues are discussed in relation to the conversational

[1] We thank the anonymous reviewers and the editor for their comments; and Sebastian Hoffmann, Vaclav Brezina, Agneta Svalberg, and Julie Norton for useful discussions.

components of the *British National Corpus* (BNC)[2] from 1994 and 2014, they could potentially apply to any pair or set of corpora separated in time.

Both iterations of the BNC offer countless opportunities for anyone interested in the English language (researchers, teachers, students, or laypeople) to explore patterns of authentic British English speech, variation across registers and speakers, and changes over time. Each corpus has been planned and documented in detail (see Burnard 2007, for BNC1994, and Love *et al.* 2017 or Brezina *et al.* 2021, for BNC2014). The demographically sampled conversation components of the corpora, widely known as BNC1994DS and BNC2014S, are among the largest collections of naturally-occurring spoken discourse currently available for free public use, and they support investigations into associations between language use and speakers' social characteristics.[3] Studies exploiting the affordances of BNC1994DS include Anderwald (2002) and Rühlemann (2007), while comparisons between BNC1994DS and BNC2014S appear in McEnery *et al.* (2017) and Brezina *et al.* (2018b).

While both conversational BNCs offer impressively large quantities of material from across the social spectrum (e.g., by region, gender, age, occupation), their compilers have readily admitted that the representation of these groups is somewhat uneven (Burnard 2007; Love *et al.* 2017; Love 2020). This is understandable given the unprecedented nationwide scale of each corpus, yet the limited resources to build them, the difficulty of implementing a strict sampling procedure at the outset, and the prioritization of each project to represent users and uses of British English at their respective time, and to adopt the latest standards of corpus categorization (cf. Crowdy 1993 on BNC1994DS and Love *et al.* 2017 on BNC2014S). However, anyone using these corpora needs to be aware of their limitations in terms of comparability and representativeness. We are not the first to raise this point (see Axelsson 2018; Love 2020), but we expand on these comparability issues and highlight new ones, alongside proposals for mitigating them.

Corpus comparability is, on the face of it, much simpler than representativeness, but no less important. As Gablasova *et al.* (2017: 137) state,[4] it refers to "the degree to which two corpora are similar… [in] represent[ing] different genres of a language or

[2] http://www.natcorp.ox.ac.uk/
[3] The Spoken BNC1994 also contains a 5–million-word non-conversational, context-governed component (Burnard 2007), which we leave aside for comparability reasons.
[4] Gablasova *et al.* (2017) treat 'genre' as equivalent to what other studies, including ours, call 'register'.

speakers." The extent to which we can meaningfully interpret differences in results between corpora clearly hinges on an equal footing in these parameters. Comparability is arguably a more acute issue for spoken than written corpora since, in general, spoken language changes more rapidly than written language (Leech *et al.* 2009). As to representativeness, Biber's (1993: 243) much-cited definition —"the extent to which a sample includes the full range of variability within a population" — is non-controversial, but open to different implementations. It can refer, for example, to situationally-defined registers (e.g., job interviews or job advertisements), speaker demographics, or the distributions of linguistic features (see section 2.1). To instill more confidence in any reported results, Egbert *et al.* (2022) exhort corpus compilers and users to be more explicit in estimating the extent to which their corpus represents what it purports to represent. Moreover, concerns for representativeness and comparability can easily clash. For instance, in striving to make a corpus representative of the registers or demographics of its time or in infusing it with state-of-the-art design features and standards, the risk grows of limiting opportunities for direct comparison with a corpus from another point in time.

Fortunately, the compilers of BNC2014S have taken measures to support backward compatibility with BNC1994DS, such as issuing a list of mappings between the respective social class categories and age categories (see section 2.1). Moreover, the size of each spoken BNC and the detailed metadata provided on speaker and register characteristics afford innumerable ways to subsample and suit different research purposes (see Love 2020 and Brezina *et al.* 2021), including boosting comparability. In this paper, we evaluate one such subsample, the BNClab subcorpus (Brezina *et al.* 2018a), and present some proposals for enhancing diachronic comparability within it.

Two further issues affecting comparisons between the two spoken BNCs, which (to our knowledge have not been discussed, are differences in a) transcription quality and, potentially at least, b) grammatical annotation. With alignment of the transcriptions in BNC1994DS to its speech recordings now available (Coleman *et al.* 2011), a surprisingly large number of transcription errors in this corpus can be found. Quality control measures in BNC2014S inspire greater confidence in its transcription accuracy, despite its audio files being publicly unavailable (see section 2.2). One therefore has to decide whether discarding false positives from BNC1994DS will undermine comparisons with BNC2014S. In this paper, we present a strategy that mitigates the impact of this problem. Regarding grammatical annotation, although both corpora use the same part-of-speech

(POS) tagging software, the final output may differ in the delicacy of the tags displayed, depending on the version being used.

To illustrate challenges in spoken corpus comparability and analysis and ways to overcome them, we provide a case study on the past perfect, as in (1):

(1) That's the first time you *'d met* her? (BNC2014 S6HP:S0303)[5]

The past perfect may seem an odd choice for a case study, as it is relatively uncommon in the tense and aspect system of English, as well as particularly infrequent in conversation (Mindt 2000). Yet recent studies report that the past perfect is undergoing a dramatic change, with significant declines in both spoken (Bowie *et al.* 2013; Smith and Waters 2019) and written English (Yao and Collins 2013). However, currently, there is no specific evidence of change in the register of conversation. Bowie *et al.*'s (2013) study examines the registers of the *Diachronic Corpus of Present-Day Spoken English* (DCPSE)[6] collectively rather than individually, and is limited to the late twentieth century. Using a corpus of biographical interviews from the popular BBC Radio programme *Desert Island Discs*, Smith and Waters (2019) find a small but significant decline of the past perfect between the 1980s and the early 2000s. Moreover, they note that the construction is socially stratified, with older and more highly educated speakers being more conservative in its use. An investigation of the past perfect in the conversational BNCs also illustrates typical challenges in retrieval and quantification of grammatical constructions in spoken discourse, notably:

a) Determining the unit of frequency measurement for the target construction and accounting for competitor constructions (see section 4.1).

b) Designing corpus queries for acceptable recall and precision of the target construction in sometimes 'messy' spoken data (see section 2.3).

c) Filtering out superficially similar vernacular forms. For instance, in informal British English, *have got* is commonly used with stative meaning, as in *she'd got a family* (BNC1994: PS25A), and this is easily mistaken for a past perfect (see section 4.4).

d) Addressing errors in corpus transcription (see section 2.2).

[5] Corpus references are to the parent BNC filename and speaker identifier.
[6] https://www.ucl.ac.uk/english-usage/projects/dcpse/

Our methodology addresses general areas of change and social variation in the past perfect. The research questions underpinning the case study are:

1. Has the frequency of the past perfect changed in recent everyday conversational British English?
2. What patterns of sociolinguistic change and variation are evident in recent British conversational use of the past perfect?

By offering an up-to-date picture of the frequency of the construction in British English, our research also has potential implications in applied linguistics, particularly in English language teaching (ELT). If, for example, our results support earlier studies on spoken British English by finding a substantial decline in conversational use, there is arguably a case for reducing attention to the past perfect in teaching materials and curriculum development. Conversely, if the past perfect is found to be dramatically expanding in contemporary use, it would seem worthwhile to share this discovery with ELT publishers, teachers, and learners (cf. Curry *et al.* 2022). Learners might also benefit from awareness-raising of the prevalence of alternatives to the past perfect (see section 4.1) and how to locate them in spoken corpora.

The paper is organized as follows. We first expand on key concepts, including representativeness, comparability of corpora, precision, and recall (section 2). We then describe how we negotiated the challenges summarized in points i) to iv) above, in pursuit of a more level playing field to compare the two conversational BNCs (sections 3 and 4). Finally, we discuss preliminary corpus findings on the past perfect, including their implications and limitations, and comparison to previous studies (section 5).

2. Representativeness, comparability, precision, and recall

In this section, we review theoretical and practical aspects of representativeness and comparability, including previous attempts to address them in the two conversational BNCs. We then consider concepts relevant to retrieval of linguistic features from corpora, namely precision and recall, and technological means to boost them.

2.1. Representativeness and comparability

As suggested above, representativeness and comparability are key concepts that need to be considered in any cross-time comparison of corpora, not just the spoken BNCs. In corpus linguistics, where written data has generally been prioritized, representativeness typically refers to the extent to which the corpus reflects *situational* variation (e.g., in communicative purpose or level of interactivity) within and/or across its component text registers (Biber 1993). It tends to be only in the register of conversation that corpus linguistic studies shift focus to *demographic* representativeness (Smith and Waters 2019). In sociolinguistics, by contrast, demographic representation is a prime concern, with sampling of speakers designed to reflect the social diversity in the community investigated (Sankoff 2005), but typically on a local rather than a national scale. Thus, both disciplines use a form of stratified sampling, one focused on texts as the sampling units, the other on speakers. One further kind of representativeness to note is *linguistic* (or distributional) representativeness, that is, the extent to which the corpus "includes the range of linguistic distributions in the population" of texts or speakers (Biber 1993: 243). Egbert *et al.* (2022) lament that many corpus linguistic projects fail to evaluate the representativeness of their corpus relative to their research goals. At the same time, they argue that representativeness is a matter of degree rather than an all-or-nothing construct, and that full representativeness is an idealized target and unattainable in practice (Egbert *et al.* 2022: 12).

Comparability is less explicitly discussed than representativeness in either the sociolinguistic or the corpus literature. Gablasova *et al.* (2017) identify comparability as a major issue in corpus-based second language acquisition research, where direct comparisons have been routinely drawn between the language use of L2 and L1 speakers but without paying attention to potentially confounding factors, such as type of elicitation task and L1 speakers' language proficiency. The tension between representativeness and comparability has arguably received more attention in diachronic corpus studies. Leech and Smith (2005) describe the challenge of extending the design model of the Brown and the LOB corpora (sampling date: 1961; Hofland *et al.* 1999) back to the 1930s and earlier, when genres such as science fiction and academic subdisciplines, e.g., sociology, were far less established. Baker (2023) encounters the reverse challenge in extending the Brown corpus model to British English published in 2021 and, to optimize comparability, he excludes new genres such as horror fiction, which did not exist in the 1960s.

In diachronic corpus studies, a balance also needs to be struck between comparability and contemporaneity of standards. Advances in computer hardware and corpus software, and improved standards of metadata, can lead the creators of a newer corpus to depart from the best practice of an earlier corpus. Annotation standards, e.g., the kinds of distinctions used in part-of-speech (POS) tagging, can also vary from one corpus to another, sometimes even when the same annotation software is used (see section 2.3 and section 4.2).

The representativeness and comparability of BNC1994DS and BNC2014S are discussed in Love (2020: 186–189). He notes, for example, that neither corpus uses strict stratified sampling. In BNC1994DS, only the speakers with recording responsibility were sampled in advance (by a random method), but the other speakers, like all speakers in BNC2014S, were selected opportunistically. Constraints of budget and time made it impossible to obtain balanced representation of social groups. For example, male speakers in BNC1994DS are over-represented in relation to females and to their proportions in the UK population as a whole. Conversely, females in BNC2014S are over-represented. Regarding age, in BNC1994DS speakers aged 25–59 are over-represented relative to the UK population, while in BNC2014S, those aged 19–29 proliferate. In both corpora, speakers from England are represented far better than those from Scotland, Wales, and Northern Ireland. Clearly, there are significant differences in the composition of the two spoken BNCs, and if due attention is not given to these differences, there is a risk of drawing naive conclusions from cross-time comparisons.

Love *et al.* (2017) describe measures taken to support backwards compatibility, and therefore comparability, of metadata categories between the newer and older corpus. For example, they provide a mapping list between the more fine-grained age bands of BNC2014S into those of BNC1994DS. Similarly, they translate the nine socioeconomic class categories (NS-SEC) in BNC2014S into the four social grade categories used in BNC1994DS. In terms of annotation, however, the corpora differ in that BNC2014S was tagged using a more fine-grained set of POS-tags than BNC1994DS. Further differences in POS-tagging are described in section 4.2.

2.2. Previous comparative studies of the conversational BNCs

In two edited collections of papers on the conversational BNCs (McEnery *et al.* 2017 and Brezina *et al.* 2018b), several contributors discuss comparability issues between BNC1994DS and an early sample release of BNC2014S, including age and class regroupings. Axelsson (2018), for instance, highlights a possibly greater awareness among speakers in BNC2014S of being recorded (because of more stringent ethical requirements for prior consent), which may have led to the conversations acquiring a more focused and formal character. If this is indeed the case, it is potentially problematic for diachronic comparison, yet difficult to see how it can be overcome.

The BNClab subcorpus (Brezina *et al.* 2018a) seeks to enhance comparability by deriving a judgment sample from the two parent BNCs. Judgment sampling involves, as Schilling-Estes (2007: 169) states,

> using one's judgment to decide in advance what types of speakers to include in the study and then obtaining data from a certain number of each type.

Using the BNClab subcorpus, Reichelt (2021) uncovers changing patterns in the pragmatic markers *kind of* and *sort of* across time and social groups. Given its potential for investigating spoken language change and variation in relatively controlled conditions, we evaluate the comparability and representativeness of the BNClab subcorpus (in section 3.1), and a modified version of it (in section 3.2) used in our own study.

To our knowledge, no studies comparing the two BNCs have yet addressed the issue of transcription quality in the 1994 corpus. The issue is more pervasive and concerning than the several instances of incorrect speaker assignment noticed by Axelsson (2018). It includes numerous cases where the content roughly matches the audio but the linguistic forms are incorrect, as illustrated in (2), and cases where neither content nor form match the recording, as shown in (3). Such anomalies have come to light following a project to align the BNC1994DS transcriptions with the original sound files (Coleman *et al.* 2011), and the subsequent implementation of audio playback of concordance hits in the *BNCweb* tool (Hoffmann and Arndt-Lappe 2021).

(2) Then I **bought**, yeah, I said I feel as if I've gone deaf [Correction: Then I **thought**…]. (BNC1994 KB2:PS01U).

(3) **but, it was a pity** he was able to speak on the telephone [Correction: **well considering** he was able...]. (BNC1994 KBW:PS087).

User access to recordings of BNC2014S is not yet possible. However, there are good reasons to believe that transcription of this corpus was done far more carefully. Each transcription went through careful rounds of checking (Love 2020) and transcribers were thoroughly trained on the transcription protocols and formally registered their level of confidence in identifying the speaker of a given utterance. Likewise, the recording devices used in the mid-2010s (smartphones) were far superior to the devices used in BNC1994DS.[7]

2.3. Precision, recall, POS-tagging, and retrieval software

When querying a corpus for a linguistic feature, an important consideration is finding the right balance between recall and precision. Recall is a measure (expressed as a percentage) of the extent to which a query retrieves all valid instances of a target item in the data. In practice, recall is difficult to quantify since it requires a fully hand-edited dataset. Precision, which is also expressed as a percentage, refers to the proportion of retrieved instances that are actually valid (see Jucker *et al.* 2008). In corpus studies, it is generally agreed that low precision is more tolerable than low recall, since automated results are likely to be hand-checked, and having a near full set of examples is key to a thorough analysis (see Hoffmann *et al.* 2008). While there is no consensus as to what constitutes acceptable precision thresholds, Jucker *et al.* (2008: 277) suggest that

> precision errors are not a serious problem, until the number of hits exceeds what is possible to scan manually, and until precision falls below a certain threshold: one tends to overlook positive examples if precision is much lower than 1%.

An excessive number of hits is an important issue in our study, not for the past perfect, but for the far more prevalent past non-perfect (e.g., *took*) with which it competes (see section 4.2.3). As for precision, we typically boost precision scores well beyond one percent (even in corpora of spontaneous speech) by using a grammatically annotated version of each corpus and sophisticated corpus query tools to exploit the annotations. In our study of the past perfect in the spoken BNCs, using this combination of tools obviates the need to manually sift through 45,087 cases of *had/'d* for instances containing a trailing participle.

[7] BNC2014S also allows users to investigate individual transcriber consistency by identifying them in the metadata.

Nevertheless, two factors need to be acknowledged as affecting recall under these conditions. The first one is that automated POS-taggers make errors. While error rates are generally reassuringly low, at around 3–4 percent, they will be higher for multiply ambiguous words (e.g., *left* is a noun, an adjective, a past tense verb, or a past participle, depending on context). The second consideration is that phrasal constructions like the past perfect can be used discontinuously (Trask 1993), that is, between the auxiliary and the participle, one or more words may intervene (e.g., *she had erm already gone*). It is therefore imperative to work out a strategy for optimizing recall in discontinuous uses of a construction, particularly in the unpredictable environment of spontaneous conversation. We address this issue in section 4.2.1.

3. Obtaining a sociolinguistically-balanced dataset from the BNCs

3.1. Our starting point: The BNClab subcorpus

Developed at Lancaster University, the BNClab subcorpus samples 250 speakers from BNC1994DS and 250 speakers from BNC2014S. Unlike most sociolinguistic judgment samples, the assignment of speakers to groups for BNClab was performed *post-hoc*, after the BNCs themselves had been created, which could make the judgment harder than when selecting during data collection. The subcorpus covers all nations of the UK, an unusually large area for a judgment sample. It has near-equal gender balance (126 females and 124 males in each of 1994 and 2014), and good representation across age cohorts (Brezina *et al.* 2018a; Reichelt 2021), as also shown in the Appendix. The age groups are sufficiently fine-grained to allow studies of apparent-time change within each period.

The basis for determining social class in the BNClab subcorpus (as in the full BNCs) is the speaker's occupation. This is in line with typical sociolinguistic research (Milroy and Gordon 2003), although classifying occupations is notoriously complex, and a different approach is taken in each parent corpus. In BNC1994DS a social grade scheme is used, based on categories in the UK's *National Readership Survey* (see Love *et al.* 2017). BNC2014S, meanwhile, uses the UK government's official *National Statistics Socio-economic Classification* (NS-SEC) scheme (Love *et al.* 2017), which broadly relates to the type of contract the occupation typically involves (Atkinson 2015).[8] Love

[8] Notably, jobs involving higher specificity (and scarcity) of skills and lower ease of monitoring are found at the top of the scale (Atkinson 2015).

et al. (2017: 332) helpfully provide a mapping table between the two classification schemes. In BNClab, speakers with occupations rated as NS-SEC classes 1 to 4, or social grade AB to C1, are assigned to middle class, while speakers in NS-SEC classes 5 to 8, or social grades C2 to E, are treated as working class, although some manual adjustments were made to improve consistency.

As can be noticed in the Appendix, the BNClab subcorpus has a class imbalance, with the 2014 component markedly under-representing working-class speakers and over-representing middle-class speakers relative to the 1994 subcorpus. To some extent, these differences reflect the changing nature of British society. By the mid-2010s, an increasing portion of the UK population was university-educated and engaged in higher-status professional occupations than in the 1990s. Even so, the middle-class numbers in BNClab exaggerate the scale of this upward mobility.

Likewise, an arguably important social variable for representativeness that is not included in the BNClab subcorpus is ethnicity. An increase in ethnic diversity is another major area of change in UK society.[9] However, the lack of data about ethnicity in BNC1994DS makes the omission from BNC2014S understandable.

Finally, the creators of the BNClab subcorpus excluded all speakers who produce fewer than 1,000 words. This is the minimum that Biber (1993) reports as sufficient to profile most grammatical features in a given text, suggesting adequate distributional representativeness (see section 3.2).

3.2. Modifications to the BNClab subcorpus

While the BNClab subcorpus provides a very promising platform for sociolinguistic inquiry, the fairly balanced numbers of speakers for single social variables (e.g., gender) sometimes mask sizable differences at the granular level, for instance, at the intersection of categories such as region and social class. In the 2014 data, for example, just one of the eight speakers from Scotland is categorized as working class, with six being middle class, and one unknown. Wales has nine middle-class but just two working-class speakers from 2014. By including all four component countries of the UK, plus Ireland, the chances of getting balanced representation at the granular level are greatly reduced. This can be

[9] https://www.ons.gov.uk/peoplepopulationandcommunity/culturalidentity/ethnicity/articles/2011censusanalysisethnicityandreligionofthenonukbornpopulationinenglandandwales/2015-06-1

problematic when investigating linguistic features known or suspected to exhibit marked social stratification, such as the past perfect. To mitigate this and other comparability issues and to balance the three social dimensions of gender, class, and age, we adapted the BNClab subcorpus into a new modified version, hereafter BNClab-M. Six main changes were made.

Firstly, we limited the speakers to those from England. Naturally, excluding Scotland, Wales, and Northern Ireland makes this revised sample unrepresentative of the UK as a whole, but the more focused geographical selection concentrating on the largest national population of speakers gives us more scope to balance the social variables. Unfortunately, this modified sample is no more capable than the original BNClab subcorpus of evenly reflecting regional variation in England (see Beal 2010), although wherever possible we included speakers across the five English regions recognized by BNClab (North, Midlands, Southeast, London, and Southwest).

Secondly, only speakers categorized in BNClab as either working class or middle class were considered. Those with uncategorized social class were discarded. Retired people were omitted because they constitute a socially opaque group, with little in common beyond their senior age. In view of the importance of age for stratification of the past perfect (see section 5), it would be desirable to find a way to incorporate retirees at a future point, but ideally incorporating their former occupations. Students were also omitted because they are a similarly problematic group for social class assignment since, in most cases, they are not in the labor market. Similarly, trainees (e.g., trainee engineers, nurses, and typists), whose incorporation into the labor market is unknown, were discarded.

Thirdly, regarding age, we excluded children (i.e., under-18s). The number of children in the BNClab subcorpus is patchy across regions (e.g., just two children from northern England). As for adult speakers, these were categorized into two age cohorts, namely under-45s and over-45s (the latter including 45-year-olds). This is based on observations in the sociolinguistic literature, that "the speech of middle-aged adults tends to be highly conservative, often more conservative even than that of older speakers" (Milroy and Gordon 2003: 39).

Fourthly, we reclassified cases that we considered to be errors and removed speakers whose classifications seemed uncertain. To do this, we drew on various sources of information, notably the BNC1994 and BNC2014 header files (showing speaker

occupations and relationships between speakers), and a set of social grade reclassifications of BNC1994 prepared at Oxford University.[10] We moved some speakers from working class to middle class where we judged their occupation to be similar to other, well-established, middle-class roles. Examples include a chartered engineer (speaker PS1BT) and a consultant engineer (S0179). Less clear-cut but still arguably middle class are clerks, administrators, and other office workers, who in the BNC metadata tend to be assigned social grade C1 or higher, that is, (lower) middle rather than working class. Our exclusions included those in an occupation typically placed as working class but who hold a degree (e.g., a graduate chef, speaker S0603), and those with a close family member assigned to a different class (e.g., a childminder, speaker PS14B, originally recorded as working-class but whose husband is a teacher).[11] However, it must be acknowledged that it is difficult to be fully consistent across 1994 and 2014 in applying such exclusions, since the earlier BNC did not record speakers' educational level.

Fifthly, in a few cases, information from the transcription files themselves helped inform a decision on inclusion and categorization of a speaker. For example, in BNC2014S, speaker S0463 is listed as a taxi driver, with social grade C1 and NS-SEC class 4, but as middle class in the BNClab documentation. In the transcription, the speaker refers to his previous work in finance, casting sufficient doubt on his class designation for him to be excluded.

Sixthly, we targeted at least five speakers for each combined set of social characteristics, as this is a common minimum target in sociolinguistic studies (Horvath 2013: 12), although we did not always manage to reach this. For cells in short supply, we turned to the full versions of BNC1994DS and BNC2014S, specifying the relevant demographic criteria to locate nine additional speakers.

The composition of the BNClab-M sample is detailed in Table 1.

[10] We gratefully acknowledge Katie Henley's work in this area. http://www.phon.ox.ac.uk/files/docs/SpokenBNCoccupationsubgroups.xlsx

[11] Both speakers are assigned social grade AB in the BNC metadata.

			1994		2014	
Gender	**Class**	**Age**	**Speakers**	**Words**	**Speakers**	**Words**
Female	Working class	Under 45	11	39,424	5	14,852
		Over 45	4	48,052	7	49,896
	Middle class	Under 45	16	69,150	24	166,296
		Over 45	4	15,159	17	141,724
Male	Working class	Under 45	12	25,943	6	12,526
		Over 45	4	10,032	5	52,381
	Middle class	Under 45	13	37,881	21	136,705
		Over 45	10	27,262	15	121,092
Total			**74**	**272,903**	**100**	**695,472**

Table 1: Speaker numbers and word counts in the BNClab-M sample

Reflecting on the comparability and representativeness of the BNClab-M subcorpus, we are aware that we have made significant compromises to the latter in order to boost the former. Our attempts to rebalance the social group sizes has yielded some success, bringing us closer to our target minimum of five speakers in each cell. The numbers of working-class speakers are now more balanced across the periods (31 in 1994, 23 in 2014), although middle-class speakers are still significantly over-represented in 2014 versus 1994 (77 and 43 respectively). We still have near-parity of females and males (88 and 86 respectively). Age groups are reasonably balanced, the biggest discrepancy being the relatively low figure of 22 over-45s in 1994, versus 44 in 2014.[12] However, our binary classification of age, as below or above 45, hinders analysis of change in apparent time. Overall, speaker numbers are somewhat low for analyzing individual variation within social factor groups (cf. Brezina and Meyerhoff 2014). Moreover, as Sönning and Krug (2022) observe, marked differences in word count between individual speakers may skew feature frequencies by social group. The latter issue was mitigated by retaining the minimum threshold from BNClab of 1,000 words per speaker, although we did not apply an upper limit. These characteristics need to be kept in mind when reviewing the results.[13].

[12] Average ages in these groups are comparable: under-45s' mean age is 35 in 1994 and 32 in 2014; for over-45s, the corresponding means are 51 and 55.

[13] The list of speakers in BNClab-M, and their characteristics, is openly available at https://doi.org/10.25392/leicester.data.25594368.

4. GAUGING THE FREQUENCY OF THE PAST PERFECT

Accurately establishing the frequency of the past perfect in the BNClab-M sample involves several steps, notably, determining an appropriate unit of measurement, setting an effective search strategy to retrieve occurrences of the past perfect (and any relevant competing constructions) from the corpus, and manually correcting the results. We describe each of these steps in turn.

4.1. Frequency and competitors of the past perfect

In corpus studies and variationist sociolinguistic studies, the choice is whether to relativize the raw number of occurrences of a linguistic feature to:

a) the corpus size (in words), i.e., normalized frequency: for example, past perfect instances per million words; or

b) the superordinate category the construction belongs to: for the past perfect, this would be the total number of finite verb phrases; or

c) the set of variants (choices) available for conveying the same or a similar discursive function (e.g., past perfect and other expressions of past time). In variationist sociolinguistics, this set of choices is called the 'linguistic variable' (Tagliamonte 2006).

For several reasons, the third type of measurement is preferable. As Bowie *et al.* (2013) and Smith and Waters (2019) point out, it is the measure that reflects the opportunity of occurrence of the past perfect most accurately. While the first metric is relatively easy to compute, since only the past perfect needs to be counted, it is clear that not every word in the corpus provides an opportunity for the construction to occur (Ball 1994). Also, the past perfect is a multi-word (rather than a single-word) construction, and makes normalized frequencies problematic. Using the superordinate category (finite verb phrases) reduces these problems by narrowing the field to more plausible contexts but misses the fact that the past perfect is restricted to past time. The third metric addresses this problem. It also handles the problem of transcription errors in BNC1994S better (see section 2.2), since any losses of examples due to faulty transcription should affect the past perfect and its competitors equally. At the same time, it must be acknowledged that circumscribing the variable context is far from straightforward, particularly at the level

of syntax (see Lavandera 1978). It entails reviewing the functions and uses of the construction in question and its putative competitors, as well as determining their degree of functional equivalence (Tagliamonte 2006).

The basic function of the past perfect is to express the anteriority of a past situation to a reference time earlier in the past, the latter being mentioned explicitly —as in (4)— or recoverable from the context —as in (5)— (cf. Declerck 2006). Example (4) illustrates that the verb may additionally be marked for progressive aspect.

(4) By the time I arrived, everyone else *had* already *left*. (Depraetere and Langford 2019: 198)

(5) Jane got that job she interviewed for. – I*'d been wondering* about that. (Depraetere and Langford 2019: 198)

In certain conditions it is possible to replace a past perfect with a past non-perfect (that is, a past simple or a past progressive), provided that the anteriority relationship can still be inferred, as illustrated in (6).

(6) After we *finished/had finished* the meeting, we all went out for a drink. (Depraetere and Langford 2019: 198)

However, the past perfect is by no means always substitutable by the non-perfect, and because the former has the specialized meaning of anteriority built in, substitution in the opposite direction is far less feasible. Ideally, we would test the acceptability of replacing each corpus example with a non-perfect but, given the many thousands of past non-perfects in BNClab-M, this would be prohibitively time-consuming. As in Bowie *et al.* (2013), we have pragmatically opted for a looser notion of the linguistic variable than in typical variationist studies, namely all past-marked verbs. Another pragmatic decision was to exclude the present perfect from this set of choices, as shown in (7). In British English, the present perfect tends not to occur in narrative past situations. The example in (8), with a definite time specifier, is a rare exception.

(7) I*'ve given* up smoking.

(8) (…) and then Saturday I*'ve put* that one up again. (BNC2014 SF8D:S0152)

However, in the specific context of unreal past conditionals introduced by *if* or *wish*, we do need to account for non-standard perfects, for instance, the double perfect, as in (9), which is alleged to be spreading in British English (Huddleston and Pullum 2002: 151).

(9) (…) if he *had*n't *have left* our command she was gonna make a formal complaint. (BNC2014 SVD6:S0256)

Other variants of the non-standard perfect include *would* or the elided form *'d*, as in *if he wouldn't have left*, *if he'd have left*. A survey by Ishihara (2003) suggests these forms are increasingly acceptable in colloquial American English.

4.2. Retrieving a comprehensive, comparable, and manageable set of examples

We considered three corpus tools, *BNClab* (Brezina *et al.* 2018a), *CQPweb* (Hardie 2012), and *BNCweb* (Hoffmann *et al.* 2008). The latter is very similar to *CQPweb* but hardwired to BNC1994. Each of these tools permits queries based on POS-tags, either individually or in sequence. The source of those tags in each case is Lancaster's CLAWS4 automatic tagger (Garside and Smith 1997), which in spoken texts has an estimated precision rate of 97 per cent and a recall rate of 98.8 per cent (Leech and Smith 2000). However, there are some differences in the tagging implementation and search software functionality that might affect equivalent retrieval of linguistic features from the spoken BNCs. The differences are summarized in Table 2.

	BNC1994DS in a) *CQPweb* b) *BNCweb*	BNC2014S in *CQPweb*	Subcorpora of BNC1994DS, BNC2014S in *BNClab* tool
Corpus scope	Full corpus or customized subcorpora	Full corpus or customized subcorpora	BNClab subcorpus only
CLAWS tagset	C5	C6	C6
Ambiguity tags	Yes	No	No
CLAWS Spoken mode	Yes	Yes	No
Template Tagger used	Yes	No	No
Flexibility of queries	High	High	More limited
Audio playback	a) No b) Yes	No	No

Table 2: Tagging implementation across the spoken BNCs and retrieval tools

The POS-tags in BNC1994DS are from the CLAWS C5 tagset, whereas those in BNC2014S are from the more granular CLAWS C6 tagset. Personal pronouns, for instance, are represented in C5 by just one tag (PNP), whereas in C6 they have ten tags, depending on person, number, and case (e.g., PPIS2: first person plural, nominative).[14] BNC1994DS also includes ambiguity tags (Burnard 2007), which signal where the probabilities of two competing tags were estimated by CLAWS to be too close to call.

[14] A mapping list between C6 and C5 is available at https://ucrel.lancs.ac.uk/claws/mapC7toC5.txt.

The ambiguity tag VVD-VVN, for example, is non-committal about whether a given word (e.g., *looked*) is past tense (VVD) or a past participle (VVN), although the order indicates that VVD is more probable. Moreover, both BNC1994DS and BNC2014S in *BNCweb*/*CQPweb* were tagged with CLAWS run in spoken mode. This means that the disambiguation of POS-tags was improved by the use of training data extracted from previous, hand-corrected spoken corpora. In the case of BNC1994DS, though not BNC2014S, a supplementary software named *Template Tagger* (Smith 1997) was used, affording marginal improvements in tagging accuracy.

As for retrieval capabilities, *CQPweb* supports queries with flexible pattern-matching, including optional and repeatable elements. At the time of writing, the *BNClab* tool has less advanced search functionality than *CQPweb*/*BNCweb*, but does support queries based on POS. Finally, only *BNCweb* currently supports audio playback of concordance hits.[15]

Given these circumstances, and the need to optimize comparability of the two spoken BNCs, we used the *BNClab* tool with its more consistent POS-tagging for the retrieval of less complex structures in the BNClab-M sample – that is, adjacent (i.e., non-discontinuous) past perfects (with *had/'d* immediately followed by a past participle), as well as past non-perfects (e.g., *took*, or *was* in *was sleeping*). Meanwhile, for the more complex types of retrieval (i.e., discontinuous past perfects and non-standard perfects), we set up the BNClab-M subcorpus in *BNCweb* and *CQPweb*.[16] *BNCweb* additionally allowed us to verify the transcription of most examples from 1994 by listening to the audio recordings. Thus, a combination of tools helped us to optimize the recall of these structures, as we detail below.

4.2.1. Non-discontinuous past perfects

A simple *BNClab* query sufficed for retrieving straightforward past perfects, where the auxiliary (*had*/*'d*) and past participle (tagged V*N) are adjacent, as illustrated in (10).

[15] Space prevents exhaustive coverage of BNC-compatible retrieval tools. However, *Lancsbox X* (Brezina and Platt 2024) now offers a promising alternative by tagging BNC1994DS in the same (C6) POS-tags as BNC2014S. The issue remains of inability to play back audio to verify transcriptions.

[16] More precisely, we made a 1994 BNClab-M subcorpus in the *BNCweb* area of the Lancaster server, and a 2014 subcorpus in the *CQPweb* area. To maximize recall, any differences in results noticed between the tools were added to the pool of hits.

(10) VHD V*N

One difficulty in retrieving the past perfect in speech is elliptical uses, where the trailing past participle is understood but does not appear, as shown in (11).

(11) Had you done it? – Yes, I *had* (invented example).

We could not devise a query to find such cases, and so none are included in our results. For consistency, even though our query for past non-perfects picked up elliptical cases (e.g., *yes, I did*), we discarded them.

4.2.2. Discontinuous past perfects

Since we did not know all the forms of intervening material in spoken past perfects in advance, our retrieval strategy had two heuristic steps: a) determine the maximum interval between the auxiliary and the participle, and b) list all POS sequences that appear within that interval. For the first step, by repeated experiments, we found that queries containing *had/'d* and a participle separated by more than five words (and up to ten words) yielded no valid cases of the past perfect in either period.[17] Note that the query in the 1994 data allowed for ambiguity tags in either order, VVN-VVD (i.e., ambiguous, but past participle more likely) and VVD-VVN (ambiguous, but past tense verb more likely). In the data retrieved, the participle tended to be part of a separate verb phrase from the one containing *had/'d*, as exemplified in (12)

(12) (…) if they **had** an accident <pause> the people would get **killed.** (BNC1994 KCT: PS0FP)

For the second step, we ran a pair of queries in *BNCweb*/*CQPweb* (on the 1994 and 2014 parts, respectively, of the BNClab-M sample) with a five-word maximum interval between auxiliary and participle and used the *Frequency Breakdown* tool to list the types of intervening POS-tag sequence that occur. The top ten POS-sequences in each period are shown in Table 3. Note that the POS-tags listed under 1994 are the simpler C5 tags, while those under 2014 are the finer-grained C6 tags. The only ambiguity tags under 1994 have tag VVN listed first (i.e., ambiguous, but past participle deemed more likely). This

[17] The queries we ran to check this were:
a) for BNC1994: [pos="VHD"] [pos!="PUN|TO0|V.*N"]{6,10} [pos="V[BDHV]N.*|.*V[BDHV]N"]
b) for BNC2014: [pos="VHD"] [pos!="PUN|TO|V.*N"]{6,10} [pos="V[BDHV]N.*|.*V[BDHV]N"].

reassures us that the 2014 recall rate is not disadvantaged by the omission of ambiguity tags.

	BNClab-M sample: 1994				BNClab-M sample: 2014		
Rank	**POS sequence type**	**Cases**	**% of types**	**Rank**	**POS sequence type**	**Cases**	**% of types**
1	VHD XX0 VVN	69	24.0%	1	VHD XX VVN	208	23.8%
2	VHD AV0 VVN	49	17.0%	2	VHD RR VVN	111	12.7%
3	VHD AV0 VBN	15	5.2%	3	VHD XX VBN	42	4.8%
4	VHD XX0 VBN	14	4.9%	4	VHD XX RR VVN	26	3.0%
5	VHD PNP VVN	9	3.1%	5	VHD RR VBN	24	2.7%
6	VHD AV0 VHN	7	2.4%	6	VHD RR VHN	22	2.5%
7	VHD UNC VVN	5	1.7%	7	VHD RR RR VVN	17	1.9%
8	VHD PNP VDN	5	1.7%	8	VHD XX VHN	16	1.8%
9	VHD XX0 VHN	5	1.7%	9	VHD XX VDN	14	1.6%
10	VHD XX0 AV0 VVN-VVD	5	1.7%	10	VHD PPH1 VVN	11	1.3%

Table 3: Query breakdown for past perfect candidates separated by up to five words (top ten items)

To optimize recall of discontinuous past perfect, concordances of all candidate cases were hand-checked. In what follows we discuss three notable patterns.

The dominant pattern is that of negation and/or adverb modification, i.e., items containing tag XX0 and AV0 in the 1994 subcorpus, and items containing XX and RR in the 2014 subcorpus. Examples are provided in (13) and (14), the latter also including the discourse marker *like*, which is treated by CLAWS as an adverb.

(13) He probably *had***n't** *paid* that much anyway. (BNC1994 KBX:PS1DW)

(14) (…) she *had* **like literally just** *pressed* submit on her assignment. (BNC2014 SUVL:S0598)

Another pattern is that of inverted subject-verb word order in questions and conditionals, e.g., the sequences ranked 5th and 8th under the 1994 subcorpus, and 10th under the 2014 subcorpus. Typically, the subject is pronominal, as in (15).

(15) (…) *had* **I** *had* more time yesterday. (BNC2014 SGMM:S0483)

A further pattern relates to spoken disfluency features which include hesitation markers, generally transcribed as either *er* or *erm* in the BNC but are tagged differently in BNC1994DS (with C5 tag UNC, for unclassified item) and BNC2014S (C6 tag UH, for

interjection). Examples are (16) and (17).[18] Another disfluency feature the POS-tag breakdown finds is truncated words, where the speaker breaks off mid-word). These are tagged as unclassified items (UNC in C5, FU in C6), and illustrated in (18) and (19).[19]

(16) when we *had* **er <pause>** *ordered* it (BNC1994 KBW:PS08A)

(17) (…) we*'d* **erm** *exchanged* some Tesco vouchers on a couple of occasions (BNC2014 S64H:S0257)

(18) (…) cos the jacket *had* **s-** <pause> *fallen.* (BNC1994 KBF:PS04V)

(19) (…) she*'d* actually **w-** *gone* past the turning. (BNC2014 SA69:S0262)

4.2.3. Past non-perfects

Specifying a query for the past non-perfect is relatively simple in that only one item, a past-marked verb form (e.g., *took*, *was*), need be identified. The *BNClab* query in example (20) below retrieves all such verbs by conflating the C6 tags VBD, VHD, VDD, VVD, VBDM and VBDR.

(20) (V*D OR VBD*)

However, the vast number of hits this query returned (99,698) from the BNClab-M subcorpus necessitated a reduced sample for manual analysis.[20] We opted to sample one in 50 non-perfects from each period, selecting them systematically to minimize bias. The precision rate for the queries was 71 per cent for 1994 and 80 per cent for 2014.

4.2.4. Non-standard perfects

Based on the literature (e.g., Denison 1993; Huddleston and Pullum 2002), the general structure of non-standard perfects appears to be the one shown in Figure 1.

1	**2**	**3**	**4**	**5**
If/	Subject	*had*	*have*	Past participle
wish		*would*	*'ve*	
		'd	(*of*)	

Figure 1: Structure of non-standard perfect

[18] Adding to this complexity, hesitation markers are assigned the C6 tag FU (unclassified) rather than UH (interjection) in the BNClab interface.
[19] In the BNClab interface, queries skip over truncated items.
[20] We acknowledge the help of Loveen Dyall in extracting and verifying the query results.

As with the discontinuous past perfect, we can use test queries to discover the types of interpolating POS-sequences within this structure in conversation and the maximum length of the subject in words. By this process, we found that the maximum interval between *if/wish* and *had/'d* in slot 3 was four words, as illustrated in (21), while for *would* it was just two words, as shown in (22).

(21) (…) **if** Bolton and Blackburn yesterday ***had***n't *have been* such a high-profile game. (BNC2014 S9B9:S0152)

(22) (…) **if** the police **would've** raided your nan's. (BNC2014 S4YQ:S0253)

Also, it is quite common in BNC1994DS for the auxiliary verb *have* to be transcribed as *of* in representing spoken British English. Our queries took account of this, as can be seen in (23).

(23) If it *had*n't **of** *been* for Steve's car breaking down … (BNC1994 KCX:PS1FC)

Despite the non-standard character of these perfects, our queries on BNClab-M achieved good precision: 96.8 per cent on the 1994 section and 87.9 per cent on the 2014 section.[21]

4.3. Managing transcription inaccuracies

To address the issue of transcription quality in BNC1994DS (see section 2.2), we checked all retrieved examples of past perfect, past non-perfect and non-standard perfect for which audio is available (approximately 80% of cases). We removed all false positives, which included a) clear errors, where something other than the target structure can clearly be heard or the utterance is clearly attributed to the wrong speaker, and b) inaudible cases, where the target structure could not be heard in repeated listening.[22] For the putative past perfect, illustrated in example (24), what looks in the transcription like a hesitation marker (*er*) sounds on closer listening far more likely to be a reduced auxiliary (*'ve*), and therefore part of a double perfect. In (25), the official transcription has *wasn't the a big one* and is attributed to a female, yet the voice is almost certainly that of a male, and *it mustn't be a big one* is clearly audible.

[21] Query for 1994: "if|wish.*"%c []{1,4} [word="\'d|had|would"%c] [pos="XX0|AV0|UNC|ITJ"]{0,} [pos="VHI|VHB|PRF"] [pos="XX0|AV0|UNC|ITJ"]{0,} [pos="V[BDHV]N|.*V[BDHV]N"]
Query for 2014: "if"%c []{1,4} [word="\'d|had|would"] [pos="XX|R.*|FU|UH"]{0,} [pos="VHI|VH0|FU"] [pos="XX|R.*|FU|UH"]{0,} [pos="V[BDHV]N"]

[22] We did not discard cases where the audio and the transcription were irrecoverably misaligned and impossible to verify. These are equivalent to the BNC2014S cases, which lack audio.

(24) (…) if our Margaret had **er** been working. (BNC1994 KB1:PS01B)

(25) I'm almost frightened to put a crescendo in because it **wasn't the,** a big one. (BNC1994 KBH:PS05B)

Errors attributable to transcription issues in the 1994 part of BNClab-M totaled 64 for the past perfect (7.3% of candidate cases) and 28 for the past non-perfect (5.2% of candidate cases). While it is disconcerting to find so many transcription errors, it is reassuring that they affect the two target constructions in similar proportions.

4.4. Exclusions

Among the categories of hits that we excluded were errors resulting from speech disfluencies. If the target construction (past perfect/past non-perfect/non-standard perfect) verb phrase was deemed to be incomplete because of a false start, we excluded it. This is shown in (26), where the false start is highlighted.

(26) (…) **they*'d booked*** they*'d booked* a four-wheel drive (BNC2014 SMEB:S0238)

Another common source of error is the stative, simple past use of *had got*, illustrated in (27), which is actually more frequent than the use of past perfect *had got*. Such cases were discarded.

(27) Yeah but was she a woman living on her own or *had* she *got* a husband? (BNC1994 KCT:PS0FX)

5. RESULTS AND DISCUSSION

With the adjustments to sample selection, transcription, retrieval, and frequency calculation described above, we now present provisional results from the BNClab-M sample, beginning with the overall frequency of past perfects and past non-perfects. Table 4 extrapolates figures for the past non-perfect by multiplying the total number of non-perfect hits or candidates (filtered for working/middle-class speakers from England) by the precision rates calculated by manual analysis of the one in 50 subsets (see section 4.2.3).

1994			2014			Change
Past perfect	Past non-perfect (extrapolated)	Past perfect %	Past perfect	Past non-perfect (extrapolated)	Past perfect %	Significance
634	19,053	3.2%	2,224	57,977	3.7%	**

Table 4: Overall frequencies of past perfect and past non-perfect (* p<.05; ** p<.01; *** p<.001)

We see that overall usage of the past perfect in the BNClab-M data significantly increases from 3.2 percent to 3.7 percent of past-marked tense forms. At the same time, we observe social differentiation and patterns of change according to age, gender, and social class, although they are not necessarily significant, as illustrated in Table 5.

			1994		2014		Change
Gender	**Class**	**Age**	**Past Perfect**	**Non-perfect[a]**	**Past Perfect**	**Non-perfect[a]**	**Significance**
Female	Working class	U-45	113 3.5%	3076	17 1.9%	895	**
		O-45	89 3.2%	2699	61 2.7%	2194	n. s.
	Middle class	U-45	163 3.4%	4695	629 3.4%	17881	n. s.
		O-45	26 3.1%	818	989 5.2%	18131	**
Male	Working class	U-45	74 2.4%	3045	20 2.9%	674	n. s.
		O-45	15 3.2%	451	22 1.5%	1410	*
	Middle class	U-45	86 3.3%	2519	233 2.5%	9069	*
		O-45	68 3.7%	1749	253 3.2%	7723	n. s.

Table 5: Results across social groups in the BNClab-M sample ([a] extrapolated figures). * p<.05; ** p<.01; *** p<.001, n.s.=not significant

At this granular level, a more complex picture emerges. The overall increase of past perfects seems mainly attributable to female middle-class speakers over 45 years of age. Recall, however, that this is the group with the highest discrepancy in speaker numbers between 1994 and 2014 (see Table 1). Meanwhile, declining proportions of past perfect are found in five of the eight social groups, including both age cohorts of female working-class speakers and male middle-class speakers. But only a few of these changes are statistically significant.

Our conclusions about sociolinguistic variation are similarly tentative. At the aggregate level, there is a general tendency for the past perfect to be used more by middle-class than working-class speakers and, in 2014, by females than males. At the granular level we find, for instance, that rates of past perfect in female working-class and female middle-class are higher among younger speakers in the 1994 corpus, but in the 2014 corpus they are higher among older speakers. These figures need further investigation and corroboration.

Regarding past unreal contexts, we also see a proportional increase of the past perfect, with non-standard variants becoming less popular, as shown in Table 6.

	1994		**2014**		**Change**
	Cases	**% of total**	**Cases**	**% of total**	**Significance**
Past perfect	75	55.1%	195	76.2%	*
Non-standard perfect	61	44.9%	57	22.3%	***
Double	9	6.6%	14	5.5%	n.s.
Would	5	3.7%	5	2.0%	n.s.
Elided (*'d*)	47	34.6%	42	16.4%	***
Total	**136**	**100.%**	**256**	**100%**	

Table 6: Unreal past conditionals in BNClab-M subcorpus (* p<.05; *** p<.001, n.s.=not significant)

Further work is needed to understand why our overall results do not support those of Bowie *et al.* (2013) and Smith and Waters (2019), who both found a significant decline of the past perfect in recent spoken British English. Recall that the DCPSE corpus used in Bowie *et al.*'s (2013) focuses on the late twentieth century (1960s-1990s) and is not stratified by sociodemographic variables. Moreover, their results do not specify distributions across spoken registers, making direct comparisons with our results problematic. The *Desert Island Discs* BBC radio study by Smith and Waters (2019) is closer in timeframe to the present study, and includes similar social variables (e.g., age, gender, occupation, and education) for speakers from England. Smith and Waters (2019) found that age and education, rather than gender and occupation (as in the present study), correlated most significantly with the frequency of the past perfect. However, they operationalized occupation not as a socioeconomic index but as an estimate of the speaker's occupational need to use standard English (cf. Sankoff and Laberge 1978).

Another factor worth recalling is the possible effect of speakers in BNC2014S being more aware of being recorded (cf. Axelsson 2018). This may have led some speakers to be more careful to a) use the past perfect in contexts where the non-perfect would suffice

(e.g., with temporal clauses), and b) avoid non-standard perfects in unreal past conditionals. Further investigation of this possibility seems appropriate.

6. CONCLUSION

Our paper has methodological implications for the diverse users of spoken corpora, in particular users of the two conversational BNCs, primarily researchers, but increasingly also teachers and students in linguistics, language teaching, and beyond. Like the creators of these corpora, we wish the affordances to be embraced widely. Our concern has been to increase awareness of issues that arise when comparing these corpora across time, especially regarding changes in grammar.

Addressing the need for closely comparable data, we took advantage of Brezina *et al.*'s (2018a) BNClab subcorpus. This judgment sample affords a closer balance of social group representation in the two periods than that of the parent BNCs, although its geographical scope (all nations of the UK) affects representativeness and comparability at the granular level. We partially improved comparability by limiting the subcorpus to working-class and middle-class speakers from England but narrowed the demographic representativeness in the process. Further refinement of the sample is no doubt possible.

The case study on the past perfect navigated challenges in investigating grammar in spontaneous conversation, namely deciding on the field of competition and units of measurement, and devising queries to retrieve constructions with satisfactory recall and precision. A heuristic approach allowed us to capture variability among discontinuous past perfects and non-standard perfects, facilitating recall. The mixed quality of transcriptions in BNC1994DS can also affect comparability with BNC2014S. Our decision to discard hits from BNC1994DS containing inaccurate transcription should not disrupt comparability with BNC2014S, since we made equivalent corrections for rival constructions and used proportional frequencies in both periods.

The implications of our results to date are less clear-cut. We find consistently higher frequencies of the past perfect by female and middle classes speakers. The finding that the past perfect is spreading in conversational use appears to contradict previous reports of a decline in spoken British English. While the discrepancy may be related to unbalanced recruitment of some categories of speaker in the two periods, it seems premature to suggest that the past perfect is on the wane, at least in English conversation.

Our investigation has highlighted some contemporary spoken usage characteristics associated with the past perfect that receive scant attention in English language teaching materials and curricula. One example, occurring in unreal past conditionals, is what we have labeled the 'non-standard' perfect', although it seems widely accepted in colloquial contexts (Huddleston and Pullum 2002; Ishihara 2003). These forms are almost entirely overlooked in ELT resources (Ishihara 2003) and, despite a proportional decline, their use in informal conversation seems frequent enough to draw the attention of advanced learners, at least as receptive knowledge (cf. Timmis 2005). Similarly, the simple past, stative use of *had got* in British English is almost absent from ELT coursebooks, and yet it is more common than its past perfect homograph.

In the future, we aim to follow up this study through detailed, selective exploration of the parent BNCs where social categories —particularly age— have more differentiated breakdowns.

References

Anderwald, Lieselotte. 2002. *Negation in Non-standard British English: Gaps, Regularizations and Asymmetries.* New York: Routledge.

Atkinson, Will. 2015. *Class*. Cambridge: Polity Press.

Axelsson, Karin. 2018. Canonical tag questions in contemporary British English. In Vaclav Brezina, Robbie Love and Karin Aijmer eds, 96–119.

Baker, Paul. 2023. A year to remember? Introducing the BE21 corpus and exploring recent part of speech tag change in British English. *International Journal of Corpus Linguistics* 28/3: 407–429.

Ball, Catherine. 1994. Automated text analysis: Cautionary tales. *Literary and Linguistic Computing* 9: 295–302.

Beal, Joan. 2010. *An Introduction to Regional Englishes: Dialect Variation in England*. Edinburgh: Edinburgh University Press.

Biber, Douglas. 1933. Representativeness in corpus design. *Literary and Linguistic Computing* 8/4: 243–257.

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 and Miriam Meyerhoff. 2014. Significant or random? A critical review of sociolinguistic generalisations based on large corpora. *International Journal of Corpus Linguistics* 19/1: 1–28.

Brezina, Vaclav and William Platt. 2024. *#LancsBox X*. http://lancsbox.lancs.ac.uk/ (accessed 5 May 2023.)

Brezina, Vaclav, Dana Gablasova and Susan Reichelt. 2018a. *BNClab*. http://corpora.lancs.ac.uk/bnclab (accessed 5 May 2023.)

Brezina, Vaclav, Robbie Love and Karin Aijmer eds. 2018b. *Corpus Approaches to Contemporary British Speech: Sociolinguistic Studies of the Spoken BNC2014*. New York: Routledge.

Brezina, Vaclav, Abi Hawtin and Tony McEnery. 2021. The written *British National Corpus* 2014 – design and comparability. *Text and Talk* 41/5–6: 595–615.

Burnard, Lou. 2007. *Reference Guide for the British National Corpus (XML edition)*. http://www.natcorp.ox.ac.uk/docs/URG/ (accessed 5 May 2023.)

Coleman, John, Mark Liberman, Greg Kochanski, Lou Burnard and Jiahong Yuan. 2011. Mining a year of speech. In *Proceedings from the Workshop of New Tools and Methods for Very-Large-Scale Phonetics Research*, 16–19. http://www.phon.ox.ac.uk/jcoleman/MiningVLSP.pdf (accessed 5 May 2023.)

Crowdy, Steve. 1993. Spoken corpus design. *Literary and Linguistic Computing* 8/4: 259–265.

Curry, Niall, Robbie Love and Olivia Goodman. 2022. Adverbs on the move: Investigating publisher application of corpus research on recent language change to ELT coursebook development. *Corpora* 17/1: 1–38.

Declerck, Renaat. 2006. *The Grammar of the English Verb Phrase. Volume 1: The Grammar of the English Tense System: A Comprehensive Analysis*. Berlin: Mouton de Gruyter.

Denison, David. 1993. *English Historical Syntax: Verbal Constructions*. London: Longman.

Depraetere, Ilse and Chad Langford. 2019. *Advanced English Grammar: A Linguistic Approach*. London: Bloomsbury.

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

Gablasova, Dana, Vaclav Brezina and Tony McEnery. 2017. Exploring learner language through corpora: Comparing and interpreting corpus frequency information. *Language Learning* 67/1: 130–154.

Garside, Roger and Nicholas Smith. 1997. A hybrid grammatical tagger: CLAWS4. In Roger Garside, Geoffrey Leech and Anthony McEnery eds., 102–121.

Garside, Roger, Geoffrey Leech and Anthony McEnery eds. *Corpus Annotation: Linguistic Information from Computer Text Corpora*. London: Longman.

Hardie, Andrew. 2012. CQPweb – Combining power, flexibility and usability in a corpus analysis tool. *International Journal of Corpus Linguistics* 17/3: 380–409.

Hofland, Knut, Anne Lindebjerg and Jørg Thunestvedt. 1999. *ICAME Collection of English Language Corpora*. Bergen: The HIT Centre.

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

Hoffmann, Sebastian and Sabine Arndt-Lappe. 2021. Better data for more researchers: Using the audio features of BNCweb. *ICAME Journal* 45: 125–154.

Horvath, Barbara. 2013. Ways of observing: Studying the interplay of social and linguistic variation. In Christine Mallinson, Becky Childs and Gerard Van Herk eds. *Data Collection in Sociolinguistics: Methods and Applications.* New York: Routledge. https://doi.org/10.4324/9780203136065

Huddleston, Rodney and Geoffrey Pullum. 2002. *The Cambridge Grammar of the English Language*. Cambridge: Cambridge University Press.

Ishihara, Noriko. 2003. “I wish I would have known!”: The usage of *would have* in past counterfactual *if*- and *wish*-clauses. *Issues in Applied Linguistics* 14/1: 21–48.

Jucker, Andreas, Gerold Schneider, Irma Taavitsainen and Barb Breustedt. 2008. Fishing for compliments: Precision and recall in corpus-linguistic compliment research. In Andreas Jucker and Irma Taavitsainen eds. *Speech Acts in the History of English*. Amsterdam: John Benjamins, 273–294.

Lavandera, Beatriz. 1978. Where does the sociolinguistic variable stop? *Language in Society* 7/2: 171–82.

Leech, Geoffrey and Nicholas Smith. 2000. *Manual to Accompany the British National Corpus (Version 2) with Improved Word-class Tagging*. https://ucrel.lancs.ac.uk/bnc2/bnc2postag_manual.htm (accessed 5 May 2023.)

Leech, Geoffrey and Nicholas Smith. 2005. Extending the possibilities of corpus-based research on English in the twentieth century: A prequel to LOB and FLOB. *ICAME Journal* 29: 83–98.

Leech, Geoffrey, Marianne Hundt, Christian Mair and Nicholas Smith. 2009. *Change in Contemporary English: A Grammatical Study*. Cambridge: Cambridge University Press.

Love, Robbie. 2020. *Overcoming Challenges in Corpus Construction: The Spoken British National Corpus 2014*. New York: Routledge.

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.

McEnery, Tony, Robbie Love and Vaclav Brezina. 2017. Introduction: Compiling and analysing the Spoken British National Corpus 2014. *International Journal of Corpus Linguistics* 22/3: 311–318.

Milroy, Lesley and Matthew Gordon. 2003. *Sociolinguistics: Methods and Interpretation*. Oxford: Blackwell.

Mindt, Dieter. 2000. *An Empirical Grammar of the English Verb System*. Berlin: Cornelsen.

Reichelt, Susan. 2021. Recent developments of the pragmatic markers *kind of* and *sort of* in spoken British English. *English Language & Linguistics* 25/3: 563–580.

Rühlemann, Christoph. 2007. *Conversation in Context: A Corpus-driven Approach*. London: Continuum.

Sankoff, David. 2005. Problems of representativeness. In Ulrich Ammon, Norbert Dittmar, Klaus Mattheier and Peter Trudgill eds. *Sociolinguistics: An International Handbook of the Science of Language and Society*. Berlin: Walter de Gruyter, 998–1002.

Sankoff, David and Susan Laberge. 1978. The linguistic market and the statistical explanation of variability. In David Sankoff ed. *Linguistic Variation: Models and Methods*. New York: Academic Press, 239–250.

Schilling-Estes, Natalie. 2007. Sociolinguistic fieldwork. In Robert Bayley and Ceil Lucas eds. *Sociolinguistic Variation: Theories, Methods, and Applications*. Cambridge: Cambridge University Press, 165–190.

Smith, Nicholas. 1997. Improving a tagger. In Roger Garside, Geoffrey Leech and Anthony McEnery eds., 137–150.

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.

Sönning, Lukas and Manfred Krug. 2022. Comparing study designs and down-sampling strategies in corpus analysis: The importance of speaker metadata in the BNCs of 1994 and 2014. In Ole Schützler and Julia Schlüter eds. *Data and Methods in*

Corpus Linguistics: Comparative Approaches. Cambridge: Cambridge University Press, 127–160.
Tagliamonte, Sali. 2006. *Analysing Sociolinguistic Variation*. Cambridge: Cambridge University Press.
Timmis, Ivor. 2005. Towards a framework for teaching spoken grammar. *ELT Journal* 59/2: 117–125.
Trask, R.L. 1993. *A Dictionary of Grammatical Terms in Linguistics*. New York: Routledge.
Yao, Xinyue and Peter Collins. 2013. Recent change in non-present perfect constructions in British and American English. *Corpora* 8/1: 115–135.

Corresponding author
Nicholas Smith
University of Leicester
School of Education
21 University Road
LE1 7HR
Leicester
United Kingdom
Email: ns359@leicester.ac.uk

received: July 2023
accepted: April 2024

APPENDIX: COMPOSITION OF THE ORIGINAL BNCLAB SUBCORPUS, REPRESENTING ALL UK REGIONS AND CLASSES (BASED ON BREZINA *ET AL.* 2018A)

		1994	**2014**
Gender	Female	126	126
	Male	124	124
Age	0–14	16	12
	15–24	47	56
	25–34	50	60
	35–44	47	31
	45–59	46	50
	60–74	31	31
	75–95	13	10
Social class	Middle class	62	113
	Working class	63	36
	Retired	35	27
	Student	41	49
	Unknown	49	25
Region	England: London	31	28
	England: Midlands	41	28
	England: North	47	62
	England: Southwest	30	25
	England: Southeast	34	45
	Scotland	13	8
	Ireland	22	1
	Wales	18	14
	Other	14	39
Total		**250**	**250**

RiCL Research in Corpus Linguistics

Rethinking interviews as representations of spoken language in learner corpora

Pascual Pérez-Paredes[a] – Geraldine Mark[b]
University of Murcia[a] / Spain
Cardiff University[b] / United Kingdom

Abstract – Following the call to examine the role of learner corpora in SLA research (Bell and Payant 2021), this paper discusses spoken learner corpora —specifically those collected through interviews— and considers the aspects of spoken learner language that they represent. The interview is both an elicitation technique and a complex genre. The overlapping of the two conceptualisations under the same term may give rise to problems of definition about the nature of the language collected and, as a consequence, to difficulties in interpretation when assessing the characteristics of spoken learner data. In this paper, we use original research to exemplify some of the areas that need some rethinking in terms of future reconceptualisation about how spoken data are collected and analysed. This research shows the potential impact of the degree of interviewer/interviewee engagement with the task, suggesting that not enough attention has been paid to the genre of interview in learner corpus research.

Keywords – learner corpus research; spoken language; task; interview; representativeness

1. INTRODUCTION[1]

This paper touches broadly on the ubiquity of the 'interview' as a means of gathering learner language and the challenges of using interviews to represent everyday spoken language. It seeks to show the limitations of some standard practices of gathering spoken learner data, focusing particularly on the practice of interviewing and the language produced in interview

[1] The authors would like to thank the anonymous reviewers and the editors for their invaluable insights and constructive feedback. Their thoughtful comments have significantly contributed to the enhancement and clarity of this manuscript.

tasks. In doing so, it considers the challenges of defining the interview as a genre. Our research asks whether the data resulting from interview tasks offer a valid representation of spoken learner language and draws attention to its use as a spoken learner language benchmark. It argues that, if corpus linguists are to claim that data collected through interviews are indeed representative of a given mode (e.g., written vs. spoken), we need to pay specific attention to ensuring that the interactive features of everyday spoken language are represented. The research further contends that, if the data are deemed not to be representative of spoken language, we need to understand the language in use that these data do represent (Crawford 2022) and must be wary of using them to investigate the spoken interactional proficiency of learners. We offer some evidence from spoken learner corpora about how the role of interviewers may impact the final product used by researchers when discussing L2 spoken data.

The paper is a relevant contribution to the field of learner language research, as it affects the ways in which the scope of the findings and claims that derive from the analysis of learner corpora are conceptualised and framed. We seek to contribute to methodological innovation in the field by recommending further reconceptualisation of the use of interview data and interviewers' role in collecting such data, as well as pointing to other innovative means for gathering spoken learner language.

In what follows, we first look at definitions and representations of the interview as a genre, and how it is used as a tool in data collection and analysis in L1 and L2 spoken corpora (Section 2). We discuss what we understand by 'spoken language', particularly in the field of language learning and teaching, and the interview as a representation of this. We then turn our attention to the use of the interview in collections of learner language, looking specifically at the role of the interviewer and the effect on interaction. In Section 3, we demonstrate this through original research looking at examples from existing L2 and L1 corpora —such as the *Louvain International Database of Spoken English Interlanguage* (LINDSEI; Gilquin et al. 2010)[2] and the *Louvain Corpus of Native English Conversation*

[2] https://uclouvain.be/en/research-institutes/ilc/cecl/lindsei.html

(LOCNEC; De Cock 2004)[3]— and conclude that the interview offers a narrow representation of spoken language that does not necessarily allow for evidence of the interactional features of everyday spoken language (Section 4).

2. The interview in corpus linguistics: Definitions and representations

In this section, we show how interview data are represented in L1 corpora (Section 2.1) and contrast this with the ways in which interviews are used in learner spoken data collection (Section 2.2). We consider both the impact of the task and the interviewer on the final L2 product (Section 2.3).

2.1. Interviews and their use in L1 spoken corpora

The super-genre 'interview' subsumes diverging assumptions of how speakers construct interaction within the boundaries of the communicative situation where the interview takes place. Despite the apparent simplicity with which one might think that the interview genre can be defined, the concretisation of the genre in corpus linguistics (henceforth CL) presents some challenges. McCarthy and Carter (1994: 191) have defined the interview as a genre that is

> sufficiently broad to take in a variety of sub-types from minimally interactional, maximally transactional events (e.g., formal political interviews) to maximally interactional personal encounters (e.g., chat shows, therapeutic interviews).

At one end of the scale, the 'maximally interactional' is typically focused on relationship creation, where shared involvement and reciprocity and, for example, expression of stance (Biber *et al.* 1999), is evident in the discourse. At the other end, the 'maximally transactional' focuses principally on information provision and exchange. This distinction allows corpus linguists to conceptualise interaction as a key variable to be factored into the data collection

[3] https://corpora.uclouvain.be/catalog/corpus/locnec

process, the ways in which interviews are designed during the corpus building process, and the analysis of the genre.

Terminological problems arise when it comes to categorising interviews in CL. Let us take, for example, the *British Component of the International Corpus of English* (ICE-GB).[4] In ICE-GB, around three per cent of the spoken data (20,000 words) are contributed by broadcast interviews (Lee 2002), which are texts that display minimal interaction. In turn, face-to-face conversations represent around 15 per cent of the spoken data in the corpus. Broadcast interviews are classified as 'public dialogue', while face-to-face conversations are conceptualised as 'private dialogue' (Lee 2002). According to Lee (2002), both broadcast interviews and face-to-face conversations belong to the same medium (spoken) and interaction (dialogue), but they represent different super-genres and subgenres, as shown in Figure 1.

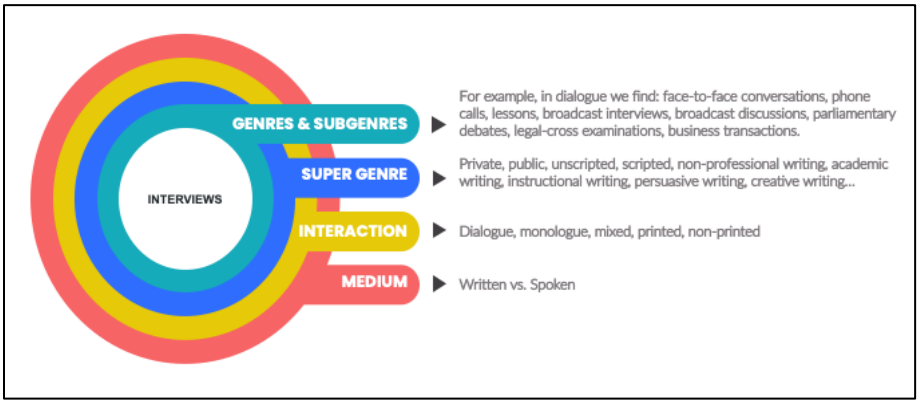


Figure 1: Text classification in ICE-GB. Adapted from Lee (2002)

Broadcast interviews and face-to-face conversations represent, therefore, instances of different dialogic subgenres. In the *Spoken Component of the First British National Corpus* (Spoken BNC1994),[5] the super genre 'interview' mainly includes job interviews, history interviews, and narrative broadcast interviews. Interviews are either institutionally situated or are reflective of professional practices. In the *Spoken Component of the Second British National Corpus* (Spoken BNC2014; Love 2020),[6] which comprises ten million words of

[4] http://ice-corpora.net/ice/index.html
[5] http://www.natcorp.ox.ac.uk
[6] http://corpora.lancs.ac.uk/bnc2014

spoken English gathered from the UK public between 2012 and 2016, the data are classified into 992 activity types, representative of everyday spoken interactions (e.g., chatting with friends or colleagues chatting in coffee breaks). Interviews feature in only two of these activity types and constitute 0.1 per cent of the Spoken BNC2014. As we write this paper, the *Lancaster-Northern Arizona Corpus of American Spoken English* (LANA-CASE), which is the American English counterpart to the BNC2014, is being compiled. This corpus aims to collect 1,000 hours of self-recorded conversation between two or three adults (see Hanks *et al.* 2024). Interviews, however, are not represented in LANA-CASE. Judging from the relative scarcity of interviews in the spoken corpora outlined here, we may conclude that the interview genre is perceived as marginal in representing a benchmark for everyday spoken language in English language corpora, whereas the conversation genre is pervasive and perceived as highly representative of spoken language.

The interview genre, when represented in English corpora, seems not to rely exclusively on interlocutors' shared interpersonal context, but rather favours the construction of texts that depend on public formal interactions. For example, Biber (1995) reports that, from a multi-dimensional analysis perspective, Dimension 1 shows how involved and interactional interviews score in comparison with other registers. A register with a high score on this Dimension exhibits frequent occurrences of private verbs such as *think*, omissions of *that*, present tense verb phrases, contractions, and second person pronouns (Biber 1995: 117). As Biber (1995: 151) notes, high scores show evidence of "highly interactive, affective discourse produced under real-time constraints, whether spoken or written." Telephone conversations, together with face-to-face conversations, show the highest score on this Dimension, while interviews fall below, close to personal letters and spontaneous speeches. In other words, Biber shows that the interview data are not necessarily representative of an involved, interactional register, whereas conversations are.

For Biber *et al.* (1999) and Biber *et al.* (2021), conversations are grounded in a shared context where specification of meaning is avoided and the contextual background provides the backdrop for many of the exchanges that take place between members of the family, friends, or fellow workers. These exchanges are characterised, among others, by the pervasive use of non-clausal elements, personal pronouns, so-called inserts, and low lexical

density. Interlocutors in a conversation dynamically co-construct the discourse, taking turns and adapting, as Biber *et al.* (1999: 1039) state,

> their expression to the ongoing exchange [...] the to-and-fro movement of conversation between speaker [...] the occurrence of utterances which [...] either form a response or elicit a response [...] known as adjacency pairs.

In McCarthy's (2010: 7) terms, there is a "shared responsibility" among interlocutors in face-to-face interactions to maintain a "flow across turn boundaries […] captured by the metaphor of confluence, reflecting the jointly produced artefact which constitutes an efficient and successful interaction." This adaptation to the flow of discourse is achieved additionally through clause co-construction, back-channelling, hedging, and discourse marking, often exemplified through the use of what Hasselgreen (2004) labels 'smallwords' (e.g., *you know, sort of, really, just, well*), which are —as McCarthy (2010: 11) points out— "interactive and flow-sustaining" in everyday conversation.

So far, we have seen that the interview super genre is likely to subsume different assumptions and approaches to how speakers construct interaction within the boundaries of the communicative situation where the interview takes place. A crucial consideration here is that if the role of interaction is not thought out in advance, corpus compilers may collect data which, under the same umbrella term, may represent very different types of subgenres. This can be partly explained by the fact that the field of CL has a long tradition of unobtrusively collecting data (Stubbs 2007) that has already been produced by members of a community, and where the use of interviews as an elicitation method is not particularly frequent. As Koester (2022) argues, interviews are not usually associated with corpus studies other than in the learner corpus domain. In fact, Koester sees interviews as a complement to corpus data rather than as one of the main genres in L1 corpora.

The debate about interviews in CL contrasts with the attention paid to the interviewer's role in social sciences and in sciences related to education or applied linguistics (Mann 2011). In social and education sciences, the literature abounds on the kind of interaction and dynamics that interviewers need to foster during interviews to collect data that are dense enough to support a qualitative analysis (Cohen *et al.* 2017). In fact, there exists a variety of conceptualisations about the nature of interviews in social science research. Mann (2011),

for instance, has discussed two metaphors which are revealing from an epistemological perspective. In one of them, the interviewer is a traveller, which evokes a post-modern constructivist position that contrasts with the interviewer as "the positivist miner extracting nuggets of raw truth" (Mann 2011: 7). For Mann, all interviews are unavoidably meaning-making ventures where the interviewer's contribution to the co-construction of the interview content must be explicitly acknowledged and may thus become a topic for analysis itself. With this in mind, here we argue that interviews as elicitation instruments for data collection do seem to align with interviews as a genre. The reason for the lack of debate about the role of the interviewer in eliciting language may well lie in 1) the positivist nature of the data collection procedure followed by most researchers in CL (mining nuggets), and 2) the different conceptualisations about the nature of what interview data stand for. The invisibility of the interviewers in co-constructing interview data (Jones 2022) is, perhaps, partly explained by a view in CL stressing that texts can be collected but cannot be manipulated directly by the corpus designers, as this would alter the nature of the observed phenomena.

2.2. Interviews in learner corpus research

In learner corpus design, in stark contrast to L1 corpora, interviews are one of the main genres (Tracy-Ventura *et al.* 2021) that, together with argumentative essays, academic writing, and narratives, are most widely represented (Gilquin 2021). Of the 201 learner corpora identified in the *Learner Corpora around the World* database (LCW),[7] around a quarter (49) are categorised as spoken. Of these, 37 per cent are described as containing interviews. For Tracy-Ventura *et al.* (2021: 414),

> oral corpora often consist of interviews (primarily between a researcher and the participant), narrative retells based on pictures or silent films, or monologues based on a prompt.

These tend to be tasks that EFL learners are familiar with. Both LINDSEI (Gilquin *et al.* 2010) and the *Trinity Lancaster Corpus* (TLC; Gablasova *et al.* 2019)[8] —to cite two of the

[7] https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html
[8] https://cass.lancs.ac.uk/trinity-lancaster-corpus/

most relevant spoken learner corpora— represent some of these spoken tasks that learners typically engage in. These two corpora differ, at least, in one key area of corpus design. The LINDSEI tasks were designed to gather, observe, and analyse learner language from a contrastive interlanguage analysis perspective (Gilquin *et al.* 2010), allowing for comparisons between different L1 groups (as well as L1 speakers through the LOCNEC corpus). In turn, TLC is the result of the language produced by exam candidates, comprising tasks from the Graded Examinations in Spoken English (GESE), developed and administered by Trinity College London.

A third important learner corpus, widely used in learner corpus research, is the spoken dialogue component of the *International Corpus Network of Asian Learners of English* (ICNALE SD),[9] which includes "approximately 270-hour videos of oral interviews conducted with 405 college students in ten regions in Asia" (Ishikawa 2019: 154). The TLC L2 data, as reported in Gablasova *et al.* (2019), include over 2,000 L2 speakers from different cultural and L1 backgrounds and contain 4.2 million words (tokens) of transcribed spoken interaction between exam candidates (L2 speakers of English) and examiners (L1 speakers of English), which makes it the largest corpus of spoken L2 English at the time of writing. The TLC tasks show a combination of controlled tasks and an explicit task description available to the exam candidates. According to Gablasova *et al.* (2019: 133), the interaction "develops dynamically between the L1 and L2 speaker." As regards topics in TLC, Gablasova *et al.* (2019: 154) state that

> in two tasks (presentation and discussion) the topic is selected freely by the candidate, while in the other two tasks (interactive task and conversation), the topics are selected by the examiner.

For Gablasova *et al.* (2019: 147), comparability in terms of linguistic setting and speaking tasks is key when collecting the learner data during the interview, as "all interviews are conducted by trained examiners from Trinity College London following the same principles as in the L2 interviews."

[9] https://language.sakura.ne.jp/icnale/

LINDSEI contains oral data produced by upper-intermediate and advanced learners of English from different L1 backgrounds. The LINDSEI CD-ROM (Gilquin *et al.* 2010) comprises 11 L1 components: Bulgarian, Chinese, Dutch, French, German, Greek, Italian, Japanese, Polish, Spanish, and Swedish. All components follow the same structure. The interview is made up of the same three tasks: a set topic, a free discussion, and a picture description. The corpus includes 554 interviews, totalling 1,080,232 words. The interviews were transcribed and marked-up according to the same conventions by the different national teams. In the first LINDSEI task, the interviewees were given three topics and were asked to choose one of them, think about it for a few minutes without taking notes, and then talk about it.[10]

This first task of the interview is therefore predominantly monologic. After this, the interviewer goes on to ask questions and interacts with the learner. The questions address the topic chosen in the first task and then other subjects, including life at university, hobbies, or travels abroad (Gilquin *et al.* 2010). In the third task of the interview, the interviewees were asked to look at four pictures making up a story and to describe what they saw. For Friginal *et al.* (2017: 43), the LINDSEI interviews illustrate "how learners shift their use of various linguistic features, covering a range of discourse domains" and provide "a wealth of information on how learners actually use language in interviews."

Alongside LINDSEI, LOCNEC —a comparable corpus of interviews with L1 speakers of English, designed to represent L1 conversation— was compiled to provide a baseline for L1-L2 comparisons. LOCNEC mirrors the tasks and interview approach in LINDSEI and is made up of 50 interviews. Aguado *et al.* (2012) completed an additional 28 interviews at Manchester Metropolitan University in 2006 following the same design criteria.

Friginal and Polat (2015) conducted a multi-dimensional analysis of LINDSEI to identify the dimensions of English learner talk and interpret the resulting dimensions,

---

[10] In Gilquin *et al.* (2010), the first topic is 'an experience you have had which has taught you an important lesson. You should describe the experience and say what you have learned from it'. The second topic is 'a country you have visited which has impressed you. Describe your visit and say why you found the country particularly impressive'. Finally, the third topic is 'a film/play you have seen which you thought was particularly good/bad. Describe the film/play and say why you thought it was good/bad'.

comparing how they are distributed across the different L1 backgrounds. One of the new identified dimensions shows that the picture description task is functionally distinct from the other two tasks. Pérez-Paredes and Sánchez-Tornel (2019) support this finding in their multi-dimensional analysis investigation of the extended LOCNEC, observing a statistically significant difference between the interactive part and the picture description. However, in the LINDSEI interview, the different tasks do not align with one particular type of interaction, or with any of the subgenres seen in Figure 1, as they arguably represent EFL classroom tasks that spread over different types of interactions and super genres (i.e., narratives and descriptions).

Not all spoken learner corpora, however, prioritise interviews as a means of eliciting data. In the *Michigan Corpus of Academic Spoken English* (MICASE),[11] which comprises 1.8 million words from lectures and classroom discussions, 12 per cent of the speakers have an L1 other than English. Likewise, the *TOEFL 2000 Spoken and Written Academic Language* (T2K-SWAL)[12] was designed to provide a basis for test construction and validation of spoken and written registers in US universities (Biber *et al.* 2004) and captures the language as used by students and lectures across study groups, service encounters, or class sessions. Similarly, the *British Academic Spoken English Corpus* (BASE)[13] represents language used in academic contexts such as seminars or lectures and includes a small amount of L2 learner output (Friginal *et al.* 2017). Also, the *Vienna-Oxford International Corpus of English* (VOICE)[14] captures interactions of spoken English as a lingua franca and also includes interviews, although other elicitation techniques such as seminar discussions, panels, or meetings are more frequent. The VOICE compilation criteria emphasise the lingua franca status of the interactions represented in the data but the L2 learning dimension is not an explicit focus in its design.

The interview genre and the roles of interviewers have not received much attention in specialised CL literature. When considered in the context of their scarcity in L1 spoken

[11] https://quod.lib.umich.edu/m/micase/
[12] http://universal.elra.info/product_info.php?cPath=42_43&products_id=1497
[13] https://www.reading.ac.uk/acadepts/ll/base_corpus/
[14] https://voice.acdh.oeaw.ac.at/

corpora, as described above, this is not entirely surprising. In the second edition of *The Routledge Handbook of Corpus Linguistics* (O'Keeffe and McCarthy 2022), the term 'interviewer' is used in only two occasions, and neither the role of the interviewer nor their ability to influence language during interviews is discussed. In *A Practical Handbook of Corpus Linguistics* (Paquot and Gries 2021), the term 'interviewer' is simply not found, while interviews are regarded as genres and interview topics are occasionally referred to, in the context of L1 corpora, as a source of bias that may lead to the overrepresentation of linguistic features (Gut 2012) such as, for instance, the past tense in the *Freiburg English Dialect* corpus (FRED; Anderwald and Wagner 2007). In *The Routledge Handbook of Second Language Acquisition and Corpora* (Tracy-Ventura and Paquot 2021), the term 'interviewer' is found only once, when discussing the design of LINDSEI (Gilquin *et al.* 2010).

Despite the abovementioned absence of reference in the literature and the sparse representation of the interview genre in L1 spoken corpora, the interview itself is ubiquitous in L2 spoken data. For example, Bell *et al.* (2021: 218) elected to use an interview task for the spoken component of their corpus study on L2 grammatical development because of 1) its popularity with previous studies and researchers and 2) "potential influences that task condition (monologic vs dialogic) can have on language production." The reference to "potential influences" is important because, as we have seen, both LINDSEI and TLC include dialogic tasks where interaction might be expected to emerge. However, as Gráf (2017: 29) points out, in LINDSEI, the execution of these tasks, is "left very much to the coordinator's own experience or initiative." While Friginal *et al.* (2017) have highlighted the usefulness of interview data to investigate interaction, the nature of the interaction, as pointed out by Gráf (2017), is not totally clear in terms of the linguistic and functional characteristics of the corpus data collected. Gráf (2017) has in fact highlighted some open questions and debates in the research design of standard spoken learner corpora, including the lack of concrete research questions and the lack of specific guidelines in terms of how the interviews are carried out regarding their communicative content. Gráf wonders whether interlocutors are expected to produce certain grammatical or lexical patterns and, if so, how this is supposed to be achieved. Similarly, he has also expressed doubts about the weight of monologic tasks in the design of oral corpora and the interviewers' active/passive role in the conversational construction of the interview. This leads us to consider the role of the interviewer in the

collection of L2 data and the impact this has on the shape and profile of the data we collect. What role does the interviewer play? Is the interviewer an interlocutor, a conversant, a facilitator, a co-ordinator, a passive listener? As Bell *et al.* (2021: 218) argue, we return to the "potential influences that task condition (monologic vs. dialogic) can have on language production," and the effect of (degrees of) engagement from the interviewer in the discourse.

2.3. The interviewer in learner corpus research

McCarthy and Carter (1994) explore issues related to integrating discourse and conversational practices into the language learning classroom. They highlight the question of whether learner performance or output engage with the discourse process in a learning context. They explore this by investigating the differences in the same interview tasks firstly undertaken between two L1 speakers and secondly between an L1 and an L2 speaker in a learning context. Both interviewers are given the same brief for the tasks. McCarthy and Carter (1994: 189) note that in the L1:L1 interview the speakers orient themselves towards a more relational-style interview sub-genre, in which their joint goals are interactional and not essential to the "transactional structure of the encounter," whereas the L1:L2 interview follows a transactional question-answer structure and gets the job of the interview done efficiently, but with less involvement between participants. McCarthy and Carter (1994: 191) point out that the non-intimate interview as a genre "is not well attuned to interactional features: reciprocity and affective convergence are not at all among its goals." They argue that the interviewer may wish to behave in a more 'human' way, but the restrictions from the setting up of the task as a transactional encounter may be a barrier to this. This may result in less interactional output, where features of everyday spoken discourse are not necessarily represented. Unlike in everyday conversation, in learner data collection, the interviewer manages time and topic shifts.

We now return to the pioneering LINDSEI corpus to explore the issues pointed out in McCarthy and Carter (1994), and the effect of these on the data collected. The LINDSEI compilation guidelines (Gilquin *et al.* 2010) specify that the objective of the LINDSEI project was to collect spoken interlanguage during informal interviews. As outlined in Section 2.1, the interviews had to follow a pattern in which the interviewers had some freedom as regards

the actual questions they asked the learners. The guidelines recommend minimal interruption from the interviewer. This is an elicitation task during which the interviewer facilitates the goal of interlanguage collection through a series of questions and answers, not necessarily interactions. The LINDSEI guidelines make this clear and the interviewer and learner turns and tokens are itemised separately, since the learners' turns will be of the utmost importance. The guidelines also refer to the relationship between the interviewer and the interviewee and acknowledge that the status of the interviewer in relation to the interviewee may impact the progress of the interview and its formality. Here we suggest that hedging the impact underestimates the effect of the interviewer on the proceedings and presents one understanding of spoken language in which the monologic or dialogic nature of the data is not of primary importance. Depending on the consistency of the involvement of the interviewer and the degree to which reciprocity is encouraged across the data, the result may, on the one hand, tend towards a representation of spoken language which potentially orients towards written norms (characterised by monologic, transactional responses to questions) or, on the other, towards a representation of spoken language which reflects co-construction, and is dialogic and interactional. For the purposes of learner corpus research, it is problematic if 1) the data purport to contain one representation of language in use in a given context but does not, and 2) the data are used to make judgements about learner language proficiency. This introduces variables which are potentially ignored in traditional learner corpus research.

In practice, the variation in the freedom to interact by interviewers and in the degree to which interviewee and interviewer understand the needs of the genre is critical to the data produced and to its interpretation and comparability (see Gráf 2017). This becomes increasingly important in two research contexts: 1) when interview-elicited data are used to investigate features of L2 spoken interactional language, and 2) when interview-elicited data are used as benchmarks for spoken learner data. To date, the LINDSEI data are frequently used for both these representative and comparative purposes. For example, Larsson *et al.* (2023) use LINDSEI to represent learner speech when exploring development of grammatical complexity in writing, and investigating whether learners move away from speech-like production towards more advanced written production. They state that they "use LINDSEI to represent a benchmark for speech and ICLE to represent a benchmark for

writing" (Larson *et al.* 2023: 8). Similarly, Friginal *et al.* (2017: 45) select LINDSEI because it

> is especially well suited to investigations of learner talk because of its large size, representativeness (as noted earlier, 11 L1 backgrounds with approximately 50 interviews each), and the consistency of its implementation.

Castello (2023) also uses LINDSEI and LOCNEC to represent spoken interaction when investigating stance adverbials in discourse and conversation management from spoken English interaction. Likewise, in a study exploring the use of *well* as a discourse marker, Aijmer (2018) uses the Swedish component of LINDSEI and L1 LOCNEC to investigate uses of *well* and finds that the L1 speakers use *well* more frequently than the L2 speakers to signal turn-taking. As Aijmer herself states, she uses the two corpora to examine similarities and differences between the L1 and L2 spoken English and encourages use of the differences as a target for remedial classroom work. Aijmer acknowledges the possible effect of the interview format and notes that other types of interaction (e.g., conversation) may have given different results. This acknowledgment, we believe, gives credence to our argument that equating L2 speakers' performance in a subset of spoken English (i.e., in this case the LINDSEI interview) with overall L2 speaking performance can be questionable and worthy of further investigation. Aligning with McCarthy and Carter's (1994) distinction between the kind of interactions that take place in a learner corpus style interview and the same task between two L1 speakers, Crawford (2022: 93) points out that, as examples of dialogic discourse between a learner and an interviewer, LINDSEI is "of limited use for those interested in investigating how learners manage face to face conversations." By any means, this is not to undermine LINDSEI or other similar data. Well-designed corpora, such as LINDSEI, are highly representative of the language used in a given context (Crawford 2022). However, what we are exploring here is the need to be aware of the limitations and variables at play when using the output from interview tasks as a broad representation of spoken learner language.

In the next section, discussing examples from LINDSEI, we exemplify some of the challenges and limitations of using the interview format for data collection and, generally, for the use of interviews as representations of spoken language. We point to an interviewer

effect on the learner data and demonstrate how considerations of the interaction between the interviewer and the interviewee may inform future protocols for corpus design, collection, and analysis of learner data. We noted above the role of smallwords in everyday conversation and their discourse function in co-construction spoken language (cf. Section 2.1). Since among these, adverbs play an important part and perform multiple roles in the discourse, we have chosen to close in on adverb use and its role in interactivity between participants and across different tasks to exemplify our argument.

3. LIMITATIONS IN THE COLLECTION OF LEARNERS' OUTPUT: SOME EXAMPLES

Adverb functions and their roles in spoken communication have been well documented in corpus studies (Carter and McCarthy 2006; Waters 2013; Beeching 2016; Aijmer 2018; Curry *et al.* 2022, among others). In this section, we demonstrate adverb use across the data. For the purposes of this paper, we have extracted three examples using *really, well,* and *maybe* to demonstrate the potential impact of the degree of interviewer/interviewee engagement with the task and the effect on the data produced (see rationale below). We show 1) how varying degrees of interactivity differ in their opportunity for turn-taking, co-construction, and discourse management, 2) how power relations between participants might affect the data collected, and 3) how participants may be struggling to understand the interview genre within this pedagogical interaction. In terms of methodology, we compared three of the LINDSEI learner subcorpora —Spanish, German, and Chinese L1s— alongside a parallel extended version of the L1 English LOCNEC corpus, enlarged with 28 additional interviews (Aguado *et al.* 2012). We arrived at the adverb selection identified above by first using the *Sketch Engine* corpus search tool to extract adverb frequencies using the POS tag for adverbs (RB).[15] The RB tag produced a wide spectrum of forms, not all of which are generally categorised as adverbs (e.g., *yeah, not, n't, but*) in widely considered essential reference grammars (cf. Biber *et al.* 1999 or Carter and McCarthy 2006). We filtered the results using these criteria, resulting in the following highest frequency ranking items: *so,*

[15] https://www.sketchengine.eu/

very, well, just, really, quite, and *maybe*. We then examined the different tasks as separate entities comparing them across datasets and documented the differences in use between each L1 and each task. Alongside statistical tests examining the frequency of the individual adverbs across tasks and L1, we analysed the collocational and colligational patterns of usage for the adverbs across task and L1 and provided a qualitative in-depth view of their functional and positional use.

We first looked at the quantitative differences in adverb contributions between interviewer and interviewee and the degree to which this varied among datasets. We analysed the breakdown of all token counts between the interviewer and interviewee content. In some datasets, the interviewer plays a more (inter)active role than in others. For instance, we found varying degrees of participation from the interviewers, ranging from 19.7 per cent of all tokens in the L1 Chinese data to a 30.5 per cent share in the L1 English data. We then considered the effect of the status of the interviewer on the kind of interactions that took place. In 96.35 per cent of the interviews, the interviewer is an L1 English speaker (sometimes the participants' teacher or their language support assistant). We were also interested in the effect of the interviewer directly on the potential mirroring of linguistic choices from the interviewee. Such considerations may have implications for 1) a valid comparability of the four datasets, and 2) representations of spoken learner language.

The three examples below showcase instances where co-construction of meaning and interaction is not always in place. This approach may seem to favour a conceptualisation of the interview as a data elicitation technique or a method with minimum involvement or implication on the side of the interviewer.

3.1. Example 1: Really *as interactional device*

We analysed the distribution of *really* across the four groups and found that, in our study, *really* was used more frequently by German L1 and English L1 speakers than by Chinese L1 and Spanish L1 speakers. Table 1 shows the raw frequency mean of *really* per speaker per task for the four groups of language.

	Number (individual tasks)	Mean scores	Standard deviation	Standard error
Chinese L1 speakers	159	0.72	1.688	.134
English L1 speakers	228	4.19	5.145	.341
German L1 speakers	149	4.30	4.242	.348
Spanish L1 speakers	144	1.78	3.278	.274
Total/Average	**680**	**2.89**	**4.252**	**.163**

Table 1: Frequency of *really* per speaker and task in the LINDSEI data

German L1 speakers displayed the highest average frequency per speaker and task while English L1 speakers showed the highest standard deviation. Chinese L1 speakers displayed the lowest average frequency across all speakers and tasks. An ANOVA test confirmed that the overall frequency differences were significant (Welch's $F(3, 336.458) = 53.76$, $p = .001$). In the German L1 group, we found a significant difference (post hoc Bonferroni pairwise comparisons) between the picture description task and the set topic ($p = .001$) and between the picture description and the free discussion ($p = .001$). In the L1 Chinese group, post hoc Bonferroni pairwise comparisons revealed that there was no significant difference between the picture description task and the set topic ($p = .194$), the set topic and the free discussion task ($p = .061$), and the free discussion and the picture description ($p = 1.00$). In the L1 English group, post hoc Bonferroni pairwise comparisons showed a significant difference between the picture description task and the set topic ($p = .001$), and the picture description and the free discussion ($p = .001$). No differences were attested between the set topic and the free discussion tasks. In the L1 Spanish group, with a Greenhouse-Geisser correction for sphericity, no significant main effect for task type [$F (1.666, 81.650) = .295$, $p = .705$, partial eta 2 = .006] was found.

As described in Section 2.1, the three tasks in LINDSEI vary in terms of their potential for interactivity, with the first task being predominantly monologic, the second predominantly dialogic, and the third a description of a series of pictures. Figure 2 shows the normalised distribution per one million words for *really* across the four datasets and the three tasks. We note low usage among the Chinese L1 speakers, particularly in task 2 (e.g., CH2), a dip in German L1 use in task 2, high usage in the English L1 group for task 1, and a reduction in tasks 2 and 3. We also note the lowest usage among Spanish speakers, though there is a rise in task 3, which is an opposite trend to what happens in the English L1 group.

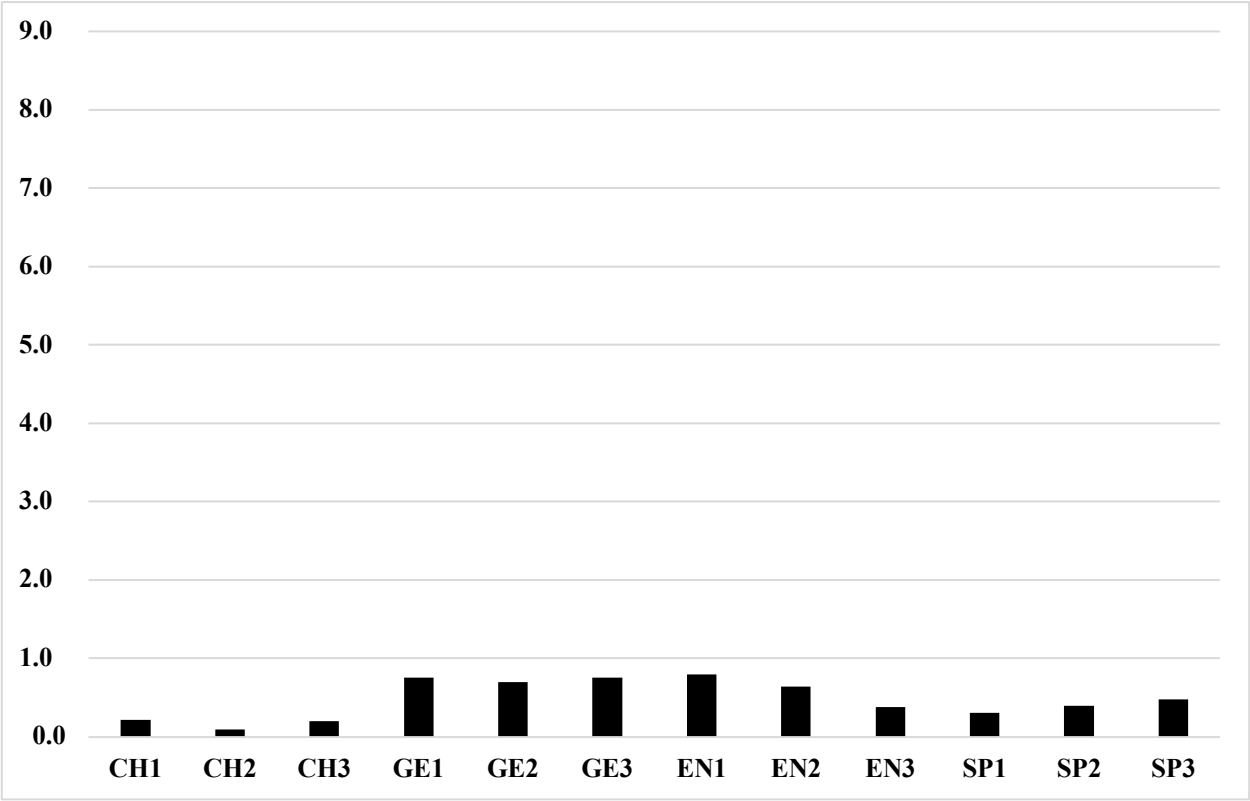


Figure 2: Normalised frequencies (per 1,000 words) of *really* across L1 groups and tasks in the LINDSEI interview

However, an in-depth understanding of the occurrences cannot simply rely on frequency counts of word forms. A qualitative, manual analysis of the functional and positional use of *really* was carried out across the four datasets, using the functional taxonomy illustrated in Table 2.

	Function	Example
1	Booster, emphatic, degree	*It was really good; I really wanted to see you*.
2	Sceptical response	*Really? I think that's unlikely*.
3	Response token	*Oh really*.
4	Factually true, actually	*It doesn't look like she is really*.
5	Hedging	*She doesn't really like it*.
6	Concessive / summative	*It's a bit disappointing really*.

Table 2: Functional categorisation of *really* (after Myers 2010)

In all groups, *really* was used most frequently as a booster in a range of positions (e.g., *It was really good; I really wanted to see you*). Our analysis shows that there are no instances of *really* functioning as sceptical response (function 2) and very few instances of it functioning

as a response token (function 3) in the sampled data. Both functions are highly interactive and a feature of back-channelling behaviour in everyday spoken interaction (O'Keeffe and Adolphs 2008) and yet they are not attested in the LINDSEI interview data. We might attribute their absence to 1) the fact that interviewers and interviewees are not encouraged to interact in the tasks in all the sections of the interview, and 2) the possible effect of the relationship between interviewee and interviewer. This may have arguably impacted turn-taking management. Firstly, the opportunity to interact does not always present itself, the transactional nature of the classroom interview genre being a barrier where the interviewer asks a question, and the interviewee provides relevant information by way of response. Secondly, both functions 2 and 3 might require a degree of contradiction of the interviewer by the interviewee which might not be considered appropriate.

By way of contrast, *really* is found 48,492 times in the Spoken BNC2014. Here it occurs ten per cent of the time as a sceptical response or a response token (functions 2 and 3), in the types of contexts illustrated in Figure 3:

want any more scrambled eggs you can have them	really	? I want you to have them okay because then cos
k it's like a hu three hundred and twenty days of rain	really	? a year oh dear we're talking a wet place no you'
e extender that when I arrived a a it didn't recognise	really	? or maybe it was since the extender but we do ha
er had any problem with it so I'd be even tempted to	really	? okay is that the same make? it's the same one r
sarcastic you have no chance in mine at the moment	really	? I've er I've had a temporary kitchen for the last t
had a temporary kitchen for the last twelve years oh	really	? and now why's it temporary? cos you were plan
k my worktops are pardon? my worktops are full real	really	? ye you know it's just I mean it yeah you know I h
Christmas cake I've still got Christmas cake have you	really	? yes wow it's going on forever it's never will neve
s yeah been trying to get them to come out to see us	really	? yes I mean just really easy just get on the ferry y
s actually submitted to the British Museum mus yeah	really	? mm yeah and and and do you just go out in you
n and he was in that and she said it was terrible mm	really	? oh well you c can't you know you can't win them
mm oh she's downstairs well I'll wait for her then oh	really	? you always and I just wai I just er yeah and I go
eah they they're always saying can we go and see ?	really	? yeah you know and every really? month or som
n we go and see ? really? yeah you know and every	really	? month or something yeah yeah you know she's
ald's he would absolutely yeah love it wolf it down oh	really	? yeah he absolutely loves you know Felix and W

Figure 3: Examples of *really* as a sceptical response or a response token in the Spoken BNC2014

Studies on everyday conversation indicate that response tokens perform functions of listenership, ranging from a continuer function —which maintains the flow of discourse using minimal types of responses (i.e., *mhm, yeah*)— to a more convergent response, using markers of agreement (*oh right, did you*?), or to a more engaged response, such as *really, absolutely, exactl*y (O'Keeffe and Adolphs 2008).

In LINDSEI the lack of engagement between the interviewer and the interviewee is often noticeable through the consistent use of a minimal response such as *mhm*, whose function is to keep the interviewee talking. This is illustrated in extract (1), where Speaker 1 (S1) is the interviewer and Speaker 2 (S2) the interviewee.

Extract (1): CH0104P1

S1: How are you . mm

S2: Fine I'm a little nervous

S1: That's okay . it's not a test . erm so you've chosen a topic

S2: Mm I want choose the topic number one but I wonder

S1: Mm

S2: If . I can change a little bit

S1: Okay

S2: Okay

S1: Mhm

S2: Then I want to talk about my experience in the summer holiday . but . I don't think . it will . teach me a lesson . it just very impressive it make me think a lot

S1: Mhm

S2: Yeah last holiday and I (mm) I stay uni in the university . for . maybe . a month . and because I have a student . and his mother asked me to come here to give her (eh) tutor . tutor jo (eh) ask him ask me . sorry . ask me to teach him . and to improve his English . actually my student is good at the other objects subjects

S1: Mhm

S2: Such as his physics and chemistry's very good but he's (eh) he

In extract (1), there are many opportunities for interaction which are not acted upon by the interviewer. In everyday conversation, S2 might well interpret S1's response as disinterest, as a dispreferred response, but, in this task, the human social element has been removed. In other words, the participants are performing a transaction (McCarthy and Carter 1994) for a pedagogical goal. The only interactional aspect in extract (1) is in the task set-up when S1 asks S2 how they are, to which S2 replies that they are nervous. S1 telling them it's not a test ends the relational interactional element and reverts to the task in hand, marking the transition with *so*. The LINDSEI design does create opportunities for the use of *really* but, in the L2 data, these occur predominantly with a booster or factual meaning function, at a clausal or

phrasal level, rather than at a discourse level with an interactional function. For example, in the Spanish L1 data, 20 per cent of the occurrences use *really* with a factual function (see Table 2), and all of these occur in the third task (the picture description), while 72 per cent of the sampled occurrences were used as a booster, with an equal distribution between phrasal and mid-clause position, as shown in extracts (2) and (3).

Extract (2): SP0107P1
But then if you *really* look into her you can't find anything at all

Extract (3): SP124P2
But I remember it was *really* nice

In summary, if the interviewer does not typically offer anything other than minimal response within the discourse, there are no opportunities for the interviewee to use interactive devices, such as response tokens, as there is nothing to respond to.

3.2. Example 2: Power roles

As has already been discussed, the degree to which the interviewer co-constructs meaning and interacts is variable. In this second example, we do see interactional features emerging in some of the interviews. However, we do not see equal opportunity for their use. Our study shows that use of *maybe* might suggest an interviewer/interviewee relationship effect. For example, the L1 Chinese speakers showed a strong preference for the use of *maybe* to express uncertainty and imprecision. They also used *maybe* to offer possible options or explanations in response to questioning from the interviewer, a use which was predominantly favoured by Spanish and German speakers. This was also favoured in the L1 English sample as a means to give a non-committal response to the interviewer, hence avoiding a contradiction. This is illustrated in extract (4), retrieved from the Chinese data, where S1 is the interviewer and S2 the interviewee. S2 responds to S1's assertion by avoiding a direct disagreement (with the use of *maybe*) while continuing to say the opposite of what S1 has asserted.

Extract (4): CH0115P2

S1: So it's okay to have a little fun as long as you . don't neglect your responsibility

S2: Maybe so . (mm) like a student I should study very hard to . (er) learn more and to . make . my (er) myself more pro like a professor . (uhu)

Extract (5) taken from the German data in part 3 of the interview also demonstrates this non-committal use to avoid disagreement.

Extract (5): GE0144P3

S2: He actually drew her the way she was .. (er) with… all her failures and . (erm) .. blessin no not blessings

S1: . (Erm)

S2: With all failures and

S1: Positive characteristics

S2: Maybe positive characters … or negate the negative aspects as well

However, extract (5) is an example in which the interviewer (S1) does try to co-construct S2's turn with *positive characteristics*. Rather than disagreeing and rejecting the co-construction, S2 prefaces the partial repetition of S1's turn with a *maybe*, followed by a hesitation and then the opposite reframing of S1's contribution *or negate the negative aspects as well*. As pointed out above, examples such as these, which contain interactional discourse such as co-construction, are not common in the LINDSEI L2 data. Where they do occur, they may be indicative of an unequal power role between interviewer and interviewee and the degree to which the interviewer engages in interaction. This is a variable that may not be consistently applied throughout the data collection but one which —as we have discussed above— will have an effect on the language used and the opportunity to interact. Extract (6), below, is an L1:L1 example retrieved from the LOCNEC data from part 1, which is the task designed to be more monologic. The exchange shows a greater degree of co-construction of meaning between the interviewer and the interviewee, resulting in the occurrence of interactive features such as evaluation, back-channelling, and responses tokens.

Extract (6): LOCNEC53P1

S2: I was gonna actually do that for my project but I looked at it and thought no too much no way it scared me

S1: So it it would have been interesting

S2: Yeah it it's been done since apparently but er

S1: Yeah

S2: Yeah I I opted to go for Billy Joel instead

S1: Mhm

S2: He's a lot more down to earth

S1: Mhm

S2: Really film

S1: But so is the whole book written in that language or

S2: Yeah

S1: Even the descriptions and

S2: Yes it's all from his point of view so he's saying oh yeah he was a bolshy with with and I was saying but occasionally he'll give a translation in brackets just one word

S1: Oh that's nice

In summary, if power roles vary from interview to interview and/or from L1 subcorpus to L1 subcorpus, opportunities of use and issues of comparability between data sets arise.

3.3. Example 3: Making sense of the genre

In this example, we show evidence that speakers do not necessarily understand the demands or purpose of the task and how to engage with it, and that this has an effect on the output. We have already seen that, in the L1 data, there is a greater proportion of participation from the interviewer (30%). For example, in this data, the L1 English speakers are more likely to engage with the exchange, for example, by using adverbs for attitudinal effects to express opinion or soften disagreement, trying to make sense of the task as a conversation. We first illustrate this with the use of *well*. While many of the examples of *well* were found to be used by L1 English speakers as a speech management tool —namely, for pausing, reformulating, and introducing a new turn— there were more examples in the sample of *well* used for attitudinal effect than in the other L1 groups. In extract (7), where the speaking roles are occupied by L1 speakers, the interviewee (S2) uses *well* to contradict the interviewer (S1) and soften the following *no*, followed by an explanation softened with *just*. Similarly, in

extract (8), there is a combination of hesitation and *well* to soften the *no,* followed by an explanation, which is hedged by *I mean*:

Extract (7): LOCNEC35P2

S1: What do you do when it rains when it pours

S2: I get wet

S1: Oh so you you you don't take the bus you

S2: No not usually no

S1: Oh it's very brave

S2: Well no I just don't like the waste of time hanging around for the bus and eh hanging around for the bus

Extract (8): LOCNEC55P2

S1: So it's erm since you you want to do forensics I dunno what that's why you you decided to: er do biology or

S2: Em well no er I mean I'm I'm doing biology because that's it's the one subject I've I've always found easy and I enjoy it

In extract (9), the speakers attempt to construct the interview as a conversation. With the use of *really* at the end of the extract, S2 takes up the interviewer's initial question and threads their answer through the discourse with a final summarising answer ending in *so ... really* in order to answer the question.

Extract (9): LOCNEC1P2

S2: Erm I I'm doing a linguistics minor erm as part of er

S1: And what are you doing

S2: Oh actually it's I don't know if it counts as a minor itself it's part of English literature erm

S1: Ah so you're doing literature and you're doing some courses in linguistics

S2: Yes yeah

S1: Mhm

S2: Er just the one in fact er

S1: Just okay

S2: Yeah

S1: Uhu and er why did you choose literature

S2: Erm well

S1: Good question

S2: I I've always been erm very keen on reading

S1: Mhm

S2: And and in my first year I did English literature and language and French **so** there was reading involved in most of my courses **really**

In contrast, extracts (10) and (11) illustrate the interviewees (S2) explicitly referring to the demands of the task.

Extract (10): CH0105P3

S1: Now could you start now

S2: Okay okay see if I can talk for three or five minutes .. okay may I start now

S1: Yeah

S2: Okay . (mm) I'd like to talk about a film I I I have seen

Extract (11): CH0126P3

S2: Shall I make a make up a story or just tell what happen in this picture

S1: Make up a story

S2: okay ...

In summary, there is variability in the way both interviewers and interviewees orient towards the genre. Some speakers struggle to make sense of what is being asked of them, and whether to engage in co-construction of an interactional nature or to pursue a more transactional question and answer approach. The L1:L1 interviews show evidence of orientation towards the interactional opportunities in the task, whereas the L1:L2 interviews appear to orient towards the transactional.

4. Discussion

Under the term 'interview' we find at least two different conceptualisations: 1) an elicitation technique, and 2) a distinct, albeit complex genre. The overlapping of both conceptualisations under the same term may give rise to problems of definition about the nature of the language collected and, therefore, problems of interpretation when assessing the characteristics of spoken learner data. While learner corpus research may have favoured a miner approach (Mann 2011) to spoken data gathering, it may have inadvertently contributed to the

underrepresentation of some substantial sub-genres in spoken communication, such as face-to-face conversation, where discourse co-construction is key. We argue that, to recreate the communicative situation that takes place during a conversation, it is necessary to rethink the way in which spoken data are collected. We suggest that the way the context and task are set up establishes a particular type of 'pedagogical interaction'. Discourse Act theory (Allwood 2000; Bunt 2022) argues that spoken communication is multidimensional and complex, relying on a range of activities that, among others, involve task movement, allo-feedback, turn-management, contact management, discourse structuring, partner communication management, or social obligations management. It is unclear how these dimensions shape communication in L2 interchanges (Bunt 2022) where, as in LINDSEI or TLC, L1 speakers are in charge of some of the time and discourse management dimensions. However, the examples in Section 3.3 have shown that the tasks included in LINDSEI did afford, for example, L1 English speakers' uses of *well* for attitudinal effects. Whether this is the result of power imbalance/balance between speakers of different or the same L1s or a more individual 'chatty' approach to the interactions, some of the tasks may facilitate different approaches to participation as speakers in conversations perform a series of functions such as turn grab, turn keep, or turn release (Bunt 2022). The examples in Sections 3.1, 3.2, and 3.3 suggest that for different speakers the interview may draw on different assumptions about the nature of the task and their self-perceived role in the task (i.e., L1 interviewer vs. learner who, following a request from their lecturer, is taking an examination or has volunteered to take part in an interview in her university). Coming back to the use of *really*, in conversations, feedback may refer to different levels of communication such as attention, perception, understanding, evaluation, or execution. We wonder to what extent learners are comfortable trying to engage with the interviewer to give more than a simple answer to a question. Similarly, we may wonder to what extent interviewers feel comfortable facilitating learners' repetition, co-constructing the discourse, clarifying, offering puzzled faces, nods, or verbal back-channelling, which are common features of real-life spoken interactions. These may create opportunities of use for words such as *really, well*, or *actually*, to name but a few.

The examples in Section 3 demonstrate further methodological challenges in collecting and analysing spoken learner data in general. We suggest that some of these challenges are

related to the specific features found in the LINDSEI data that we have explored in this paper but which, we believe, are generalisable to other spoken learner data:

1. Some of the data are more dialogic than others, which has an effect on the degree to which the language produced is more or less transactional or interactional. This has implications for representativeness and corpus design, for instance, in scoping to what degree the resulting data need to display interactional and/or transactional features.
2. The opportunity for interaction is not equal across different interviews and therefore there may be issues of comparability between datasets.
3. The power roles or relationships between the interviewer and the interviewee are likely to influence the language used.
4. The participants' perception of the task is open to interpretation. Language learners in instructed contexts may be more or less familiar with the 'classroom interview' genre. L1 interviewers may be grappling with the restrictions of the task.
5. The data elicited by the interview as a genre presents a limited representation of spoken language. For research that is interested in spoken learner language (broadly defined), and not solely in the performance of learners in spoken language assessments, other types of learner data need to be added in to the mix to complement interview data and present a more comprehensive picture of spoken learner language.

Such considerations must have direct implication not only in terms of comparability but also in the adequate representation of spoken learner language. Learner corpus research is a relatively new field and, while huge strides have been made in understanding learner language so far, there is always further work to be done.

We turn our attention to future research and ask how we can gather a more inclusive broader representation of language learning and learner language (Pérez-Paredes and Mark 2022), particularly in relation to spoken language. We note that the object of focus of previous learner corpus studies can be categorised into discrete features (e.g., lexis, parts of speech, and tenses), composite features (e.g., measures of lexical sophistication and clausal complexity), and constructs (e.g., metadiscourse features and involvement). We point out

that not all areas receive equal attention, and that according to Paquot and Plonsky (2017), only 30 per cent of studies are concerned with discourse and ten per cent with pragmatics. Currently available learner corpora offer a window on a narrow conceptualisation of language learning and language use, which may explain some of the challenges to making corpora more representative of spoken learner language.

LINDSEI and TLC represent different learner language products. While the former shows spontaneous spoken communication, the latter represents a highly practised language test that is familiar to the students before the interview. In this sense, TLC offers a highly contextualised experience which is mediated by the testing nature of the interview, representing the type of language which is fostered by the testing culture of the certificate awarding institution which, in turn, has a trickle-down effect on the types of tasks carried out in classrooms (McCarthy 2020). The former type of corpus, the one represented by LINDSEI, perhaps offers a wider choice of opportunities in terms of use and interaction not necessarily linked to testing practices, and which may offer a more diverse representation of different types of face-to-face interaction. The range of activities used for the collection of the Spoken BNC2014 (Love 2020) such as chatting about general stuff, talking over lunch, academic colleagues chit-chat over coffee, watching TV, discussing fashion, evening catch-up with housemate, talking at book club, dinner conversation about fixing computer, acquaintances having a chat, family advising, etc. may inspire the design of new tasks that may complement our current findings about spoken L2 use. However, as we have seen, the interviewer —who could facilitate unscripted, spontaneous interaction— is encouraged to be absent or is removed from the analysis, for fear of them getting in the way or influence the learner product. As pointed out by Tracy-Ventura *et al.* (2021: 414), "interaction corpora that consist of conversations between learners, or informal conversations between learners and L1/expert target language speakers, are sorely needed."

There are therefore broad areas where corpus researchers can improve. The first concerns the type of data we are collecting and analysing. As suggested by Friginal *et al.* (2017: 274), future spoken learner corpora may need to address design considerations such as "register coverage [that integrate] more situational contexts, interview questions, and peer response topics or paired activities," and the role and effect of interaction in the collection

process. Methods of data collection play a pivotal role in shaping the authenticity of language samples within learner corpora. To enhance the representation of authentic interaction, we can make the most of new and developing technologies. This involves designing tasks that mirror real-life interactions and reflect power dynamics present in natural discourse across a variety of contexts and learning scenarios. As suggested in Pérez-Paredes and Mark (2022: 323),

> well-designed corpora allow researchers to understand monologic and dialogic communication as they reveal aspects of frequency, collocation, colligation, function and speaker variation that would otherwise remain hidden.

Harnessing technological advancements, such as mobile devices and data collected 'in the wild', can provide unfiltered, unscripted language samples, enabling a more genuine and alternative portrayal of spoken language that offers alternatives to the role of interviewers as data collection managers. One such use of technology has been developed by Knight *et al.* (2021) in the construction of the *National Corpus of Contemporary Welsh* (CorCenCC).[16] As Knight *et al.* (2021: 798) point out, the data are gathered through a mobile crowd-sourcing app which is designed to align methods of collection with the Web 2.0 age, and "enables 'live' user-generated spoken data collection via crowdsourcing." A crowdsourcing approach was also taken in the development of the Spoken BNC2014 (McEnery *et al.* 2017; Love 2020). These advancements offer opportunities to bridge the gap between controlled data collection and the intricacies of unscripted, spontaneous linguistic exchanges. Collection methods may allow us to represent a broader conceptualisation of language use both inside and outside of the classroom. This in turn lends us a more inclusive perspective both on language learning product and learning process. Interviewer-led data could be combined with methods that, while remaining ethical and transparent to language learners, can favour the collection of longitudinal data and different types of interaction that are representative of the wealth of turn management options available in conversations which are not staged as interviews or led by L1 interviewers. However, this leads us into a final area of improvement, delving into the applied pedagogical dimension. How do we truly integrate meaningful social

[16] https://corcencc.org/

interactions into the classroom? Here we call for a return to the seminal question posed by McCarthy and Carter (1994): How effectively do transactionally oriented tasks represent interactive language use and self-presentation in the classroom? Alongside this, as Curry and Mark (2023) discuss, there is a need to consider how spoken language is represented in educational materials and classroom settings and the subsequent circularity of effect this has on the language used in instructional contexts. By way of example, Fung and Carter (2007: 433) have suggested that the frequency of discourse markers in learner English

> reflect[s] the unnatural linguistic input ESL learners are exposed to and the traditional grammar-centred pedagogic focus [on] the literal or propositional (semantic) meanings of words rather than their pragmatic use in spoken language.

Exposure to naturalistic sampling is limited in EFL classroom contexts and more awareness of the spoken register is needed (Mukherjee 2009; Aguado *et al.* 2012). Pérez-Paredes (2019) has suggested the exploitation of annotated spoken pedagogic corpora for secondary school learners to teach pragmatic and lexico-grammatical features of spoken language. EFL textbooks in primary and secondary education levels tend to under-represent spoken interaction (Curry and Mark 2023). In many mainstream coursebooks, task needs are blurred under the generic vague heading ‘speaking’, which often involves monologues and can range from role play to discussion, opinion giving, or response to input. Dialogues on the page may be used to present grammatical or lexical content, rather than attending to dialogic features. Carter and McCarthy (2017) point out that even though spoken grammar has come of age, it is still under the influence of a pedagogy derived from written language. They offer a host of suggestions for future exploration of spoken language including “increased exploitation of spoken learner corpora” as well as “a challenge to the ways in which grammatical and discourse patterns and socio-cultural context are captured” (Carter and McCarthy 2017: 11).

5. Conclusion

In this paper, we have highlighted some of the challenges in using the interview as a research tool for the collection of learner spoken data, in an attempt to learn from data collection and analysis thus far and to provide a platform for future work. Our paper contributes to increasing methodological reflection in applied linguistics and CL in what McKinley and

Rose (2017) have identified as a need for researchers to create the spaces in which to discuss not only results but also the methodological considerations that affect their praxis. In spoken corpus research, the design and the construction of corpora, as well as the vagaries of recording, transcription, coding, and marking of spoken data, have received considerable focus (Knight and Adolphs 2022). However, a discussion of the unrealised potential of what the analysis of existing spoken corpora can offer researchers —in the way of insights into the collection and investigation of future learner corpora— is of continued interest and relevance.

Some of the main takeaway messages from our research are the following:

1. The terms 'interview' and 'spoken learner language' are perhaps too broad and may give rise to a variety of conceptualisations about the nature of the learner data collected. Interpretations of existing corpus data need to consider the complexity involved in defining the nature of interviews (McCarthy and Carter 1994), the nature of the tasks, and the degree of interaction and involvement present.
2. We may be in danger of drawing erroneous conclusions about learner proficiency and the ability to use or not use the interactional features that are characteristic of spoken language when, in reality, the opportunity to do so does not present itself. The interview as a task may not provide the learner the opportunity to use their spoken repertoire.
3. The inconsistency of the occurrence of interaction between the participants in the interview is key to the learner data quality, the resulting genre, and comparability issues with overlapping subgenres across learner corpus research. We have examined the use of a selection of adverbs of relevance in spoken communication in some subsets of LINDSEI and the extended LOCNEC, showing the potential impact of interviewer/interviewee engagement with the task. The role of interviewers is key in terms of the quality and the nature of the collected data. More attention to and awareness of their influence on the resulting data is needed in learner corpus research.
4. We argue that data collected through interview tasks is not fully representative of spoken learner language as discussed. Future spoken corpora will need to explore avenues to represent oral interaction that complement the existing interviewer-led

collection methods, particularly in non-testing situations. This will undoubtedly benefit our insights into the nature of language use, language acquisition (Tracy-Ventura and Myles 2015; Tracy-Ventura *et al.* 2021), and conversational pragmatics (Bunt 2022), and may inform L2 pedagogy with interactional data that may contribute to represent a wider range of sites of L2 learner engagement in language learning and teaching (Carter and McCarthy 2017; Tyler and Ortega 2018).

REFERENCES

Aguado-Jiménez, Pilar, Pascual Pérez-Paredes and Purificación Sánchez. 2012. Exploring the use of multidimensional analysis of learner language to promote register awareness. *System* 40/1: 90–103.

Aijmer, Karin. 2018. Intensification with very, really and so in selected varieties of English. In Sebastian Hoffmann, Andrea Sand, Sabine Arndt-Lappe and Lisa Marie Dillmann eds. *Corpora and Lexis*. Leiden. Rodopi, 106–139.

Allwood, Jens. 2000. An activity-based approach to pragmatics. In Harry Bunt and William Black eds. *Abduction, Belief and Context in Dialogue: Studies in Computational Pragmatics*. Amsterdam: John Benjamins, 47–80.

Anderwald, Lieselotte and Susanne Wagner. 2007. The Freiburg English Dialect Corpus: Applying corpus-linguistic research tools to the analysis of dialect data. In John C. Beal, Karen P. Corrigan and Hermann L. Moisl eds. *Creating and Digitizing Language Corpora Volume 1: Synchronic Databases.* London: Palgrave Macmillan, 35–53.

Beeching, Kate. 2016. *Pragmatic Markers in British English: Meaning in Social Interaction*. Cambridge: Cambridge University Press.

Bell, Philippa, Laura Collins and Emma Marsden. 2021. Building an oral and written learner corpus of a school programme: Methodological issues. In Bert Le Bryn and Magali Paquot eds. *Learner Corpus Research Meets Second Language Acquisition*. Cambridge: Cambridge University Press, 214–242.

Bell, Philippa and Caroline Payant. 2021. Designing learner corpora: Collection, transcription, and annotation. In Nicole Tracy-Ventura and Magali Paquot eds, 53–67.

Biber, Douglas. 1995. *Dimensions of Register Variation: A Cross-Linguistic Comparison*. Cambridge: Cambridge University Press.

Biber, Douglas, Susan Conrad, Randi Reppen, Pat Byrd, Marie Helt, Victoria Clark, Viviana Cortes, Eniko Csomay and Alfredo Urzua. 2004. *Representing Language Use in the University: Analysis of the TOEFFL 2000 Spoken and Written Academic Language Corpus*. Princeton: Educational Testing Service.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English.* Harlow: Longman.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 2021. *Grammar of Spoken and Written English*. Amsterdam: John Benjamins.

Bunt, Harry. 2022. The multifunctionality of utterances in interactive discourse. In Zihan Yin and Elaine Vine eds. *Multifunctionality in English: Corpora, Language and Academic Literacy Pedagogy*. London: Routledge, 11–29.

Carter, Ronald and Michael McCarthy. 2006. *Cambridge Grammar of English: A Comprehensive Guide*. Cambridge: Cambridge University Press.

Carter, Ronald and Michael McCarthy. 2017. Spoken grammar: Where are we and where are we going? *Applied Linguistics* 38/1: 1–20.

Castello, Erik. 2023. Stance adverbials in spoken English interactions: Insights from corpora of L1 and L2 elicited conversations. *Contrastive Pragmatics* 4/2: 243–273

Cohen, Louis, Lawrence Manion and Keith Morrison. 2017. *Research Methods in Education*. New York: Routledge.

Crawford, William J. 2022. Corpora and speaking skills. In Reka R. Jablonkai and Eniko Csomay eds. *The Routledge Handbook of Corpora and English Language Teaching and Learning*. New York: Routledge, 89–101.

Curry, Niall, Robbie Love and Olivia Goodman. 2022. Adverbs on the move: Investigating publisher application of corpus research on recent language change to ELT coursebook development. *Corpora* 17/1: 1–38.

Curry, Niall and Geraldine Mark. 2023. Using corpus linguistics in materials development and teacher education. *Second Language Teacher Education* 22: 187–208.

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

Friginal, Erik, Joseph J. Lee, Brittany Polat and Audrey Roberson. 2017. *Exploring Spoken English Learner Language Using Corpora: Learner Talk*. London: Springer.

Friginal, Eric and Brittany Polat. 2015. Linguistic dimensions of learner speech in English interviews. *Corpus Linguistics Research* 1: 53–82.

Fung, Loretta and Ronald Carter. 2007. Discourse markers and spoken English: Native and learner use in pedagogic settings. *Applied Linguistics* 28/3: 410–439.

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. 2021. Learner corpora. In Magali Paquot and Stefan Th. Gries eds, 283–303.

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.

Gráf, Tomáš. 2017. *The Story of the Learner Corpus LINDSEI CZ*. Karlova: Univerzita Karlova, Filozofická fakulta. https://dspace.cuni.cz/bitstream/handle/20.500.11956/97524/1541592_tomas_graf_22-35.pdf?sequence=1&isAllowed=y

Gut, Ulrike. 2012. The LeaP corpus: A multilingual corpus of spoken learner German and learner English. In Thomas Schmidt and Kai Wörmer eds. *Multilingual Corpora and Multilingual Corpus Analysis.* Amsterdam: John Benjamins, 3–24.

Hanks, Elizabeth, Tony McEnery, Jesse Egbert, Tove Larsson, Douglas Biber, Randi Reppen, Paul Baker, Vaclav Brezina, Gavin Brookes, Isobelle Clarke and Raffaella Bottini. 2024. Building LANA-CASE, a spoken corpus of American English

conversation: Challenges and innovations in corpus compilation. *Research in Corpus Linguistics* 12/2: 24–44.

Hasselgreen, Angela. 2004. *Testing the Spoken English of Young Norwegians: A study of Test Validity and the Role of 'Smallwords' in Contributing to Pupils' Fluency*. Cambridge: Cambridge University Press.

Ishikawa, Shin'ichi. 2019. The ICNALE spoken dialogue: A new dataset for the study of Asian learners' performance in L2 English interviews. *English Teaching* 74/4: 153–177.

Jones, Christian. 2022. What are the basics of analysing a corpus? In Anne O'Keeffe and Michael McCarthy eds, 126–139.

Knight, Dawn and Svenja Adolphs. 2022. Building a spoken corpus? In Anne O'Keeffe and Michael McCarthy eds, 21–34.

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 and Evaluation* 55/1: 789–816.

Koester, Almut. 2022. Building small specialised corpora. In Anne O'Keeffe and Michael McCarthy eds, 48–61.

Larsson, Tove, Tony Berber Sardinha, Bettany Gray and Douglas Biber. 2023. Exploring early L2 writing development through the lens of grammatical complexity. *Applied Corpus Linguistics* 3/3: 100077. https://doi.org/10.1016/j.acorp.2023.100077

Lee, David. 2002. Genres, registers, text types, domains and styles: Clarifying the concepts and navigating a path through the BNC jungle. In Bernhard Kettemann and Georg Marko eds. *Teaching and Learning by Doing Corpus Analysis: Proceedings of the Fourth International Conference on Teaching and Language Corpora.* Leiden: Rodopi, 245–292.

Love, Robbie. 2020. *Overcoming Challenges in Corpus Construction: The Spoken British National Corpus 2014*. London: Routledge.

Mann, Steve. 2011. A critical review of qualitative interviews in applied linguistics. *Applied Linguistics* 32/1: 6–24.

McCarthy, Michael. 2010. Spoken fluency revisited. *English Profile Journal* 1. https://doi.org/10.1017/S2041536210000012.

McCarthy, Michael. 2020. *Innovations and Challenges in Grammar.* London: Routledge.

McCarthy, Michael and Ronald Carter. 1994. *Language as Discourse: Perspectives for Language Teaching*. Routledge: London.

McEnery, Tony, Robbie Love and Vaclav Brezina. 2017. Compiling and analysing the Spoken British National Corpus 2014. *International Journal of Corpus Linguistics* 22/3: 311–318.

McKinley, Jim and Heath Rose eds. 2017. *Doing Research in Applied Linguistics: Realities, Dilemmas and Solutions*. London: Routledge.

Mukherjee, Joybrato. 2009. The grammar of conversation in advanced spoken learner English. In Karin Aijmer ed. *Corpora and Language Teaching*. Amsterdam: John Benjamins, 203–230.

O'Keeffe, Anne and Svenja Adolphs. 2008. Response tokens in British and Irish discourse. In Klaus P. Schneider and Anne Barron eds. *Variational Pragmatics*: *A Focus on Regional Varieties in Pluricenctric Languages.* Amsterdam: John Benjamins, 69–98.

O'Keeffe, Anne and Michael McCarthy eds. 2022. *The Routledge Handbook of Corpus Linguistics*. London: Routledge.
Myers, Greg. 2010. Stance-taking and public discussion in blogs. *Critical Discourse Studies* 7/4: 263–275.
Paquot, Magali and Stefan Th. Gries eds. 2021. *A Practical Handbook of Corpus Linguistics*. New York: Springer International Publishing.
Paquot, Magali and Luke Plonsky. 2017. Quantitative research methods and study quality in learner corpus research. *International Journal of Learner Corpus Research* 3/1: 61–94.
Pérez-Paredes, Pascual. 2019. The pedagogic advantage of teenage corpora for secondary school learners. In Peter Crosthwaite ed. *Data Driven Learning for the Next Generation: Corpora and DDL for Pre-tertiary Learners*. London: Routledge, 67–87.
Pérez-Paredes, Pascual and Geraldine Mark. 2022. What can corpora tell us about language learning? In Anne O'Keeffe and Michael McCarthy eds, 312–327.
Pérez-Paredes, Pascual and María Sánchez-Torne. 2019. The linguistic dimension of L2 interviews: A multidimensional analysis of native speaker language. *Focus on ELT Journal* 1/1: 4-26.
Stubbs, Michael. 2007. On texts, corpora and models of language. In Michael Hoey, Michaela Malhberg, Michael Stubbs and Wolfgang Teubert eds. *Text, Discourse and Corpora: Theory and Analysis.* London: Bloomsbury, 127–161.
Tracy-Ventura, Nicole and Florence Myles. 2015. The importance of task variability in the design of learner corpora for SLA research. *International Journal of Learner Corpus Research* 1/1: 58–95.
Tracy-Ventura, Nicole and Magali Paquot eds. 2021. *The Routledge Handbook of Second Language Acquisition and Corpora*. London: Routledge.
Tracy-Ventura, Nicole, Magali Paquot and Florence Myles. 2021. The future of corpora in SLA. In Nicole Tracy-Ventura and Magali Paquot eds, 409–424.
Tyler, Andrea and Lourdes Ortega. 2018. Usage-inspired L2 instruction: An emergent, researched pedagogy. In Andrea Tyler, Lourdes Ortega, Mariko Uno and Hae In Park eds. *Usage-Inspired L2 Instruction : Researched Pedagogy.* Amsterdam: John Benjamins, 3–26.
Waters, Cathleen. 2013. Transatlantic variation in English adverb placement. *Language Variation and Change* 25/2: 179–200.

Corresponding author
Pascual Pérez-Paredes
University of Murcia
Department of English Philology
Plaza de la Universidad s/n
30001 Murcia
Spain
E-mail: pascualf@um.es

received: August 2023
accepted: June 2024

RiCL Research in Corpus Linguistics

Developing a coding scheme for annotating opinion statements in L2 interactive spoken English with application for language teaching and assessment

Yejin Jung – Dana Gablasova – Vaclav Brezina – Hanna Schmück
Lancaster University / United Kingdom

Abstract – Evaluative meanings are known to be difficult to identify and quantify in corpus data (Hunston 2004). The research in this area has largely drawn on the annotating schemes offered by the frameworks of Appraisal (Read and Carroll 2012; Fuoli 2018) or stance (Simaki *et al.* 2019). However, these annotation schemes have been applied predominantly to written production and to first language use. This study, therefore, proposes an annotation scheme for identifying and classifying linguistic expressions of opinion with particular application for second language (L2) language teaching and language assessment contexts. In addition, the coding scheme also specifically deals with spoken interactive communication, with particular attention paid to aspects such as the co-construction of opinion statements (Hovarth and Eggins 1995). The paper outlines the components of the coding scheme along with their theoretical underpinning, addresses some of the challenges in applying the codes and annotating real-life data, and discusses future possibilities and considerations related to the application of the coding scheme.

Keywords – evaluative language; linguistic expression of opinion; coding scheme; L2 pragmatic ability; spoken corpora; learner corpora

1. INTRODUCTION[1]

Expressing evaluative meanings, defined as language which serves to express "a speaker's attitude, stance, viewpoint, or feelings on entities or propositions" (Hunston and Thompson 2000: 5), is an integral part of human communication. Indeed, as Thompson and Alba-Juez (2014: 5) point out, "finding a text or even a sentence without any trace of evaluation is a very challenging, if not impossible task." Evaluative language has been analysed in different genres, such as legal communication (Goźdź-Roszkowski 2018), academic writing (Jiang and Hyland 2015), and media discourse (Bednarek 2006),

[1] We would like to thank the ESRC Centre for Corpus Approaches to Social Science (CASS) at Lancaster University and the Trinity College London, for permitting data access for this study. We are also very grateful to the anonymous reviewers and the editors, Robbie Love and Carlos Prado-Alonso, for their valuable comments.

and across different modes of communication (written, spoken, and online), see Greenberg (2000) or Mullan (2010). Due to its central role in communication, the ability to express views also features prominently in the domains of language teaching and assessment, with a number of courses, textbooks and exams highlighting this language function. For example, the ability to state and support an opinion is included as an indicator of L2 communicative competence in widely used language proficiency frameworks such as the *Common European Framework for Reference* (Council of Europe 2020) and the *American Council on the Teaching of Foreign Languages* oral proficiency guidelines (ACTFL 2024). In addition to this, many standardised language proficiency tests —such as the *International English Language Testing System* (IELTS 2019), the *Graded Examinations in Spoken English* (GESE; Trinity College London 2024), the *Test of English as a Foreign Language* (TOEFL; Educational Testing Service 2018)— ask test takers to express opinions in order to evaluate their linguistic ability.

Linguistic expression of opinion represents a complex language function, which draws on interaction of social, cognitive and linguistic resources, making it challenging to identify and classify different types of opinion statements. Several coding frameworks have been developed (e.g., Martin and White 2005; Wiebe *et al.* 2005; Gray and Biber 2012) to operationalise different types of evaluative language, including expressions of opinion. Building on this research, the current study proposes an annotation scheme for the identification and classification of linguistic expressions of opinion with particular application in second language (L2) teaching and assessment contexts. Such research is crucial for setting curricular goals in the teaching of communicative skills as well as when assessing different stages of communicative and interactional competence in language tests (Roever 2011; Galaczi 2014). In addition to contributing to a better understanding of L2 pragmatic ability, the coding scheme also addresses interactive spoken communication, a genre characterised by frequent exchange of views between interlocutors (Biber *et al.* 2021). This study, therefore, seeks to develop a coding scheme that is applicable in contexts characterised by a high degree of turn-taking and co-construction of discourse between two or more interlocutors. To reflect these aims, the data used in the study were selected from the *Trinity Lancaster Corpus of spoken interactive L2 English* (TLC; Gablasova *et al.* 2019). The study will first present the broader framework and rationale for the coding scheme and then introduce the specific components of the scheme along with examples of the coding. It will next focus on an

empirical evaluation of the coding scheme and discuss the challenges of applying it to L2 data and interactive spoken production.

2. Theoretical framework for the annotation scheme

2.1. Principles of defining linguistic expressions of opinion

As a first step in the coding scheme development, it is necessary to identify the major principles for defining opinion statements, which will then serve as a general framework for the annotation scheme. The construct of evaluative language has been investigated and/or operationalised in a number of widely-used frameworks, representing different theoretical and methodological approaches. For example, the Appraisal framework (Martin and White 2005), based on Systematic Functional Linguistics, has been influential in identifying and categorising different types of evaluative (emotive and attitudinal) statements, which include evaluation of people's characteristics and the qualities of objects/entities. In addition to Appraisal, Wiebe *et al.* (2005) created a coding scheme for application in an NLP context which focuses on automatic identification of internal mental and emotional states such as opinions, beliefs, thoughts, emotions, sentiments, and speculations, with particular attention paid to the classification of intensity and polarity of these attitudes. Using corpus linguistic methodology, research on stance such as Gray and Biber (2012) categorised different types of speaker position adopted towards a statement or an entity, with a particular focus on the lexical and grammatical resources used for indexing perspective. Beyond this, Hyland (1998, 2005) developed a categorisation for different expressions indicating different types and degrees of speaker/writer stance and engagement (e.g., hedges and boosters).

Drawing on this body of research from different fields, several principles for defining opinion and distinguishing it from other types of evaluative language can be formulated. First, when investigating language related to expressing opinions, it is important to distinguish between two phenomena (Bednarek 2009): 1) an expression of opinion that refers to the psychological reality of forming and expressing an evaluative judgment, and 2) the linguistic expression of opinion by speakers/writers. While acknowledging the complexity of cognitive, psychological, social, and linguistic processes involved in the formation and expression of opinions, this study focuses only on the second category.

Second, in order to identify an instance of evaluative language, previous frameworks relied on a combination of two key approaches: 1) the presence of explicit linguistic expressions that mark values, subjectivity, and stance, and 2) contextual clues. For example, Wiebe *et al.* (2005) detect private states using three types of linguistic markers: explicit mentions of private states, speech events expressing private states, and expressive subjective elements. The markers that have been closely related to evaluative language include value-laden words (e.g*., great, horrible*) and stance markers indicating (un)certainty towards a proposition (e.g., *maybe, really*), see Biber *et al.* (1999). While the presence of such words may provide explicit clues to the occurrence of evaluative statements, evaluative judgments can be also implied via other words and contexts (see Section 3.2.1 for examples).

Finally, it is necessary to distinguish expressions of opinion from other related types of evaluative language, especially that of linguistic expression of affect, since the two constructs have been interlinked, to some degree, in several coding frameworks (e.g., Martin and White 2005; Wiebe *et al*. 2005). These approaches reflect the complex relationship between emotions and opinions, and the fact that many evaluative statements (e.g., *Violeta is a fine person* or *I like Violeta*) can be linked to a speaker's positive/negative emotions toward an entity/proposition. However, a distinctive feature of an opinion statement is that it is not a purely emotional reaction to an entity/proposition, but it also involves a cognitive process (Bednarek 2006). It has been argued that an opinion involves a value judgment, which entails an (implicit or explicit) comparison between the object of evaluation and a norm (Labov 1972; Martin and White 2005). For example, the evaluative statement *Violeta is a fine person* involves the process of comparing Violeta against the normative principles of 'being fine'. On the contrary, an emotional state or process does not necessarily involve a comparison against (implied or perceived) norms. As a result, statements which include an explicit reference to what is traditionally considered an emotion (e.g., *I love it*), see Bednarek (2008) or Mackenzie and Alba-Juez (2019), can be distinguished from the expression of opinion, although it is acknowledged that a linguistic expression of opinion can be based on both a cognitive and an emotive response to the entity/proposition that is being evaluated. Examples (1)–(3) illustrate the type of evaluative statements that include linguistic cues (underlined) indexing emotion, which were thus excluded from the coding of opinion statements. All examples in this paper are taken from the *Trinity Lancaster Corpus* (TLC), described in

Section 4.1. The individual ID of each text is given in the square brackets after the extract. In the transcripts, Speaker 1 (S1) always denotes an L1 speaker, while Speaker 2 (S2) refers to an L2 speaker.

(1) S2: Venice is my favourite city [6_SP_1]

(2) S2: and and I I love the culture of New York because there's a lot of s= every of people [2_7_SP_47]

(3) S2: I would like to design modern architecture because I like how the architects er er play with er geometric shapes and forms [2_7_SP_8]

2.2. Defining linguistic expression of opinion in language teaching and assessment contexts

The ability to express opinions is a key part of L2 users' communicative and interactional competence, reflecting the stage of their linguistic/pragmatic development (Galaczi 2014). This ability, therefore, represents a central concern in many language teaching and testing contexts with a focus on L2 communicative strategies. For example, the task specification for the spoken component in the IELTS exam states that "the ability to communicate opinions and information on everyday topics and common experiences and situations" is one of the main skills assessed in the task (IELTS 2019: 5). In the spoken part of the Aptis General exam, "the candidate gives opinions and provides reasons and explanations" (O'Sullivan and Dunlea 2015: 22). In the GESE test, the test takers are expected to "communicate facts, ideas, opinions and attitudes about a chosen topic" (Trinity College London 2024: 38). In these speaking tests, L2 users are typically asked to state and support their opinions as well as engage with the opinions expressed by other interlocutors.

Despite the role played by this particular language function (i.e., the expression of opinion), there seems to be only a limited body of research on how to reliably identify and evaluate opinion statements in these contexts. Previous studies addressed different aspects of evaluative language use by L2 speakers such as stance-taking and expressions of (dis)agreement (Iwasaki 2009; Fordyce 2014; Galaczi 2014; Bardovi-Harlig *et al.* 2015; Gablasova *et al.* 2017; Fogal 2019; Pérez-Paredes and Bueno-Alastuey 2019). However, this research mostly focused on forms related to linguistic evaluation and provided only a limited insight into the nature of communicative strategies related to opinion stating by L2 speakers at different proficiency levels. The annotating scheme

presented in this study has therefore been developed to enable researchers and practitioners to capture and assess different aspects of communicative strategies employed by L2 speakers in interactive communication. In particular, it allows researchers to 1) measure the frequency of opinion statements expressed by the speakers, 2) measure the complexity of their expressions of opinion, and 3) record whether L2 speakers are able to state their opinions independently in the course of interaction or whether additional support from the teacher/examiner may be required to elicit the views. The components in this scheme allow researchers to distinguish, for example, L2 speakers who may have the linguistic resources to formulate a simple opinion statement, but who may not have appropriate pragmatic knowledge to manage intersubjective relations (e.g., employ politeness strategies to mitigate the impact of their expressed views) or who may struggle to express their opinions in an ongoing, fast-paced conversation. Such findings may support previous research outcomes that showed that highly proficient L2 speakers engage in opinion exchange in more collaborative and reciprocal manners than intermediate-level speakers (Galaczi 2014).

3. CODING SCHEME

3.1. General approach to the annotation methodology

The main aim of this paper is to introduce and evaluate a scheme that pays special attention to L2 production in an interactive context, with a direct application in language teaching and assessment. The current scheme builds on previous research on annotation of evaluative language and linguistic expression of opinion, while taking into consideration the specific pedagogical and assessment concerns discussed above. To reflect this purpose, the scheme addresses three dimensions: 1) stating of an opinion, 2) providing a support for the opinion statement, and 3) the interactional pattern in which the opinion statement/support was produced.

3.2. Annotation categories

3.2.1. Opinion statements

A linguistic expression of speaker opinion, or an opinion statement, is defined in this study as a speech act/language function that informs the listener about a speaker's opinion

expecting no specific response or action (as opposed to a question or a directive), see Biber *et al.* (2002). To code an utterance as an opinion statement, the following three conditions must be satisfied:

1. An opinion statement involves a marker or markers of social values, subjectivity, or comparison (Hunston and Thompson 2000).
2. An opinion statement is attributed to the speaker, not someone/something else (Sinclair 1986; Hunston 2000).
3. An opinion statement is syntactically declarative (Biber *et al.* 2002).

First, an opinion statement should include a value judgement on the object or situation being evaluated; that is, it includes an evaluation in terms of good/bad, positive/negative, important/unimportant, reliable/unreliable, certain/uncertain (Hunston and Thompson 2000; Bednarek 2006). An opinion statement thus must contain at least one grammatical or lexical item that indexes values (e.g., *great*), subjectivity (e.g., *my*) or comparison (e.g., *just*, *never*), see Hunston and Thompson (2000). There is no exhaustive list of such items and thus the interpretation may vary depending on the context. While some lexical and grammatical items may be more explicitly evaluative in meaning (e.g., *best*, *beautiful*), others may be more context-dependent such as modal verbs (e.g., *may* and *could*). (Bednarek 2006; Thompson and Alba-Juez 2014). Example (4) demonstrates the use of the word *special*, which in some contexts can serve an evaluative purpose (in the sense of 'extraordinary' or 'remarkable'), but in the context of this example fulfils a descriptive function without implying a judgement (i.e., it refers to something with a specialised purpose).

(4) S2: in my free time I work in a shop of jumping clay […] okay jumping clay is a special clay which you can mould and create a lot of things like objects for the kitchens er would you like to see something made with jumping clay? [2_7_SP_4]

Example (4) illustrates the fact that the presence of a lexical item by itself may not be sufficient to identify an opinion statement and that the context of the utterance needs to be analysed for further evidence of evaluation, such as markers of comparison to an implied norm or a reference point.

Second, an opinion is inherently personal (Myers 2004) and, therefore, to distinguish an expression of opinion from other types of (evaluative) statements, only the evaluative statements whose communicative context allows them to be attributed to the

speaker as the source of the judgement will be considered an opinion statement. There are various ways of (not) indicating that a value judgment is from the speakers themselves, described using the constructs of 'evidentiality' (Chafe and Nichols 1986), 'subjectivity' (Traugott 2010), or 'epistemic stance' (Biber *et al.* 1999). For example, speakers may mark the source of information (e.g., *according to x*) when making an evaluative comment to indicate how certain the speaker is about the truth or validity of the comment (Biber *et al.* 1999). Speakers may also choose to adjust the extent to which they appear responsible for their value judgment by attributing it to themselves (e.g.,*from my perspective*) or to other sources (e.g., *the government said*), see Hunston (2000) or Sinclair (1986). Thus, if a speaker makes an evaluative statement but attributes the value judgement or perspective to a different entity, the comment may be classified as an expression of value, but not as a linguistic expression of the speaker's opinion (e.g., *Spanish people think that the new law has a positive influence*). A range of linguistic markers can be used to determine whether a statement satisfies the second criterion and whether the speakers are expressing their own opinion rather than reporting views attributable to another source. For example, these are linguistic items such as first-person pronoun (*I*, *my*) and/or epistemic stance markers (e.g., *I think*, *I believe*, *from my perspective* and *in my opinion*). When the statement explicitly identifies a source of view other than the speaker, it is not coded as an opinion statement. Examples (5)–(6) demonstrate such use, with the source of the view underlined.

(5) S2: they say is really difficult to pass [2_7_SP_8]

(6) S2: oh er in Galicia it's er it's said that er recycling's not er an important question [2_7_SP_49]

However, it should be also noted that it is relatively common for speakers not to explicitly indicate the source of information. Without any explicit marking, the opinion statement is therefore attributed to the speaker. This applies also in the cases when the opinion statement occurs as a direct response to the other interlocutor's question/request for an opinion as in such situations, the speakers are considered to express their own opinions.

Finally, the last condition requires the opinion statement to be realised in the form of a declarative sentence (Biber *et al.* 2002). As a result, the expression of opinion potentially implied in different types of questions will not be coded as opinion statements.

Examples (7)–(9) illustrate such uses in different types of yes-no questions (7), wh-questions (8), and tag questions (9).

(7) S2: Do you think advertising is necessary? [SP_112]

(8) S2: What do you think of advertising? [SP_108]

(9) S2: Lovely isn't it? [2_6_SP_66]

In addition, an evaluative comment realised in a directive/imperative manner (e.g., *be careful when you buy clothes*) is not considered as an expression of opinion.

3.2.2. Opinion supporting statements

A supporting statement is defined as a statement that provides a supportive or background information for an opinion statement. A speaker may make a supportive move in consideration or anticipation of a listener's response (Edmondson 1981). The discourse function of these supporting statements is two-fold. First, they provide additional information about the nature of the opinion statement and, second, they play a role in the intersubjective and interactional dimension of the communication. For example, they can serve as face-saving devices used to mitigate the social impact of an opinion statement (e.g., to weaken or strengthen it), see Blum-Kulka *et al.* (1989).

In the current study, two conditions are required for a statement to qualify as a supporting statement. First, it has to satisfy the condition of providing additional information about the opinion statement that preceded it. Second, a functional link between the two statements —the opinion statement and the supporting statement— has to be identifiable (e.g., from the presence of an explicit marker or contextual clues). It is also possible for a supporting statement to contain an opinion statement. However, if the two conditions listed above are fulfilled (i.e., the second opinion statement provides additional information about the preceding statement and there is evidence of a functional link between the statements), the second statement will be coded as a supporting statement.

Based on the previous literature (Carlson and Marcu 2001; Galaczi 2014) and on a small-scale grounded analysis of a sample of the data, five main types of supporting statements have been included in the coding scheme. These are supporting statements

expressing 1) reason, 2) elaboration, 3) contrast, and 4) evidence. The fifth category, 'other', is used where none of the previous types of supporting statement can be applied.

The category of 'reason' involves supporting statements in which the speaker provides a cause, motivation, or background for the opinion previously stated. Many supporting statements in this category are explicitly marked by the use of *because*, as illustrated in example (10).

(10) S1: yeah okay so it was list A designer goods okay so what do you think about erm designers who spend a lot of energy and time into designing beautiful expensive clothes and then having them copied and sold in the streets?

S2: like fakes

S1: fake yeah mm

S2: I think that it's not fair because someone is spending a lot of time doing an exclusive product for people who can afford it and then if er if other persons make fake of tho-those product they increase at no er the value of those first er products will become lower [2_SP_1]

The second major category, 'elaboration', involves the speaker providing additional information, or specific details about the opinion statement. In this category, the links between the information included in the opinion statement and the supporting statement can reflect different relationships (e.g., general-specific, whole-part, or object-attribute). This type of a supporting statement usually provides an example of what is stated in the opinion statement, as shown in example (11), or paraphrases it with the aid of different lexical items, as in example (12).

(11) S1: ah yeah yeah and what about the alphabet?

S2: er that's very difficult

S1: yeah

S2: they have for example four As [6_SP_31]

(12) S1: you have to trust the site yeah

S2: yes it's er er secure like buying on Zara or shops like that you know it's safe to do it but not er in a strange er shop [2_SP_5]

If the supporting statement involves ‘contrast’, the third major category, it includes a statement that contradicts the opinion statement or offers a different perspective on the evaluative judgement from the opinion statement, as demonstrated in example (13).

(13) S1: she had a little bit of fat on her knee so she had an operation on her knees

S2: I think that’s silly but I think <unclear text="this"/> kind of people have a lot of pressure and because they are all the time er in the media and everyone is looking at them I think they want to feel that I’m perfect [2_SP_9]

The next category involves providing ‘evidence’; in these cases, the supporting statement indicates the source of information that the evaluation in the opinion statement is based on. Evidence differs from ‘reason’ in that it explicitly states external evidence such as data, sources, or numbers that demonstrate the validity of the statement. This use is shown in example (14).

(14) S2: because alcohol really damage your health it’s prove by medicine [SP_107]

Finally, the supporting statements that do not fall into these four major categories are coded as ‘other’. These include statements with some evidence that the speaker is attempting to provide a support for an opinion statement, but it is impossible to determine the nature of the support (e.g., as the statement remained incomplete). Reasons for incomplete supporting statements include interruption from the other interlocutor who then shifts the interaction to a different topic, as illustrated in example (15).

(15) S1: It’s very difficult to explain this question because er it's e-er-ex=

S2: because

S1: <voc desc="laugh"/> <unclear text="bus driver"/>

S2: to to to for me the this situation is out of control the military are going to come in [2_SP_32

Following the definition of an opinion statement (see Section 3.2.1) and a supporting statement (Section 3.2.2), this study distinguishes two main types of opinion statements: 1) an opinion statement without a supporting statement, referred to as ‘simple opinion

statement', and 2) a 'complex opinion statement', which consists of an opinion statement followed by a supporting statement.

3.2.3. Interactional context of the opinion/supporting statements

The third annotation category relates to the interactional context of the opinion or supporting statements. The main communicative function recognised in this category is whether an opinion statement/supporting statement was expressed spontaneously by the speaker (referred to as 'unprompted') or whether it followed directly from the other interlocutor's question or request (referred to as 'prompted'), see Degoumois *et al.* (2017). This dimension recognises that, while the frequency and type of opinion/supporting statements provide insights into L2 proficiency, the interactional context in which these statements occurred can provide additional understanding when interpreting the nature of these expressions of opinion. Previous studies have suggested that overall speaking proficiency might have an effect on L2 speaker's ability to respond to interlocutors in ongoing interactions, with advanced speakers demonstrating a greater tendency to build on the utterances of previous speakers appropriately (e.g., Watanabe 2017; Abe and Roever 2019). This dimension is therefore crucial 1) to further our understanding of the nature of L2 speakers' proficiency and ability to express opinions, and 2) to describe the nature of an opinion expression in an interactive communication better. In addition to providing insights into L2 communicative abilities, this dimension is important in order to understand the co-constructed nature of opinion expression in language assessment contexts. For example, it can contribute to a better understanding of Oral Proficiency Interviews as a genre of language assessment discourse and its communicative features related to the test takers at different proficiency levels.

The prompted opinion/supporting statement is identified if the other interlocutor directly asks the speaker to express their views (e.g., *what do you think about this?)* or to express a support for an opinion statement (*why do you think so?*) and the speaker provides such statement as a response, as shown in example (16). When no such preceding prompt is identified in the conversation, the opinion/supporting statement is coded as unprompted.

(16) S2: very difficult

S1: yeah so why why is it difficult? how is it difficult?

S2: because it's completely different to Spanish [2_6_SP_31]

It should be noted that in some contexts it may be difficult to determine whether an opinion/supporting statement is prompted or not. First, a request for an opinion can be stated in syntactically varied ways (Green 2014). For example, while it most commonly takes the form of an interrogative (e.g., *do you think it's an important skill?*), it can also occur as an imperative (e.g., *tell me about Euro*), a declarative (e.g., *we'll talk about tourism and particularly the negative side of tourism okay some people say that tourism can do more harm than good*), or a declarative with a tag question (e.g., *very hard, isn't it?*). Thus, for example, in the case of the declarative sentence above, it remains uncertain whether the speaker is asking for an opinion or just stating a proposition. In the example including an imperative, it is unclear whether the speaker is asking for an opinion about Euro or for something else (e.g., facts).

The coding scheme had to consider also the instances where one speaker asks for an opinion related to a particular topic, but the other speaker offers an opinion about a different topic, as shown in example (17):

(17) S1: how im= how important is it for you to earn a good salary?

S2: It's a good question [2_6_SP_67]

In example (17), S1 is asking for an opinion about a specific proposition (i.e., earning a good salary) but S2 responds with an opinion statement about the question itself. The coding schemes regards such responses as an unprompted opinion statement. An opinion/supporting statement is considered prompted if the object of evaluation (either an entity or a proposition) is introduced by the first speaker. On the other hand, the statement is considered unprompted if a new object of evaluation is introduced by the second speaker in their response.

As stated above, one of the aims of developing this coding scheme was to make it applicable in spoken interactive communication. Considering that it can be quite challenging to segment streams of L2 spoken data due to repetitions, ellipses, or false starts (Foster *et al.* 2000), the current coding scheme clarifies how to deal with a number

of challenges related to syntactic or interactional realisations of opinion such as elliptical responses and coordinated phrases.

First, a common pattern that occurs specifically in interactive communication is that of elliptical responses (Chia and Kaschak 2023). In the coding scheme, affirmative expressions such as *yes*, *no*, *of course*, and other similar variants are considered (and coded as) instances of opinion statements if they occur as a direct response to an interlocutor's prompt, as illustrated in S2's response in example (18).

(18) S1: but do you do you think that it's a good idea to have pocket money maybe whe= maybe in the future?

S2: yes because for example if you go with your friends or with any other people to anywhere and you have to for example buy something or take a taxi to … [2_6_SP_31]

Second, in terms of a range of syntactical types of opinion/supporting statements, these can be realised by noun phrases (example 19) and relative clauses (example 20) if it is clear from the context what entity or proposition is being evaluated.

(19) S2: and okay problems with the prices problems with the mortgages [SP_35]

(20) S2: you have to be able to talk to the audience, which is very difficult [2_6_SP_5]

Next, a common issue in interactive communication is presented by coordinated clauses or words. In utterances where evaluative meanings are coordinated (usually by a connector *and*), an opinion statement should evaluate a single entity or proposition. When two entities or propositions are evaluated in a coordinated clause, they are considered as two instances of opinion statements, as shown in examples (21) and (22):

(21) S2: I I think that is a job very interesting (okay) (opinion statement 1) and I think that I a= I am a a person qualificate for this job (opinion statement 2) [6_SP_29]

(22) S2: so that it's difficult to ask a question to the teacher (opinion statement 1) and and speak to the teacher alone (opinion statement 2) [2_6_SP_62]

Finally, in interactive communication, it is important for the scheme to systematically address co-constructed meaning. This is a situation where two speakers collaboratively

develop an evaluative statement, which thus cannot be attributed to only one of the speakers. Example (23) demonstrates such pattern, which is not treated as an opinion statement in the proposed scheme.

(23) S1: oh yes

S2: that

S1: it's safer

S2: is safer

S1: isn't it??

S2: than than other ways to

S1: okay okay

S2: to pay and with children er I think that's people have to to be

S1: be a bit more careful

S2: to be care

S1: don't they?

S2: yes [6_SP_60]

4. Evaluation of the coding scheme: An empirical study

4.1. Data

The study used a subset of the *Trinity Lancaster Corpus* (Gablasova *et al.* 2019), reflecting the aim to make the scheme applicable to teaching/testing contexts. The corpus consists of 4.1 million words from the transcriptions of over 2,000 dyadic interactions recorded as part of the GESE, an international exam of spoken English, developed and administered by Trinity College London (Trinity College London 2024). The corpus contains data from L2 speakers from different L1 backgrounds and three main proficiency levels of the Common European Framework of Reference: B1 (pre-intermediate), B2 (intermediate), and C (advanced, comprising C1 and C2 levels).

For the evaluation of the coding scheme 29 transcripts from the TLC were selected. Each speaking exam in the GESE involves one L2 speaker (the test taker) and one L1

speaker (the examiner). Data from two speaking tasks were used in the study: 1) the conversation task, in which the interlocutors exchange and discuss their views on general topics selected by the examiner, and 2) the discussion task, in which the topic is selected and introduced by the test taker. Both tasks are highly interactive and thus offer a communicative environment in which opinions on a number of topics are stated and discussed by the two speakers. Only the data from the L2 speakers were used in this study, even though the contributions from the L1 speakers were taken into consideration in the interpretation and coding of the production of L2 speakers. The 29 transcripts represent data from three proficiency levels: ten transcripts at the B1, ten at the B2, and nine at the C level of proficiency. The transcripts were also selected to represent L2 English speakers from two L1 backgrounds —Chinese (14) and Spanish (15)— in order to include L2 production from typologically different L1s and cultural backgrounds, which could affect communicative preferences and strategies of L2 speakers.

4.2. Procedure

4.2.1. Coding procedure

Initially, two coders independently coded the sample to identify opinion statements. Between the two coders, 320 opinion statements were identified in the utterances produced by L2 speakers in 29 transcripts. The first coder was an expert coder involved in the development of the scheme; the second coder was then trained to apply the scheme. The training involved multiple rounds of practice coding sessions, in which the second coder applied the scheme and compared the results to the first coders' results, and differences between the codes were discussed. During these sessions, both typical (model) examples of each component of the scheme were used as well as more borderline cases. During each stage, the second coder compared their annotations to the first coder. The differences were discussed with reference to the coding scheme and the criteria were further clarified. After this process, the second coder proceeded to code the 29 transcripts independently. Inter-rater agreement statistics (agr, AC1, and Cohen's kappa) were calculated to evaluate the consistency of the application of the coding scheme.

The coding procedure first involved each coder identifying an instance of an opinion statement and, second, deciding whether it is a case of a simple or a complex opinion statement. The simple opinion statement consists only of the main opinion

statement (example 24), while a complex opinion statement consists both of the opinion statement and a supporting statement (example 25).

(24) S2: that was horrible [2_SP_1]

(25) S2: it was very tiring we had to go like seven hours a week [2_6_SP_5]

The instances of complex opinion were then coded further for the type of supporting statement following the categories presented in Section 3.2.2. As a final step, the context in which both simple and complex opinions were produced, was further explored and both types of opinion expressions were coded as prompted or unprompted. Figure 1 provides an overview of the main categories of the scheme; an overview of all types of opinion and supporting statements involved in the scheme is available in the Appendix.

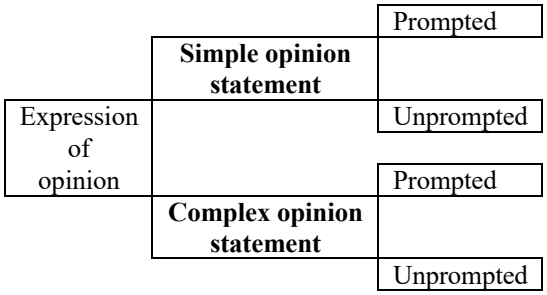


Figure 1: Coding scheme for expression of opinion: Main categories

4.2.2. Inter-rater agreement: general considerations

Identifying and classifying expressions of opinion is a challenging task due to the complexity and, in many cases, also the subjectivity of the decision-making involved. Identifying opinion statements is considerably different from the coding schemes where the choices are limited (e.g., when the annotators select one option from a closed set) or when the coding is applied to entities (units) that have already been identified (e.g., the annotators are asked to select a category to which a certain word belongs). When identifying opinion statements in a continuous interactive discourse, annotators have to make decisions based on a potentially unlimited list of expressions that can act as markers of evaluative judgements in given contexts. Further, there are no pre-set boundaries or a finite number of entities to code (e.g., an utterance could contain several opinion statements or none). Such coding, therefore, involves high-inference decisions and has been observed to result in a greater difference in agreement than the situations when

annotators deal with well-defined (or pre-defined units) and closed sets of attributes to apply in coding (Allwood *et al.* 2007; Read and Carroll 2012).

Statistically, the agreement measure agr seems to be the most appropriate to quantify inter-rater agreement in these situations. Agr is a directional measure of agreement, using one rater as a baseline for the other rater's performance (Wiebe *et al.* 2005: 196). Agr is particularly suitable in cases such as pragmatic annotation of corpora, where raters are tasked with the identification of linguistic features in the flow of discourse rather than applying a coding scheme to a given set of cases (e.g., Wilson *et al.* 2006). The measure allows evaluating consistency of application of a particular coding scheme in the following way from the perspective of different raters:

Rater A taken as a baseline:

$$agr_1 = \frac{A\ matching\ B}{A}$$

Rater B taken as a baseline:

$$agr_2 = \frac{B\ matching\ A}{B}$$

Two values of agr are therefore typically reported and sometimes interpreted as precision and recall. They can be combined in a single F1 measure (Van Rijsbergen 1979; Fuoli and Hommerberg 2015) as follows:

$$F1 = \frac{2 \times agr_1 \times agr_2}{agr_1 + agr_2}$$

In addition, inter-rater agreement about further classification of opinion statements, once identified, can be measured using standard statistics such as Cohen's kappa (κ) and AC1 (Brezina 2018: 90–91). Kappa and AC1 are appropriate in this case, as once the raters identified an opinion statement, they are then selecting from a closed set of two options: presence or absence of a supporting statement/prompt. While AC1 represents a recent and sophisticated measure of inter-rater agreement, estimating agreement by chance more precisely in extreme cases (Brezina 2018), Cohen's kappa has been used in this study as well in order to allow for comparison with other earlier studies that opted for this measure. In Cohen's kappa, values close to 0 indicate that the agreement is most likely due to chance, while the values close to 1 indicate a very strong to full agreement between the

raters. In order to interpret the level of agreement, we follow Rietveld and van Hout (1993), who interpret the values in the following way: 0–0.20 indicate a small degree of agreement, 0.21–0.40 means fair agreement, 0.41–0.60 are considered to represent a moderate agreement, while 0.61–0.80 suggest a strong agreement between the raters. AC1 operates on the same scale and has therefore a similar interpretation.

4.3. Empirical results

4.3.1. Agreement on the identification and type of opinion statements

Table 1 provides an overview of the results, showing the per cent agreement, agr, Cohen's kappa, and AC1 based on the rating of 320 cases identified by two coders.

	Inter-rater agreement				
	% agreement	**Agr**	**F1**	**K**	**AC1**
Presence of opinion statement	N/A	0.52/0.64*	0.57	N/A	N/A
Presence of opinion statement (intersecting cases counted as agreement)	N/A	0.59/ 0.74*	0.66	N/A	N/A
Presence of a supporting statement	79.1	N/A		0.5 (p<.001)	0.64 (p<.001)
Presence of a prompt	83.58	N/A		0.66 (p<.001)	0.068 (p<.001)

Table 1: Inter-rater agreement on three dimensions of the coding scheme for expression of opinion. (*The first value refers to the agr metric results based on the first coder taken as the baseline and the second value uses the second coder as the baseline)

We will first examine the level of agreement regarding whether a stretch of text was considered to qualify as an opinion statement. Overall, in the strictest interpretation, the agr metric shows that, taking the first rater as the baseline, over 50 per cent agreement between the raters (0.52) was reached; taking the second rater as the baseline, even a stronger agreement (over 0.64) was reached between the raters. The combined F1 score for this situation is 0.57. While there was a relatively good agreement between the raters, in addition to the opinion statements identified by both raters, each rater also identified additional statements that went unnoticed by the other rater. A closer inspection of the results revealed that there were two systematic issues that affected inter-rater agreement. The first involved distinguishing the nature of the evaluative statement (i.e., an opinion statement vs. a supporting statement) when dealing with interconnected expressions of opinion. The second issue involved dealing with boundaries of opinion statements, that

is, a situation where both raters identified the presence of evaluative language but differed in how they determined the boundaries of an opinion statement. Several coding schemes that dealt with classifying utterances or expressions in larger stretches of naturally-produced discourse reported similar issues that affected the rate of agreement regarding the identification and counting of the target units (e.g., Wiebe *et al.* 2005; Allwood *et al.* 2007; Read and Carroll 2012). These studies approached the issue at the level of coding and considered "intersecting text as agreement" (Read and Carroll 2012: 433; see also Wiebe *et al.* 2005). When intersecting cases were considered as agreement (following Wiebe *et al.* 2005 and Read and Carroll 2012), the agr metric reported that, taking the first rater as the standard, a fairly strong agreement between the raters (nearly 0.6) was reached; taking the second rater as the standard, a substantial agreement (over 0.7) was reached between the raters. The combined F1 score was 0.66 in this case. The issues of establishing the boundaries and distinguishing between interconnected opinion statements are further discussed in Section 4.3.2 below.

Regarding the presence of supporting statements, a relatively high per cent agreement (nearly 80 per cent) was achieved between the raters, with kappa and AC1 indicating a fairly strong agreement. When inspected more closely, some of the disagreement between the raters was related to the issue of interconnected opinion statements described above. In these cases, both raters recognised a stretch of discourse as being related to an opinion statement; however, while one of the raters coded the two utterances as separate opinion statements, the other rater coded the following utterance as a supporting statement.

Finally, focusing on the presence of a prompt for an expression of opinion, the results showed a high per cent agreement (above 80 per cent), while kappa and AC1 also indicated a substantial agreement between the raters. Further analysis revealed that the disagreement between the raters was mostly due to the distance of a prompt from the expression of opinion statement, with prompts separated by several turns from the expression of opinion being most often coded differently by the coders.

4.3.2. Discussion of results and implications

Overall, the study identified the strengths as well as challenges to be further addressed when applying the coding scheme for identifying expressions of opinion for pedagogical and language testing purposes.

First, it demonstrated the challenges of identifying instances of opinion statements in continuous, interactive discourse, characterised by chaining of utterances and dynamic topic development. In addition to dealing with the typical features of spoken production such as unfinished or rephrased utterances, a further difficulty was presented by some features of learner language (e.g., utterances with semantic, lexical, or grammatical errors). These in some cases created additional challenges for understanding the meaning of an utterance, as demonstrated in example (26) taken from an L2 speaker in the data.

(26) well that I I that I I'm <pause > well I am a very late very late person

In some cases, therefore, these features affected the ability of the raters to understand and interpret the meaning of the utterances and to apply the coding scheme with confidence. It is possible that these challenges —present in spoken interactive communication— led each rater to identify a set of additional cases in line with their specific interpretation of the flow of the discourse. Many of these were, upon a review by the other rater, considered to be acceptable instances of opinion statement. This pattern suggests that it would be beneficial for several coders to code each transcript highlighting 'candidate' opinion statements according to the scheme criteria. As a second stage, either an expert coder could review these candidate statements and make a final decision to accept/reject, or the two (or more) raters could further review each other's candidate statements and, if needed, discuss the rationale behind their decision. This approach would likely result in a higher recall rate of opinion statements marked by explicit as well as more subtle clues.

Another systematic issue that affected the inter-rater agreement on the identification of the opinion and supporting statements was related to the ability to determine the boundaries and relationships between individual opinion statements. Given the nature of the discourse in this study (i.e., interactive communication), it was relatively common for a speaker to express an opinion and follow it by a supporting statement, which would then lead to another opinion statement, often on a related topic. This pattern appeared especially typical of advanced L2 speakers in the data, whose turns tended to be longer and more complex, making it more challenging to distinguish between individual

opinion/supporting statements. This issue is illustrated in example (27), which is characterised by the proximity of several opinion statements as part of connected speech.

(27) S1: okay thank you very much now let's talk about erm money =

S2: money er a money in my life it is important but not much but is important because I I have to buy everything with money it's necessary

S1: okay do you have a job? [6_SP_1]

In this example, both adjacent statements satisfy the criteria for identifying an opinion statement, as discussed in Section 3.2.1: 1) *a money in my life it is important but not much but is important* and 2) *it's necessary*. The first of these statements is also followed by a supporting statement of reason (*because I I have to buy everything with money*). While the annotators agreed on the first statement and the supporting statement, disagreement tended to appear when dealing with statements such as *it's necessary*. A possible difference was related to the question of whether this statement represents a continuation of the preceding utterances (i.e., of the supporting statement, expressing justification for the opinion statement) or a new opinion statement, closely related to the previous topic. On the one hand, it could be considered as a new opinion statement, as it satisfies the relevant criteria; on the other hand, it could be viewed as part of the supporting statement in that it further elaborates on the supporting statement and provides additional justification.

Classifying such adjacent evaluative statements is challenging as they are often related to the same conversational topics, as shown in example (27) above. When two propositions evaluate the same entity, they often contain pronouns that refer back to the entity mentioned previously (e.g., *it*) or semantically related lexis (e.g., *important* and *necessary*). As a result, such stretches of thematically connected utterances make it difficult for annotators to determine at what point one instance of an opinion statement ends and another one starts. Example (28) further demonstrates the issues involved in distinguishing between different evaluative comments about the same entity.

(28) really because they have an singing teacher okay er is called [name] that she has a very beautiful voice she's really a she's a really good person and she helps us to sing really really well [2_6_SP_71]

In this example, the speaker is offering three propositions that qualify as an opinion statement according to the coding scheme: 1) *she has a very beautiful voice*, 2) *she is a really good person*, and 3) *she helps us to sing really really well*. These statements express value judgements about a teacher which are coordinated (e.g., they use the connector *and*). What is notable is that each statement evaluates the entity according to different attributes, for example, the voice quality, personality, and profession. This type of discourse raises the question of whether —if different statements evaluate an entity from three different perspectives— they should be viewed as three different instances of opinion. In its current form, the scheme does not explicitly instruct the coders on how to deal with such inter-connected statements. However, as the study demonstrated, such guidance is necessary in order to reliably address such utterances that are relatively common in interactive speech, where the meaning of utterances develops dynamically.

In general, the results suggested that the scheme can be applied with sufficient degree of reliability to the target communicative settings (i.e., L2 interactive production). However, following the findings of this study, it is recommended that the areas that represent systematic difficulties are further addressed in the coding scheme and in the guidance for the raters. In particular, the guidance for the raters should include several examples of evaluative statements that demonstrate stretches of discourse with several inter-connected opinion/supporting statements as well as examples with multiple opinion statements related to the same entity. Such guidance will likely result in a more reliable application of the coding scheme and greater agreement between the raters. However, it should be acknowledged that the high degree of subjectivity involved in investigating evaluative language will naturally result in a certain degree of difference between individual raters.

While the study brought encouraging results in terms of investigating expression of opinion in L2 production, there are also some limitations involved in the development and application of the scheme. First, the coding was restricted to the expression of opinion produced by L2 speakers (test takers), excluding the language produced by the L1 speakers (examiners) in the study, although their production was taken into consideration when classifying the L2 utterances, reflecting the co-constructed nature of the discourse. Second, the coding system was only evaluated in relation to the dyadic production involving two speakers taking turns, in a semi-formal environment. Its usability in different environments (e.g., with multiple L2 speakers, in informal conversations) should

be further investigated. Finally, the coding was based mainly on the linguistic information available in the transcribed speech, without access to paralinguistic (e.g., pauses, laughs) or non-linguistic features (e.g., gestures indications of speaker turn). Such information would be valuable in interpreting the nature of the utterances. It should be noted, however, that working with transcribed speech may reflect the resources available in the testing and teaching context.

5. Conclusion

This study described and evaluated a coding scheme for the expression of opinion, which was specifically designed for the identification of this pragmatic feature in spoken interactive discourse, although its use can be extended to other settings including written communication. The scheme thus complements and further broadens the resources available for investigating evaluative language in different contexts and for different purposes. A particular contribution of the scheme lies in its application to L2 production and to pedagogical/assessment settings, where expressing views plays a major role. The annotation scheme proposed in the study can lead to a better understanding of L2 pragmatic knowledge by providing a tool for systematically recording different types of opinion expressions and relating them to L2 proficiency (e.g., Jung 2024). The findings based on the scheme can be then applied in the development of teaching materials and for evaluating the validity of (speaking) tasks that focus on eliciting and assessing expressions of opinion.

References

Abe, Makoto and Carsten Roever. 2019. Interactional competence in L2 text-chat interactions: First-idea proffering in task openings. *Journal of Pragmatics* 144: 1–14.

ACTFL. 2024. *ACTFL Proficiency Guidelines 2024*. American Council on the Teaching of Foreign Languages. https://www.actfl.org/uploads/files/general/Resources-Publications/ACTFL_Proficiency_Guidelines_2024.pdf

Allwood, Jens, Loredana Cerrato, Kristiina Jokinen, Costanza Navarretta and Patrizia Paggio. 2007. The MUMIN coding scheme for the annotation of feedback, turn management and sequencing phenomena. *Language Resources and Evaluation* 41: 273–287.

Bardovi-Harlig, Kathleen, Sabrina Mossman and Heidi E. Vellenga. 2015. The effect of instruction on pragmatic routines in academic discussion. *Language Teaching Research* 19/3: 324–350.

Bednarek, Monika A. 2006. Evaluation and cognition: Inscribing, evoking and provoking opinion. In Hanna Pishwa ed. *Language and Memory: Aspects of Knowledge Representation.* Berlin: Mouton De Gruyter, 187–221.

Bednarek, Monika A. ed. 2008. *Emotion Talk Across Corpora*. London: Palgrave Macmillan.

Bednarek, Monika A. 2009. Language patterns and attitude. *Functions of Language* 16/2: 165–192.

Biber, Douglas, Susan Conrad and Geoffrey Leech. 2002. *Longman Student Grammar of Spoken and Written English*. London: Longman.

Biber, Douglas, Jesse Egbert, Daniel Keller and Stacey Wizner. 2021. Towards a taxonomy of conversational discourse types: An empirical corpus-based analysis. *Journal of Pragmatics* 171: 20–35.

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. 1999. *Longman Grammar of Spoken and Written English*. London: Longman.

Blum-Kulka, Shoshana, Juliane House and Gabriele Kasper eds. 1989. *Cross-cultural Pragmatics: Request and Apologies*. New York: Ablex.

Brezina, Vaclav. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge: Cambridge University Press.

Carlson, Lynn and Daniel Marcu. 2001. *Discourse Tagging Reference Manual.* https://web.archive.org/web/20170808131213id_/https://www.isi.edu/~marcu/discourse/tagging-ref-manual.pdf

Chafe, Wallance and Johanna Nichols eds. 1986. *Evidentiality: The Linguistic Coding of Epistemology*. New York: Ablex.

Chia, Katherine and Michael P. Kaschak. 2023. Elliptical responses to direct and indirect requests for information. *Language and Speech* 67/1: 228–254.

Council of Europe. 2020. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment: Companion Volume*. https://rm.coe.int/common-european-framework-of-reference-for-languages-learning-teaching/16809ea0d4

Degoumois, Virginie, Cécile Petitjean and Simona Pekarek Doehler. 2017. Expressing personal opinions in classroom interactions: The role of humor and displays of uncertainty. In Simona Pekarek Doehler, Adrian Bangerter, Geneviève de Weck, Laurent Filliettaz, Esther González-Martínez and Cécile Petitjean eds. *Interactional Competences in Institutional Settings: From School to the Workplace*. Berlin: Springer, 29–57.

Edmondson, Willis. 1981. *Spoken Discourse: A Model for Analysis*. London: Longman.

Educational Testing Service. 2018. *The Official Guide to the TOEFL Test.* New York: McGraw-Hill Education.

Fogal, Gary G. 2019. Tracking microgenetic changes in authorial voice development from a complexity theory perspective. *Applied Linguistics* 40/3: 432–455.

Fordyce, Kenneth. 2014. The differential effects of explicit and implicit instruction on EFL learners use of epistemic stance. *Applied Linguistics* 35/1: 6–28.

Foster, Pauline, Alan Tonkyn and Gillian Wigglesworth. 2000. Measuring spoken language: A unit for all reasons. *Applied Linguistics* 21/3: 354–375.

Fuoli, Matteo. 2018. A stepwise method for annotating appraisal. *Functions of Language* 25/2: 229–258.

Fuoli, Matteo and Charlotte Hommerberg. 2015. Optimising transparency, reliability and replicability: Annotation principles and inter-coder agreement in the quantification of evaluative expressions. *Corpora* 10/3: 315–349.

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.
Gablasova, Dana, Vaclav Brezina, Tony McEnery and Elaine Boyd. 2017. Epistemic stance in spoken L2 English: The effect of task and speaker style. *Applied Linguistics* 38/5: 613– 637.
Galaczi, Evelina D. 2014. Interactional competence across proficiency levels: How do learners manage interaction in paired speaking tests? *Applied Linguistics* 35/5: 553–574.
Goźdź-Roszkowski, Stanislaw. 2018. Values and valuations in judicial discourse. A corpus- assisted study of (dis)respect in US Supreme Court decisions on same-sex marriage. *Studies in Logic, Grammar and Rhetoric* 53/1: 61–79.
Gray, Bethany and Douglas Biber. 2012. Current conceptions of stance. In Ken Hyland and Carmen Sancho Guinda eds. *Stance and Voice in Written Academic Genres*. London: Palgrave Macmillan, 15–48.
Green, Anthony. 2014. *Exploring Language Assessment and Testing: Language in Action*. London: Routledge.
Greenberg, Joshua. 2000. Opinion discourse and Canadian newspapers: The case of the Chinese "Boat People." *Canadian Journal of Communication* 25: 517–537.
Horvarth, Barbara and Suzanne Eggins. 1995. Opinion texts in conversation. In Peter H. Fries and Michael Gregory eds. *Discourse in Society: Systemic Functional Perspectives*. New York: Ablex, 29–45.
Hunston, Susan. 2000. Evaluation and the planes of discourse: Status and value in persuasive texts. In Susan Hunston and Geoff Thompson eds. *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford: Oxford University Press, 176–207.
Hunston, Susan. 2004. Counting the uncountable: Problems of identifying evaluation in a text and in a corpus. In Alan Partington, John Morley and Louann Haarman eds. *Corpora and Discourse*. Berlin: Peter Lang, 157–188.
Hunston, Susan and Geoff Thompson eds. 2000. *Evaluation in Text: Authorial Stance and the Construction of Discourse*. Oxford: Oxford University Press.
Hyland, Ken. 1998. Boosting, hedging and the negotiation of academic knowledge. *Text* 18/3: 349–382.
Hyland, Ken. 2005. Stance and engagement: A model of interaction in academic discourse. *Discourse Studies* 7/2: 173–192.
IELTS. 2019. *Information for Candidates Introducing IELTS to Test-takers*. Cambridge: University of Cambridge Local Examinations Syndicate.
Iwasaki, Noriko. 2009. Stating and supporting opinions in an interview: L1 and L2 Japanese speakers. *Foreign Language Annals* 42/3: 541–556.
Jiang, Feng and Ken Hyland. 2015. 'The fact that': Stance nouns in disciplinary writing. *Discourse Studies* 17/5: 529–550.
Jung, Yejin. 2024. *Examining L2 Speakers' Expression of Opinion in the Trinity Lancaster Corpus*. Lancaster: Lancaster University Dissertation.
Labov, William. 1972. *Language in the Inner City: Studies in the Black English Vernacular*. Pennsylvania: University of Pennsylvania Press.
Mackenzie, Lachlan and Laura Alba-Juez eds. 2019. *Emotion in Discourse*. Amsterdam: John Benjamins.
Martin, James and Peter White eds. 2005. *The Language of Evaluation: Appraisal in English*. London: Palgrave Macmillan.
Mullan, Kerry. 2010. *Expressing Opinions in French and Australian English Discourse*. John Amsterdam: John Benjamins.
Myers, Greg. 2004. *Matters of Opinion*. Cambridge: Cambridge University Press.

O'Sullivan, Barry and Jamie Dunlea. 2015. *Technical Report: Aptis General Technical Manual.* Version 1.0. English Language Assessment Research Group: British Council.

Pérez-Paredes, Pascual and M. Camino Bueno-Alastuey. 2019. A corpus-driven analysis of certainty stance adverbs: Obviously, really and actually in spoken native and learner English. *Journal of Pragmatics* 140: 22–32.

Read, Jonathon and John Carroll. 2012. Annotating expressions of Appraisal in English. *Language Resources and Evaluation* 46/3: 421–447.

Rietveld, Toni and Roeland van Hout. 1993. *Statistical Techniques for the Study of Language and Language Behaviour*. Berlin: Mouton De Gruyter.

Roever, Carsten. 2011. Testing of second language pragmatics: Past and future. *Language Testing* 28/4: 463–481.

Simaki, Vasiliki, Carita Paradis and Andreas Kerren. 2019. A two-step procedure to identify lexical elements of stance constructions in discourse from political blogs. *Corpora* 14/3: 379–405.

Sinclair, John M. 1986. Fictional worlds. In Malcolm Coulthard ed. *Talking about Text: Studies Presented to David Brazil on His Retirement*. University of Birmingham: English Language Research, 43–60.

Thompson, Geoff and Laura Alba-Juez eds. 2014. *Evaluation in Context*. Amsterdam: John Benjamins.

Traugott, Elizabeth C. 2010. (Inter)subjectivity and (inter)subjectification: A reassessment. In Hubert Cuyckens, Kristin Davidse and Lieven Vandelanotte eds. *Subjectification, Intersubjectification and Grammaticalization*. Berlin: Mouton De Gruyter, 29–71.

Trinity College London. 2024. *Examination Information: Graded Examinations in Spoken English (GESE)*. https://www.trinitycollege.com/resource/?id=5755

Van Rijsbergen, Cornelis. 1979. *Information Retrieval*. London: Butterworth and Co.

Watanabe, Aya. 2017. Developing L2 interactional competence: Increasing participation through self-selection in post-expansion sequences. *Classroom Discourse* 8/3: 271–293.

Wiebe, Janyce, Teresa Wilson and Claire Cardie. 2005. Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation* 39: 165–210.

Wilson, Theresa, Janyce Wiebe and Rebecca Hwa. 2006. Recognizing strong and weak opinion clauses. *Computational Intelligence* 22/2: 73–99.

Corresponding author
Yejin Jung
Lancaster University
Department of Linguistics and English Language
County South
LA1 4YL
Lancaster
United Kingdom
Email: y.jung@lancaster.ac.uk

received: August 2023
accepted: April 2024

APPENDIX: OVERVIEW OF THE DIFFERENT TYPES OF OPINION STATEMENTS (OS) AND SUPPORTING STATEMENTS IN THE CODING SCHEME

	Opinion Statements Categories		
	Presence of prompt	**Supporting statement**	**Type of Opinion Statements**
1	Not present	Not provided	Unprompted Simple Opinion Statement
2	Not present	Provided	Unprompted Complex Opinion Statement
3	Present	Not provided	Prompted Simple Opinion Statement
4	Present	Provided	Prompted Complex Opinion Statement
	Supporting Statements Categories		
1	Giving a reason		
2	Giving elaboration		
3	Giving a contrasting idea		
4	Giving evidence		
5	Others		

RiCL Research in Corpus Linguistics

Corpus as a slice of life: Representing naturally occurring language and its speakers

Giorgia Troiani[a/b] – John W. Du Bois[b] – Andrey Filchenko[a]
Nazarbayev University[a] / Kazakhstan
University of California, Santa Barbara[b] / United States

Abstract – Discourse is subject to numerous forces that shape its form. One force that is underestimated is the interactional dynamic among interlocutors. In devising the criteria that inform data selection for a corpus of spoken discourse, designers may end up prioritizing the collection of spontaneous discourse and overlook the fact that this type of discourse can still display artificial interactional dynamics. We propose an approach to spoken corpus compilation that aims at preserving naturally occurring interactional dynamics by choosing as focus of the corpus the representation of participants' lives. Through the analysis of speech events collected in different projects, we demonstrate the advantages of sourcing naturally occurring discourse over spontaneous data. We then discuss a series of practices that the authors implemented in different contexts to ensure the collection of naturally occurring data. We argue that this framework yields the construction of corpora that are representative not only of a language, but also of the lives of its users.

Keywords – corpus design; naturally occurring discourse; spontaneous discourse; recording practices

1. INTRODUCTION[1]

When creating a corpus of everyday spoken discourse, designers must balance their theoretical assumptions about language with time and resource constraints. They do so by establishing strict criteria that guide the decision over which data to include in the corpus. One productive approach is to prioritize informal spontaneous discourse (e.g., conversation), either exclusively (Chui and Lai 2008; Raso and Mello 2012; Love *et al.* 2017; Goria and Mauri 2018) or in combination with other genres (Greenbaum 1991; Burnard 2002; Kucera 2002; Oostdijk 2002). While this approach guarantees the collection of everyday conversational data, it places the corpus designers' attention on

[1] This project is funded by the *Nazarbayev University Collaborative Research Projec*t (grant #021220CRP1422). The paper has been improved by the helpful comments of Chloe Willis, Guillem Belmar Viernes, and Jordan Douglas-Tavani. We thank an anonymous reviewer for their comments that helped improve our work.

structural features of language, promoting its treatment as a series of constructions disembodied from the interactional context in which they are produced. An alternative to this approach is the 'cast the wide net' model (Du Bois and Troiani 2022), a framework that integrates corpus linguistics with anthropology, first implemented in the design of the *Santa Barbara Corpus of Spoken American English* (SBCSAE; Du Bois *et al.* 2000). Within this framework, the focus of corpus design is shifted from the representation of language to the representation of the participants' behavior and their lives.[2]

In this paper, we present the theoretical foundations that underpin the 'cast the net wide' framework. The second author originally developed the framework in the construction of the SBCSAE. The first and third authors implemented and adapted this framework in the compiling of the *Multimedia Corpus of Modern Spoken Kazakh Language* (MULTICORSKL; Filchenko *et al.* 2023), the first spoken corpus of Kazakh. Our experience with corpus construction has been informed by disciplines like anthropology and language documentation. Drawing on the discourse and conversational analysis of different types of speech events, we illustrate an approach to spoken corpus design that prioritizes the role of individual speech events in participants' lives as criterion for data collection. We demonstrate how speech events that are selected according to this criterion (i.e., naturally occurring data) can differ from events that prioritize different criteria. We particularly stress how 'spontaneous' data are not necessarily naturally occurring. We argue that naturally occurring data preserve interactional social dynamics between participants to a speech event, which in turn results in the construction of a corpus which aims to be representative of the lives of participants, rather than of linguistic structures. In the rest of this section, we show how spontaneous and naturally occurring types of discourse differ from each other by analyzing the interactional dynamics of similar speech events collected with different protocols. We then illustrate the shortcomings with restricting data selection to spontaneous discourse (Section 3) and the adjustments introduced to our data collection protocol to ensure the collection of naturally occurring discourse (Section 4). We base this paper on the analysis of data from the MULTICORSKL and of recordings produced for previous projects (Section 2).

Within the 'cast the net wide' framework, the goal of a corpus is to capture the context in which language arises, i.e., the life of participants, rather than language itself.

[2] While our expertise is on spoken discourse, many of these considerations are valid for sign languages as well.

We can visualize these two different approaches in metaphorical terms by considering archaeological artifacts. If one is interested in Roman art, they can observe it in a museum, where curators have selected representative pieces, reconstructed their cultural functions, and established connections for the visitors. Alternatively, they can visit the archaeological site of Pompeii, a southern Italian town destroyed by a volcanic eruption in 79 AD, walk through perfectly preserved neighborhoods, and observe an artifact in the spatial and cultural context where it originated. In this way, one can notice that Roman mosaics are found not on walls, but on the floors of busy areas (pools and halls), and one may deduce that they were not intended to be used solely as decorations, but also as means of protecting floors and supporting foot traffic. Once the function of Roman mosaics is clear, it becomes evident why they are made of sturdy stone tesserae and lack the color vibrancy of their Byzantine counterparts, which were produced mainly as decorative pieces and crafted from glazed glass.

In a similar way, corpora can represent linguistic structures (corpus as exhibition) or the life events in which structures are used (corpus as archaeological site). Depending on the goal that is prioritized, the same speech event (e.g., a conversation) can end up looking different. As discussed above, we present an approach to corpus construction named ‘cast the wide net’ which yields naturally occurring discourse, i.e., discourse produced by the participants to fulfill life needs that are independent of the researcher’s agenda. Providing examples sourced using different data collection protocols we show how naturally occurring data differ from elicited discourse, i.e., discourse prompted by a researcher, and does not overlap with spontaneous discourse, i.e., discourse that was not planned before the moment of production.

The consequences of data collection approaches on the linguistic structures of an event can be observed in Text 1 that showcases distribution and use of backchannel and features a naturally occurring and spontaneous conversation between two spouses. Backchannels are verbal or non-verbal messages used to acknowledge that the interlocutor holds the floor, and that the interaction can proceed (Drummond and Hopper 1993; Heinz 2003). Different languages exhibit different forms of backchannel.

TEXT 1

1	ADILET;[3]	*jeŋgem qaytıs bolğanda otıramın dedi me?*	Did my sister-in-law say she will sit when she is dead?
2		*Birdeŋe dep söylep.*	She said something.
3		*Endi bar,*	Anyway,
4		*Qanatqa ayt degen şıqtı Qanatqa bilmeymin?*	To Qanat – she said – "don't I know that they went to Qanat?"
5		*Birew aparmasa.*	Unless someone will catch him.
6	AISHA;	*Ïä.*	Yes.
7	ADILET;	*Endi sonımen ägi aqşanı köpirtip jür ğoy ol.*	'At the moment he is making money.'
8		*Mïlliöndı.*	Millions.
9	AISHA;	*Ägi eki jüz jïr- eki mïlliöndı.*	Two hundred and two millions
10	ADILET;	*Eki jüz teŋgege şıt et #,*	Two hundred tenge for fresh meat

In this excerpt, one of the speakers (Adilet) is dominating the conversation, while the other (Aisha) limits her contribution to backchannel (intonation unit (IU) 6) and occasionally provides details to co-construct a narrative (IU 9). In the text, backchannel is used with the typical function of acknowledging speakership and signaling engagement in the exchange.

Let us now compare Text 1 with a conversation (Text 2) with similar characteristics (e.g., it is spontaneous, contains the same gender dynamics, and features a similar unbalanced distribution of turns among interlocutors), but was collected through elicitation. Text 2 is sourced from a recording produced during the training phase of the MULTICORSKL. It was excluded from the corpus, as it does not feature naturally occurring discourse.

TEXT 2

129	QAIRAT;	*mağan jalpı osı*	to me in general
130		*arab tili degende*	when they [Arabic speakers] say 'Arabic'
131		*yağnï olar nege*	I mean why do they
132		*sonşama ülken*	in a territory that covers such a
133		*anday awmaqta ornalasqan jer*	large region like that
134		*nege onı bir til dep*	why do they consider it as one
135		*sanaydı desem*	language
136		*ol negizi sayasï turğıdan ğoy bılayşa aytqanda ïä*	it is basically from a political stand point so to speak right
137		*(1.2)*	
138	BOTAGOZ;	*ımm*	uhmm
139		*(0.5)*	

[3] All names used in the texts are pseudonyms, except for the first author and for the speakers who asked to be recognized.

140	QAIRAT;	*jalpı*	in general
141		*ne sebepti*	for what reason
142		*olar özderi sanay ma sonı*	they themselves consider
143		*bir tilde söylesemiz dep*	that they are speaking a single language

In this excerpt, where Qairat dominates the exchange, the topic of conversation was proposed by the researcher who collected the event and Botagoz does not associate with Qairat beyond classes. This suggests that Botagoz, differently than Aisha above, does not have a reason to engage with this event besides complying with a request from a researcher. The lack of a reason for engagement is reflected in the structure of the exchange. Qairat produces long turns without encountering backchannel, and even when backchannel is present, it is triggered by a long moment of silence (IU 137). Because in naturally occurring interactions speakers tend to minimize gaps and silence between turns (Pomerantz 1984; Stivers *et al.* 2009), Botagoz's silence in Text 2 is indicative of some sort of interactional trouble. In the absence of a gap between turns, the backchannel would indicate engagement with the interlocutor (Tottie 1991), but, in this case, the silence suggests that backchannel is not fulfilling this function. In fact, backchannel is here offered in reparation to the silence so as to make up for lack of engagement.

The linguistic structures featured in the two excerpts are nested within different interactional dynamics that are a consequence of the method used to collect the data. For this reason, the form, function, and position of a linguistic element (backchannel) is subject to changes. In Text 1, both interlocutors are ultimately interested in conveying and receiving information that are consequential to their lives outside of the exchange itself. For example, Aisha provides a detail to Adilet's claim that their acquaintance Qanat made money and clarifies the exact amount of money that was earned because both people associate with Qanat and knowledge of this detail may inform their future exchanges with him. In Text 2, Botagoz and Qairat seem uninterested in the exchange which was elicited by a researcher. The event was not planned beyond the prompting of the conversation from the researcher, as such, the language of the event is spontaneous, but Botagoz has no reason to engage with Qairat's opinion beyond the moment of the specific exchange. The different levels of engagement of interlocutors with the current event is visible in the linguistic and interactional structures of the event itself. In these specific cases, it is visible in the length of turns (shorter in Text 1) and in the different distribution and function of backchannel in both excerpts.

The differences between naturally and non-naturally occurring data can be noticed at the level of discourse and interactional dynamics of an event. Consider, for example, question-answer sequences in different speech events. Question-answer sequences vary as a consequence of the degree of formality of a speech event and on the number of speakers involved, as well as on the function they fulfill in the exchange (Stivers *et al.* 2010). Table 1 shows the frequency of questions and the question-answer turn distribution of different speech events. Results in Table 1 are based on the analysis of randomly selected ten-minute-long excerpts within different Kazakh speech events. The (non-)naturally occurring conversations come from the MULTCORSKL and the podcast interview comes from content created by the Kazakh-language media group *Salem Social*. All events are dialogic.[4]

	Questions per 1,000 words	**Mean of words in a response turn (by any interlocutor)**	**Mean of words in any turn (by dominant responder)**	**Questioner dominance**
Natural conversation	18.4	4 (SD = 3.9)	5 (SD = 5.3)	0.56
Podcast interview	14.3	21 (SD = 17.5)	22 (SD= 27.3)	0.60
Non-natural conversation	15.8	19 (SD = 30.2)	35 (SD = 20)	0.60

Table 1: Structuring of questions in speech events with different interactional expectations

We followed the coding scheme proposed by Stivers and Enfield (2010) to identify questions and answers in the excerpts. An utterance is regarded as a question if it relies on lexico-morpho-syntactic or prosodic marking (formal question) or if it sought to elicit information, confirmation, or agreement, even in the absence of formal markers (functional question). Requests for physical actions and questions inside reported speech were excluded from the analysis. For this analysis, we used a broader understanding of answers that conflates categories of 'non-answer' and 'answer' responses (Stivers and Enfield 2010: 2624). We consider an utterance to be an answer when it engages with the question as put either directly or indirectly (e.g., instances such as *I don't know*, *maybe*, requests for repetition, etc.).

The mean of words in a response turn (column 3) gives information about the average length of a turn with the function of response. This information does not distinguish between interlocutors. The mean of words in turns produced by the dominant

[4] The podcast interview featured three participants, but the role of interviewer is shared by two of them.

responder (column 4) gives information about the length of any turn —regardless of its function— produced by the interlocutor that produced more turns (i.e., that spoke the most). We computed the standard deviation for the previous measures: higher values correlate with exchanges where there are few long turns and many shorter turns. Finally, we computed the questioner dominance by using as numerator the number of questions asked by the participants that produced more questions and as denominator the total number of questions in the exchange (column 5). Values close to 0.5 index a dynamic where participants produce the same number of questions (e.g., a conversation), while values close to 1 index a dynamic where the production of questions is allocated to one participant (e.g., a typical interview).

The average response length in a naturally occurring conversation is strikingly lower than in the podcast interview and non-naturally occurring conversation (4 vs. 21/19 words). Moreover, the turns in the naturally occurring conversation are of a rather homogeneous length (SD = 3.9), while the turns in other events combine short turns with longer turns (SD = 17.5 for the podcast and SD = 30.2 for the non-naturally occurring conversation). This distribution is not unusual for the podcast interview, where responses can consist either of one-word answers to polar yes/no questions or of long uninterrupted sequences that are prompted by the interviewer as a mean to elaborate on the one-word answers. This behavior is maintained by the main responder throughout the entire interview: the speaker produces many one-word turns (backchannel and polar answers) and fewer long uninterrupted turns (mean = 22; SD = 27.3). Again, this is expected in an interview. In the non-naturally occurring conversation the main responder mirrors this behavior producing many one-word turns interspersed with uninterrupted turns that occasionally reach above 100 words of length (mean = 35; SD = 20). Moreover, questions tend to be slightly unidirectional, with one of the speakers asking questions more than the other (questioner dominance = 0.6), as it is the case for the podcast interview.

The behavior of the non-naturally occurring conversation can be explained by the research protocol used to collect the event. Participants to dialogic events that have been elicited do not have a reason to engage with their interlocutor other than to comply with the researcher's requests. In the absence of a reason to engage with the interlocutors, they may not have an interest in negotiating the role of speaker in the conversation.[5]

[5] In conversation, the visible signs of this negotiation are elements like overlap, backchannel, disfluencies, and other devices that are used to either claim or give the floor.

Interactional dynamics that distribute the conversational roles in a rigid fashion could be a convenient fallback to maintain the exchange enough to fulfill the researcher's request. In other words, one of the participants takes over the role of main questioner and directs the exchange, prompting the responder to produces responses that are quite long for a conversation. This interactional dynamic results in an exchange where both speakers produce relatively long uninterrupted turns that receive no signals of engagement from the other participant. This situation is unmatched both in the naturally occurring conversation, where the two speakers visibly engage with each other and, in the podcast, where the main responder occasionally shifts out of their role to engage the interlocutors in the co-constructions of narratives (De Fina and Perrino 2011).

This analysis shows that spontaneous and naturally occurring conversations are non-orthogonal categories and that interactional dynamics contribute to the structuring of discourse in ways that cannot be explained solely in terms of the genre of a speech event. Given this, it is worth considering whether different approaches to the collection of spoken discourse equally respect the naturally occurring interactional dynamics of an event. We argue that focusing on representing the lives of speakers, rather than language structural features alone, reduces the risk of introducing artificial dynamics to the data. For this reason, we suggest an approach to corpus compiling that prioritizes the representation of the lives of its participants. In the following sections, we demonstrate how this goal can be achieved through the 'cast the net wide' framework.

2. DATA

Data for this paper consist of recordings from different varieties collected by the authors in the context of projects with different goals and a diverse range of recording protocols. All the recordings have been transcribed into IUs according to the Discourse Functional Transcription system (Du Bois *et al.* 1993). These data are complemented by excerpts from the SBCSAE (Du Bois *et al.* 2000) and the *Switchboard Corpus* (Godfrey *et al.* 1992).

Kazakh data are extracted from the MULTCORSKL (Filchenko *et al.* 2023). As of February 2024, the corpus is composed of 150 hours of recordings (80 of which have been transcribed), produced by over 100 participants. The transcribed portion of the

corpus comprises approximately 23 thousand IUs and 600 thousand words.[6] The MULTCORSKL is a corpus of video and audio recordings of a diverse range of naturally occurring social interactions taking place in Kazakh-speaking communities in Kazakhstan and China. Among these interactions there are conversations, lectures, traditional events featuring storytelling (*aitys*), interviews, task-oriented exchanges (e.g., food preparation and games), social, and religious events. Further data will be collected and annotated through the end of 2024. We complement these data with recordings collected in the training phase of the project, which were excluded from the corpus for not fully complying with the criteria for inclusion.

Data featuring Italian and Bustocco (Western Lombard) were collected in Busto Arsizio (northern Italy) by the first author during summer 2018 as part of a documentation project. A total of three hours was collected, distributed across two conversations (one among women friends in their 70s and one between a grandmother and her nephew), an elicited narrative, and poems.

Data featuring varieties of Mixtec come from two sources. The first narrative features Jeremías Salazar speaking in Sà'án Sàvĭ ñà Yukúnanĭ, a language of the Mixtec family from Oaxaca (Mexico). The narrative was collected by the first author in the context of a collaborative documentation effort and took place in California (Salazar *et al.* 2021). The second narrative features Juan Miranda speaking in Tù'un Na Ñuu Sá Matxí Ntxè'è, another language of the Mixtec family from Oaxaca (Mexico). The event was collected in San Martin Duraznos (Oaxaca) as part of a documentation effort (Auderset and Hernández Martínez 2021). Both projects are part of a larger endeavor to document varieties of Mixtec spoken in Oaxaca, Puebla, and Guerrero (Hernández Martínez *et al.* 2021).

3. Beyond spontaneous discourse

To meet corpus users' demand for everyday language data, one particularly productive approach employed by corpus designers is that of prioritizing the collection of spontaneous discourse; see Raso and Mello (2014) for a discussion of the reasons for the choice. Spontaneity is either identified through the analysis of the structural features of the language used in an event (Pitt *et al.* 2005) or it is assigned as a feature to specific

[6] Kazakh is a highly agglutinative language, so a metric like number of words is only partially informative.

genres, for example, one can collect only conversation (Love *et al.* 2017) or only events that do not display features typical of written language (Čermák 2009). In this section, we raise issues with the notion of 'spontaneity'. Firstly, we demonstrate how structurally-based definitions of spontaneity fail to capture events when participants deviate from the expected standard for their own interactional motivations. Secondly, we show how lack of spontaneity is not an inherent quality of specific events, but rather a feature that speakers can mobilize for communicative purposes. Finally, we prove how speech events of the same type can end up displaying different interactional qualities as a consequence of the methodological protocol employed in data collection. In general, we argue that fitting speech events into aprioristically determined definitions of spontaneity is often done at the expense of the recognition of speakers' agency.

3.1. Spontaneity as a structural feature

When spontaneity is defined in structural terms, it is not clear which feature (if any) is necessary and sufficient to uniquely identify speech as spontaneous. One could try to derive a list of these features by contrast with lab-produced speech, but phoneticians point out that, even in controlled experimental environments, speakers rarely exhibit features traditionally associated with contrived data (Xu 2010). Hence, features of contrivance are not enough to identify elicited speech events, but the definition could still maintain discriminating power in the opposite direction, that is, it could be that their presence is enough to qualify a speech event as non-spontaneous. To check whether this is the case, let us consider the conversational behavior of two of the most widely-recognized (Xu 2010) hallmarks of non-spontaneous speech: 1) slow speech rate and 2) hyperarticulation.

Isolated occurrences of hyperarticulation and slow speech rate are used in conversation to signal errors and repairs (Biro *et al.* 2022). This function can be maintained in extended instances, in specific cases like the one presented in Text 3. In this excerpt, Luca (native in Italian) is visiting his grandmother Pina. Pina employs hyperarticulation to repeat and correct the words her grandson mispronounces in Bustocco. Luca employs hyperarticulation to accommodate his grandmother, who has suffered from hearing loss, by repeating portions of speech that she has not heard. In Text 3, Luca asks Pina about the last time she visited a mountain peak she used to visit with her late husband.

TEXT 3

1	PINA;	*E qui*	And here
2		*la pazienza*	(we need) patience
3		*sem qui [inscì]*	we are here like this
4	LUCA;	*[L'ultima volta] che te s'è andà su al Tornion qua-*	the last time you went up the Tornion wh-
5		*quando l'è stata*	When was it
6		*(0.2)*	
7	PINA	*eh*	uh
8	LUCA;	*%L'ultima volta*	The last time
9		*Che te s'è andà su*	That you went up
10		*Al Tornion%*	The Tornion
11	PINA;	*Ah il nonno è morto nel dieci*	Ah your grandfather died in (twenty-)ten

As Pina is recounting her impossibility to go hiking because of health issues (IUs 1–3 are the ending portion of the turn), Luca asks about the last time she has been on a specific mountain peak (IUs 4–5). Pina does not understand the question and asks for repetition (IU 7). At this point, Luca raises his voice, slows down the speech rate, and hyperarticulates the elements he is producing. He splits the original content of IU 4 (*the last time you went up the Tornion*) into three IUs (8–10), each one containing a syntactic constituent (*the last time*, *that you went up*, *to the Tornion*), which he hyperarticulates. This is done to facilitate Pina's comprehension and, when she provides an answer (IU 11), the exchange resumes.

Text 3 features an unplanned informal speech event that routinely takes place in the participants' lives, and yet, it extensively presents hallmark features of non-spontaneous speech. Though there are cases in which these features are in fact encountered in non-spontaneous speech, they can also occur as a consequence of accommodating to the communicative conditions of the interaction. This strategy is used by adults interacting with children (Kuhl *et al.* 1997; Uther *et al.* 2007), caregivers with elders (Kemper 1994), and native speakers with L2 speakers (Kangatharan *et al.* 2021). In other words, employing hyperarticulation and slow speech rate is a productive interactional strategy adopted in situations with an unbalance of linguistic competence. Eliminating these linguistic features from the corpus may lead to a reduced visibility of the speakers who produce them, a limitation which corpus compilers are aware of and deal with in different ways. Ultimately, if we deem linguistic accommodation to be a usual occurrence in the life of many language users, then it is of value to include it in a corpus of spoken

discourse. This was a particularly pressing issue in the case of the MULTCORSKL because of the linguistic landscape in which we are operating. Because Kazakhstan displays a large-scale institutionalized multilingualism with age stratification (Agbo and Pak 2017), cases of linguistic accommodation are a frequent occurrence of life. This is by no means unique to Kazakhstan, yet representation of L2 speakers tend to be confined to specialized corpora.

3.2. Spontaneity as a genre feature

An alternative definition of spontaneity conceptualizes it in terms of lack of planning of an event. In this approach, spontaneity is evaluated along a continuum and treated as a feature of the genre (or register) of a specific speech event (Swales 1990; Blackwell and White 2018). Social activities are arranged in relation to each other according to the level of spontaneity that is allowed by the norms regulating them. These norms maintain a certain degree of stability within a culture and can vary cross-culturally. For example, religious rituals in the Catholic world are rigorously planned both in the structural order of the sequences to be performed and in the words to be used (Szuchewycz 1994). Shamanic rituals in Buryatia (or elsewhere in Russian Siberia) may include relatively unplanned exchanges between the petitioner and their ancestor' spirits (Quijada *et al.* 2015; Nagy 2016).

In addition to cross-cultural differences, the association between degree of planning of an individual speech event and genre is not guaranteed even within the same cultural context. Within the same genre, different traditions may have an impact over the shaping of the single event. For example, a conference presentation for a historian entails the reading of a written document prepared in advance, a practice that, at most linguistics conferences, would signal someone as either a novice or an outsider. Even variations at the individuals' level can have an impact over the degree of planning of an event. Recent years saw a rise in social media content about mental health (Haq *et al.* 2022) that led terms and tools from psychology to sip into the everyday vocabulary of certain demographics (Scherlis 2023). Among these tools, there is a script for conflict management called 'I-statements' (Rogers *et al.* 2018), for instance*, I felt ignored* instead of *you ignored me*. The popularization of this communicative script makes it so that, even within the same demographic, an instance of the 'couple fight' can either play out as a potentially volatile improvised event or as a carefully planned exchange featuring pre-

ordered sequences of grammatical structures that have been selected in advance.

Furthermore, degree of planning can vary even within the course of the same speech event, as there is no guarantee that what started as an event of a specific genre will continue as such (Schegloff 1988). This is especially visible inside conversation, which functions as an interactional matrix where participants can embed texts of different nature. For example, in Text 4, extracted from the SBCSAE, two women are discussing the impeachment of former US president Bill Clinton.

TEXT 4

216	MAUDE;	Okay,
217		Here it is,
218	LONI;	[Oh].
219	MAUDE;	[I] got it a little backwards.
220		(...) (H) (%)
221		Uh:,
222		This is Article Two,
223		Section Four.
224		<READ (...) (H) The President Vice-President and all civil officers of the United States,
225		shall be removed from office on impeachment for,
226		(H) a:nd conviction o:f,
227		(.) (%) treason,
228		(.) bribery,
229		(H) or,
230		(.) other high crimes,
231		(...) and misdemeanors. READ>

In the midst of the conversational exchange, Maude retrieves a law handbook and moves from a stretch of unplanned speech (IUs 216–223) into reading the definition of *impeachment* to her interlocutor (IU 224 onwards). The written excerpt was planned and produced way before the moment of the interaction, and it is used by a participant to construct an argument in favor of her position. In this case, we can observe a movement from unplanned to planned discourse. A movement in the opposite direction can be seen in Text 5, where the speaker inserts a chunk of unplanned speech into an otherwise planned event.

TEXT 5

1	GINETTO;	*<RECITE tuta sta beleza,*	all this beauty,
2		*la ma muisna ul coeui,*	makes my heart tender,
3		*E men,*	and I,
4		*Ca iu soi da buen ora RECITE>,*	who have been here for an hour,
5		*(.)*	(.)
6		*<L2 C'è un errore qui scusa >?*	Excuse me there is a mistake here?
7		*(.)*	(.)
8		*Ma muisna ul coeui?*	Makes my heart tender?
9		*/a mən/.*	/a mən/.
10		*<L2 io ho detto >,*	I said,
11		*/e mən/.*	/e mən/.
12		*<L2 Possiamo [interro][₂mpere:],*	Can we stop,
13	GIORGIA;	*[mhm],*	mhm
14		*[₂si certo].*	sure.
15	GINETTO;	*ripartire:,*	restart,
16		*C'- c'è [₃una pa]usa,*	there is a pause,
17	GIORGIA;	*[₃assolutamente]>.*	absolutely.

In this case, the speaker, Ginetto Grilli, is a renowned poet writing in Bustocco. During what the first author originally intended as a recording session of a conversation, Ginetto offers instead to recite a series of poems. During the declamation, he moves between planned recital speech and conversation to comment on language issues. In this excerpt, we can observe one instance where he interrupts the recital to repair his performance (IU 5). Same as Maude, Ginetto is merging material with different levels of planning for communicative purposes.

A potential objection to our criticisms to the unreliability of the concept of spontaneity is that corpus designers can always suspend the application of structural definitions in cases like the exchange between Pina and Luca (Text 3) and ignore local instances of planned speech in the case of Maude (Text 4). We do not doubt that this is true, but we contend that the issue with a structurally-based or a genre-static definition of spontaneity lies in the number of 'exceptional' cases that one needs to account for, in the nature of those 'exceptional' cases (Hall 2008), and in what they reveal about which speakers and genres get to set the standard of what counts as linguistic data of interest. Moreover, we contend that definitions which leave up to the individual researcher the interpretation of (frequent) deviations from the norm require more work to explain the outliers than they provide heuristics for the selection of spontaneous speech.

3.3. Spontaneity and data collection protocols

The final issue with the concept of spontaneity lies in the fact that it can be impacted by the methodology employed for data collection, which makes it difficult to compare events of the same type across corpora. For example, consider two instances of spontaneous phone conversations. Text 6 is taken from the *Switchboard Corpus* (Godfrey *et al.* 1992). In this corpus, participants chose some topics of conversation from a list of prompts and were matched by a robot operator with a stranger that had selected the same interests.

TEXT 6

Topic 303:
THE TOPIC IS CLOTHING. PLEASE FIND OUT HOW THE OTHER CALLER TYPICALLY DRESSES FOR WORK. HOW MUCH VARIATION IS THERE FROM DAY TO DAY? HOW MUCH VARIATION IS THERE FROM SEASON TO SEASON?

B: okay hi
A: hi um yeah I'd like to talk about how you dress for work and and um what do you normally what type of outfit do you normally have to wear
B: well i work in uh corporate control so we have to dress kind of nice so i usually wear skirts and sweaters in the winter time slacks i guess and in the summer just dresses
A: um-hum
B: we can't even well we're not even really supposed to wear jeans very often so it really doesn't vary that much from season to season since the office is kind of you know always the same temperature
A: and is right right is there is there um any is there a like a code of dress where you work do they ask

In Text 6, both interlocutors display typical features of unplanned informal interaction such as disfluencies (*um*), floor holders (*well*, *you know*), and contractions (*doesn't*, *I'd like*, *can't*). In terms of conversational organization, backchanneling is present (*um-hum*), but turns are rather long, and overlapping is limited. Similarly to the non-naturally occurring conversation in Table 1, participants coalesce around fixed interactional roles: A takes on the role of questioner and B that of responder. This strategy gives participants a structure to maintain engagement. Compare this dynamic with the exchange in Text 7, extracted from the SBCSAE, which features spontaneous naturally occurring discourse.

TEXT 7

1 >ENV: ... ((RING)) ... ((RING))
2 JILL: ... Hello=,
3 JEFF: .. How's my favorite girl in the world.
4 JILL: (H) Hey .. ba=by.
5 JEFF: .. Who's —
6 (H) Who's the girl that .. I love?
7 JILL: @@@@[@]

8 JEFF: [Who's] the girl that I'll do anything for?
9 JILL: @@@@ .. [2(H)]
10 JEFF: [2I'll] wash her feet with my [3mouth].
11 JILL: [3@@]@ @Aw= @honey,
12 [4(H) how] are you[5=].
13 JEFF: [4@@@]
14 [5after] a ten mile run.
15 JILL: Ew.
16 .. [@@]
17 JEFF: [I would].
18 JILL: .. Ew.
19 JEFF: I would.
20 JILL: Ew.

In Text 7, turns are considerably shorter than in Text 6. Overlap between participants is common and engagement is supported not by interactional roles, but by dialogic resonance, the repetition of linguistic structures across different intonation units (Du Bois 2014). Here, Jeff and Jill have similar hierarchical standing in the exchange, and this results in an event with different interactional features than Text 6.

The excerpts in Text 6 and Text 7 are both spontaneous but sourced through different research protocols. Consequently, participants display different levels of engagement in the event. Participants in the *Switchboard Corpus* excerpt have been exposed to structures and lexical choices in the written prompt, which may have primed them to use these structures. As it was for Text 2, participants will likely not engage with the opinions of their interlocutor beyond the current exchange, which results in them falling back onto rigid interactional structures to sustain engagement. Lack of engagement from participants has also been cited as the reason why the *Switchboard Corpus* displays different dynamics of information flow when compared to other corpora of English (Wasow 2002). Despite these considerations, data from the *Switchboard Corpus* are still routinely employed in studies of spontaneous discourse in a wide variety of disciplines because it is considered that they are "natural (appeared in spontaneous speech), can form an intonational phrase, and [their] frequency [is] not driven by a specific and uncommon genre" (Arnon and Snider 2010: 71). The sustained popularity of the *Switchboard Corpus* suggests that corpus users are keen on using conversational data and that there is indeed a need for more data of this kind. However, conversational data should not be sourced at the expenses of interactional dynamics.

4. Recording naturally occurring discourse

Instead of focusing on linguistic structures, one may want to place their focus on the users of a language instead. Speech events do not exist outside of speakers' practices (Duranti and Goodwin 1992) and, as such, they do not exist outside of speakers' lives. We propose that representation of participants' lives can be achieved by building a corpus of naturally occurring discourse. For a speech event to be considered 'naturally occurring' it must happen for the motivations of the participants, have consequences on their lives beyond the moment of the recording, and take place even if it was not going to be recorded (Du Bois 2003). Adopting this framework, only recordings that are culturally, socially, and interactionally relevant to the participants can be included in the corpus.

The 'cast the net wide' model was first successfully implemented by the SBCSAE (Du Bois *et al.* 2000). In this section, we discuss the ways in which the team of the MULTCORSKL adjusted its own original protocol to ensure that the corpus would contain only naturally occurring discourse. We mainly had to reconsider: 1) the level of agency that we wanted to grant participants, 2) the role of researchers in the community and in the events, and 3) the place of conversation in the corpus. Other adjustments related to the use of remote data collection protocols due to the impact of COVID-19, as well as smaller adjustments that are specific to the Kazakhstani geography, will not be discussed here (Troiani *et al.* 2022). We structured this discussion around the points that required a reconsideration of our theoretical assumptions about the nature of discourse, because we think they can offer elements of reflection to teams setting out to build a corpus.

4.1. Discourse and the agency of participants

The largest innovation introduced by the 'cast the net wide' framework is the notion that the only criterion for inclusion of speech events in a corpus should lay in the participants' motivations. The reason for this choice has been explained in Section 1, namely, speech events that are inconsequential to participants' lives yield artificial interactional dynamics that, in turn, have effects on grammar. In the multilingual context of Kazakhstan, one of the immediate ways in which lack of motivation became visible in grammar was the erasure from the data of the otherwise daily occurrences of code-switching with Russian, Mandarin Chinese, Uzbek, and English. Participants tended to control the amount of code-switching they produced and explicitly ascribed the reason to the fact that the team recruited them to participate in a corpus of Kazakh language. Assuming that the research

team was after exclusively ‘pure Kazakh’, they restrained themselves from producing code-switching. In response to this, we began introducing the project to potential participants as an endeavor to capture instances of life in Kazakhstan, rather than instances of Kazakh language. Far from being trivial, this shift resulted in the preservation of everyday multilingualism in the speech events.

A somewhat technically demanding adjustment was the decision to relinquish control over the speech event to the participants themselves. Wherever possible, we trained the participants to handle cameras and recorders and left it up to the individual to record events happening in their life. This protocol had been successfully employed in the SBCSAE. The implementation of such a protocol requires researchers to spend time training participants, to engage in consistent communication with participants to ensure the smooth operation of the recordings, and an availability of time (to train participants in the handling of equipment) and material resources (to distribute equipment across participants). When delegating the recording in its entirety was not possible, we let participants propose to us the events they wanted to have recorded. In our case, enlisting the help of participants in the planning and recording of events was worth the effort, especially in cases of cultural significance, culture-specific restrictions, or religious value that the presence of researchers would have disrupted, and which could easily be attended by members of the community that usually participate in these events.

Perhaps the most complex adjustment to be implemented in order to ensure the full agency of participants was accepting that they could bring and impose their own expectations of the event in the recording session. For example, consider the following interaction, which features the first author and four speakers of Bustocco, all women above 60. The first author had originally intended to collect some sociolinguistic information about the participants, set up the recorder, and leave them alone. Consent had been collected days prior to this event. After one of the participants mentions that she was acquainted with a relative of the first author, the whole event is re-casted by the participants as an interview, as can be seen in Texts 8 and 9.

TEXT 8

1	CINZIA;	*io ho vent’anni in piu’?*	I am 20 years older?
2		*Hai visto [la mia data],*	you saw my date (of birth),
3	GIORGIA;	@@	@@
4	BARBARA; > CINZIA	*[oh ascolta],*	hey listen,
5		*Lascia parlare –*	let (-) speak –

6		*Com'è che si chiama.*	What's your name.
7	GIORGIA;	*Giorgia.*	Giorgia.
8	BARBARA;	*Eh lascia parlar la Giorgia adesso.*	Let (the) Giorgia speak now.
9	GIORGIA;	@@@	@@@
10		*No,*	No,
11	CINZIA;	*eh.*	what.
12	BARBARA;	*lascia parlar [la Giorgia adesso].*	Let (the) Giorgia speak now.
13	GIORGIA;	*[va benissimo],*	It's fine,
14	BARBARA;	*e' lei che deve farti le domande,*	she is the one that has to ask you the questions,
15	ANNA;	*se deve dirci qualcosa ce lo dica?*	If you have to ask something let us know?

In Text 8, as Cinzia is in the middle of a complaint (IUs 1–2), Barbara interrupts her to demand that the first author be allowed to speak (IUs 4–6). Barbara repeats the request a few times (IUs 5, 8, 12), while the first author overtly states her intention not to be part of the interaction (IUs 9–10 and 13). At this point, Barbara explicitly details her expectations about the conversational roles in the event, and namely that the first author is the person in charge of asking questions (IU 14). Anna affiliates with Barbara by assuring the first author that she is in charge of directing the interview as she sees fit (IU 15). Barbara, Anna, and Cinzia frame the event as an ethnographic interview, and they hold each other accountable for their role as interviewees in the exchange. The research tools (sociolinguistic questionnaire) that the first author brought into the event are also actively re-purposed to fit the participants' interactional goals. For example, consider Text 9.

TEXT 9

BARBARA;	*C'è anche un cimitero che è stato- –*	There was also a cemetery that has been –
ANNA;	*ecco e allora lì c'era una casetta,*	Right and then there was there a little house,
CINZIA; > GIORGIA	*cià andiamo avanti,*	Come on let's move on,
	<L2 ndem innanzi >,	<L2 Let's move on >,
ANNA;	*c'era qualcosa?*	Was there something?
	Lì dove han fatto –	There where they did –
CINZIA;	*<READ usa il bustocco >,*	< READ do you use Bustocco>,
GIORGIA;	*quanto spesso usa il bustocco.*	How often do you use Bustocco.
CINZIA;	*spesso spesso,*	Often often,
	Spessissimo,	Very often,
GIORGIA;	*tutti i giorni?*	everyday?

As other participants are still engaging the previous question, Cinzia grows impatient,

leans in on the form that the first author is holding in hand, and requests her to continue with the next question. Cinzia reads the question herself and provides a short response that she will later expand into a narrative. At this point in the exchange, the sociolinguistic questionnaire has become a prop for the interlocutors to continue their conversation and the first author adjusted to meet the participants' expectations by asking questions when provided with the opportunity.

The event in Texts 8 and 9 would have not been recorded, had the first author not given up on her expectations about how a conversation between friends should look like. After the recorder is set up and one of the participants instructed to push the button wherever comfortable, as the first author is getting ready to leave, Cinzia explicitly motions her to sit and fill out the questionnaire. The first author's positionality leads the participants to assume that her interactional role in the exchange is that of interviewer. Specifically, the fact that a participant is acquainted with the grandmother of the first author contributes to the re-casting of the event in terms that are familiar with the participants.[7] As Cinzia and the first author move through the form, they adhere to the 'school interview' expectation that Cinzia has in mind. In response to the questions, Cinzia offers long narratives that the other participants end up co-constructing through commentaries, addition of details, banter, and comparisons to their own life experience. About 20 minutes into the event, the first author finally accepts that the event that is unfolding is indeed an instance of naturally occurring conversation and records the following 3.5 hours. The data that resulted from this session is a conversation among four friends embedded into an overarching sociolinguistic interview. The sociolinguistic interview is not by itself a genre that happens for the sake of participants' motivations, but this speech event shows how participants and researchers can co-construct the naturalness of an event by letting participants' agency take over the event. In this case, participants engaged in the interaction with the goal of telling their life stories to a researcher, member of the community, who could have easily been one of their grandchildren. Adjusting to these expectations resulted in the collection of a naturally occurring event.

[7] Participants refer to the doctoral program in which the first author is enrolled at the time of the research as 'school', and they mention having already participated in assignments where students are asked to interview older members of their family.

4.2. Discourse and the positionality of the researcher

Adopting the motivations of participants as base criterion for data collection also requires reconsidering the figure of the researcher and the role of the researcher in the speech event. Scientists are traditionally encouraged not to participate in recordings to preserve data authenticity (Potter 2002). This observation makes sense in a context where the researcher is not a member of the community, because the interactional dynamics of the exchange are artificially constructed. The consequences of merging the roles of community members, researchers, and participants can be particularly well appreciated in elicited data. Consider for example the following two excerpts featuring two varieties of Mixtec. In Text 10, Jeremías Salazar is retelling a childhood memory to the first author. The narrative is in Sà'án Sàvĭ ñà Yukúnanĭ, a language the first author does not speak nor understand.[8]

TEXT 10

1	JEREMIAS;	*kuĕ níkùu tavă-kue-yì-tí cha ñàà*	they could not take it out
2		*Tsà'ă ñà-ka*	the reason is
3		*tí luu-ni kìsi luu-ka ra nìkee viĭ-ni chùún-ka cha tá nchò'ŏ-tí ra*	the pot was so little that the chicken barely fit so when it cooked
4		*ñàà kuĕ níkùu tavă-kue-yì-tí cha saán kúu-ñà ñàà*	they could not take it out that is when
5		*Ntsà'àn-kue-yì ra ñàà*	they went and
6		*nìntà... nì*	he
7		*nìkăni-à mátsá'nù tavà ñàà*	he called my grandma so (then)
8		*nchìnchiĭ-kue-yì nixi să'a-kue-yì ñàà távî chùún-ka*	she would help them get the chicken out
9		*cha tíí saán ntsàĭ ra ñàà*	probably she just got there
10		*ntsă'nchĭ chùún-ka tíí ñàà*	she cut the chicken

This narrative was elicited in the context of a documentation project. Jeremías chose the topic in advance, but he did not script the content, which resulted in a somewhat unplanned speech. This excerpt features disfluencies (Belmar and Salazar 2023) and repairs (IU 6), but no backchannel from the first author. Jeremías does not attempt to include the first author in the exchange. The elicitation protocol presented here is not uncommon among documentary linguists and it yields a considerably different result when the event is elicited by a community member. In Text 11, Juan Miranda is telling a narrative in Tù'un Na Ñuu Sá Matxí Ntxè'è to a young woman member of his community

[8] Materials are accessible at https://sites.google.com/view/saansavi-yucunani (accessed 25 September 2023).

and speaker of the variety.[9]

TEXT 11

1	JUAN;	*xa kasantxeena xa sa'ana lucha ñi'i ñi'itana shu'un*	they are already sowing they are already struggling they are finding finding money
2		*ta chixin madre nte'i chi koo mi ña koo ña'a vi*	and before there were no mothers who were poor there are no well
3		*unkivi koo mi modo kakio*	we cannot save ourselves in any way
4		*ta vitxi*	and now
5		*ta kachi ji'un*	as I told you
6		*Ntxe'e*	look
7		*yo'o*	you
8		*Kunte'ivo ñaa*	we(.incl) are poor because of this

As in the case with the first author in Text 10, Juan's interlocutor never claims the floor and provides only occasional backchannel. But in this case, she is a plausible audience for the event and the lack of overt verbal engagement does not prevent Juan from involving her in the event through various strategies. In IU 5, he recalls a previous point in the conversation to establish that he and the interlocutor share joint knowledge (*as I told you)*. In IU 6, he uses an attention-grabbing expression (*look*) and in IU 7 he overtly addresses the interlocutor indicating her as a recipient of the narrative through the use of a second person pronoun (*you*). Finally, in IU 8 Juan shifts to an inclusive first-person plural (*-o* in *kunte'iv-o*) and recruits the interlocutor in the narrative (*this is why we are poor*) acknowledging that the interlocutor has experience on the topic of the narrative, namely, the economic conditions of the Pueblo of San Martin Duraznos.

The narratives in Texts 10 and 11 have both been elicited, but Text 11 has been elicited enlisting community members as recorders and creating a condition where participation in the event would be plausible. This gives Juan a motivation to ensure the interlocutor understands the story, which results in a significant alteration of grammatical structures employed in the event. The data resulting from the elicitation in Text 11 are not 'naturally occurring', yet they can be considered naturalistic, that is, elicited data that attempt at recreating the interactional dynamics that would have been at play in naturally occurring discourse (e.g., by ensuring that data are elicited by a member of the speaker's community). While accessing naturally occurring data may not always be possible (e.g.,

[9] The original transcription was translated into Spanish. The English translation was done by the first author. Interactional elements, disfluencies, and truncation are not transcribed in the original data.

in communities with a small number of speakers), a valid alternative is to employ elicitation protocols which make an explicit point about maintaining interactional dynamics. One such model is the Pear Story elicitation protocol in its original formulation (Chafe 1980) or in the modified version featuring a three-party interaction (Kibrik and Fedorova 2018).

If the recorder is recognized as a member of the community, their refusal to intervene could alter the dynamics of the event more than letting themselves be recruited into the participants' plans. This recruitment can play out in extreme forms like the case of the first author being involved in the entire event of Texts 8 and 9 are taken, or it can materialize in shorter occurrences like the one in Text 12. Here, the researcher (Serik) is operating the camera and recording a conversation between his grandparents. The event unfolded for a couple of hours. Aisha and Adilet have been conversing in a room by themselves, Serik occasionally checks on the camera.

TEXT 12

1	AISHA;	*Taksïmen mına jazwım bitse qaytıp ketemiz dep otır*	He is saying he will go back to the city by taxi as soon as he is done with his recordings
2		*Qalağa (H)*	To the city
3		*endi voobşçe ketemin deydi ğoy*	He is saying he will be gone for good
4		*Sen oqwğa ketesiŋ be*	Are you leaving for school?
5		*Astanağa*	To Astana
6	ADILET;	*Oqwın bitte emes pe ey*	Don't you know he has already finished his studies
7	SERIK;	*Altısında ketemin*	I'll leave on the sixth
8	AISHA;	*A.*	Come again?
9	SERIK;	*Altısında ketemin*	I'll leave on the sixth
10	AISHA;	*Altısında*	On the sixth?
11	ADILET;	*Astanağa*	To Astana
12	AISHA;	*Ïä ket- ketedi Astanğa*	Yes, he will leave for Astana

This excerpt comes after a period of silence in which Aisha is knitting and Adilet is resting. Upon noticing Serik close to the camera, Aisha tells Adilet that their grandson is about to return to leave town (IUs 1–3). Aisha addresses Serik directly and asks whether he is returning to school (IUs 4–5). Serik responds (IU 7 and 9) and Aisha and Adilet resume their conversation. In this case, Serik's involvement in the recording is not only plausible, but also necessary to maintain a natural dynamic. Serik is the researcher, but he is also Aisha's grandson, and it is in this capacity that his intervention is required. Participation of researchers into recordings is a practice that was not actively encouraged

within the MULTICORSKL, but it was not prohibited when the researcher's presence in the event is usual.

4.3. Discourse and conversation

The final adjustment needed to obtain a corpus of naturally occurring discourse was the reconsideration of the role of conversation within the corpus. The importance of conversation cannot be understated. Conversation is the most basic form of communication available to people and a universal of human language (Schegloff 2015). It is the vehicle through which speakers acquire, maintain, and modify grammar (Thompson *et al.* 1996; Lytle and Kuhl 2017). Moreover, the status of conversation as a standalone genre is contested by many scholars (see Warren 2006 for an overview of the debate). As seen in Section 3.2, conversation is best understood as an interactional matrix within which other genres are embedded as needed. For these reasons, conversation is a non-negotiable component of our corpus, a position that has been explicitly supported by other corpus designers, whether with the intention to cater to users' needs (Love *et al.* 2017: 324), highlight the role of spoken discourse in language acquisition (Raso and Mello 2014), or maximize resource efficiency by selecting data that are as different as possible from written discourse (Čermák 2009).

While the fundamental role of conversation in human life should not be understated, a corpus representing life cannot limit itself to conversation alone. Corpora like the SBCSAE and the MULTICORSKL record language in use for the motivations of the participants and source speech events that are consequential to the lives of participants. Though grammar is acquired in conversation, many life goals are achieved by genres other than conversation or otherwise spontaneous discourse. This includes goals that are not concretely measurable, such as the competent performance of culture and identity (Dingemanse and Floyd 2014). For this reason, the MULTICORSKL does not limit itself to conversation and includes a series of recordings featuring poem recitals, ritualistic animal slaughters, political speeches, interviews to culturally relevant figures, guided visits to historical sites, etc. We hope in this way to gain a representation of language in use that is as faithful as possible to the lives of speakers of Kazakh.

5. Conclusions

In this paper, we presented the theoretical assumptions underlying the 'cast the net wide' framework, an approach to data collection that aims at the compiling of corpora of naturally occurring spoken discourse. This framework posits as the goal for a corpus to be representative of the life of participants rather than of the linguistic structures in use. The reason for this decision is demonstrated in Section 1, where we present two case studies (backchannel and question-answer sequences) showing how the frequency and distribution of interactional features, as well as the structuring of interaction, varies as a result of whether the speech event under analysis is naturally or non-naturally occurring. The result of this analysis suggests that differences between speech events go beyond their genre and presence or lack of spontaneous speech.

We turn to demonstrate the issues with relying on the notion of spontaneity in Section 3. First, we show that the structurally based definition of spontaneity is not a reliable heuristic for the detection of unplanned speech, in particular when considering interactions that feature linguistic accommodation in situations of power unbalance (e.g., native speakers talking to L2 speakers, caregivers to elderly or children). Then we demonstrate how spontaneity is not a fixed quality of a genre, as language users can mix genres and degrees of planning to achieve their goals. The data presented suggest that spontaneous conversation overall acts as an interactional matrix where interlocutors can embed less spontaneous genres for their own reasons. Finally, we demonstrate how the same speech event can result in an exchange of varying spontaneity as a consequence of the research protocol employed in recording it. All these observations motivated us to define an alternative criterion for the selection of data in a corpus of spoken discourse.

As an alternative to the selection of structural criteria such as spontaneity, we propose an approach to corpus design that prioritizes the collection of naturally occurring discourse. We define naturally occurring speech events as events that happen for the social and interactional goals of the participants and would take place even if there were not going to be recorded, and, as such, are consequential to the participants' lives beyond the moment of the recording. We suggest that there are a few adjustments corpus designers can introduce to their workflow if they are interested in sourcing naturally occurring discourse. In particular, we suggest that participants are recruited as recorders and are given the opportunity to bring their understanding of the event. We also suggest that researchers who are present to an event comply with the understanding of the interactional

dynamics displayed by participants and involve community members in the data recording phase. We argue that these adjustments can help researchers approximate, as best as possible, the representation of language as it occurs in the lives of its speakers.

References

Agbo, Seth A. and Natalya Pak. 2017. Globalization and educational reform in kazakhstan: English as the language of instruction in graduate programs. *International Journal of Educational Reform* 26/1: 14–43.

Arnon, Inbal and Neal Snider. 2010. More than words: Frequency effects for multi-word phrases. *Journal of Memory and Language* 62/1: 67–82.

Auderset, Sandra and Carmen Hernández Martínez. 2021. Documenting Tù'un Na Ñuu Sá Matxí Ntxè'è, a mixtec language of Oaxaca, Mexico. *Endangered Languages Archive*. http://hdl.handle.net/2196/a3085a77-687a-48b9-9caf-a48c3c1f1f1f.

Biro, Tifani, Annie J. Olmstead and Navin Viswanathan. 2022. Talker adjustment to perceived communication errors. *Speech Communication* 138: 13–25.

Blackwell, James W. and Peter R. R. White. 2018. The building blocks of speech: Spontaneity, pre-packaging and the genre structuring of university lectures. *Text & Talk* 38/3: 267–290.

Burnard, Lou. 2002. Where did we go wrong? A retrospective look at the British National Corpus. In Bernhard Ketteman and Georg Marko eds. *Teaching and Learning by Doing Corpus.* Amsterdam: Rodopi, 51–70.

Čermák, František. 2009. Spoken corpora design: Their constitutive parameters. *International Journal of Corpus Linguistics* 14/1: 113–123.

Chafe, Wallace L. 1980. *The Pear Stories: Cognitive, Cultural, and Linguistic Aspects of Narrative Production.* Westport: Praeger.

Chui, Kawai and Huei-ling Lai. 2008. The NCCU corpus of spoken Chinese: Mandarin, Hakka, and southern Min. *Taiwan Journal of Linguistics* 6/2: 119–144.

De Fina, Anna and Sabina Perrino. 2011. Introduction: Interviews vs. 'natural' contexts: A false dilemma. *Language in Society* 40/1: 1–11.

Dingemanse, Mark and Simeon Floyd. 2014. Conversation across cultures. In N. J. Enfeild, Paul Kockelman and Jack Sidnell eds. *The Cambridge Handbook of Linguistic Anthropology*. Cambridge: Cambridge University Press, 447–480.

Drummond, Kent and Robert Hopper. 1993. Back channels revisited: Acknowledgment tokens and speakership incipiency. *Research on Language & Social Interaction* 26 2: 157–177.

Du Bois, John W. 2003. Discourse and grammar. In Michael Tomasello ed. *The New Psychology of Language: Cognitive and Functional Approaches to Language Structure.* London: Lawrence Erlbaum Associates, 61–102.

Du Bois, John W. 2014. Towards a dialogic syntax. *Cognitive Linguistics* 25/3: 359–410.

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.

Du Bois, John W., Stephan Schuetze-Coburn, Susanna Cumming and Danae Paolino. 1993. Outline of discourse transcription. In Jane A. Edwards and Martin D. Lampert *Data: Transcription and Coding in Discourse Research.* London: Lawrence Erlbaum *Talking*, 45–89.

Du Bois, John W. and Giorgia Troiani. 2022. *Cast the Net Wide: Corpus as a Slice of Life*. (Presentation, 25 February 2022*)*. Bologna: Italy.

Duranti, Alessandro and Charles Goodwin. 1992. *Rethinking Context: Language as an Interactive Phenomenon*. Cambridge: Cambridge University Press Cambridge.

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).

Godfrey, John J., Edward C. Holliman and Jane McDaniel. 1992. SWITCHBOARD: Telephone Speech Corpus for research and development. In *Proceedings of the International Conference on Acoustics, Speech, and Signal Processing.* San Francisco: IEEE Computer Society, 517–520. https://doi.org/10.1109/ICASSP.1992.225858

Greenbaum, Sidney. 1991. The development of the International Corpus of English. In Karin Aijmer and Bengt Altenberg eds. *English Corpus Linguistics: Studies in Honour Svartvik*. London: Longman, 83–91.

Hall, Kira. 2008. Exceptional speakers: Contested and problematized gender identities. In Janet Holmes and Miriam Meyerhoff eds. *The Handbook of Language and Gender*. New York: Wiley Blackwell, 353–371.

Haq, Ehsan-Ul, Lik-Hang Lee, Gareth Tyson, Reza Hadi Mogavi, Tristan Braud and Pan Hui. 2022. Exploring mental health communications among Instagram coaches. In Nitin Agarwal, Zongmin Ma and Jon Rokne eds. *Proceedings of the 2022 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining.* New York: IEEE Press, 218–225.

Heinz, Bettina. 2003. Backchannel responses as strategic responses in bilingual speakers' conversations. *Journal of Pragmatics* 357: 1113–1142.

Hernández Martínez, Carmen, Griselda Reyes Basurto and Eric W. Campbell. 2021. MILPA (Mexican Indigenous Language Promotion and Advocacy): A Community-centered linguistic collaboration supporting indigenous Mexican languages in California. In Justyna Olko and Julia Sallabank eds. *Revitalizing Endangered Languages: A Practical Guide.* Cambridge: Cambridge University Press, 216–217.

Kangatharan, Jayanthiny, Maria Uther and Fernand Gobet. 2021. The effect of hyperarticulation on speech comprehension under adverse listening conditions. *Psychological Research* 86: 1–12.

Kemper, Susan. 1994. Elderspeak: Speech accommodations to older adults. *Aging, Neuropsychology, and Cognition* 1/1: 17–28.

Kibrik, Andrej A. and Olga V. Fedorova. 2018. An empirical study of multichannel communication: Russian pear chats and stories. *Psychology. Journal of the Higher School of Economics* 15/2: 191–200.

Kucera, Karel. 2002. The Czech National Corpus: Principles, design, and results. *Literary and Linguistic Computing* 17/2: 245–257.

Kuhl, Patricia K., Jean E. Andruski, Inna A. Chistovich, Ludmilla A. Chistovich, Elena V. Kozhevnikova, Viktoria L. Ryskina, Elvira I. Stolyarova, Ulla Sundberg and Francisco Lacerda. 1997. Cross-language analysis of phonetic units in language addressed to infants. *Science* 277 (5326): 684–686.

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.

Lytle, Sarah Roseberry and Patricia K. Kuhl. 2017. Social interaction and language acquisition: Toward a neurobiological view. In Eva M. Fernández and Helen Smith

Cairns eds. *The Handbook of Psycholinguistics*. New York: Wiley Blackwell, 615–634.

Nagy, Zoltán. 2016. *The Khanty of Vasyugan. Change of the Religious System in XIX-XXI Centuries*. Tomsk: Tomsk State Pedagogical University Publishing House.

Oostdijk, Nelleke. 2002. The design of the spoken Dutch corpus. In Pam Peters, Peter Collins and Adam Smith. *New Frontiers of Corpus Research*. Amsterdam: Rodopi, 105–112.

Pitt, Mark A., Keith Johnson, Elizabeth Hume, Scott Kiesling and William Raymond. 2005. The Buckeye Corpus of conversational speech: Labeling conventions and a test of transcriber reliability. *Speech Communication* 45/1: 89–95.

Pomerantz, Anita. 1984. Agreeing and disagreeing with assessments: Some features of preferred/dispreferred turn shapes. In J. Maxwell Atkinson and John Heritage eds. *Structures of Social Action: Studies in Conversation Analysis*. Cambridge: Cambridge University Press, 57–101.

Potter, Jonathan. 2002. Two kinds of natural. *Discourse Studies* 4/4: 539–542.

Quijada, Justine B., Kathryn E. Graber and Eric Stephen. 2015. Finding 'their own': revitalizing buryat culture through shamanic practices in Ulan-Ude. *Problems of Post-Communism* 62/5: 258–272.

Raso, Tommaso and Heliana Mello. 2012. The C-ORAL-BRASIL I: Reference corpus for informal spoken Brazilian Portuguese. In Vládia Pinheiro, Pablo Gamallo, Raquel Amaro, Carolina Scarton, Fernando Batista, Diego Silva, Catarina Magro and Hugo Pinto eds. *Computational Processing of the Portuguese Language*. New York: Springer 362–367.

Raso, Tommaso and Heliana Mello. 2014. Spoken corpora and linguistics studies: Problems and perspectives. In Raso, Tommaso and Heliana Mello eds. *Spoken Corpora and Linguistic Studies*. Amsterdam: John Benjamins, 1–24.

Rogers, Shane L., Jill Howieson and Casey Neame. 2018. I understand you feel that way, but I feel this way: the benefits of I-language and communicating perspective during conflict. *PeerJ* 6: e4831. https://doi.org/10.7717/peerj.4831.

Salazar, Jeremias, Guillem Belmar, Catherine Scanlon, Giorgia Troiani and Eric W. Campbell. 2021. Bridging diaspora: Technology in the service of the revitalization of Sà'án Sàvĭ ñà Yukúnanĭ. In Eda Derhemi ed. *Endangered Languages and Diaspora.* Berkshire: Foundation for Endangered Languages, 176–185.

Schegloff, Emanuel A. 1988. From interview to confrontation: Observations of the bush/rather encounter. *Research on Language & Social Interaction* 22/1–4: 215–240.

Schegloff, Emanuel A. 2015. Conversational interaction the embodiment of human sociality. In Deborah Tannen, Heidi E. Hamilton and Deborah Schiffrin eds. *The Handbook of Discourse Analysis*. New York: Wiley Blackwell, 346–366.

Scherlis, Lily. 2023. Boundary issues. *Parapraxis.* https://www.parapraxismagazine.com/articles/boundary-issues

Stivers, Tanya, N. J. Enfield, Penelope Brown, Christina Englert, Makoto Hayashi, Trine Heinemann, Gertie Hoymann, Federicoi Rossano, Jan Peter, Kyung-Eun Yoon and Stephen C. Levinson. 2009. Universals and cultural variation in turn-taking in conversation. In *Proceedings of the National Academy of Sciences* 106/26: 10587–10592. https://doi.org/10.1073/pnas.0903616106.

Stivers, Tanya, Nick J. Enfield and Stephen C. Levinson. 2010. Question-response sequences in conversation across ten languages: An introduction. *Journal of Pragmatics* 42: 2615–2619.

Stivers, Tanya and N.J. Enfield. 2010. A coding scheme for question–response sequences in conversation. *Journal of Pragmatics* 42/10: 2620–2626.
Swales, John M. 1990. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge university press.
Szuchewycz, Bohdan. 1994. Evidentiality in ritual discourse: The social construction of religious meaning. *Language in Society* 23/3: 389–410.
Thompson, Sandra A., Emanuel A. Schegloff and Elinor Ochs. 1996. *Interaction and Grammar*. Cambridge: Cambridge University Press.
Tottie, Gunnel. 1991. Conversational style in British and American English: The case of backchannel. In Jan Svartvik, Karin Aijmer and Bengt Altenberg eds. *English Corpus Linguistics: Studies in Honour of Jan Svartvik*. London: Longman, 254–271.
Troiani, Giorgia, John W. Du Bois, Gulnar Sarseke, Andrey Filchenko, Ilya Salimzianov, Nikolay Mikhailov, Fatima Moldashova, Akyl Akanov, Moldir Bizhanova, Dameliya Koishybayieva, Aigerim Khamitova, Tomiris Nurgalyieva, Aigerim Seiilbek, Bybaris Seitak, Bota Tursunova and Aruzhan Yelubay. 2022. Remote workflow as educational opportunity: The experience of the Multimodal Corpus of Spoken Kazakh language. *Coyote Papers*: 11–18.
Uther, Maria, Monja A. Knoll and Denis Burnham. 2007. Do you speak E-NG-LI-SH? A comparison of foreigner-and infant-directed speech. *Speech Communication* 49/1: 2–7.
Warren, Martin. 2006. *Features of Naturalness in Conversation*. Amsterdam: John Benjamins.
Wasow, Thomas. 2002. *Postverbal Behavior*. CSLI Stanford: The University of Chicago Press.
Xu, Yi. 2010. In defense of lab speech. *Journal of Phonetics* 38/3: 329–336.

Corresponding author
Giorgia Troiani
Nazarbayev University
Block 7, office 7e.119
53 Kabanbay Batyr
01000
Astana
Kazakhstan
Email: giorgia.troiani@nu.edu.kz

received: October 2023
accepted: June 2024

RiCL Research in Corpus Linguistics

Design and construction of a social media corpus: Influencers' speech in vlogs

Hülya Mısır
University of Birmingham / United Kingdom

Abstract – This article outlines the creation of a social media corpus of Turkish vlogs on *YouTube*, aimed at analyzing the translanguaging practices and multimodal communication of Turkish social media influencers. It firstly describes the process of constructing the corpus, including transcription conventions and *ad hoc* annotation. The article then analyzes the phenomenon of translanguaging, with an emphasis on its prevalent forms and modes. Given the challenges associated with compiling a multimodally rich social media corpus, this paper provides strategies for manually transcribing and annotating linguistic and semiotic features in *ELAN* software, as well as strategies for managing tier-based annotations for vlog datasets. Additionally, the study presents approaches for handling non-standard linguistic codes and marked occurrences in language contact zones, illustrated through examples drawn from the vlog corpus where Turkish serves as the standard code.

Keywords – social media corpus; corpus design; corpus construction; vlog; influencer; translanguaging

1. INTRODUCTION[1]

Working with social media data is often challenging due to its messy nature, typically characterized by a high degree of noise and heterogeneity. Hence, data from the web as a corpus requires rigorous attention to the processes of corpus design, data collection, balance and representativeness, and pre-processing. Although various software and online platforms have emerged to streamline the collection and analysis of web-based corpora, the protocol for compiling corpora from the Internet may be more intricate than that of working with traditional written or spoken data. This is because web-based data is translocal and multimodal, which adds a layer of complexity to the decision-making process.

In the digital media landscape, communication is a multifaceted process that transcends linguistic and semiotic boundaries. Digital discourse often defies the constraints of a single

[1] I would like to thank the Scientific and Technological Research Council of Türkiye (Tübitak) for supporting my research, from 2018 to 2022, through the 2211-A National Ph.D. Scholarship.

language or mode of expression. Users have embraced the practice of combining their linguistic repertoires, engaging in what I will later term 'translanguaging' to navigate the intricate web of digital interactions (Mısır and Işık Güler 2023). Simultaneously, the integration of various resources like images, emojis, and videos has enabled users to express themselves more creatively and effectively. The digital media environment presents a complex space for communication, one that demands an in-depth exploration of the multifaceted nature of language use in this digital context.

Studies have shown that the language of social media is intertwined with an array of semiotic and material resources (Jacquemet 2005; Blommaert and Rampton 2011). The specific form of social media discussed in this paper is vlogs. Vlogs feature a combination of speech, text, and multimodal elements such as emojis, subtitles, or audio and visual components additive to videos. Therefore, focusing on these ubiquitous resources as modes and explaining the significance of their role in mediated communication, and how they co-create meaning, becomes essential.

The present study describes the process of designing and building a corpus of Turkish social media influencers' vlogs and shows non-standard language use and digital affordances used in interacting with an imagined audience by social media influencers. The corpus shows the features of real language-in-use evidenced by active content generators whose perceived influencing power is significant. However, corpora of the Turkish language, primarily spoken corpora, do not particularly take into account the expanding non-standard language use. This paper highlights a noteworthy aspect of non-standard language use, namely translanguaging, within the context of designing a corpus of contemporary language use. Translanguaging, as defined by Otheguy *et al.* (2015: 281), involves "the deployment of a speaker's full linguistic repertoire without strict adherence to the socially and politically defined boundaries of named (typically national and state) languages." In light of this diverse linguistic phenomenon, the challenge arises when devising annotation schemes for their incorporation into corpora. It is within the realm of corpus annotation that I descriptively demonstrate the pivotal role played by translanguaging. Specifically, I underline how its annotation becomes a critical concern in the context of computational analysis, as a substantial portion of social media discourse exhibits translanguaging, demanding specialized attention that can disrupt automation processes.

2. VLOGS AND MULTIMODAL DIGITAL COMMUNICATION

Vlogs (video + blogs) are audiovisual forms of blogging conventionalized on *YouTube*. Vlogs are considered self-mediated quasi-interaction, and have interactional patterns (i.e., one-to-many) that feature "mass-mediated monologue" performances (Dynel 2014: 41). Vlogs dominantly consist of user-generated content and are typically characterized by unscripted, informal monologues delivered by the vlogger, who serves as both the content creator and the vlog's subject (Frobenius 2011).

This style of content creation challenges traditional communication norms by introducing a new kind of audience —an 'imagined audience'— in the digital sphere. As a result, it transcends the boundaries of conventional language codes and communication modes, encouraging a shift away from a mere focus on "languages as distinct codes" (Zhu and Li 2020: 15). Linguistic investigations of these forms of contemporary mass communication and linguistic behavior in online social interaction have contributed to a broader understanding of language, emphasizing a global discourse of "translingual hybridity" (Kramsch 2018: 113).

In digital landscapes, individuals develop diverse multimodal ecosystems, accommodating various combinations of linguistic repertoires. Translanguaging theory is a responsive approach to this heterogeneity and superdiversity with its focus on the flexible and creative use of linguistic resources and vibrant linguistic repertoires (Li 2011). It acknowledges the interplay between people's repertoires, virtual repertoires, and general linguistic practices within communities, which organically evolve through lived experiences. Blommaert (2008: 16) emphasizes that language use in this digital age is not confined to any national or stable linguistic framework but is intimately tied to an individual's life journey, following the unique biographical trajectory of the speaker. This perspective shifts the focus from language as a rigid construct to an emphasis on the speaker's linguistic repertoire and practices.

In examining these contemporary mass communication and linguistic behaviors within online social interactions, it is essential to delve into the essence of multimodality, a concept that encompasses various resources for message composition, including textual, aural, linguistic, spatial, and visual modes (Schmidt and Marx 2019). Multimodality in social media communication can explain how users leverage this diverse range of resources to enhance the expressiveness and impact of their content, creating rich and engaging digital discourse. Scholars have increasingly recognized the importance of multimodality in digital communication, especially in content-driven environments like gaming (Schmidt and Marx

2019) or vlogs (Lustig *et al.* 2021). These studies explain how language, gaze, gestures, posture shifts, and the visual frame coordinate intermodally to make meanings by exploring people's relations with their domestic (material) environments. In the context of vlogs, a multimodal approach can foreground typical vlogging locations and settings, which play a central role in constructing the visual aspects and characteristics of vlogging with the regularized spaces and commodities. Multimodal elements are crucial for ensuring the integrity and effective presentation of the communicated content. As such, merely analyzing the transcribed texts of vlogs does not adequately capture the meaningful whole or flow of the expected turn-taking. To gain a more comprehensive understanding of the communication dynamics at play, it is essential to examine the accumulated repertoires of means and modes employed by vloggers and how they are coordinated in the communication process.

3. Corpus design and construction

In the present study, I describe the design and construction of a corpus of Turkish social media influencers' vlog content on *YouTube*. The influencers were selected through criterion sampling. The criteria include 1) speaking Turkish as their first language, 2) being based in Turkey as stated in the *YouTube* profile, 3) having a follower count of over 250,000 on *Instagram* and *YouTube*, 4) being an active content generator at the time of the data collection, and 5) having accounts open to public view. A noteworthy aim was to obtain an informed idea about the design features of the platform from which one collects data for the 'representativeness' of the corpus. Having applied this line of criteria, I aimed to represent the language and practices of macro-influencers, i.e., those who have a large number of followers and are represented by a professional agency, which indicates their established 'enterprise' status in the influencing market in Turkey.

Additionally, *YouTube* vlogs are categorized as synchronous ('go live') or asynchronous ('upload video'). The former, characterized by real-time interaction with the audience, prioritizes instant feedback, audience engagement, and unaltered content dissemination. By contrast, the latter entails the creation of pre-recorded, edited, and strategically planned video content. These two vlogging modes exhibit noticeable disparities in both structural organization and operational approaches, potentially giving rise to variables in corpus analysis attributable to the diverse characteristics inherent in synchronous and asynchronous vlogging. Hence, this study exclusively examines asynchronous vlogs to achieve the representativeness

of curated content where the content creator has complete control over what they want to share. They are more likely to reflect the creator's intended message and image.

The specialized snapshot corpus of Turkish influencers' communication contains 120,928 tokens of transcribed speech in vlogs posted between 2020 and 2021. The *YouTube* vlogs were chosen in chronological order of posting to avoid any bias in selection. However, videos that were three minutes or less, such as music clips, were not considered vlogs and were excluded irrespective of their posting order. The corpus design is presented in Table 1. A dataset of 30 videos was compiled by gathering five videos from each influencer's profile. The footage length ranges from 13 to 42 minutes. In the construction of the corpus, I considered video length and token count as critical criteria. Balancing by token count was important to ensure an equitable representation of each participant and context within the linguistic data. However, it is important to note that there exists a natural trade-off between video length and token count. Longer videos tend to contain more tokens, yet they may also include extended periods of silence or non-linguistic elements. In contrast, shorter videos, while having a smaller token count, can offer a more concentrated source of linguistic information. Recognizing this trade-off, I aimed for a balanced approach, opting to equalize both video length and token count.

ID (Number of transcribed vlogs)	**Total footage**	**Tokens**
DO (5)	120 mins 29 sec	19,056
EL (5)	113 mins 39 sec	18,179
DB (5)	130 mins 38 sec	22,834
KD (5)	124 mins 49 sec	17,549
EF (5)	140 mins 36 sec	22,851
MO (5)	127 mins 09 sec	20,459
Total (30)	757 mins 20 sec (12hs 37 mins)	120,928

Table 1: The details of the vlog corpus

The data was processed in *ELAN* (V.6.2; Wittenburg *et al.* 2006), a free tool for developing annotation and creating relationships between tiers. The software can incorporate speech segmentations in a time-aligned manner, transcriptions, part-of-speech annotations, and a limitless range of other modes annotated on different tiers. Such features facilitated surpassing the fundamental restriction of representing "all features of communication through the same mode —that of a textual record" (Knight *et al.* 2009: 2).

In *ELAN*, I initiated my workflow by importing videos and employing a predefined template that I had set up to maintain consistency across files, following the same annotation tier scheme. This standardized approach proved highly practical, particularly when applying the identical template to all videos, ensuring the comparability of tiers during the analytical phase. The annotation tiers consisted of 1) text, 2) tokenized tier, 3) translanguaging categories, 4) vlogging resources, and 5) consistent tiers for each participating speaker in the interactions. Each tier had a hierarchically sorted parent tier.

In processing the data, I worked in the annotation mode in *ELAN* to create utterance boundaries for 30 videos, deciding where the utterance began and ended. This segmentation process facilitated the transcription process and transcript alignment at the utterance level in the following steps. Upon completing the transcription on a tier called 'Text', the tokenization of this tier was performed automatically, which created a tier to place tokens individually. I created this tier to annotate translanguaging instances concerning the place of the token rather than the utterance as a complete line.

ELAN also supports metadata storage. Metadata from social media is shaped by what information is embedded in the structure of the platforms. For *YouTube*, metadata can be categorized into three types: a) automatically generated metadata (URL, date posted), b) semi-automatically generated metadata generated by clicks (metrics of views, dis/likes), and c) self-generated metadata (channel name, caption); see Schmidt and Marx (2019: 134). For this study, metadata included categories (a) and (c) and excluded the interaction data (b). Figure 1 shows a sample vlog metadata scheme formed in CMDI format, a relatively customizable format to display metadata in *ELAN*. Apart from the descriptive information in metadata categories (a) and (c), the CMDI files contained communicative functions (confessional, informational, instructional), genre features such as setting and location (question and answer, interview room), and footage (sit-down, slice-of-life, and behind-the-scenes) of the vlogs, and corpus information, including token and type.

General information	Vlog corpus
Resource media	YouTube
Resource genre	Vlog
Resource caption	#KerimcanDurmaz Kerimcan Durmaz "Peşimde" Hikayesi
Publication date	25/03/2021
Length	22 minutes 36 seconds
Location	Interview room
Creator	KD
Communicative Function	Confessional, Informational
Genre features	Q&A, dialogue, story-telling
Footage	Sit-down footage
Modality information	Multimodal
Access	Public
Link	https://www.youtube.com/watch?v=Z39JXvUKvk4
Subject Languages	Multilingual
Token count	3,688

Figure 1: The metadata scheme in *ELAN* metadata display

While collecting data from *YouTube*, I regarded captions as integral and compositional elements. Their display on the software panel used for constructing the corpus facilitates a swift and comprehensive review of the elements contributing to the generation of meaning, thereby aiding in the subsequent analysis. The information contained in vlog captions through emojis or hashtags, such as *BU EVİN ODASI YOK* 🍉🌳 *| TINY HOUSE VLOG (THIS HOUSE HAS NO ROOM* 🍉🌳 *| TINY HOUSE VLOG)*, is relevant to the interpretation of the compositional elements of digital communication.

The metrics of views, dis/likes, or comment counts are dynamic data that showcase the audience's reaction and the content generator's popularity. They demonstrate an overview of public engagement and content dissemination, which grants impact in the market for social media users like influencers. For this study, the metrics were merely examined for viewership and subscribership of each influencer in sampling influencers.

Transcription of audio-visual data is similar to that of audio data to a large extent. Based on the purpose of transcribing, different transcription conventions can be followed or developed to represent speech in written form. For example, applying general principles of orthographic transcription of a particular language suffices when the sole purpose of the corpus compilation is to produce a corpus of transcribed texts (Love 2020). For the vlog corpus, I

largely followed the transcription system developed in the *Turkish Spoken Corpus* project.[2] However, it is important to note that, along with evolving social media affordances, language practices have evolved and changed in a way that transidiomatic usage and translingual digital lexis entangle standardized language codes, which has resulted in new considerations and unique challenges to overcome in constructing transcription conventions.

Automated transcription processes in constructing corpora from the web, like auto-captioning, may seem useful at first. For instance, *YouTube* auto-captioning provides time-aligned subtitles for videos. The transcriptions feature several languages with varying degrees of accuracy rate. Although *YouTube*'s speech recognition quality is improving through deep learning algorithms (Bokhove and Downey 2018), for languages other than English, the success rate is unsatisfactory. The complex morphology of agglutinative languages like Turkish results in a high out-of-vocabulary rate, reducing the accuracy of automatic speech recognition (Arısoy *et al.* 2009). Hence, when automated captioning does not accurately communicate the intended message, and the transcription demands extensive manual repair, using them is not time efficient. Table 2 compares auto-captions and manual transcription of a Turkish influencer's vlog, and the results clearly show that manual transcription serves better for the objectives of the present study. Each line needs multiple corrections, ranging from inaccurate morphological derivations and wrong/missing proper names and nouns to missing chunks of phrases, some of which are in a different language than Turkish. Auto-captioning fails to recognize that the short dialogue between <S1> (Speaker 1) and <S3> (Speaker 3) is English, further evidencing its limitations of approximating English+Turkish constructions to either code. Consequently, automated transcriptions seem to be of limited use, which makes manual work inevitable for building spoken corpora (Love 2020).

Auto-generated subtitles	**Researcher-generated subtitles**
25:01 artık çok yoruldum Ali domates **aldı** limon almadık **a** şu limonlar olay olay 25:08 **asparagas Lara** bak **Ama** bu ne ya **B**u nasıl **bilim o** arkadaşlar 25:16 kız **bir** şey değil mi **karaip korsanlarındaki Oh Kaptan** ahtapot 25:23 **Ya** bu nasıl **iman** Allah aşkına **eski Memories that garip garip garip sınav** 25:31 **gör Vay** inanılmaz kokuyor Ali	25:01 <S1> artık çok yoruldum Ali domates **aldık** limon almadık **aa** şu limonlar olay olay **<S2> o ne** 25:08 <S1> **asparaguslara** bak _ bu ne ya **b**u nasıl **bir limon** arkadaşlar 25:16 kız **bu** şey değil mi **K**ara**y**ip **K**orsanlar'**ı**ndaki **o**_ **k**aptan ahtapot 25:23 _bu nasıl **limon** Allah aşkına **excuse me what is this <S3> fast-forwarded speech** <S1> **got it got it got it smells good** 25:31 **yeah when they <incomprehensible>** <S1> **wow** inanılmaz kokuyor Ali

Table 2: An illustration of the accuracy of auto-generated Turkish subtitles on *YouTube*[3]

[2] https://std.metu.edu.tr/

[3] Underlining indicates deleted syllable(s).

In building a vlog corpus, addressing the transcription of colloquial speech posed a significant challenge, primarily stemming from irregular pronunciation of words and morphemes. Slang, foreign nouns, neologisms, non-standard pronunciations, deviations from standard pronunciation, discrepancies in foreign word pronunciation, and the disruptive effects of digital features like fast-forwarding, as Table 2 shows, would impede the automated search capacity of corpus tools. To overcome this problem, I made the strategic choice to establish a manual transcription system that prioritized standardization in each language code to promote accuracy and facilitate more efficient search capabilities. For searchability, I found that the best way to approach examples such as (1) was to represent English pronunciation and the agglutination in Turkish.

(1) ***Loop bant****larımızı kullanıyoruz****, bosu ball****larımızı kullanıyoruz.*
'We use our **loop bands**, and we use our **bosu balls**.'

The *ball* is an English code pronounced as /bɔ:l/, yet the speaker falsely pronounces it as /bɔl/ without the elongation. Here, ***ball****larımızı* is represented as *ball* (English code) -*lar* (plural suffixation in Turkish) -*ımız* (first person plural possessive pronoun), and -*ı* (accusative case). If I transcribed the speaker's pronunciation of *ball,* the representation in Turkish phonetics would be *bol* /bɔl/, which becomes *bollarımızı* (a non-word). However, since *bol* ('abundant'), has a different meaning in Turkish, it would be confusing for a Turkish speaker until the co-text and context become clear. This type of talk is not uncommon in this corpus, which is addressed as *trans*languaging.

Other decisions included the use of apostrophes for proper nouns of non-Turkish origin (i.e., *Wet and Wild'lar*, 'the Wet and Wild products'), the capitalization of proper names, the removal of false starts (disfluencies), and the transcription of fragmented meaningful phrases. Non-linguistic cues and phonetic representations of loan and foreign words (i.e., *aplikasyon*, 'applications') were also documented. Acronyms were capitalized as they appeared in the vlogs (i.e., *TL/Turkish Lira, PT/personal trainer*). In colloquial Turkish, morphological changes, particularly in root or ending forms, are common and can impact token counts. While colloquial usage is often conventionalized, phonetic variations like *falan (*'so and so'*)* can inflate type counts. Hence, I combined the variations in colloquial use in all represented linguistic codes for searchability, and colloquial usage, such as in *falan* for every *falan*, *felan*, or *filan*, was standardized. Ambiguities in meaning or pronunciation were clarified with phonetic transcriptions using the IPA (e.g., /rɪəlslaɾ/).

4. ANNOTATION OF NON-STANDARD ELEMENTS IN CONTEXT

Vlog data contain standard and nonstandard texts and visual materials in combination; therefore, establishing an *ad hoc* annotation scheme is critical for identifying these elements to guide the analysis. The annotation categories were created via a cyclical process in which five vlogs (16.6% of the data) were initially marked to observe the local occurrences and commonalities of translanguaging instances and construct an annotation scheme. To ensure reliability, an expert linguist was consulted to identify and evaluate the tentative categories corresponding to a set of observations of translanguaging instances in the corpus. She annotated three randomly selected vlogs using the existing *ELAN* template. Inter-annotator reliability was calculated in *ELAN*, and over 85 percent overlap was assured for each category. I identified the constituent elements to gain a thorough understanding of the factors that drive translanguaging by examining the recurring patterns of translanguaging practices. Based on the corpus data, I identified the following practices that realized the translanguaging phenomenon. The categories were annotated on the tiers labeled 'translanguaging category' by examining the parent tier tokenized in *ELAN* for each .eaf file, as shown in Figure 2.

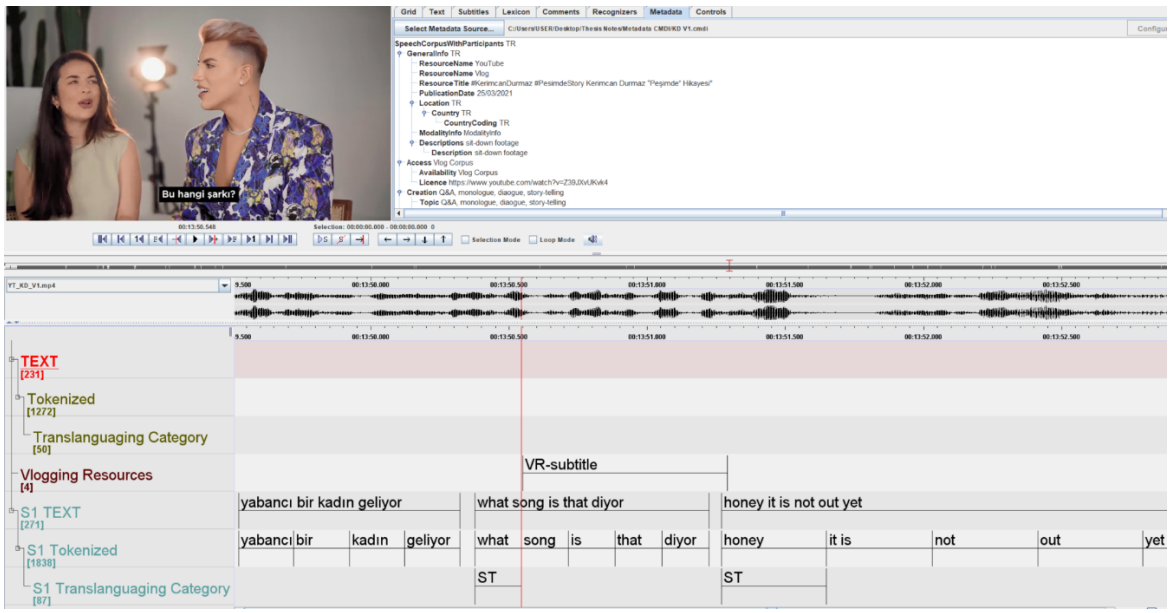


Figure 2: The segmentation and annotation of a vlog in ELAN

Translingual Insertion (TI) is a linguistic phenomenon that involves the prominent use of specific lexical items to represent or refer to global entities, such as commodities or places, that are relevant to the context of the speech event in vlogs. To illustrate this phenomenon, consider example (2), where an influencer in a decluttering vlog holds two makeup products in her hand and compares them uttering the names of the products in the original language and integrating

an ablative case marker in Turkish (*-dan 'by'* in this translation) as relevant to the syntactic construction of the meaning.

(2) *Mesela şunlar birbirine muadil*.[4]
'For example, these are equivalent [products].'
Ee Groundwork <TI> ve Color Tattoolardan Dusk Doll <TI>
'Uhm, Groundwork and Dusk Doll by Color Tattoo.'

Digital Lexis (DL) includes born-digital lexical items and English vocabulary with extended meaning in the digital context. Example (3) illustrates a digital lexis *like* with an extended meaning indicating netspeak and Turkish suffixes. The transformation of form, function, and meaning of such global items is what translanguaging explains.

(3) *Videomu beğenip likelayıp <DL> abone olmayı lütfen unutmayın*.
'Please do not forget to like [like] my video and subscribe.'

Different from TI, where speakers keep the original code in their speech, phonetic transliterations (TL) represent words from one code that uses approximate phonetic or spelling equivalence of another code. The annotation includes both commonly used phonetic TLs that function as borrowing and idiosyncratic phonetic TLs that emerged in interaction, as illustrated in (4)–(5).

(4) *Kovitten <TL> önce uzunca bir zaman evimiz yoktu*.
'We did not have a home for a long time before Covid.'

(5) *Ee crispy coconut rolls aa kıtır kokonat <TL> ruloları*.
'Uhm, crispy coconut rolls uhm crispy coconut rolls.'

Slang is a part of the speakers' linguistic repertoire and represents exclusive usage regarded as idiolects and idiosyncratic expressions. The slang in the analysis includes new forms or meanings as illustrated in (6) and (7). *Popi* is a newly constructed lexical item where the word *popular* in English and *-i* —which is an adjective-making suffix in Turkish— are combined. Similarly, *lubunya* (a Lubunca word)[5] and *-lar* (a Turkish plural marker) are far from a monolingual construction and thus annotated for their slang character.

(6) *Ya bence benim değil herkesin en beğendiği ve en popi <SL> serumu.*
'Well, I think this is not just my favorite serum but everyone's favorite serum.'

(7) *Lubunyalar* <SL>
'Lubunyas.'

[4] All translations are mine.
[5] Lubunca is an anti-language primarily used among gay male and trans-female populations in İstanbul.

Interpreting as a resource for translanguaging (IN) refers to the co-occurrences of language equivalents. This annotation is theoretically based on Baynham and Lee's (2019) dynamic account of translation that manifests as activity and practice in translanguaging space. IN shows interpreting activities that partake in the flow of translanguaging, especially in cases where interpreting occurs in the co-text of the source text, as shown in (8).

(8) *İşte mesela kajudan fermente böyle bir peynirimle yine **plant based** bitki temelli <IN> böyle bir İtalyan I am nut OK more daring than dairy diye bir peynirim var.*
'For example, I have this fermented cashew cheese and a plant-based Italian cheese called I am nut OK more daring than dairy.'

The spontaneous translanguaging (ST) category comprises discontinuous and unplanned language codes, encompassing instances ranging from isolated single token expressions in multiple languages, as in (9), to more extended turns, or expressions embedded in the context or co-text of Turkish, as in (10).

(9) *Bir adet tişört. Thanks. <ST>*
'A tshirt. Thanks.'

(10) S1 *Senin ağzına malzeme veriyorum.*
'I have put the words in your mouth.'

S2 *Whatever. <ST*
'Whatever.'

S2 *I will call you later. <ST>*
'I will call you later.'

S1 *Fine. <ST>*
'Fine.'

S1 *Film zaten benim hayat boyu yapmak istediğim şeydi.*
'Making a film was already what I wanted to do all my life.'

Translanguaging is a multimodal phenomenon, yet as Blackledge and Creese (2017) point out, translanguaging studies have paid little attention to multimodality. In this study, I identified the existence of multimodal elements contextualizing the annotated translanguaging instances. This annotation required paying attention to the resources in the digital space and visuals embedded in the segment of the instance. The annotation tier is Vlog Resource (VR). Speakers use these digital resources to recontextualize communication and achieve specific social goals. To illustrate the prevalence and implications of these multimodal resources, commonly occurring vlogging resources in the corpus are annotated. Example (11) showcases subtitles as a semiotic element, which improves the interaction for better viewer engagement, as the speaker assumes a linguistic gap between her and the imagined audience.

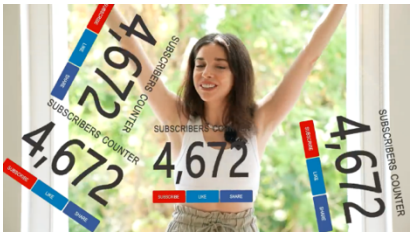
(11) 1 S1 *Dansçıların kendine özel makyaj sanatçısı var.*
'Dancers have their own make-up artist.'
2 *Taya Shawki, who was ee Ariana Grande's best friend tour partner (#1)*
'Taya Shawki, who was, uhm, Ariana Grande's best friend tour partner (#1)'

<VR- subtitle>
Ariana Grande'nin en iyi arkadaşı
'Ariana Grande's best friend'

#1

The subtitle is only one of the multimodal ways of making meaning in a digital context. Another example is (12), which illustrates multimodal meaning-making with textual and visual modes. The digital element is the real-time *YouTube* subscriber counter displaying numerical increments, and the textual elements in English such as *subscribers counter*, *subscribe*, *like*, and *share*. Using the global semiosis of the social media signifiers, S2 borrows credibility from *YouTube* with the red colored button *subscribe*, *Twitter* with the light blue colored button *like,* and *Facebook* with the darker blue colored button *share*. This semiotic element of social media discourse exhibits intertextuality with speech, as evidenced by the semantic frame that underlies its meaning.

(12) 1 S2 *Şu an böyle.*
'Now it goes like.'
2 *(#1) Tak tak tak tak tak tak tak tak tak aboneler böyle artıyor da olabilir.*
'(#1) The subscribers may also be snowballing like tak tak tak tak tak tak tak tak tak.'
3 *Umarım öyle olur.*
'I hope it goes like that.'
<VR-text+visual>



#1

In this study, the adoption of a bottom-up approach to annotating translanguaging instances led to creating *ad hoc* categories that can guide researchers working with social media corpora to investigate multilingual speakers' language use and communication practices, since messy datasets such as multimodal and multilingual communication can be quite challenging in determining where to begin. Therefore, these categories can be regarded as temporary indicators of a macro phenomenon, and not as a taxonomy, and are used to catalyze an interpretation of the complexities of current language practices regarding contextual factors. In addition, *ELAN* facilitated the preparation of diverse output formats. For instance, CSV files

enabled the automatic extraction of annotations within *ELAN* across multiple files. The integration of filters within this format enabled the precise extraction of specific tags. The Time-aligned Interlinear Text format demonstrated versatility and enhanced the presentation of transcriptions and annotations, making it the chosen format for sharing this vlog corpus, which is freely and publicly accessible.[6]

5. Conclusions

Translanguaging in vlogs serves as a reflection of real-world communication practices in digital spaces, presenting researchers with an opportunity to explore the evolving nature of online communication that transcends language boundaries. The diversity present in vlogs provides rich data with which researchers may explore the use of multiple languages within a single discourse. However, while the prevalence of translanguaging offers valuable insights, it also introduces challenges in natural language processing tasks. These challenges include difficulties in part-of-speech tagging, syntactic analysis, and parsing, due to blurred language boundaries and non-conventional sentence structures. Despite these challenges, the identified categories of translanguaging instances present researchers with a valuable resource for understanding and addressing these complexities. By transforming these challenges into opportunities for computational and corpus analysis, researchers can make more effective use of online content and in the context of digital sociolinguistics.

The vlog corpus represents a rich medium of communication that extends beyond spoken language, incorporating a range of semiotic and multimodal elements that are central to communication in social media. The visual framing of the vlog, which encompasses edits and mode-mixing facilitated by modern technological tools and editing techniques, plays a critical role in exploring the phenomenon of multimodal deixis within the spoken corpus. It is where vlog creators combine spoken language with visual cues, gestures, and other non-linguistic elements to convey meaning and engage their audiences. Hence, the findings of this study underscore the importance of taking a multimodal perspective when analyzing online communication practices.

In addressing the ethical considerations of using publicly available *YouTube* vlogs for research, I ensured that all content used was explicitly public, adhering to *YouTube*'s terms of

[6] https://github.com/hulyamsr/Social_Media_Influencer_Corpus

service.[7] I informed the influencers and their agencies about the research through emails and social media, explaining that their publicly available content would be used for academic purposes, without any scraping of follower data or commercial use. While I did not seek explicit consent, I suggest that using publicly available content does not typically require consent if the interaction is intended as a public performance and invites wider engagement and visibility. Vlogs form a promotional genre on *YouTube*, especially in the influencer market, with branding and business-promoting activities. Since influencers intend their work to be public, protecting autonomy, privacy, and confidentiality is less likely to be an obligation. Hence, they inhabit a less controversial ground since they publish in the public sphere and exercise advertisement-oriented content dissemination for work-related purposes as self-employed adults. Furthermore, I ensured that no sensitive content was displayed to avoid any risk of harm. Based on this stance, I made time-aligned transcriptions that contained both the translanguaging annotations as well as the URLs to the original *YouTube* videos, such that researchers interested in (Turkish) spoken corpora of social media may facilitate a broader spectrum of research endeavors. I emphasize that the publicly accessible version of the corpus contains my transcriptions, annotations, and *YouTube* URLs, but not media components (e.g., video files, screenshots). This decision primarily stems from practical considerations, particularly on the challenges associated with hosting large video files on alternative platforms, given that *YouTube* serves as the primary hosting service for the content.[8]

Several implications of building and using this vlog corpus may be drawn. Firstly, the vlog corpus includes asynchronous vlogs with beauty, fashion, and lifestyle content that are narrations of mundane activities, often monologic and informational in nature and promotional in content (i.e., a tour of a wellness center). The corpus can be studied for marketing discourse analysis and the reevaluation of interaction and monologue patterns in online communication, particularly in social media culture and language. This includes examining authentic language and linguistic styles. Secondly, in comparison to other corpora, the corpus may be used to analyze the effect on authenticity and style in language use caused by differences between synchronous and asynchronous vlogging modes. These differences are significant, as the structural and operational variations in the content creation process may lead to distinct linguistic patterns; the ways in which virtual interaction and audience engagement are

[7] https://www.youtube.com/static?template=terms

[8] In *YouTube* 's terms of service, under *Content on the Service* it is stated that "Content may be provided to the Service and distributed by our users and *YouTube* is a provider of hosting services for such Content." Based on these terms of service, the content remains public once it is uploaded for public view and disseminated by the users.

engineered differently on each platform have implications for analyzing language used in real-time interaction versus edited content. Thirdly, using *ELAN* for annotation and organizing data in a multimodal fashion allows for a more holistic analysis of social media communication. Moreover, I argue that transcription conventions need to evolve to keep up with changing language practices, especially in the context of social media and translanguaging practice; we should evaluate how transcription conventions have traditionally been used and how they can adapt to the evolving nature of language in digital communication, and explore the challenges of automated transcription and its implications for researchers relying on these tools for building corpora or analyzing spoken language.

REFERENCES

Arısoy, Ebru., Doğan Can, Siddika Parlak, Hasim Sak and Murat Saraçlar. 2009. Turkish broadcast news transcription and retrieval. *IEEE Transactions on Audio, Speech, and Language Processing* 17/5: 874–883.

Baynham, Mike and Tong King Lee. 2019. *Translation and Translanguaging*. New York: Routledge.

Blackledge, Adrian and Angela Creese. 2017. Translanguaging and the body. *International Journal of Multilingualism* 14/3: 250–268.

Blommaert, Jan. 2008. *Grassroots Literacy*. New York: Routledge.

Blommaert, Jan and Piia Varis. 2011. Language and superdiversity. Diversities 13/2: 3–21.

Bokhove, Christian and Christopher Downey. 2018. Automated generation of 'good enough' transcripts as a first step to transcription of audio-recorded data. *Methodological Innovations* 11/2: 1–14.

Dynel, Marta. 2014. Participation framework underlying YouTube interaction. *Journal of Pragmatics* 73: 37–52.

Frobenius, Maximiliane. 2011. Beginning a monologue: The opening sequence of video blogs. *Journal of Pragmatics* 43/3: 814–827.

Jacquemet, Marco. 2005. Transidiomatic practices: Language and power in the age of globalization. *Language and Communication* 25/3: 257–277.

Knight, Dawn, David Evans, Ronald Carter and Svenja Adolphs. 2009. HeadTalk, HandTalk and the corpus: Towards a framework for multi-modal, multi-media corpus development. *Corpora* 4/1: 1–32.

Kramsch, Claire. 2018. Trans-spatial utopias. *Applied Linguistics* 39/1: 108–115.

Li, Wei. 2011. Moment analysis and translanguaging space: Discursive construction of identities by multilingual Chinese youth in Britain. *Journal of Pragmatics* 43/5: 1222–1235.

Love, Robbie. 2020. *Overcoming Challenges in Corpus Construction: The Spoken British National Corpus 2014.* New York: Routledge.

Lustig, Andrew, Gavin Brookes and Daniel Hunt. 2021. Social semiotics of gangstalking evidence videos on YouTube: Multimodal discourse analysis of a novel persecutory belief system. *JMIR Mental Health* 8/10: e30311. https://doi.org/10.2196/30311

Mısır, Hülya and Hale Işık Güler. 2023. Translanguaging dynamics in the digital landscape: Insights from a social media corpus. *Language Awareness* 32/3: 1–20.

Otheguy, Ricardo, Ofelia García and Wallis Reid. 2015. Clarifying translanguaging and deconstructing named languages: A perspective from linguistics. *Applied Linguistics Review* 6/3: 281–307.

Schmidt, Axel and Konstanze Marx. 2019. Multimodality as challenge: YouTube data in linguistic corpora. In Janina Wildfeuer, Jana Pflaeging, John A. Bateman, Ognyan Seizov and Chiao-I Tseng eds. *Multimodality: Disciplinary Thoughts and the Challenge of Diversity*. Berlin: Mouton De Gruyter, 115–144.

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.

Zhu, Hua and Wei Li. 2020. Translanguaging, identity, and migration. In Jane Jackson ed. *The Routledge Handbook of Language and Intercultural Communication*. New York: Routledge, 234–248.

Corresponding author
Hülya Mısır
University of Birmingham
School of English, Drama and Creative Studies
Edgbaston
Birmingham
B15 2TT
United Kingdom
Email: h.misir@bham.ac.uk

received: May 2023
accepted: June 2024

RiCL Research in Corpus Linguistics

Review of Gillings, Mathew, Gerlinde Mautner and Paul Baker. 2023. *Corpus-Assisted Discourse Studies*. Cambridge: Cambridge University Press. ISBN: 978-1-009-16815-1. DOI: https://doi.org/10.1017/9781009168144

Tamsin Parnell
University of Nottingham / United Kingdom

Corpus linguistics and discourse analysis have long been said to exhibit a methodological synergy (Baker *et al.* 2008). The combination of approaches allows researchers to achieve both breadth and depth in their analysis while countering the criticism that discourse studies are prone to 'cherry-picking'. The burgeoning field of Corpus-Assisted Discourse Studies (CADS) is testament to the power of combining the two approaches to address contemporary social issues. However, as Gillings, Mautner, and Baker recognise, the uptake of the approach outside of linguistics "has not been as enthusiastic as might be expected" (p. 1). *Corpus-Assisted Discourse Studies* aims to redress this. The book offers a delicate balance between theoretical and empirical insights, peppered with relevant case studies that demonstrate how to conduct a corpus-assisted discourse study. Spanning seven chapters, the Cambridge Element provides beginners with a clearly explained introduction to the research area. As such, it would be appropriate not only for undergraduate and postgraduate students within linguistics, but anyone interested in the relationship between language and society.

Chapter 1 begins by explaining that CADS research explores discourse by examining corpora. It highlights areas of interest for the CADS researcher, including social representation, ideology, diachronicity, and institutional discourses, acknowledging that these are tied together by social questions rather than purely linguistic ones.

Chapter 2 —entitled 'The Rationale for CADS'— outlines the trajectory of CADS research, from the early linguistic interest in social questions (as pioneered by Firth) to Baker's (2006) seminal monograph *Using Corpora in Discourse Analysis*. According to the authors, part of the rationale for CADS is that it "puts analyses on more reliable empirical foundations" (p. 6). Going deeper, corpus linguistics and discourse analysis are united by a focus on linguistic patterning: combining the two approaches allows researchers to reveal the "incremental effect" of discourse (Baker 2006: 13). Noting that corpus linguistics allows both quantitative and qualitative insights, Gillings, Mautner, and Baker explain that the CADS researcher should oscillate between quantitative and qualitative components and can combine corpus linguistics and discourse studies in "any number of ways" (p. 8), as it is the combination of approaches that enables triangulation.

The third chapter leads the reader through the process of building a corpus for CADS research, starting by underscoring the importance of 'representativeness'. A distinction is made between reference corpora (typically representative of a broad language variety such as British English in the early 2010s) and specialised corpora (which represent a smaller language variety such as the works of Charles Dickens). Reference corpora, as the chapter explains, "provide an important benchmark against which [the discourse analyst] can interpret the evidence gleaned from their specialised, purpose-built corpora" (p. 9). In building the specialised corpus, what matters is that "the volume and the nature of the data are 'appropriate' for the research question" (p. 9). As the authors state, CADS researchers often work with newspaper data, with each article constituting a single text saved in txt format. Newspaper articles are popular texts with corpus linguists because they are not only politically significant but are easy to collect. Currently there are questions surrounding the collection of some other data types, including social media content. For example, is it ethical to combine a corpus of tweets when posters might not expect such public scrutiny? Here, the authors signpost to useful research, including Collins (2019) and Lutzky (2021). The final question the authors answer in this chapter is how big a corpus should be for CADS research. They explain that 'bigger' is not always better when it comes to CADS, and that the answer will depend on your research question.

Chapter 4 provides readers with a corpus toolkit, that is, a range of methods that can be used to answer your research question. The methods covered are frequency

analysis, concordance analysis, collocation analysis and keyword analysis. The explanation of frequency includes a helpful distinction between types and tokens, as well as an overview of tagging (both parts-of-speech and semantic). It elucidates the processes of creating a wordlist (ordering linguistic units either alphabetically or by frequency) and running searches for linguistic units across parts of a corpus (subcorpora). Case studies of UK Supreme Court judgements including at least one dissenting argument and a corpus of *Administrative Science Quarterly*[1] articles and book reviews usefully illustrate the power of frequency analysis to generate further questions. The importance of 'dispersion' is also touched upon, as words may be frequent in only one or two texts and therefore not be representative of the corpus as a whole, although this clustering may lead to further discourse analytical insights.

Section 4.2 covers concordance analysis, including important technical information such as how to sort, thin, and expand the concordance lines to make them easier to manage. It distinguishes CADS from other linguistic areas by explaining that, in this perspective, "discourse is the focus of analysis, and corpus assistance helps us to link large-scale social phenomena with linguistic choices at the micro level" (p. 23). To achieve this macro-level and micro-level synergy, researchers must go beyond the concordance line both in the sense of reading the co-text and considering the social context that shapes and is shaped by the corpus. The authors set out four ways to conduct a concordance analysis (pp. 23–25) along the axes of structured-unstructured and bottom-up-top-down, noting that when completing actual research, these types may overlap. They also encourage critical reflection on concordance analysis, a topic which is addressed in more detail in Gillings and Mautner (2024).

Section 4.3 addresses collocation analysis. Gillings, Mautner, and Baker reflect on how different methodological choices (such as length of collocational span) can alter results, encouraging experimentation to determine the most representative and useful set of collocates. A brief yet insightful discussion of statistical measures is offered in this subsection —an area which is a common cause of trepidation for those new to the more quantitative side to CADS. More sophisticated approaches to collocation analysis are also given a special mention, including the *Sketch Engine's Word Sketch* tool (Kilgarriff *et al.* 2014), and *#LancsBox*'s collocational network visualisations (Brezina *et al.* 2015). These are only cursory overviews, undoubtedly due to the audience and word limit.

[1] https://journals.sagepub.com/home/asq

Nevertheless, signposting to further reading that covers these areas would have been useful for those readers looking to progress to more complex approaches.

Section 4.4 introduces keyword analysis. Again, the discussion of techniques for calculating keyness is important for equipping new researchers with the confidence to choose which techniques to use. Equally enlightening is the explanation of how to group keywords and which to focus on. Perhaps the most important reflection, however, is that it is important to capture not just 'differences' between corpora, but also similarities. The authors establish ways to investigate similarity, including comparing two corpora against a third reference corpus.

Chapter 5 is titled 'CADS in Practice'. The strength of this chapter is its case study. Returning to the *UK Supreme Court* corpus explored in a previous case study, the authors present an analysis conducted on *Sketch Engine* in which they explore lexemes expected to play a part in expressing dissent. Their frequency study of *disagree* produces an "underwhelming result" (p. 40). To find a more fruitful result, they offer two approaches. The first one is via knowledge conventions about the genre of legal writing, which can be gained by reading a "fair number of texts from the corpus" (p. 40); this knowledge would lead us to the collocation *I disagree*. The second approach would be to look at the collocations of *disagree*, which reveal that the intervening adverb *respectfully* is more frequent in the dissenting subcorpus than in the majority subcorpus. The finding that "one of the characteristics of judges" framing of dissent is to buffer its impact with standardised politeness markers (p. 41) produces further research questions. Ultimately, the case study helpfully shows how different tools allow different routes into the data, and how promising paths can be distinguished from blind alleys (p. 42).

Another important facet of Chapter 5 is its recognition that CADS methods are seldom linear and orderly. Rather, "a little messiness" should be expected —albeit "without jettisoning the idea of systematic and transparent data analysis" (p. 43). To address the messiness, Gillings, Mautner, and Baker offer a musical metaphor in which each tool is regarded as an instrument. The point of the metaphor is to show that "CADS uses corpus tools flexibly, iteratively, and in a mutually reinforcing manner" (p. 44).

Chapter 6 discusses the limitations and potential pitfalls of CADS. The authors acknowledge that because CADS requires a "lexical hook," it is harder to identify

"broader discursive phenomena with multiple and unpredictable lexical realisations" (p. 45) such as argumentative strategies or extended metaphors. In this case, the researcher must return to discourse analysis proper. Equally, CADS can tell us little about "how meaning unfolds in longer stretches of text' and 'how interactants negotiate meaning in conversation" (p. 45). Thirdly, it is hard for CADS researchers to identify 'absences' in the data (although contrastive techniques can help to remedy this). Of course, there is also the issue of examining multimodal data through corpus linguistics methods, which —while increasingly taking place— is still difficult to do. Despite these issues, the authors question whether they can be referred to as limitations, since "CADS should be judged against what it was designed to do in the first place" (p. 46).

For beginners, an important aspect of the chapter is the section dedicated to pitfalls in CADS research. Drawing on their expertise as reviewers and seasoned CADS researchers, the authors remind readers that texts should not be collected just because they are easy (resulting in an 'all you can eat' approach), but because they are an integral part of the corpus. They also recommend discussing interpretations of data with colleagues to ensure it "passes the litmus test of intersubjective validity" (p. 48). Finally, the authors discuss the writing up stage and how necessary it is to strike a balance between reporting too much information about the methodological process and too little. This guidance is undoubtedly useful for those writing up a CADS project for the first time.

Chapter 7, which is the final chapter, reflects on the research journey. The authors acknowledge that their account of the research process has been selective and is necessarily incomplete (although, I would argue that it is sufficiently detailed to support beginners). In this chapter, the authors make the pertinent point that "disciplinary labels and identities ought to matter less than the commitment to unravel the mysteries of language" (p. 51), an important reminder for those who are working in interdisciplinary teams. They conclude by outlining areas in which CADS is developing, including how keywords are calculated, automating ways of categorising keywords, the use of *R*,[2] and research in languages other than English.

Overall, I would recommend *Corpus-Assisted Discourse Studies* to anyone interested in how a corpus linguistic approach to discourse analysis can strengthen research into social questions. I would encourage those tentatively reading this review

[2] https://www.r-project.org/

from outside of linguistics to take the leap and experiment with the tools outlined in the book. I would also suggest that those with more experience of CADS research read the book as a refresher, not least for the reflections on the pitfalls and potential limitations of the approach.

REFERENCES

Baker, Paul. 2006. *Using Corpora in Discourse Analysis*. London: Bloomsbury.

Baker, Paul, Costas Gabrielatos, Majid Khosravinik, Michał Krzyżanowski, Tony McEnery and Ruth Wodak. 2008. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society* 19/3: 273–306.

Brezina, Vaclav, Tony McEnery and Stephen Wattam. 2015. Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics* 20/2: 139–173.

Collins, Luke. 2019. *Corpus Linguistics for Online Communication: A Guide for Research*. London: Routledge.

Gillings, Mathew and Gerlinde Mautner. 2024. Concordancing for CADS: Practical challenges and theoretical applications. *International Journal of Corpus Linguistics* 29/1: 34–58.

Kilgarriff, Adam, Vit Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý and Vít Suchomel. 2014. *The Sketch Engine*: Ten years on. *Lexicography* 1/1: 7–36.

Lutzky, Ursula. 2021. *The Discourse of Customer Service Tweets: Planes, Trains and Automated Text Analysis*. London: Bloomsbury.

Reviewed by
Tamsin Parnell
University of Nottingham
School of Cultures, Languages and Area Studies
Room B29a Trent Building
University Park
Nottingham
NG7 2RD
United Kingdom
E-mail: tamsin.parnell2@nottingham.ac.uk

RiCL Research in Corpus Linguistics

Review of Brookes, Gavin and Luke C. Collins. 2023. *Corpus Linguistics for Health Communication: A Guide for Research.* London: Routledge. ISBN: 978-1-003-09965-9 https://doi.org/10.4324/9781003099659

Ovidia Martínez Sánchez
University of Alicante / Spain

In this book, Gavin Brookes and Luke C. Collins introduce corpus linguistics for health communication from a research perspective. The monograph is included in the *Routledge Corpus Linguistics Guides* book series and significantly enriches the resources available to researchers in the field by 1) introducing the fields of health communication and corpus linguistics while critically evaluating recent studies in corpus-based health communication, 2) outlining the procedures involved in planning a corpus linguistic investigation of health communication, including corpus design and construction, tool selection, and implementation of analysis techniques, and 3) demonstrating the application and potential of corpus linguistic methods in the study of spoken, written, and digital health communication.

In terms of structure, the monograph consists of seven chapters, each of which is divided into different sections. Notably, all chapters follow a consistent organisational pattern, with an introductory section at the beginning and a summary section at the end, except for Chapter 7, designated as the conclusion. The chapter summaries are followed by a short list of suggested further reading, the authors' notes and the bibliography. What drew my attention was how the authors called this book an introductory textbook: "We have tried to write this book without assuming any prior knowledge of corpus linguistics or health communication" (p. 28). I was surprised by the premise that readers could easily understand the topics as they were introduced —as it happens alongside the book. I also liked the additional tasks for the readers related to the content presented in each chapter; thus, the book promotes interactive learning for researchers and students

alike. On top of that, the accessibility of the material is evident in the authors' careful attention to a modest and naturalistic stylistic approach.

The journey into the contents of the book begins with a thorough examination of Chapter 1, which introduces health communication and corpus linguistics and lays the groundwork for their significance and applications in the field. As emphasised by McCullock *et al.* (2021: 28), "health communication is a multifaceted field of research, theory, and practise concerned with delivering health-related information to diverse populations." When dealing with health communication and its definition, the authors also approach the nebulous term of 'discourse', drawing upon insights from various authors. We are also told that investigations of doctor-patient interactions have primarily dominated the study of health communication as a prototypical asymmetrical type of interaction (Linell 1990), traditionally characterised by different hierarchical positions on both sides, an aspect also explored in Chapter 4. Historically, the study of health communication primarily revolved around doctor-patient interactions, but a significant diversification emerged by the mid-1990s, encompassing interactions with other healthcare professionals such as nurses. Moving to corpus linguistics, in Section 1.3, they define a 'corpus' as a "machine-readable collection of authentic language use that has been sampled to represent a language or language variety" (p. 13). The chapter underscores the advantages of corpus linguistics in dissecting language and facilitating quantitative and qualitative insights into datasets. Notably, empirical corpus linguistics methods offer researchers in healthcare a data-driven approach to studying naturally occurring language use. This approach extends its benefits to diverse linguistic contexts, as evidenced by examples from the *RareDis Corpus* (Martínez-deMiguel *et al.* 2022), the *SetembroBR Corpus* (Ramos dos Santos *et al.* 2024), and the *EasyCall Corpus* (Turrisi *et al.* 2021). Section 1.5 candidly discusses limitations in using corpus linguistics approaches, shedding light on potential challenges for researchers. For instance, creating a comprehensive corpus is time-consuming, converting original texts can alter them, and automatically identifying some aspects like pragmatic features or metaphors is difficult. However, advancements in natural language processing provide promising solutions for those mentioned tasks. The chapter offers a concise summary clarifying key insights and prompting further exploration.

Chapter 2 starts with an introduction regarding its scope and outlines the difficulties of specialised corpus design and construction. First, Section 2.2 discusses

the considerations in designing a corpus, followed by practical aspects such as text collection (Section 2.3), cleaning (Section 2.4), and annotation processes (Section 2.5). The chapter explores the concepts of authenticity, text selection, representativeness, corpus size, and balance. It further discusses collecting spoken and written texts, detailing challenges and methodologies, including transcription complexities. Additionally, it also stresses the importance of cleaning and annotating the corpus for usability and reliability. Cleaning involves refining data by removing noise, standardising formats, correcting errors, handling duplicates, and filtering irrelevant content. Annotation adds metadata or linguistic tags like part-of-speech tags or syntactic structures to facilitate analysis. These processes ensure accuracy and usability for linguistic studies and natural language processing tasks. Tools such as *Sketch Engine* (Kilgarriff *et al.* 2014) and *#LancsBox* (Brezina *et al*. 2015) are recommended for these tasks, both highly recognised and well-known in the field of corpus linguistics. Finally, the ethical considerations involved in corpus construction are addressed in Section 2.6, emphasising the importance of responsible research practices with texts of public consumption or informed consent when building corpora, such as spoken language corpora. In the summary section, the chapter offers a roadmap for researchers in health communication, advising the utilisation of existing corpora to save time and effort compared to creating new ones.

Chapter 3 ventures into the nature of corpus analysis, explaining the methodologies and tools required to glean meaningful insights from collected data, such as *AntConc* (Anthony 2022*)* and *WordSmith Tools* (Scott 2020). Beginning with an introduction, the chapter guides the landscape of software selection for corpus analysis. It then delves into various analytical techniques, including frequency, keyword analysis, collocation, cluster, and concordance analyses, each offering unique perspectives on the corpus data. Section 3.3. offers practical examples and illustrative explanations of the mentioned tasks with the *Daily Mail Dementia News Corpus* (Brookes 2023), and it demystifies the intricacies of each analytical approach. The chapter summarises the discussed methods, emphasising the indispensable role of human researchers in interpreting and positioning software tools' output, directly impacting results. Therefore, this chapter is readable for those who do not research health communication, but still want to learn and get introduced to the analysis of language in a corpus.

Chapter 4 focuses on the study of language in spoken health communication. Section 4.2 provides an in-depth analysis of data collection and analysis approaches and discusses the challenges posed by spoken language. It emphasises that the dialogic nature of spoken health communication is one of exchange, often involving the coming together of professional and patient expertise. It also introduces two case studies on the interactions between doctors and patients, from anorexia and register analysis. The approach to spoken health communication as a dialogue emphasises the sequence of turns. In response, researchers have often combined corpus linguistics with interactional approaches to analysis, such as conversation analysis. Section 4.3 then examines different forms of spoken health communication, from clinical interactions to research discourse and media representations. Section 4.4 analyses the characteristics of spoken health communication, including medical and interpersonal aspects. It highlights the importance of interactivity and the representation of different social actors in health and illness discourse. The chapter concludes with a summary to recapitulate the main points studied.

Chapter 5, the longest in the book, begins with an introduction that contextualises the main topic, namely, written forms of health communication and how these have been the subject of corpus analysis. Section 5.3 proceeds to explore the diverse forms of written health communication, including the genres of clinical documents, media texts, historical documents, and literary works. Here, I expected to find a greater variety of genres when dealing with written medical corpora. While genres like clinical guidelines, protocols, and medical reports are typically fundamental for examining communication practices across various medical fields, notably, they are neither used nor referenced in this chapter for corpus construction. The chapter analyses the distinctive characteristics of written health communication, examining the portrayal of health professionals and patients as well as the representation of illness, treatments, and solutions. It draws upon two case studies: one dealing with a longitudinal study of dementia metaphors in UK tabloids (Brookes 2023), and another revolving around examining lived experience through features of mind style (Demjén and Semino 2021). The authors' review of different types of written health communication in this chapter shows that representations of patients across texts are characterised by those affected by illness having limited agency. Furthermore, when people who are the subject of healthcare concerns are recorded in written texts, they are often depicted as lacking autonomy. The

chapter summary suggests incorporating patients' perspectives into corpus linguistics research on health texts to enhance the representation of first-person viewpoints in healthcare reports.

In the contemporary era, we have observed a remarkable proliferation of technological innovations, which has resulted in a profound transformation of how information is stored and accessed. The Internet and digital information are now frequently the first sources consulted by the public to obtain health information and for getting help with health-related questions or problems. The accessibility and quality of digital health tools may vary, resulting in disparities in health information access. Chapter 6 examines the methodological complexities of investigating medical language in the digital domain. It examines the various forms of digital health communication, including curated health information shared by professionals, interactive exchanges between individuals and health experts, and peer-to-peer health discussions facilitated by digital platforms. Section 6.3.2 introduces digital interactions with health professionals, mainly on social media platforms. The chapter provides a case study, by Hunt and Harvey (2015), about online discursive representations of eating disorders on the Internet through health queries. From my point of view, the case study is very engaging since it highlights that, in digital spaces, individuals can potentially express themselves in ways that are distinct from professional psychiatric contexts. As the dynamic between health and professional and patient shifts, the authors have observed patient perspectives that strive for autonomy and self-management regarding their health and medicalising discourses that appear to diminish individuals' sense of responsibility for their state of ill health. Moreover, a particular strength of digital spaces is their capacity for social connectedness and forming communities around shared health concerns, which can bring together members otherwise separated by time and space.

In this chapter, it has also been seen that their review of corpus studies of digital health communication is the discursive strategies through which contributors construct the persona of expert, administer advice, and provide readers with the content to extrapolate from their personal experiences. As an example, Coltman-Patel *et al.* (2022) carry out a key-word analysis of forum threads discussing vaccination, which directed them to focus on the use of insults, which were used as rhetorical devices. Their work attests to the value of extended and contextualised investigation of keywords that direct

us to the discursive aspects of deliberations around health concerns and the interpersonal dynamics of digital forums. Chapter 6 also shows how researchers have reported the potential advantages of the broader adoption of emojis in health documents, including capturing experiences of illness symptoms and health information (Lotfinejad *et al*. 2020). Section 6.4.3 deals with how natural language processing techniques may be used to process corpus and perform tasks related to sentiment analysis. In this section, however, I expected more examples of how to perform these sentiment analyses with some steps. The chapter ends with a summary, offering a comprehensive understanding of the different topics related to digital health communication.

Finally, chapter 7, which is the concluding chapter and the shortest in the book, reflects on the continued application of corpus linguistics methods to study health communication. The chapter considers how corpus linguistics can develop alongside advancements in healthcare to ensure that those using corpus linguistics methods can continue to make a meaningful contribution to the study of health communication and practice in the design and delivery of healthcare. This concise yet pivotal chapter embraces the positivist essence of both disciplines and assesses their vitality and prospects for the future. Section 7.2 considers the future trajectory of corpus-based health communication studies while acknowledging its transformative impact on the field.

Overall, *Corpus Linguistics for Health Communication: A Guide for Research* presents a thought-provoking perspective on health communication within corpus linguistics. The authors emphasise the importance of clarity and precision in research methodologies throughout the book, which are crucial considerations as the field expands quantitatively. This extraordinary volume offers a comprehensive set of case studies, tasks, website links, and an exhaustive list of further references, facilitating an understanding of the chapters' contents. It serves as an invaluable resource for those seeking to delve into the intersection of corpus linguistics and health communication. Including case studies throughout various sections provides concrete examples and practical applications of the discussed concepts. In addition, further reading sections at the end of each chapter provide additional material with brief notes from the authors on the importance of the recommended works. Thus, the reader is given a verified reference database of articles and books. Moreover, even for those unfamiliar with corpus methodological approaches to examining the language, this book can be an

excellent opportunity to read and learn the first steps to value this field. Suffice it to say that Gavin Brookes and Luke Collins have produced a comprehensive and accessible guide that will inform and inspire further research and exploration in the fields of corpus linguistics and health communication.

REFERENCES

Anthony, Lawrence. 2022. *AntConc* (version 4.2.0). Tokyo: Waseda University. https://www.laurenceanthony.net/software

Brezina, Vaclav, Tony McEnery and Stephen Wattam. 2015. Collocations in context: A new perspective on collocation networks. *International Journal of Corpus Linguistics* 20/2: 139–173.

Brookes, Gavin. 2023. Killer, thief or companion? A corpus-based study of dementia metaphors in UK tabloids. *Metaphor and Symbol* 38: 213–230.

Coltman-Patel, Tara, William Dance, Zsófica Demjén, Derek Gatherer, Claire Hardaker and Elena Semino. 2022. Am I being unreasonable to vaccinate my kids against my ex's wishes? A corpus linguistic exploration of conflict in vaccination discussions on Mumsnet Talk's AIBU forum. *Discourse, Context & Media* 48: 100624. https://doi.org/10.1016/j.dcm.2022.100624

Demjén, Zsófica and Elena Semino. 2021. Stylistics: Mind style in an autobiographical account of schizophrenia. In Gavin Brookes and Daniel Hunt eds. *Analysing Health Communication: Discourse Approaches.* Houndmills: Palgrave, 333–356.

Hunt, Daniel and Kevin Harvey. 2015. Health communication and corpus linguistics: Using corpus tools to analyse eating disorder discourse online. In Paul Baker and Tony Mcenery eds. *Corpora and Discourse Studies: Integrating Discourse and Corpora*. Houndmills: Palgrave, 134–154.

Kilgarriff, Adam, Vit Baisa, Jan Bušta, Miloš Jakubíček, Vojtěch Kovář, Jan Michelfeit, Pavel Rychlý and Vít Suchomel. 2014. *The Sketch Engine*: Ten years on. *Lexicography* 1/1: 7–36.

Lotfinejad, Nasim, Reza Assadi, Mohammad Hassan Aelami and Didier Pittet. 2020. Emojis in public health and how they might be used for hand hygiene and infection prevention and control. *Antimicrobial Resistance and Infection Control* 9/27. https://doi.org/10.1186/s13756-020-0692-2

Martínez-deMiguel, Claudia, Isabel Segura-Bedmar, Esteban Chacón-Solano and Sara Guerrero-Aspizua. 2022. The RareDis corpus: A corpus annotated with rare diseases, their signs and symptoms. *Journal of biomedical informatics* 125: 103961. https://doi.org/10.1016/j.jbi.2021.103961

McCullock, Seth, Grace M. Hildenbrand, Katie J. Schmitz and Evan K. Perrault. 2021. The state of health communication research: A content analysis of articles published in Journal of Health Communication and Health Communication (2010–2019). *Journal of Health Communication* 26/1: 28–38.

Ramos dos Santos, Wesley, Rafael Lage de Oliveira and Ivandré Paraboni. 2024. SetembroBR: A social media corpus for depression and anxiety disorder prediction. *Language Resources and Evaluation* 58: 273–300.

Scott, Mike. 2020. *WordSmith Tools* (version 8). Stroud: Lexical Analysis Software.

Turrisi, Rosanna, Arianna Braccia, Marco Emanuele, Simone Giulietti, Maura Pugliatti, Mariachiara Sensi, Luciano Fadiga and Leonardo Badino. 2021. EasyCall corpus: A dysarthric speech dataset. *Interspeech*: 41–45.

Reviewed by
Ovidia Martínez Sánchez
University of Alicante
Instituto de Lenguas Modernas Aplicadas
Edificio Institutos Universitarios II (Parque Científico)
03080 Alicante
Spain
E-mail: ovidia.martinez@ua.es

RiCL Research in Corpus Linguistics

Review of Pettersson-Traba, Daniela. 2022. *The Development of the Concept of SMELL in American English. A Usage-Based View of Near-Synonymy.* Berlin: De Gruyter Mouton. ISBN: 978-3-11079-2201. DOI: https://doi.org/10.1515/9783110792294

Daniel Granados-Meroño
University of Murcia / Spain

The aim of the monograph is to provide a comprehensive insight into the development of the concept of SMELL in American English in the period ranging from the nineteenth century until 2009. By using a corpus-based approach, as well as a thoughtful and advanced deployment of statistical analysis, the goal is to observe the semantic evolution of five near-synonyms related to SMELL, namely *fragrant, perfumed, scented, sweet-scented*, and *sweet-smelling*.

Pettersson-Traba begins by acknowledging the difficulty in providing a complete definition of 'synonymy'. Even if some dictionaries might define this notion as the linguistic phenomenon in which a word or expression means the same as another word or expression, she points out that "a partial degree of similarity is also considered for a word or expression to constitute a synonym of another term" (Cruse 2004: 157). These two views are used in the study to distinguish between 'absolute synonymy', which is found when a total similarity between two or more words takes place, and 'partial synonymy', which relates to contexts in which the similarity is not complete. Partial synonymy is much more frequent in language, while absolute synonymy is very rare (Cruse 2004: 157–158; Divjak and Gries 2006: 24; Liu 2010: 56–57; Taylor 2003: 264).

In Chapter 1, the author guides the reader through the most relevant schools and research approaches that have dealt with 'lexical semantics'. Lexical semantics is defined as a field in linguistics that has been attempting to answer whether the semantic dimension of language is a purely linguistic feature, or it is rather influenced by encyclopaedic

knowledge —being thus relevant for the theoretical background of the study carried out in the monograph. In terms of research approaches, the most important basis for Pettersson-Traba's study are 1) distributional corpus-based approaches —which combine the interest in collocations and the use of more fine-grained statistical analyses in data retrieved from representative corpora— and 2) cognitive semantics —which is the most relevant theoretical framework in current research due to the importance of concepts such as 'prototypicality' (Rosch 1973) and 'entrenchment' (Langacker 1987).

In this first chapter, Pettersson-Traba introduces the aims, scope, relevant contributions, and structure of the study. *Fragrant*, *perfumed*, *scented*, *sweet-smelling*, and *sweet-scented* are selected as representations of the semantic field SMELL, which is interesting because of its richness in terms of near synonymy. The five near-synonyms are chosen due to their low degree of polysemy, which avoids discarding instances that denote a meaning related to other semantic fields.

In Chapter 2, the reader is provided with a classification of synonymy and an exhaustive review of the most relevant literature dealing with it. The types of synonymy which are most recurrently mentioned in classifications are 1) absolute synonymy, 2) cognitive synonymy, and 3) near-synonymy. Absolute synonymy accounts for those words or word senses identical on all four dimensions of meaning, namely, denotational, stylistic, expressive and collocational meaning (Leech 1990). Cognitive synonymy concerns pairs (or groups) of words that, despite being identical on the denotational dimension and mutually entailing one another, differ in non-denotational traits, such as connotation, register, style, or the language variety where they occur. Finally, near-synonymy, which is considered the most common type of synonymy, refers to those words that differ slightly in conceptual content and are not denotationally identical. Still, these synonyms are sufficiently similar to be interchanged in many contexts of use (Cruse 2004: 159). However, the boundaries between cognitive and near-synonymy are blurred and authors such as Edmonds and Hist (2002: 116–117) or Desagulier (2014: 153) argue for a two-fold division, namely absolute vs. non-absolute synonymy, which is the classification followed in the study.

Sections 2.2 and 2.3 provide a thorough review of the literature on the topic which is based on distribution usage methods to study synonymy. The review firstly points out that Divjak and Gries (2006) were some of the first who attempted to cluster potential near-synonyms in groups, rather than studying pairs of words, and included a wider range

of factors in the analysis, which required the use of more sophisticated multivariate techniques. Similarly, Gries and Otani's study (2010) on the near-synonyms set from the semantic domain of SIZE is also discussed in depth. Their study covers two sets —one comprising *little*, *small* and *tiny* and another including *big*, *great* and *large*— and analyses several factors (such as aspect, voice, or transitivity marking of the finite verb of the adjectives) at a morphological, syntactic, and semantic level. According to Pettersson-Traba, Gries and Otani (2010) is the most comprehensive work on synonymous adjectives, and thus, one of her main inspirations for the study on the methodological level. Finally, a previous study carried out by the author (Pettersson-Traba 2021) is mentioned as one of the few studies dealing with semantic change in near-synonymous adjectives diachronically. Pettersson-Traba (2021) examines the use of the above-mentioned synonyms related to SMELL by focusing on their modified nouns, which are classified into nine different categories. Results suggest that major changes took place in the nineteenth century, and it is hypothesised that these might be due to extralinguistic factors, such as those of industrialisation and mass production, which led to the introduction of artificially scented soaps and candles in the market.

In its first section, Chapter 3 deals more exhaustively with the synonym set (*fragrant*, *perfumed*, *scented*, *sweet-smelling* and *sweet-scented*) which is analysed in the study. It also provides the motivations behind the choice of SMELL as the object of study. The author also examines reference works to provide a preliminary idea of the meanings and contexts in which the synonym set is used. Dictionaries such as the *American Heritage Dictionary of English Language* (ACDOE),[1] the *Cambridge Dictionary* (CD),[2] or the *Merriam-Webster Dictionary* (MW)[3] are used to provide insights on how these words differ between each other depending on the period of time. The study also shows the difficulties in determining the changes these words underwent and the (blurred) boundaries between their meanings.

Section 3.2 introduces the *Corpus of Historical American English* (COHA; Davies 2010) used in the analyses presented in Chapters 4–6. Pettersson-Traba grounds the selection of this database in the need of using a very large corpus due to the low frequency of the five near-synonymous adjectives under study. COHA fulfils this requirement, as it

[1] Https://www.ahdictionary.com/
[2] Https://dictionary.cambridge.org/
[3] Https://www.merriam-webster.com/

contains more than 475 million words from more than 100,000 individual texts, divided into four different genres: fiction, magazines, newspapers and non-fiction. Likewise, the corpus is suitable for a diachronic study because it covers the period 1810–2000 .

Section 3.3 explains the data annotation process, which consists in the manual revision of the POS tagging available for COHA, namely by excluding false positives of adjectives that are actually past participle verbs (*fragrant, scented*). The semantic domain is annotated using the *UCREL Semantic Analysis System* (USAS*;* Archer *et al.* 2003), together with a manual revision assisted by a more precise database, the *Historical Thesaurus of the Oxford English Dictionary* (HTOED).[4] The remaining of Chapter 3 describes the wide range of variables included in the analysis of the first dataset (language-internal semantic, language-internal non-semantic and language-external variables). These variables and their levels are presented in Table 1 below.

Variable types	**Variable**	**Variable levels**
Language-internal semantic variables	Sense	Natural
		Artificial
		Figurative
		Indeterminate
	Semantic category	ABSTRACT
		BODY AND PEOPLE
		CLEANING
		COSMETICS
		EARTH, ATMOSPHERE, AND WEATHER
		FOOD AND DRINK
		OBJECT
		PLANTS AND FLOWERS
		SENSATION
		SPACE
		SUBSTANCE AND MATERIAL
		TEXTILE AND CLOTHING
	Animacy	Animate
		Inanimate
	Concreteness rating	Average rating of concreteness from 1 to 5
	Concreteness binary	Concrete
		Abstract
	Countability	Count
		Non-count
		Other

Table 1: Variables and their levels in the first dataset (Adapted from Pettersson-Traba 2022: 95–96)

[4] Https://historicalthesaurus.arts.gla.ac.uk/articles/

Variable types	Variable	Variable levels
Language-internal non-semantic variables	Syntactic function	Attributive Predicative complement Postpositive Other
	Degree	Positive Comparative Superlative
	Collocate	Specific noun collocate (lemma)
Language-external variables	Period	Period 1 (1810–59) Period 2 (1860–1909) Period 3 (1910–59) Period 4 (1960–2009)
	Text-type	Fiction Non-fiction Periodicals

Table 1: Continuation

Chapter 4 deals with two closely related analyses: a semasiological analysis and an onomasiological one. Semasiology is the study of particular words and the sense or concepts that they designate, having a stronger interest in polysemy. The first study therefore focuses on the analysis of the near-synonymous adjectives over time to uncover potential changes in their prototypical structure. This has the purpose of determining whether any adjective within the set has a special effect on the semantic evolution of the concept of SMELL, or rather the whole set has a similar effect on it. Here, the frequency of use of the adjectives in COHA is analysed in regard to the changes caused by the variable *Period*. The second study, which analyses synonymy rather than polysemy, deals with the examination of various expressions which are used to designate a particular concept. As such, the starting point is based on the concepts or senses rather than on the words that designate them.

The results of the semasiological analysis provided in Section 4.3 show that, regarding the variable *Sense* across the four levels of *Period*, the adjective *fragrant* remains prototypical in the natural sense, but its use decreases significantly over time. We may witness a similar evolution in the figurative sense, while the indeterminate and artificial senses increase. In the case of *perfumed*, an increase in the use of three senses, namely artificial, indeterminate, and figurative is observed, while its use to denote natural aromas declines substantially. A similar trend can be appreciated for *scented*, while *sweet-scented* and *sweet-smelling* remain stable across the four periods. The latter is more prototypically used in the natural sense even in *Period 4*, despite the downward tendency.

The patterns arising from the analysis of the variable *Semantic category* are coherent with the analysis of the variable SENSE, as the levels corresponding to natural sense tend to decline, while the ones corresponding to the remaining levels increase or remain stable. For instance, in the case of *fragrant*, levels such as PLANTS AND FLOWERS or EARTH, ATMOSPHERE AND WATER, which clearly refer to the natural sense, are the most frequent.

In turn, the onomasiological analysis shows that *fragrant* is the most salient adjective across all five natural categories, and that all prototypically artificial categories, except TEXTILE AND CLOTHING, exhibit distributional changes over time. In particular, the frequency of *scented* increases at the expense of *perfumed* and *fragrant*, becoming the most salient adjective by *Period 4*. Similarly, the frequency of *perfumed* increases considerably and becomes almost as salient as *fragrant* in *Period 4* regarding the figurative category ABSTRACT. Finally, when used for semantic categories concerning indeterminate senses, *fragrant* slightly decreases over time, mainly in favour of *sweet-smelling*. These processes show that there exists some interrelation between the variables *Sense*, *Semantic Category*, and *Period*. In Chapters 5 and 6, Pettersson-Traba explores the nature, relevance and details of these interrelations.

Chapter 5 provides a comprehensive onomasiological analysis of the synonym set by means of multivariate approaches. It attempts to explain the motivations behind the patterns described in Chapter 4 and to find out whether any of the variables in Table 1 might entail proper predictors of the speaker's preference for one adjective. In this chapter, the author makes use of a statistical analysis by using multinomial regression models and a random forest analysis. Pettersson-Traba provides a detailed explanation of the statistics, which makes it easier for the reader to understand the interpretation of the results. The results from the random forest analysis show that *Semantic category*, *Sense*, and *Period* are the most important variables of predictors in a first model obtained through the multinomial regression analysis. These variables are precisely the ones included in the analyses in Chapter 4, which provides an additional ground to the idea that the pattern behind the diachronic changes might not be random. The variable *Collocate* is included later in the models and is shown to be significantly relevant, as it increases their prediction accuracy by around ten per cent.

Finally, an interesting insight in Chapter 5 is the plausible existence of a (probably still ongoing) process of substitution within the synonymy set, whereby *scented* gains

ground at the expense of *fragrant* and *perfumed*, as the semantic categories related to artificial, indeterminate and figurative senses increase, while the natural ones (which are closely related to *fragrant*) decrease dramatically.

By using a dataset of their noun collocates in an L5-R5 context window, Chapter 6 provides a more detailed discussion of the effects of the variable *Collocate* as regards the preference of speakers with the choice of adjectives in the synonym set. These are extracted automatically from COHA by using its collocates and POS-tag options. The study uses Semantic Vector Space (SVS) modelling of nouns collocates and measures the semantic (dis)similarity between the near-synonyms. The analysis draws on the collocational profiles and Pointwise Mutual Information (PMI) to identify prominent collocations of adjectives which need to comply with the criteria postulated by Baker (2017: 98–100).

The results of the SVS analysis are fed into cluster analysis to explore and interpret the (dis)similarities between the adjectives in different periods, which provide very interesting patterns. The collocational preferences of the adjectives included in the study (*fragrant*, *scented*, *and perfumed*) result in five clusters. On the one hand, we have *perfumed* and *scented* in *P1* and P2 in Cluster 1, while *perfumed* in P3 and P4 and *scented* in P3 are in Cluster 3, nearly positioned to Cluster 2 including *scented* in P4. On the other hand, *fragrant* presents different behaviour in terms of collocational preferences, with P3 and P4 in Cluster 4, and P1 and P2 in Cluster 5 respectively.

Pettersson-Traba's results do not only confirm the two-sided pattern which is observed in previous chapters —1) decrease of natural senses-related adjectives (*fragrant*) and 2) increase of the other senses-related adjectives (*scented*, *perfumed*)— but also suggest that there is a specific period in which the shift is especially dramatic: between P2 and P3, as P1 and P2 tend to group together and be separated from P3 and P4.

The results from the SVS and cluster analyses allow the author to corroborate the historical, cultural, and social changes that might explain the patterns. Important social and technological changes took place during the period examined in the monograph, in particular in the USA, as a result of the First and Second Industrial Revolutions. Pettersson-Traba argues that this could possibly constitute the underlying motivations accounting for the rise in the use of SMELL. This hypothesis is further discussed in Chapter 6.

The study in Chapter 6 aims at testing that the First and Second Industrial Revolutions account for the rise of SMELL. To do this, the author uses the most relevant semantic categories taken from previous chapters to select some noun collocates which belong to the semantic categories in question, including the collocates of the 15 nouns most frequently modified by all five near-synonyms, among others. With this new dataset, the author aims to determine whether the patterns attested in the analyses developed in previous chapters are exclusive to these synonyms or are also attested in nouns not related to SMELL. The results are enlightening: the second-order collocates of near-synonyms are examined to pinpoint whether the conceptualisation of the semantic categories changes over time and whether these changes mirror those undergone by SMELL and the near-synonyms that designate it. Based on the data, the author considers that the development undergone by noun collocates in this category is probably related to developments in chemistry that took place during the Second Industrial Revolution. In turn, the remaining semantic categories show no major changes over time.

Finally, Chapter 7 provides a summary of the most relevant contributions of the monograph to the field of semantics, as well as some limitations of the study. Pettersson-Traba also suggests some future lines of research. For example, she considers that undertaking a cross-linguistic study of the equivalent terms of the adjectives in the synonym set in other languages from societies with similar sociocultural and technological developments would be interesting to further examine the hypotheses tested in the monograph.

I recommend Pettersson-Traba's monograph not only to those interested in historical semantics, synonymy or polysemy, but also to scholars interested in sociolinguistics. Chapters 4–6, which are the core of the monograph, constitute a very valuable source of information for those interested in making use of statistical analyses in their studies, as the chapters involve well-structured and comprehensive explanations in terms of methodology. Chapters 1–3 might be considered too long for some readers, as the author provides a very detailed review of the literature. However, given the conscientious and well-structured selection of works on the topic, these three chapters are unquestionably a useful reference for readers that might not be familiar with semantics and its historical evolution as a research field.

In short, this monograph is valuable not only for of its academic relevance and interesting results, but also for its methodological explanations of the advanced statistical

analyses. Certainly, the two prestigious linguistic awards —namely, the *Book Award Aquilino Sánchez*[5] and the *Leocadio Martín Mingorance Book Award for Theoretical and Applied English Linguistics*—[6] that the monograph has received in 2023 constitute evidence of its high standard.

REFERENCES

Archer, Dawn, Tony McEnery, Paul Rayson and Andrew Hardie. 2003. Developing an automated semantic analysis system for Early modern English. In Dawn Archer, Tony McEnery, Paul Rayson and Andrew Hardie eds. *Proceedings of the Corpus Linguistics 2003 Conference*. Lancaster: Lancaster University, 22–31.

Baker, Paul. 2017. *American and British English: Divided by a Common Language?* Cambridge: Cambridge University Press.

Cruse, Alan. 2004. *Meaning in Language: An Introduction to Semantics and Pragmatics*. Oxford: Oxford University Press.

Davies, Mark. 2010. *Corpus of Historical American English (COHA)*. https://www.english-corpora.org/coha/

Desagulier, Guillaume. 2014. 'Rather, quite, fairly, and pretty: Visualizing distances in a set of near-synonyms'. In Dylan Glynn and Justyna A. Robinson eds. *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy*. Amsterdam: John Benjamins, 145–178.

Divjak, Dagmar and Stefan Th. Gries. 2006. Ways of trying in Russian: Clustering behavioral profiles. *Corpus Linguistics and Linguistic Theory* 2/1: 23–60.

Edmonds, Philip and Graeme Hirst. 2002. Near-synonymy and lexical choice. *Computational Linguistics* 28/2: 105–144.

Gries, Stefan Th. and Naoki Otani. 2010. Behavioral profiles: A corpus-based perspective on synonymy and antonymy. *ICAME Journal* 34: 121–150.

Langacker, Ronald W. 1987. *Foundations of Cognitive Grammar. Theoretical Prerequisites*. Stanford: Stanford University Press.

Leech, Geoffrey. 1990. *Semantics: The Study of Meaning*. London: Penguin Books.

Liu, Dilin. 2010. Is it a chief, main, major, primary, or principal concern? A corpus- based behavioral profile study of the near-synonyms. *International Journal of Corpus Linguistics* 15/1: 56–87.

Pettersson-Traba, Daniela. 2021. *A Corpus-Based Study on near-Synonymy: The Concept Pleasant Smelling in 19th- and 20th-Century American English*. Santiago de Compostela: University of Santiago de Compostela dissertation.

Rosch, Eleanor H. 1973. Natural categories. *Cognitive Psychology* 4/3: 328–350.

Taylor, John R. 2003. Near synonyms as co-extensive categories: "high" and "tall" revisited. *Language Sciences* 3/25: 263–284.

[5] http://www.aelinco.es/en/i-premio-investigacion-aquilino-sanchez
[6] https://aedean.org/wp-content/uploads/listado-de-premios-actualizado-2024-marzo.pdf

Reviewed by
Daniel Granados-Meroño
University of Murcia
Facultad de Letras
Campus de la Merced
Calle Santo Cristo
30001 Murcia
Spain
E-mail: daniel.granadosm@um.es

RiCL Research in Corpus Linguistics

Review of Izquierdo, Marlén and Zuriñe Sanz-Villar eds. 2023. *Corpus Use in Cross-linguistic Research: Paving the Way for Teaching, Translation and Professional Communication*. Amsterdam: John Benjamins. ISBN: 978-9-027-21430-0. DOI: https://doi.org/10.1075/scl.113

Isabel Pizarro-Sánchez
University of Valladolid / Spain

The volume *Use in Cross-Linguistic Research: Paving the Way of Teaching Translation and Professional Communication*, edited by Izquierdo and Sanz-Villar, provides an in-depth overview of the diverse applications of corpora in cross-linguistic studies. The book presents a collection of 12 studies that, through illustrative examples, emphasize the importance of parallel and comparable corpora in various linguistic fields, including translation studies, language teaching, and natural language processing.

The opening section, authored by Izquierdo and Sanz-Villar, serves as an introduction to the volume. It contextualizes the subsequent analysis and underlines the importance of cross-linguistic research within the broader framework of contrastive linguistics and translation studies. The authors acknowledge the significant value of parallel and comparable corpora in conducting empirical studies within the field. Furthermore, they include a brief overview of the 12 chapters that follow, providing a detailed account of each individual contribution to the field.

In chapter 1, Marco and Bracho Lapiedra test, and empirically validate, the Gravitational Pull Hypothesis (GPH) on Light Verb Constructions (LVCs), adopting an innovative approach in which they formulate their hypothesis at the level of LVC types rather than individual constructions. Their study focuses on how emotional states and dynamic events are represented in translations between English, French, Catalan, and Spanish. The authors analyse the predicative nouns that collocate with the light verbs

fer (Catalan) and *dar* (Spanish) and their equivalents into English (*make* and *give*) and French (*faire* and *donner*). They firstly categorize the nouns as either emotional states or dynamic events, and then examine their frequency and salience in both translated and non-translated texts. The analysis is based on collocations extracted from the *Corpus Valencià de Literatura Traduïda* (COVALT),[1] to provide observable evidence. The results largely confirm the authors' predictions, indicating an under-representation of LVCs conveying emotional states and no significant differences for those conveying dynamic events, with positive results observed in five out of the eight language pair and LVC type combinations. This research contributes to the understanding of translation practices and illustrates the value of corpus-based studies in testing linguistic hypothesis and in raising contrastive awareness of the translated language features within the context of the translation classroom. While the scope of the study is limited to specific language pairs and LVC types, it establishes a valuable foundation for future research.

The second chapter, by Rabadán, explores the challenges and solutions of translating English LVCs into Spanish. Based on data extracted from the parallel corpus *P-ACTRES 2.0*,[2] which includes fictional and non-fictional material, Rabadán investigates how the semantic features and combinatorial capabilities of LVCs influence translation choices. Additionally, she delves into register-based variations between fictional and non-fictional texts. Through a systematic approach, the study examines a sample of the concordances of five English light verbs: *have*, *take*, *make*, *do*, and *give*. The selected sample is representative of the verbs and registers under study. The use of *P-ACTRES 2.0* provides robust empirical data, thus enabling a detailed examination of translation patterns and semantic features. The results reveal five recurrent translation patterns, with preference for full lexical verbs and single correlate verbs. This corroborates the hypothesis that the semantic features of LVCs significantly influence translation choices. Additionally, variations related to register are observed, indicating different translation strategies for fictional and non-fictional texts. These findings are interpreted in the context of translation studies and semantic theory, emphasizing the importance of understanding the semantic compatibility of LVCs in order to improve translation accuracy and consistency. Furthermore, the chapter discusses the implications for machine translation and bilingual writing support tools, outlining the potential applications of its results in enhancing machine translation systems, post-

[1] https://www.covalt.uji.es/en/
[2] https://actres.unileon.es/wp/parallel-corpora/

editing aids, and authoring support applications. Rabadán's rigorous analysis of LVCs and their translations is a significant contribution to the field, offering valuable insights into LVC semantic compatibility and translation strategies.

In Chapter 3, Molés-Cases investigates the translation from Spanish into German of manner-of-speaking expressions in narrative texts. Based on a subcorpus of ten contemporary novels from the *Parallel Corpus German Spanish* (PaGeS),[3] the research describes the translation techniques used for reporting verbs that introduce direct speech and examines the differences in translation when dealing with motion and speech domains. By analysing 1,571 bilingual concordances, the author explores how translators approach the typological differences between a verb-framed language (Spanish) and a satellite-framed language (German), with a particular focus on whether manner-of-speaking expressions are preserved or adapted in translation. The results of the research indicate that manner-of-speaking is largely maintained in translations, with transference being the predominant translation technique. Interestingly, the study also reveals that manner-of-motion is frequently added to German translations from Spanish. This practice, however, is notably less prevalent for manner-of-speaking, suggesting that typological differences do not appear to be a significant factor in the translation of the speech domain. The author also observes similar diversity in the use of manner-of-speaking verbs between Spanish and German versions, which is indicative of a unique behaviour of German within satellite-framed languages. The study offers valuable insights into the translation of reporting verbs and their lexical diversity in verb-framed and satellite-framed languages, despite being constrained to narrative texts and a specific language combination.

Sánchez Nieto's chapter deals with the translation of the German dative passive, a grammatical construction that is prevalent in German but does not exist in Spanish. The study examines the extent to which translators maintain the recipient perspective and the translation techniques employed in both German to Spanish and Spanish to German translations. Using a raw sample of texts from PaGeS, Sánchez Nieto aims to analyse frequencies, semantic roles, and translation techniques of the dative passive forms *bekommen*, *kriegen*, and *erhalten.* The sample is tagged and queried in *Sketch Engine* to retrieve paragraphs that include examples of the three variants of the dative passive.[4]

[3] https://www.corpuspages.eu/corpus/about/about?lang=en
[4] https://www.sketchengine.eu/

Following manual cleaning, the examples were imported into *ATLAS.ti* for marking and further qualitative analysis.[5] Results show that the *bekommen* passive is the most frequent, while the *erhalten* passive is not present. The data also reveal that in approximately two-thirds of the translations from German into Spanish the recipient perspective is maintained, with the remaining favouring the agent perspective. Simplification is identified as the most common translation technique and, in Spanish to German translations, the dative passive is typically employed when the recipient perspective is present in the source text. The results are interpreted in the light of the Thinking-for-Translating hypothesis, suggesting that translators adapt their strategies to align with the rhetorical style of the target language. The chapter's findings have practical implications for translation training and the development of contrastive competence.

Chapter 5, by Ramón, investigates the semantic differences between the near-synonyms English ingressive verbs *begin* and *start,* using translation corpora. This is achieved through a comprehensive cross-linguistic analysis of the parallel concordances of the lemmas *begin* and *start,* all retrieved from *P-ACTRES*, which includes 4.2 million words of English-Spanish translations across various registers. The analysis of the translations provides empirical evidence of the semantic differences between the two near-synonymous verbs. The results indicate that *begin* is more frequently followed by to-infinitive clauses and is associated with the initiation of actions. In contrast, *start* is more common with intransitive patterns and often implying the commencement of activities. In terms of their translations, the Spanish verbs *empezar* and *comenzar* are the most frequently used equivalents for both *begin* and *start.* However, the data indicate that *start* shows a wider range of translational equivalents using ingressive verbs in Spanish, in comparison to the near-synonym *begin*. This observation suggests a greater diversity of its sense relations, particularly in intransitive and transitive patterns. While considering the potential impact of target language factors, Ramón presents a compelling argument that systematic differences in translation equivalents may reveal subtle semantic differences in near-synonyms. The chapter's methodological rigour, combining quantitative analysis with qualitative interpretation of the translation choices, provides a solid basis for its conclusions.

In chapter 6, Labrador explores the complexities of core vocabulary and the use of

[5] https://atlasti.com

parallel corpora to gain insight into it, with the verb *run* serving as a case study. The English verb *run,* despite its apparent simplicity, is a fundamental tier-1 word in English with a large number of meanings and uses. This complexity presents a significant challenge for non-native speakers who intent to fully understand and employ it in their language production. The study analyses 926 occurrences of *run* in the *P-ACTRES* corpus, focusing on fictional and non-fictional texts. It applies an intra-linguistical approach to classify the uses of *run* in English and an inter-linguistic approach to analyse its Spanish translations. Factors such as syntactic structures, collocational patterns, and the expression of manner and path of motion events are considered. The findings reveal that *run* is more frequently used in fiction, particularly in expressions of motion, and it has a wide range of literal and metaphorical meanings, which are reflected in its Spanish translations. Different translation techniques are also identified, including crossed transposition, density change, and copying structure, each reflecting different aspects of the verb complexity. The chapter presents practical suggestions for implementing corpus-informed teaching methods, emphasizing the importance of teaching the different uses of core vocabulary in order to improve learners' communicative skills by producing more idiomatic and natural language.

In chapter 7 Gutiérrez Lanza examines the process of synchronizing film dialogues for dubbing, with a focus on Conversational Makers (CMs). Her study deals with the adaptation process from draft translations to synchronized film scripts in the *Corpus of English-Spanish Cinema Scripts* (TRACEci),[6] comparing these with non-translated Spanish data retrieved from the *CORPES XXI subcorpus of guiones*.[7] The aim is to evaluate the influence of synchronization on CMs and determine whether the synchronization process results in a statistically significant overuse or underuse of CMs, thereby contributing to statistical dubbese. By analysing the frequency and use of CMs in different stages of translation and synchronization, Gutiérrez Lanza finds that CMs are overused in draft translations, and that there is a significant reduction in the use of CMs from draft translations to dubbed scripts, which reflects the adjustment made to meet lip-sync requirements. Some CMs such as *ehm*, *bueno*, *bien*, and *por supuesto* are overused in the dubbed scripts in comparison to the non-translated Spanish, confirming the presence of statistical dubbese (overuse). However, the overuse of certain CMs has been eliminated during the synchronization process, resulting in an overall improvement

[6] https://trace.unileon.es/es/fondos-trace/catalogos/textos-audiovisuales-cine-y-tv/
[7] https://www.rae.es/banco-de-datos/corpes-xxi

in translation quality. The challenge lies in maintaining the naturalness of the dubbed dialogues while simultaneously ensuring synchronization. These results are of interest to both the dubbing industry and translation training, as they contribute to a better understanding of the impact of CMs on translation quality.

In chapter 8, Hermosa-Ramírez investigates the linguistic characteristics of opera Audio Descriptions (AD) and Audio Introductions (AI) of opera scripts through corpus linguistics. By analysing scripts from the *Liceu Opera House* in Barcelona and the *Teatro Real* in Madrid, the research aims to situate opera AD and AI within the spoken-written language continuum. In order to achieve this objective, Hermosa-Ramírez analyses several linguistic measures, including lexical density, type-token ratio, mean word and sentence length, and the Flesch-Szigriszt readability index. The findings show that both AI and AD scripts share features with planned written language, particularly in terms of lexical density, and with spontaneous spoken language in lexical variation. However, AIs show longer mean sentence length and mean word length than ADs, which place them closer to written language. Readability scores provide further evidence to support this distinction, with AIs displaying more written language characteristics, whereas ADs, despite their formal structure, show a slight tendency towards the spoken language end due to the need for synchronization with visual elements. The findings underline the complex and multifaceted nature of these texts, which combine elements of both spoken and written language. The author concludes with valuable practical suggestions for applying her findings, including the potential for personalized AIs that may be adapted to diverse audiences. Hermosa-Ramírez's research makes a significant contribution to the audiovisual translation field, providing insights into the distinctive linguistic characteristics of opera AD and AI.

In chapter 9, Li describes the use of a multilingual parallel corpus, compiled for journalistic translation research, through a pilot case study that focuses on the national image construction in global news translation. The author employs the *New York Times Multilingual Parallel Corpus* (NYTMPC), which is a valuable resource for analysing how national images are constructed and reconstructed through the translation of English source news articles into Chinese and Spanish. The corpus consists of more than one million running words and 753 texts aligned at the paragraph level and manually annotated for headlines, leads (or subheadlines), publication date, news section, and translation shifts. To identify patterns in image construction, an analysis of

both the news headlines and the section labels is conducted. The research identifies several key topics that are more frequently associated with China in global news and contribute to the activation of specific national images, which are differently reconstructed across languages. These topics include international politics, COVID-19, economy, and technology. The work additionally analyses how these topics contribute to the activation of specific national images across languages by rephrasing news headlines and including or excluding particular news labels. This comprehensive analysis of the NYTMPC offers useful insights into the role of translation in shaping national images in global news. The findings have practical applications for the improvement of journalistic translation practice and the enhancement of the accuracy and objectivity of media representations.

Chapter 10, by Contarino and De Camillis, presents a comprehensive study on domain-adapting and assessing a machine translation engine for the unique variety of the German used in South Tyrol public institutions. The distinctive linguistic features of this variety demand the use of specialized translation tools. Previous studies on Machine Translation (MT) performance for South Tyrolean German point to significant challenges in accurately translating its legal terminology, thus, the need for adapting an MT system like ModernMT (MMT). In order to adapt the MMT, the *LEXB* corpus is used:[8] this is a parallel corpus of bilingual legal-administrative texts and Italian laws and codes translated into German. Additionally, the authors developed a customized automatic terminology evaluation tool to assess the MT quality of South Tyrolean legal terminology. The findings reveal significant improvements in the overall translation quality following the domain adaptation, as assessed by the standard quality metrics BLEU and chrF3. However, they also point to the persistence of difficulties in accurately translating specific legal terms, which suggests the limitations of on-the-fly adaptation for domains with limited parallel data. The authors conclude that further research is required to refine MT systems and terminology evaluation tools, particularly for low-resource language pairs and specialized domains. Both *LEXB* and the automatic terminology evaluation tool are accessible to the scientific community.

Chapter 11, by Politova, Bonetskaya, Dolgov, Frolova, and Pyrkova, describes the particular difficulties of aligning lexical units between two typologically distant languages such as Russian and Chinese, for which no gold standard was previously

[8] https://www.sketchengine.eu/eur-lex-parallel-corpus/

available. In particular, the authors present a rigorous methodology for the creation of a gold standard dataset for the Russian-Chinese word alignment. They provide a detailed account of their alignment guidelines and rules, based on previous research and adapted for this specific language pair. These guidelines, supported by clear and illustrative examples, address a range of linguistic phenomena, including punctuation, pronouns, classifiers, Chinese particles, and speech figures. The evaluation section introduces a comprehensive testing methodology, where two different machine learning models are used to compare their performance. The models were trained on the *Russian-Chinese Parallel Corpus* (RuZhCorp),[9] and fine-tuned on a manually annotated gold dataset. The findings demonstrate that the best results were achieved with *LaBSE*, and that fine-tuning the models on a gold dataset improves the performance of the algorithms. In addition to its specific focus on Russian and Chinese, this work provides a valuable model for the development of alignment guidelines and gold datasets for other typologically distant language pairs. The practical implications of this research are significant, potentially improving machine translation systems and enhancing corpus-based linguistic studies.

Finally, Ortego Antón's chapter presents the methodology used in the development of *GEnerador de Fichas de EMbutidos* (GEFEM),[10] a corpus-based writing tool designed for Spanish professionals to facilitate the composing dried meat product cards in English. The author establishes a prototypical rhetorical structure for both English and Spanish, classifying each rhetorical move and step according to their occurrence frequencies, from compulsory to occasional. She also identifies the model lines for each rhetorical element and creates a bilingual terminological database on dried meats. This is achieved by identifying usage patterns and extracting linguistic data from a unidirectional Spanish-English parallel corpus and a comparable English-Spanish corpus of dried meat product cards, both of which were specifically compiled for this purpose. Thus, GEFEM was developed on the basis of these three corpus-based research elements. Its user-friendly interface guides technical writers through the writing process, using colour-coded buttons and offering terminological suggestions from a database. The terms in the database are categorized by semantic fields, such as ingredients and allergens, thereby ensuring consistent and appropriate usage of terminology. Then, users can preview and download the final product card in docx

[9] http://ruzhcorp.ruscorpora.ru/en/
[10] https://actres.unileon.es/demos/generadores/applications.html#generatorsSection

format. In sum, GEFEM is an illustrative case study of the transfer of knowledge from corpus-based linguistic research to the agri-food industry. The application of this knowledge can facilitate the expansion of companies into international markets and enhance the productivity of technical writers.

The principal strength of the volume is its comprehensive coverage of the theoretical and practical aspects of the use of corpora in cross-linguistic research and the inclusion of a diverse range of domains and languages. The editors have brought together a diverse range of studies that evidence the versatility and applicability of corpora. Each chapter is based on rigorous empirical research, providing valuable insights and practical solutions for real-world linguistic issues and applications. In addition, the volume stands out for its attention to under-researched languages and domains, thus addressing an important gap in corpus linguistics. While it has notable strengths and each chapter is valuable in its own right, it could benefit from a more cohesive thematic structure. In addition, some chapters address complex technical issues that may be challenging for readers lacking expertise in corpus linguistics or statistical methods. In conclusion, Izquierdo and Sanz-Villar have succeeded in creating a valuable resource that not only advances academic knowledge in the field of cross-linguistic research, but also offers practical solutions that can be applied in language teaching, translation, and professional communication. This book is a highly recommended reference for anyone interested in the latest developments in corpus-based cross-linguistic research.

Reviewed by
Isabel Pizarro-Sánchez
University of Valladolid
Plaza del Campus Universitario s/n
Departamento de Filología Inglesa
Facultad de Filosofía y Letras
47011 Valladolid
Spain
E-mail: isabel.pizarro@uva.es

RiCL Research in Corpus Linguistics

Review of Viana, Vander ed. 2023. *Teaching English with Corpora: A Resource Book*. London: Routledge. ISBN: 978-1-032-25297-1. DOI: https://doi.org/10.4324/b22833

Gaëtanelle Gilquin
Université catholique de Louvain / Belgium

While many scholars have recognized the relevance of corpora for teaching, several of them have lamented that corpus-based applications in the classroom are few and far between (e.g., Meunier 2011; Götz and Granger 2024). One reason that has been put forward to explain this is the lack of ready-made materials for teachers (e.g., Breyer 2009; Gilquin and Granger 2022). In this respect, *Teaching English with Corpora: A Resource Book*, edited by Vander Viana, is a most welcome contribution, as it provides English language teachers with myriad lesson plans involving corpus-based activities.

The book starts with an introductory chapter by the editor, who sets the scene for the use of corpora in and for Teaching English to Speakers of Other Languages (henceforth, TESOL). This chapter offers an excellent overview of how corpora can help answer very concrete questions that teachers or students could ask, for example "What nouns does the adjective *heartfelt* usually modify?" (p. 6). It also shows what impact Data-Driven Learning (DDL) can have by letting students carry out corpus analyses, not only in terms of language learning, but also for the development of more general skills (research skills, digital skills, literacy and/or oracy skills, numeracy skills, and autonomous skills). The chapter ends with a presentation of the rationale and structure of the book, emphasizing its goal of being

> a single one-stop-shop volume containing lesson plans showcasing how current or future TESOL professionals can use corpora in their classes to suit their pedagogical goals and to cater for their students' needs (p. 23).

This introduction is followed by 70 chapters, written by "authors worldwide who have different professional experiences" (p. 24). Each chapter is structured in the same way. It starts with a box specifying the level(s) of the students for whom the lesson is designed, the main aims of the lesson (including more general skills), the duration of the lesson, the time needed to prepare it, and the resources that are necessary (computer lab, Internet access, whiteboard, handout, etc.). A short introduction then presents the topic of the lesson and explains why it is important, sometimes with references to the literature. The core of the chapter is a list of numbered steps to be taken before the class (if preparation is needed) and in the class. This is followed by points for consideration and alternative steps as well as a list of references and suggested reading. Occasionally, a chapter ends with an appendix, which may include a concordance or further instructions to run a program, for instance.

The 70 chapters are organized around two main parts: one part is devoted to English for general purposes (the first 40 chapters), whereas the other part is devoted to English for specific purposes (the next 30 chapters). However, a very useful "at-a-glance chapter taxonomy" (pp. xii-xiii) categorizes the chapters according to several other dimensions:

a) level: elementary, intermediate, upper intermediate, advanced;
b) system: discourse (which also covers pragmatics), grammar, pronunciation, vocabulary;
c) resources: online (relying on technological devices), offline (relying on print resources);
d) class time: up to 30 minutes, 35–45 minutes, 50-60 minutes, 65–90 minutes, 120 minutes;
e) and preparation time: none, 5 minutes, 10–15 minutes, 30–45 minutes.

This manifold categorization of the chapters allows readers to select lessons that might be relevant to them depending on the teaching context. In addition, an index makes it possible to select lessons that involve, say, a particular register or a specific corpus or program.

The book comes with online support materials, available from the Routledge website. Chapters that include such extra materials are identified by means of a special symbol. This is the case for 25 of the chapters. The online materials mostly consist of handouts or worksheets that can be distributed to students. Some provide teacher's notes, answer keys, or corpus samples, for instance. Most authors can be read independently of the online materials associated with them. For a few chapters,

however, like Chapter 29 by Nausica Marcos Miguel, access to the online materials is necessary to understand what the lesson is really about.

Not only does the book address the lack of ready-made materials for teaching with corpora by making resources available, but it also makes these resources accessible, in the sense of being easy to understand. As noted in the introductory chapter, "[n]o technical or prior knowledge of Corpus Linguistics is assumed or required" (p. 26). Indeed, technical terms are defined, either in the introductory chapter or in the individual chapters. This is the case for terms having to do with corpus literacy (e.g., 'concordancer' or 'n-gram') as well as some terms having to do with linguistics (e.g., 'conceptual metaphor', 'speech act', or 'hedging'). Each chapter is abundantly illustrated with screenshots of the tools used and corpus query outputs. Very often, different figures are provided to illustrate different steps of the lesson. Some of the figures are enriched with numbers corresponding to certain (sub)steps and pointing to the relevant parts of a corpus interface (e.g., a box to fill in or a button to click on). The bibliographical references cited in the chapters or suggested for further reading tend to be limited and efforts have been made to favour pedagogically oriented references. More generally, the average of four to five pages per chapter makes for an easy reading. The writing style, guided by the set structure and the numbered lists of steps, is concise and straightforward.

In addition to making its contents easy to understand, the book makes it relatively easy to implement. A majority of the lessons require no preparation time. The plans are usually for complete lessons, not limited to the corpus analysis itself, but also proposing warm-up activities, group discussions, follow-up tasks, and sometimes even ideas for follow-up lessons. Very practical considerations are included, for example the recommendation to use corpus queries on www.english-corpora.org wisely because the number of free queries per day is limited (p. 40, Chapter 3 by Robin Sulkosky), or the advice "to observe students' reactions closely and to have the contact number of a local support group that can be contacted if needed" (p. 161, Chapter 30 by Vander Viana) when discussing gender equality (or lack thereof). The book also offers ready-made handouts and worksheets (in the form of online support materials; see above) or elements that can be used in the classroom, such as corpus examples or tables to be completed by students. Suggested answers are sometimes provided. Twelve of the lessons do not require any technology (offline activities). For the other lessons, the

necessary corpora and tools, which have been chosen because they are easy to use, can almost always be accessed free of charge (although registration may be necessary).

As is evident from the "at-a-glance chapter taxonomy" mentioned earlier, the book can cater for a wide range of needs. Besides the dimensions listed in the taxonomy, the chapters deal with a large number of themes, including some topical ones (e.g., climate change in Chapter 58 by Robert Poole). They tackle many different linguistic phenomena (e.g., transition words in Chapter 46 by Nicole Brun-Mercer, reporting verbs in Chapter 53 by Joseph J. Lee, or contractions in Chapter 57 by Megan Bruce) and many registers (e.g., online reviews in Chapter 17 by Natalia Mora-López, blogs in Chapter 24 by Maristella Gatto, or obituaries in Chapter 69 by Rudy Loock). They involve quite a few different corpora —e.g., the *British National Corpus* (BNC), the *Corpus of Contemporary American English* (COCA), the *News on the Web Corpus* (NOW),[1] or the *British Academic Written English Corpus* (BAWE)[2]— although COCA is by far the most often used corpus in the book. Also, it should be underlined that a majority of the corpora are monolingual corpora of expert writing. Other types of corpora are less commonly used, e.g., learner corpora in Chapter 2 by Rosie Harvey and Irene Marín Cervantes, parallel corpora in Chapter 5 by Elen Le Foll, student writing corpora in Chapter 56 by Jenny Kemp and Laurence Anthony, multimodal corpora in Chapter 59 by Luciano Franco and Vander Viana. Spoken corpora are found in several chapters, although most of the time the analysis relies on transcriptions and does not involve the acoustic signal. Even Chapter 21 by Roger W. Gee, on pronunciation, starts from letters or sequences of letters (e.g., *th*, *ee*) to identify words with specific sounds in COCA. The use of the *Speech Accent Archive*[3] is suggested as a possible alternative to find good candidates for pronunciation activities, but this is not, strictly speaking, a corpus. As is the case with corpora, various programs are exploited in the book, but some are more frequent than others. This is the case of www.english-corpora.org (recurrent throughout the book) and *AntConc*[4] (only used in the second part of the book on English for specific purposes, especially when the activity involves collecting one's own corpus). Other tools include *KonText*,[5] *Sketch Engine for Language Learning*,[6]

[1] See www.english-corpora.org for information on the BNC, COCA, and NOW.
[2] https://www.sketchengine.eu/british-academic-written-english-corpus/
[3] https://accent.gmu.edu/
[4] https://www.laurenceanthony.net/software/antconc/
[5] https://kontext.korpus.cz
[6] https://skell.sketchengine.eu

StringNet Navigator,[7] *VocabProfilers*,[8] and *WebCorp*.[9] Variety is also visible in the corpus techniques applied (e.g., concordances, collocates, n-grams, keywords, frequency lists, and even multidimensional analysis in Chapter 40 by Vander Viana) and in the pedagogical approaches adopted (e.g., gamification in Chapter 3 by Robin Sulkosky, L1-based teaching in Chapter 4 by Natalie Finlayson, kinesthetic learning in Chapter 6 by Riah Werner, and differentiated instruction in Chapter 45 by Loretta Fung). Furthermore, the alternative steps in the penultimate section of chapters allow for possible adaptations according to students' level, degree of difficulty, time available, register investigated, etc. And while the book focuses on the teaching of English, most of the activities could be adapted to other languages, provided the necessary corpora and tools are available and similar linguistic phenomena exist in the other languages.

As already suggested with respect to the predominance of certain (types of) corpora and tools, the book does not necessarily offer a balanced mix of activities. Chapters devoted to vocabulary are far more numerous than those dealing with pronunciation, and there are fewer chapters designed for beginners than for more advanced students. It is also unclear why a few lessons are presented as being meant for trainee teachers (e.g., Chapter 34 by Jenny Kemp and Luke Timms), while most of these lessons might just as well be organized among language students. In addition, certain chapters less fully embrace the corpus linguistic approach than others. Thus, in Chapter 6 by Riah Werner, the role of corpora is very modest: the corpus-based finding that the '*if* + present simple + imperative' construction is widely used serves as a starting point for an activity that does not involve corpora at all and that includes sentences which are very unlikely to occur in naturally-occurring language (e.g., *If you like bananas, jump*). In some chapters, invented sentences are shown to students, e.g., *I can see the book in front of me* (to illustrate the physical meaning of the verb *see*), while (simple) corpus examples would have worked perfectly well. Since all lesson plans have arguably been tested, one would have liked to see more statements of the following type, displaying a more personal take on the lesson: "I have used this task successfully on several occasions on a university pre-sessional course, including students at the weaker end of the ability range" (p. 76, Chapter 11 by John Williams). Similarly, one would have expected to find more warnings such as "Sometimes students find the idea

[7] http://nav.stringnet.org/
[8] https://www.lextutor.ca/vp/
[9] https://www.webcorp.org.uk/live/

of using technology daunting at first" (p. 267, Chapter 51 by Jenny Kemp and Laurence Anthony) or "As FLAX is a large online library, students might get lost initially" (p. 236, Chapter 44 by Eman Elturki). Many chapters tend to present an overly rosy picture of corpus-based teaching. This leads to contradictions in the book where most authors describe www.english-corpora.org or *AntConc* as student-friendly tools, but some point to the challenges that they may present for students (see p. 129 on www.english-corpora.org and p. 227 on *AntConc*). Warnings about the difficulty of collecting one's own corpus or about the amount of noise generated by certain automatic searches (e.g., *NOUN is (a/an) NOUN* to retrieve similes of the type *time is money* in Chapter 18 by Natalie Finlayson), for example, would have been desirable too —not to discourage teachers, but to reassure them that the difficulties that they are bound to encounter sometimes are part of the normal process of using corpora. The book also presents some weaknesses that are understandable given the large number of chapters that it brings together. It can thus be inconsistent in the information that is provided, in the sense that, say, the first occurrence of a term in the book is not defined, but a later occurrence is (the term 'register', for instance, is defined on p. 254, but most of its earlier occurrences are not). Another example has to do with registration to access corpora on www.english-corpora.org: while Chapters 1 and 3 refer to this corpus interface, it is only in Chapter 4 that the requirement to register is mentioned for the first time. Finally, cross-references, as very occasionally found in the book (for Chapters 32 and 33 and for Chapters 53 and 67), would have been welcome for some other chapters too (e.g., Chapters 2, 14 and 31 on phrasal verbs or Chapters 25 and 35 on conceptual metaphors).

Despite these limitations, the book offers what many have been waiting —and hoping— for: a varied collection of lesson plans that teachers can easily implement to incorporate corpora in their teaching. To paraphrase O'Keeffe *et al.* (2007: 248), cited in the introduction, the authors of the chapters do not "stop at the classroom door", but enter the classroom head-on. They do so with so many inspiring ideas that the book should have both short-term and long-term impacts. In the short run, it should lead teachers to organize some of the corpus-based activities described in the different chapters. In the long run, it should encourage them to create their own materials and develop their own lesson plans involving corpora. This, in turn, could have repercussions on textbooks and other published resources, the current weaknesses of

which are highlighted in several chapters (e.g., p. 193 and p. 217). If more and more teachers use and want to continue using corpora in their teaching, publishing houses might become less reluctant to include corpus activities in pedagogical publications. Ultimately, a book like *Teaching English with Corpora* might be just what is needed to finally give corpora the place that they deserve in the educational world.

REFERENCES

Breyer, Yvonne. 2009. Learning and teaching with corpora: Reflections by student teachers. *Computer Assisted Language Learning* 22/2: 153–172.

Gilquin, Gaëtanelle and Sylviane Granger. 2022. Using data-driven learning in language teaching. In Anne O'Keeffe and Michael J. McCarthy eds. *The Routledge Handbook of Corpus Linguistics.* London: Routledge, 430–442.

Götz, Sandra and Sylviane Granger. 2024. Learner corpus research for pedagogical purposes: An overview and some research perspectives. *International Journal of Learner Corpus Research* 10/1: 1–38.

Meunier, Fanny. 2011. Corpus linguistics and second/foreign language learning: Exploring multiple paths. *Revista Brasileira de Linguística Aplicada* 11/2: 459–477.

O'Keeffe, Anne, Michael McCarthy and Ronald Carter. 2007. *From Corpus to Classroom: Language Use and Language Teaching*. Cambridge: Cambridge University Press.

Reviewed by

Gaëtanelle Gilquin
Université catholique de Louvain
Collège Erasme
Place Cardinal Mercier 31, bte L3.03.33
B-1348 Louvain-la-Neuve
Belgium
E-mail: gaetanelle.gilquin@uclouvain.be