v. 25, n. 1, e50026, 2025
Corpus Linguistics: Studies and Applications
ISSN: 1984-6398 · DOI: 10.1590/1984-6398202550026

Revista Brasileira de Linguística Aplicada

Thematic Dossier

Using a Learner *Corpus* of Oral Interview Data? Get to Know Thy Data!

Usando um corpus de alunos com dados de entrevistas orais? Conheça seus dados!

Sylvie De Cock*
*University of Louvain (UCLouvain)
Louvain-la-Neuve | BE
sylvie.decock@uclouvain.be
https://orcid.org/0000-0001-9572-4287

Abstract: Learner corpus studies have increasingly been conducted on spoken learner interlanguage thanks to the growing availability of spoken learner corpora, some of which comprise interviews with L2 learners. This paper seeks to highlight the importance of carefully considering the specific make-up and context of collection of learner interview corpora. Drawing on insights from using the Louvain International Database of Spoken English Interlanguage (LINDSEI), the paper focuses on the following three key features of interview spoken learner corpora that cannot be overlooked in learner corpus research: an interview is not a naturally occurring conversation, an interview is led by an interviewer and an interview is often made up of more than one speaking task. Possible future spoken learner corpora developments are discussed.

Keywords: spoken learner corpora; oral interview data; LINDSEI.

Resumo: Os estudos de *corpora* de aprendizes têm sido cada vez mais realizados sobre a interlíngua oral de alunos, graças à crescente disponibilidade de *corpora* orais de aprendizes, alguns dos quais incluem entrevistas com alunos de L2. Este artigo procura destacar a importância de se considerar cuidadosamente a composição e o contexto específicos da coleta de *corpora* de entrevistas de alunos. Com base nos *insights* obtidos com o uso do Louvain International Database of Spoken English Interlanguage (LINDSEI), o artigo se concentra nos três principais recursos dos *corpora* de entrevistas orais com aprendizes que não podem ser ignorados

na pesquisa desse tipo de *corpus*: uma entrevista não é uma conversa que ocorre naturalmente, uma entrevista é conduzida por um entrevistador e uma entrevista geralmente é composta por mais de uma tarefa de fala. São discutidos possíveis desenvolvimentos futuros de *corpora* orais de aprendizes.

Palavras-chave: corpora de aprendizes; dados de entrevistas orais; LINDSEI.

1 Introduction

Computerised learner corpora have been defined as systematic "electronic collections of texts produced by language learners" (Granger, 2008, p. 259) that represent "written and/or spoken 'interlanguage'" (Gilquin, 2020, p. 283). To qualify as learner corpora, collections of learner texts need to be gathered and compiled according to explicit design criteria related to a series of learner and situational/task variables, such as learners' age, L1 or proficiency in the target language (TL) and educational context, degree of planning, task type, or duration (see Granger, 2008, and Gilquin, 2020, for a thorough discussion of learner corpus design criteria). In addition, learner corpus compilation needs to go hand in hand with the careful collection and recording of detailed information concerning the data collected and the data collection itself, as this information is crucial when conducting learner corpus research, "from study design to corpus selection/compilation, result interpretability and cumulative research" (Paquot *et al*., 2024, p. 280). Paquot *et al*.'s (2024) standardized Core Metadata Schema for learner corpora is a recent initiative that aims to foster collaboration among learner corpus compilers and researchers, and provides guidance on the collection of metadata pertaining to learners, texts, situational/task characteristics and annotations, among others.

As can be seen from the Learner Corpora around the World webpage,[1] although written learner corpora still outnumber and outsize their spoken counterparts, given the specific challenges associated with the compilation of spoken corpora, electronic collections of spoken learner productions have been catching up and learner corpus research has, as a result, increasingly been conducted on spoken learner interlanguage. A glance at the corpora listed on the webpage reveals that the spoken data they contain can be quite diverse in nature depending on, among other things, the targeted learners, the tasks included and the setting of collection (exam/language test situations; classroom activities, as in MAELC; activities that learners participate in on a voluntary basis, as in CEDELE2, COREFL, LINDSEI). Spoken learner corpus diversity can be illustrated by two of the recent spoken additions to the learner corpus family: the Corpus of Collaborative Oral Tasks (CCOT, Crawford and McDonough, 2021), which features 775 peer interactions based on no fewer than 24 different paired tasks between ESL learners, and the Trinity Lancaster Corpus (TLC, Gablasova; Brezina; McEnery, 2019), which

[1] UNIVERSITÉ CATHOLIQUE DE LOUVAIN. Learner Corpora Around the World. [S. l.], c2025. Available at: https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html. Accessed on: 2 Sept. 2024.

consists of 4.2m words of interaction around four different speaking tasks (presentation, interactive task, discussion and conversation) between L1 examiners and L2 speakers of English.

This paper sets out to highlight the importance of not overlooking the specific make-up and context of the spoken language data used when conducting learner corpus research, and when drawing pedagogical implications from the research. The spotlight is on learner corpora made up of interview data and, more particularly, on one of the large-scale spoken learner corpus pioneers, namely the Louvain International Database of Spoken English Interlanguage (LINDSEI, Gilquin; De Cock; Granger, 2010).

After a brief discussion of the spoken nature of oral learner corpora and of some of their key task and situational variables, this paper turns its attention to spoken learner corpora, made up of interview data, zooming in on insights from the use of LINDSEI in learner corpus research. The discussion revolves around three main key interview features: interviews are not naturally occurring conversations; they are led by an interviewer; and they are often made up of more than one speaking task. This paper aims to show how these three features have a considerable impact on the language that is used in interviews, which sets them apart from spontaneous conversations. Some of the linguistic phenomena commonly found in conversations may well be absent from interviews, serve different functions or exhibit different patterns because of the specific communication situation and task(s). This has important methodological implications for learner corpus research design when selecting the phenomena to be investigated, when deciding on a suitable reference corpus within the framework of Contrastive Interlanguage Analysis (controlling for genre) and when interpreting the findings. The specific nature of interviews also needs to be taken into consideration when designing pedagogical applications derived from research on a learner interview corpus to ensure relevance and appropriateness. The paper ends with considerations about possible future developments in spoken learner corpora collection.

Table 1 includes the main learner corpora mentioned in the first part of this paper so the reader can easily match the acronyms used throughout this text with the full names of the corpora and the publications where the corpora are described. However, an exhaustive review and discussion of spoken learner corpora lies outside the scope of this paper. This article mainly focuses on the corpora of English interlanguage, especially that produced in a foreign language context.

Table 1 – The spoken learner corpora mentioned in this paper

Corpus	Target language(s)	Reference
CCOT – Corpus of Collaborative Oral Tasks	English	Crawford and McDonough (2021)
CEDELE2 – Corpus Escrito del Español como L2	Spanish	Lozano (2022)
COREFL – CORpus of English as a Foreign Language	English	Lozano; Díaz-Negrillo; Callies (2021)
GLBCC – Giessen-Long Beach Chaplin Corpus	English	Müller (2005)
ICNALE – International Corpus Network of Asian Learners of English	English	Ishikawa (2023)

LeaP corpus – Learning Prosody in a foreign language	English, German	Gut (2012)
LINDSEI – Louvain International Database of Spoken English Interlanguage	English	Gilquin; De Cock; Granger (2010)
MAELC – Multimedia Adult ESL Learner Corpus	English	Reder; Harris; Setzler (2003)
NICT JLE – National Institute of Information and Communications Technology Japanese Learner English	English	Abe (2019)
PAROLE corpus – Parallèle, Oral, en Langue Etrangère	English, French, Italian	Osborne (2013)
TLC – Trinity Lancaster Corpus	English	Gablasova; Brezina; McEnery (2019)

Source: the author.

2 The Spoken Nature of Oral Learner Corpora and Some Key Task and Situational Variables

While spoken corpora of learner language all originate in language productions that use "the medium of 'phonic substance', typically air-pressure movements produced by the vocal organs" (Crystal, 1995, p. 291), the majority are actually what Ballier and Martin (2015, p. 111) call "mute spoken corpora" in that they include transcripts, i.e. written representations of recordings of speech, without easy or ready access to the audio files (for most researchers). "Speaking corpora" (Ballier; Martin, 2015, p. 110) have been time-aligned and make it possible for researchers to actually hear learners' turns (*e.g.* COREFL), while "phonetic corpora" feature syllable or segment alignment and allow for the visualisation of smaller units of speech (*e.g.* LeaP corpus). Despite recent technological developments, speaking corpora and phonetic corpora are still rather few and far between, as are multimodal learner corpora featuring video recordings, which also give researchers access to aspects of non-verbal communication (cf. Reder; Harris; Setzler, 2003). Having said this, it is not uncommon for studies based on mute spoken corpora to go beyond the transcriptions that are accessible to the research community by going back to the audio files to examine and annotate specific pronunciation and fluency features using PRAAT, ELAN or EXMARaLDA in view of the specific focus of their research (Dumont, 2018; Aas; Nacey, 2019; Ghanem, 2021).

As pointed out by Chafe and Danielewicz (1987, p. 84), spoken language is not a unified phenomenon with the spoken mode/medium "allow[ing] a multiplicity of styles". Texts produced using the spoken medium can differ a great deal from one another because of differences linked to communicative purpose(s) and situational features, such as planning, level of formality and interactiveness (Biber *et al.*, 1999). In other words, spoken texts can exhibit a wide range of genres, from informal face-to-face conversations to formal presentations, for example. This multiplicity of genres is reflected in the diversity of data contained in existing spoken learner corpora, which is also closely connected to the various monologic or dialogic

tasks that the learners perform. Monologic tasks include descriptions of people (*e.g*. COREFL), short presentations (*e.g*. TLC), narratives based on pictures or a video (*e.g*. COREFL, GLBCC), or learners agreeing or disagreeing with a statement and providing reasons for the (dis) agreement (*e.g*. ICNALE). Dialogic tasks involve some form of interaction, generally between two learners or between a learner and another person, who may or may not be an L1 speaker of the target language and who often acts as an examiner or as an interviewer. Examples of dialogic tasks are decision-making or persuasion role plays (*e.g*. CCOT, ICNALE, NICT JLE), discussions about a wide range of topics (*e.g*. TLC) or informal interviews about the learners' personal lives (*e.g*. LINDSEI).

The majority of the corpora in Table 1 are made up of a series of monologic and/or dialogic tasks (*e.g*. ICNALE, GLBCC, LINDSEI), which may correspond to the different parts of an exam or test (*e.g*. CCOT, TLC). Learners are sometimes allowed some preparation time (from a few seconds to a few minutes (*e.g*. ICNALE, LINDSEI)) often before embarking on a monologic task (so they can feel more at ease). The monologic data set in ICNALE features quite a unique arrangement in terms of preparation: the learners were given the opportunity to express agreement or disagreement with a statement on the same topic on two subsequent occasions during the interview. This decision was made following findings from

> pre-experiments […] which showed that even when given enough time for preparation, some [participants] felt very nervous and could not begin speaking immediately. […] Repeating the same points was not prohibited, but most of the participants used somewhat different lexical items and phrases even when trying to convey a similar idea (Ishikawa, 2023, p. 46).

According to Nesselhauf (2004), the more controlled the data collection procedure, the less prototypical and the more peripheral the learner corpus data. It is important to note that some spoken learner corpora contain both prototypical and peripheral learner corpus data. One example is the LeaP corpus, which is made up of tasks involving free speech (prototypical learner corpus data) and highly controlled reading aloud tasks where learners are not at all free to choose the words they use (peripheral learner corpus data). These tightly controlled tasks are nevertheless highly relevant given the focus of the LeaP project on the phonetic and phonological exploration of spoken interlanguage (Gut, 2012).

Also worth mentioning are bimodal learner corpora, which include both oral and written data produced by the same or similar learners performing (some of) the same tasks or slightly different tasks about the same topic to allow for interlanguage comparisons across the two mediums (*e.g*. COREFL, CEDEL2, ICNALE; see also Abe, 2019).

3 Spoken Learner Corpora Containing Interview Data: Insights From Research Into LINDSEI

This section aims to capture and explore some of the key features of interview spoken learner corpora that need to be carefully considered when they are used in learner corpus research. The discussion is limited to spoken learner corpora that are explicitly labelled as collections

of interviews with learners (*e.g*. ICNALE, LINDSEI, NICT JLE) and mainly draws on insights gained from research into LINDSEI.

The LINDSEI project was launched in 1995 with the aim of compiling a large corpus of spoken productions by foreign language learners of English from a series of mother tongue backgrounds. Each component of the multi-L1 learner corpus contains about 50 informal interviews, each lasting around fifteen minutes, between EFL learners and interviewers, who are L1 English speakers in 64% of the interviews (Gilquin; De Cock; Granger, 2010). The interviews follow the same set pattern and are made up of three main tasks: (1) a warming-up activity in which the learners are asked to talk about one of three topics for a few minutes (an experience that has taught them something, a country that impressed them, a film or play they thought was particularly good or bad); (2) an interviewer-led free discussion mainly about university life, hobbies, foreign travel or plans for the future; and (3) a picture-based story telling activity. While the first and third tasks are predominantly monologic, they often involve dialogic interaction, minimally in the form of backchannelling from the interviewer. Interviewers can also be seen to step in early on during the first task, as learners tend to dry up rather quickly. In other words, a monologic task can become more dialogic in situations where an interlocutor is present.

The key features highlighted in the following subsections are: an interview is not a naturally occurring conversation, an interview is led by an interviewer and an interview is often made up of more than one speaking task.

3.1 An Interview Is Not a Naturally Occurring Conversation

While face-to-face interviews share some of Clark's (1996) typical features of face-to-face conversation, including copresence, visibility, audibility and instantaneity, they also differ markedly from conversations. Clark's features of self-determination and self-expression, which are central features of conversation, do not or do not fully apply in interviews. The free exchange of turns is a fundamental organising factor of conversations. Interviews differ from conversations in this respect, as the participants in interviews do not determine for themselves what actions to take and when to take them (self-determination). Interviews are organised according to a question-answer format: the interviewer asks questions and the interviewee answers them. As noted by Lazaraton (1992, p. 383), instead of being "locally managed" as in conversations, the turn-taking system in interviews is pre-specified. This is a key defining feature of interviews. Like the interlocutors in a conversation, the participants in an interview take actions as themselves (self-expression) and are not pretending to be someone else (as in a role play for example). In addition, the participants take on the roles of 'interviewer' (questioner) or 'interviewee' (responder), which implies that they have rights and obligations as 'interviewer' or 'interviewee' (Fiksdal, 1990): the interviewer has the right and obligation to ask questions, and the interviewee has the obligation to answer these questions.

Questions are instrumental in fulfilling the central communicative purpose of an interview, which is to gather information about and/or from the interviewee. Interviewers steer the interaction in terms of topics through the questions they ask. There are different interview subtypes (*e.g*. political interview, recruitment interview, chat-show) and, depending on the specific communicative purpose of the subtype, the speech event can be characterised

as more or less transactional (or informational, i.e. conveying "factual or propositional information", Brown; Yule, 1983, p. 1) and more or less interactional (McCarthy; Carter, 1994). Research has shown that even largely transactional interviews, like recruitment interviews, also exhibit interactional language (Van De Mieroop; Schnurr, 2018), with the use of small talk and humour to construct co-membership, for example.

In the LINDSEI corpus, the learners are invariably the interviewees, and they are thus expected to answer the interviewers' questions. De Cock (2022) conducted an exploratory study into the use of interrogative clauses by the learner interviewees in the Chinese, Dutch, French and Polish components of LINDSEI. The learners were not expected to use many of these clauses, given that the interviewers are the questioners. The study reveals that the learners do actually use interrogative clauses but far less frequently than in conversational corpus data, which indicates that the pre-specified Q&A format applies. Most of the interrogative clauses used by the interviewees (1) are found in reported speech and thought; (2) are used for speech management, as in (1) and (2); perform an interview-oriented metadiscursive function, as in (3); or elicit information about the interviewers to establish common ground, as in (4). Instances like (4) may incidentally seem rather unexpected, given that it is the interviewers who are normally associated with these questions (Q&A format). However, the example also shows that interrogatives can double as rapport building devices even in the hands (or mouths) of learner interviewees. 'A' is the interviewer and 'B' is the learner interviewee in the examples below.

(1) <B> because I didn't have a penny and I couldn't buy anything I only bought. (eh) **what did I buy** I bought (erm) some… pudding I <overlap/> bought some pudding Harrods pudding you know </B> (LINDSEI_PL)

(2) but (er) she has the big problem that is (er) she has the (er) train sick **may I say that** </B> <A> (erm) suffers from train sickness yeah you can say that </A> <B> yeah yeah <laughs> </B> <A> oh actually travel sickness (LINDSEI_CH)

(3) <h nt="DU" nr="DU002"> <S> <A> good </A> <B> **can I start** </B> (LINDSEI_DU)

(4) I've I've heard a lot of people say that it was too: complicated or: <overlap/> not easy to watch or </A> <B> <overlap/> (mm)… <overlap/> **did you watch it** </B> <A> <overlap/> they don't have to be. I've never <overlap/> seen it no </A> (LINDSEI_FR)

The pre-specified Q&A format and the interviewer and interviewee roles tend to lead to longer contributions from interviewees, which Troiani, Du Bois and Filchenko (2024) have found to be typical of non-naturally occurring discourse. The longer interviewee contributions are confirmed by figures included in the handbook published along with the release of the first version of the LINDSEI corpus (Gilquin, De Cock, Granger, 2010): on average the learner interviewees in the eleven LINDSEI subcorpora in the release contribute over 73% of the words uttered in the interviews.

It is interesting that researchers sometimes refer to the TLC as an interview corpus (Ishikawa, 2023; Pérez-Paredes; Mark, 2024) when the term 'interview' is never actually used by the compilers of the corpus (Gablasova *et al*., 2017; Gablasova; Brezina; McEnery, 2019). The corpus consists of more than four million words of interaction collected within the framework of the Graded Examinations in Spoken English organised by Trinity College London. Depending on the learners' proficiency level, the exam includes up to four speaking

tasks, three of which are dialogic, namely an interactive task, a discussion and a conversation. Given the examination context, the participants act as examiners (L1 users of English) and candidates (L2 speakers from a variety of L1 backgrounds). It is important to stress that the three dialogic tasks are never described as examiner-led; instead, they are presented (Gablasova *et al*., 2017) as either jointly-led (discussion and conversation) or as candidate-led (interactive task), which sets them apart from interviewer-led data, like that found in LINDSEI. While in jointly-led speaking tasks the interlocutors are expected to "co-manage the flow of discourse" (Gablasova; Brezina; McEnery, 2019, p. 145), in the candidate-led task, the candidate is expected to be "proactive in leading the conversation using a range of means such as asking questions and offering their own experience as well as a broad range of pragmatic devices for managing the interpersonal relationships in this interaction" (Gablasova; Brezina; McEnery, 2019, p. 144). This is also highlighted in the online Trinity exam booklet, where it is made explicit that the conversation task is not "a formal 'question and answer' interview" (Trinity College London, 2009, p. 7). More generally, the use of questions by candidates is presented as key in candidates initiating turns. Interestingly enough, an investigation into the use of interrogative clauses in the TLC (De Cock, 2022) reveals that it is quite common for examiners to prompt the candidates to ask them questions, as in (5) and (6) below ('E' is the L1 English examiner and 'C' is the L2 learner exam candidate).

(5) E: okay alright *do you want to ask me a question about your topic?* **C: yes erm have you ever you play hockey?** E: mm I have yeah long time C: yes (learners' L1: Spanish; B1 proficiency level)

(6) E: mm okay okay *can you ask me a question about your topic?* C: have you ever read this book? E: no I 'm it 's no but it sounds interesting so C: okay yes it is very interesting (learner's L1: Italian; B1 proficiency level)

The interviews in LINDSEI, which are interviewer-led, are characterised as informal. The informal nature of the LINDSEI interviews is anchored in a number of elements. First, the data were collected outside language testing or exam situations (unlike in the NICT JLE). Contrary to what happens in job interviews or in interviews that take place as part of language tests or exams, in the LINDSEI interviews, the interviewer has no clear power over the interviewee. The interviewer nonetheless has a directive role in that they are the questioner. Second, the LINDSEI interviewers were instructed to make the participants feel at ease. Finally, the topics discussed during the interviews tend to be everyday topics, in many ways quite similar to what would be talked about in conversations. The interviewers also adapt the topics and the questions depending on the interviewees' responses and interests. In other words, interviewers were largely given free rein, unlike interviewers in a corpus like ICNALE, where a strict protocol had to be adhered to and a script with a set of pre-defined questions had to be followed (Ishikawa, 2023).

Although the informal interviews in LINDSEI are largely unplanned, as in face-to-face conversations, they do involve some degree of preparation, and hence planning. The interviewees were given a few minutes to choose a topic and gather their thoughts about it just before the interview (task 1). The amount of preparation involved is, however, very limited, as it is restricted to what the learners say during the first few minutes of the interview (and they tend to dry up quite quickly as mentioned above). The interviewees were specifically asked not to make any written notes.

The LINDSEI interviews cannot be regarded as 'naturally occurring' discourse as they are not "discourse produced by the participants to fulfil life needs that are independent of the researcher's agenda" (Troiani; Du Bois; Filchenko, 2024, p. 176). As pointed out in the LINDSEI handbook (Gilquin; De Cock; Granger, 2010, p. 6), the LINDSEI data are not authentic or "fully natural" because they "were not produced for real communicative purposes". Unlike naturally occurring conversation among friends, the LINDSEI interviews would not have taken place had the learners not been recruited and had the data not been recorded as part of the data collection. That said, the handbook suggests that the data in the first two parts of the interviews arguably display a fairly high degree of naturalness "as the learners were free to choose both the ideas they wanted to express and the language in which to express them" (Gilquin; De Cock; Granger, 2010, p. 6). Although the third task was more constrained because of the use of pictures, the learners were still free to choose not only how they went about completing the task, but also the actual wording they used. The decision to go for informal interviews within the framework of the LINDSEI project was motivated by the fact that this spoken genre "has the advantage of imposing few constraints on language production" (Gilquin; De Cock; Granger, 2010, p. 3), which was seen as essential in order to ensure that the corpus could be of use to researchers interested in many different aspects of spoken interlanguage. A different approach is taken in some other interview learner corpora, where tasks were designed so as to make learners use certain language forms and phenomena. For example, the learners who participated in the ICNALE spoken data collection had to start the serial picture description task with "two weeks ago" (Ishikawa, 2023, p. 50). Another example concerns the NICT JLE corpus (Tono, 2007, p. 167): two of the speaking tasks are described as meant to elicit "simple present tense narration" and "simple past tense narration", respectively.

In short, informal interviews were collected within the framework of the LINDSEI project in an attempt to gather free and informal discussions. Although they may be edging towards conversations because of their informal and free character, they are clearly interviewer-led and do not qualify as naturally occurring conversations.

In view of the specific genre collected from the learners in LINDSEI, the Louvain Corpus of Native English Conversation (LOCNEC) was also compiled to act as a comparable corpus containing the same type of interviews (the name LOCNEC is a misnomer!) but with L1 English students. LOCNEC was thus designed to make it possible for researchers to guarantee genre comparability when they carry out Contrastive Interlanguage Analysis (CIA) (Granger, 2015), which involves contrasting a reference variety (or reference varieties) with an interlanguage variety (or interlanguage varieties). A study exploring the use of pragmatic prefabs in the spoken productions by the learner interviewees in the LINDSEI French component and in face-to-face conversations in the British National Corpus (BNC) revealed that some of the prefabs used in the BNC are never or hardly ever used by the learners (*e.g.* 'come on', 'I see', 'hang on'). An exploration of LOCNEC shows that the English L1 students in the corpus also never or hardly ever use these prefabs, which can be explained by the fact that interviewees are unlikely to perform a series of pragmatic functions in an interview context given the specificity of the communication situation (De Cock, 2002). This points to the effect of genre, and highlights the fact that the importance of controlling for genre (Granger, 2015) is not limited to learner corpus research based on written corpora.

As pointed out by Pérez-Paredes and Mark (2024), interviews can be both an elicitation technique and a genre. LINDSEI is actually a combination of both: (1) the interviews were set

up in order to elicit spoken data from learners of English and include them in a corpus and (2) the bulk of the speech events gathered can be seen to belong to the informal interview genre described above. Other spoken learner corpora, described as interview corpora, do not display this dual nature to the same extent. One example is the spoken dialogic part of ICNALE. While interviews were used to elicit the dialogic data (elicitation technique), one of the speech tasks involves role playing, where the interviewee plays the part (1) of a college student who needs to persuade their supervisor (played by the interviewer) that they should be allowed to continue working part time or (2) of a customer trying to persuade a restaurant owner (the interviewer) that they should receive a refund because people smoking in the restaurant ruined their meal (Ishikawa, 2023, p. 51). Role-plays, which tend to be mainly used in language teaching and testing, are arguably not typically associated with the interview as a genre in that the participants are made to act as someone else.

3.2 An Interview Is Led by an Interviewer

When conducting research into aspects of spoken interlanguage using an interview corpus, the focus tends to be on the language used by the learner interviewees only. The linguistic phenomena under study are normally identified in or retrieved from the learner interviewee turns. Although this is of course valid as a first step in investigations of learner speech, it is important to bear in mind that the interviews are led by an interviewer, who may also have an impact on the way the learners speak.

The following interviewer variables are recorded in the LINDSEI metadata, and they can readily be accessed when examining learners' productions or in order to select the set of data to be investigated: gender, L1, knowledge of foreign languages and status (i.e. familiar, vaguely familiar or unfamiliar, depending on the degree to which the interlocutors know each other).

Interviewer variables have so far not received a great deal of attention in learner corpus research. However, research into learners' use of communicative strategies and foreign words, i.e. words from other languages than English (which, incidentally, are predominantly, but not exclusively, from the learners' L1), has shed light on the impact of interviewers' L1 or knowledge of the learners' L1 on interviewees' use of these words (Nacey; Graedler, 2013 and especially De Cock, 2017). More qualitative analyses of learner interlanguage have also provided examples of how an interviewer's input can to some extent shape a learner's use of lexis and collocations (see for example De Cock, 2011, 2019).

A 2022 study by Götz, Wolk and Jaeschke is one of the most systematic quantitative explorations of the impact of some interviewer variables on learners' productions in LINDSEI to date. The researchers set out to investigate how learners' use of four fluencemes (filled and unfilled pauses, discourse markers and repeats) in the German, Spanish, Bulgarian and Japanese LINDSEI subcorpora is affected by a series of variables, including some that are related to the communicative behaviour of the interviewer in each of the subcorpora, namely the interviewer's frequency of use of the fluencemes under study and the "overall share of the interview (as a percentage within a particular task type) that was contributed by the interviewer" (Götz; Wolk; Jaeschke, 2022, p. 281). The analysis reveals that the interviewers' communicative behaviour has a significant effect on learners' use of fluencemes with interviewers' increased contributions leading to learners' reduced use of filled and unfilled

pauses. The authors argue that their findings indicate that the phenomenon of confluence, defined by McCarthy (2006) as speakers contributing to one another's fluency, "seems to figure prominently in learner interview data" (Götz; Wolk; Jaeschke, 2022, p. 293). They conclude that their study has methodological implications in that it shows that it is essential to control for interviewers' contributions, as they can have a significant impact on learners' use of linguistic phenomena, such as fluencemes. They also highlight that, if interviews are used to evaluate learners' oral fluency, "it is advisable to instruct the interviewers to show the same communicative behavior in all interviews", thereby pointing to the importance of consistency when compiling components of the same corpus project (Götz; Wolk; Jaeschke, 2022, p. 293).

Achieving this consistency is challenging, however. Even if the interviews in LINDSEI all follow the same set pattern, Pérez-Paredes and Mark (2024) have pointed to variability in the way interviews are conducted both across and within some subcorpora. This variability stems from the lack of strict interview protocol (cf. Section 3.1 above), which was adopted to foster more informal interaction, encourage participation and ensure the learners could "choose their own wording" (Granger, 2008, p. 261) and feel at ease. The absence of strict guidelines also means that interviewers had more freedom not only to adapt to the learners' contributions and proficiency levels, but also to bring their own interpretation of informal interviews (anchored in personal experience and culture) to the data collection, with more or less engagement with the learners. As a result, while some interviewers regularly act upon opportunities for interaction, others tend to provide only minimal backchannelling and hold back to let the learners speak as much as possible. This variability raises comparability issues between the various LINDSEI components and Pérez-Paredes and Mark's (2024, p. 141) are, like Götz, Wolk and Jaeschke (2022), right to call for more "attention to and awareness of [interviewers'] influence on the resulting data". Carefully recording interviewer metadata and documenting the corpus compilation process with as much information as possible about interviewers' take on informal interviews is essential.

3.3 An Interview Is Often Made Up of More Than One Speaking Task

Spoken learner corpora, and interview learner corpora in particular, frequently comprise a series of speaking tasks, and detailed descriptions of these corpora need to be available to the research community. It is also crucial for the researchers using the corpora to be fully aware of the exact make-up of the data so they can make informed decisions about corpus data selection during the study design phase of research and provide informed interpretations when discussing the findings.

Findings from Gráf (2019) provide concrete evidence of the care that is required. The study set out to explore the effect of elicitation-task design on speech rate (measured in words per minute) in the LINDSEI Czech learner subcorpus and in LOCNEC. A smaller corpus of interviews in Czech by 31 of the 50 same speakers as in the LINDSEI Czech component was also used to investigate similarities between the learners' speech rate in English and in their L1. The analysis reveals that speech rate is significantly slower in the picture-based story-telling task than in the other two tasks (the warming up task about one of three topics and the free discussion about everyday topics) in all three corpora. An exploration of individual learner variation confirms that the majority of the native speakers and the learners (whether they are using English or Czech) exhibit a slower speech rate in the third task. Lumping all three tasks

together would have missed this finding, which has an important and direct pedagogical implication, as picture descriptions and picture-based storytelling tasks are frequently used in the classroom and in the assessment of spoken language. As emphasised by Gráf (2019), it is essential for teachers and examiners who evaluate spoken skills, and particularly fluency, to be fully aware that picture-based story tasks are cognitively challenging (even in a speaker's L1) and that they tend to lead to slower, less fluent speech. Further evidence of task type effect can, for example, be found in Dumont (2017) and Götz, Wolk and Jaeschke (2022) in connection with a series of fluency phenomena and in De Cock (2011) in connection with the use of evaluative adjectives.

In view of the potential effect of the task type, it was decided that the three tasks in the LINDSEI corpus should be identified by using mark-up so that researchers can compare language use across the tasks. The mark-up thus makes it easy for researchers to select the speaking task(s) they would like to include in or discard from their analysis.

4 Discussion

As shown in the previous section, there is far more to a learner corpus of interview data than meets the eye. Spoken language is not a unified phenomenon and corpora of spoken interlanguage can only be representative of the type of data or genre(s) included in the corpus. It is therefore essential for corpus compilers to provide detailed information about the corpus and its compilation. It is equally crucial for researchers to carefully scrutinize the make-up of the corpus data they intend to use and the available metadata during the study design phase: they need to be able to establish the extent to which the special interaction data would be suitable given the focus of the research and the research question(s) and whether certain tasks should be discarded. As emphasised by Ädel (2020, p. 10),

> [a]nybody who wants the claims made based on a corpus to be accepted by the research community needs to show in some way that the corpus material is appropriate for the type of research done. With incomplete description of the corpus, people will be left wondering whether the material was in fact valid for the study.

The characterisation of the LINDSEI informal interviews discussed above has shown that it is important to go beyond the 'spoken language' label of a corpus of oral interviews. Such a corpus needs to be described and investigated as a collection of interviews and not just as a spoken learner corpus, and certainly not as a corpus of naturally occurring conversation data. Pérez-Paredes and Mark (2024, p. 113) state 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". This statement needs reformulating, as spoken language is not homogeneous: interview data provide a representation of **one type of** spoken language and, while these data may not "allow for evidence of the interactional features of everyday" **conversation**, they nonetheless represent one kind of spoken interaction in a specific context.

As pointed out by Friginal *et al*. (2017, p. 44), LINDSEI "provides an excellent approximation of the language that L2 learners might choose to use in real-world interview

contexts". Language learners are likely to have to use their L2 outside the classroom environment within the context of university admission interviews, recruitment interviews and oral proficiency interviews, for example. Even if the LINDSEI interviews are more informal than the aforementioned speech events, LINDSEI-based research can help examine how learners fare and use language in an interview context with an interviewer who is not a peer. Findings from this kind of research can feed into data-driven learning (DDL) exercises designed to encourage learners to identify some of the language and successful strategies the learner interviewees use when they are given the floor and need to answer questions and keep talking, when they need to buy time to think about what to say next, when they face more challenging questions, or when they seek to establish rapport with the interviewer. Research on interview data collected as part of language tests or exams can also contribute to the development of DDL pedagogical material that would help learners prepare for and handle the test or examination.

In addition to the pre-specified turn-taking system and the Q&A format, the possible impact of the interviewer and the multi-task make-up of interviews also need to be taken into account and even foregrounded during the analysis of interlanguage and the interpretation of results. This is especially important in view of the consequences for pedagogical implications, as illustrated by Gráf (2019) and Götz, Wolk and Jaeschke (2022).

Researchers who adopt a CIA approach need to be able to have access to a reference variety that exhibits the same genre as that in the learner corpus they are investigating. Comparing learner interview data with L1 speaker conversation data, for example, would be comparing apples and oranges, and would not be fair on the learners. As mentioned above, when contrasting different interlanguage varieties (*e.g*. the LINDSEI_Italian and the LINDSEI_Japanese components), researchers also need to allow for the possibility of a certain degree of variability between the different LINDSEI components as a result of the absence of a strict interview protocol and the different interpretations the interviewers may have of how an informal interview should be conducted. With the second version of the LINDSEI corpus in the pipeline (with 8 new components; 19 mother backgrounds will be represented in total in version two), it would be timely to examine the extent and nature of the variability between the various components in a large-scale investigation. Another type of corpus data should also feature on learner corpus researchers and compilers' wish list. Gráf's study (2019) showcases the added value of collecting and exploring/analysing data representing the same genre (collected in the same context) produced by the same learners in their L1. This type of data can make it possible to assess the cognitive burden of a certain speaking task or to factor in personal speaking style.

5 Concluding Remarks and Looking Ahead

Existing spoken interlanguage corpora are highly valuable, provided they come with detailed descriptions and rich metadata, and their specific nature is fully considered and embraced and not swept under the carpet or mislabelled (*e.g*. when interview data is presented as conversation data). Like L1 language users, L2 learners are likely to produce spoken texts in their TL in a wide range of contexts. A wider constellation of spoken (and written) learner corpora representing different genres should thus ideally be available to the entire learner

corpus research community. The call for more carefully compiled and described spoken learner corpora of different types (*e.g*. tasks, communication situation) is echoed in Myles (2021) and Gablasova, Brezina and McEnery (2019, p. 129), who emphasise the need for more of these corpora to be able "to complement evidence from existing resources and to build more comprehensive corpus-based models of spoken L2 use, leading to a better understanding of the interplay among social, linguistic and psycholinguistic variables involved in language communication".

A spoken genre that has long been underrepresented in spoken learner corpora is that of informal face-to-face conversation between L2 learners (Pérez-Paredes; Mark, 2024). According to Gilquin (2020, p. 284), this underrepresentation can largely be explained by the fact that "[h]aving a spontaneous conversation with a friend, for example, is more likely to occur in the mother tongue (L1) than in the target language (L2)". This is especially true in a foreign language learning context. One rare exception is the Learner of Persian Spoken Corpus (LoPSC, Daghbandan, 2023), which provides an example of how the collection of foreign learner conversation data could be implemented. Daghbandan (2023) dared to venture into less well-trodden territory when she set out to compile a corpus of informal conversations between L1 English learners of Persian at a British university. The participants, who are used to having conversations in Persian together during class time, were recruited via tutors for Persian courses and were encouraged to participate in conversations between students outside of classes in order to gain additional speaking practice in Persian. The participants were free to choose when and where the conversations would take place, who they would be having a conversation with (usually in pairs or groups of three), what would be discussed and how long they would be interacting (the conversations ended up lasting about 50 minutes on average). The learners were, however, required to use the recording equipment provided to ensure good sound quality (for the subsequent transcription work). Although the 'naturally occurring' label could not be used to qualify these conversations because the learners were 'set up' to arrange conversations for the data collection, the LoPSC project provides an excellent starting point for developing initiatives that would aim to capture fairly naturalistic informal conversations between learners beyond the classroom and fill a sizeable gap in the spoken learner corpus constellation. Allowing the learners to use their own smartphones to record the conversations may contribute to increased naturalness (this was done to compile conversation corpora involving native speakers of English in the Spoken British National Corpus, 2014; cf. Love, 2020; and in the Lancaster-Northern Arizona Corpus of Spoken American English, Hanks *et al.*, 2024).

New spoken learner corpora are bound to and should be encouraged to join the constellation following the ever-evolving teaching, learning and testing situations. An example of recent addition is the spoken component of the British Council-Lancaster Aptis Corpus, which includes computer-mediated speaking tests without a live human interlocutor. These tests have become increasingly popular since the COVID-19 pandemic (Gablasova et al., 2024) and research into the new British Council-Lancaster Aptis Corpus, as reported in Gablasova et al. (2024, p. 209), "can be used to inform ongoing testing practices".

References

AAS, H. L.; NACEY, S. Methodological Concerns for Investigating Pause Behavior in Spoken Corpora. *In*: DEGAND, L.; GILQUIN, G.; MEURANT, L.; SIMON, A-C. (ed.). *Fluency and Disfluency Across Languages and Language Varieties*. Louvain-la-Neuve: Presses Universitaires de Louvain, 2019. p. 41-64.

ABE, M. Comparing Errors Across an L2 Spoken and Written Error-Tagged Japanese EFL Learner Corpus. *In*: GÖTZ, S.; MUKHERJEE, J. (ed.). *Learner Corpora and Language Teaching*. Amsterdam; Philadelphia: John Benjamins, 2019. p. 157-174.

ÄDEL, A. Corpus Compilation. *In*: PAQUOT, M.; GRIES, S. (ed). *A Practical Handbook of Corpus Linguistics*. [*S. l.*]: Springer, 2020. p. 3-24.

BALLIER, N.; MARTIN, P. Speech Annotation of Learner Corpora. *In*: GRANGER, S.; GILQUIN, G.; MEUNIER, F. (ed.). *The Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, 2015. p. 107-134.

BIBER, D.; JOHANSSON, S.; LEECH, G.; CONRAD, S.; FINEGAN, E. *Longman Grammar of Spoken and Written English*. Harlow: Longman, 1999. 1204p.

BROWN, G.; YULE, G. *Discourse Analysis*. Cambridge: Cambridge University Press, 1983. 288p.

CHAFE, W. L.; DANIELEWICZ, J. Properties of Spoken and Written Language. *In*: HOROWITZ, R.; SAMUELS, S. J. (ed.). *Comprehending Oral and Written Language*. New York: Academic press, 1987. p. 82-113.

CLARK, H. H. *Using Language*. Cambridge: Cambridge University Press, 1996. p. 432.

CRAWFORD, W. J.; MCDONOUGH, K. Introduction to the Corpus of Collaborative Oral Tasks. *In*: CRAWFORD, W. J. (ed.). *Multiple Perspectives on Learner Interaction*: The Corpus of Collaborative Oral Tasks. Berlin; Boston: de Gruyter, 2021. p. 7-16.

CRYSTAL, D. *The Cambridge Encyclopedia of the English Language*. Cambridge: Cambridge University Press, 1995. p. 489.

DAGHBANDAN, S. *A Corpus-Based Study on the Use of Conversational Persian by Learners of Persian*. 2023. Dissertation (PhD of Philosophy in Education) – The University of Edinburgh, Edinburgh, 2023. Unpublished.

DE COCK, S. *Do You Love Me*: Interrogatives in Learner Speech in LINDSEI and in the Trinity Lancaster Corpus. Paper presented at the 6th Learner Corpus Research Conference, Padova, Italy, 2022.

DE COCK, S. Foreign Words in EFL Learner Interviewee Speech: Lending Learners a Productive Fluency Helping Hand? *In*: DEGAND, L.; GILQUIN, G.; MEURANT, L.; SIMON, A-C. (ed.). *Fluency and Disfluency Across Languages and Language Varieties*. Louvain-la-Neuve: Presses Universitaires de Louvain, 2019. p. 283-299.

DE COCK, S. Pragmatic Prefabs in Learners' Dictionaries. *In*: BRAASCH, A.; POVLSEN, C. (ed.). *Proceedings of the Tenth EURALEX International Congress, EURALEX 2002*. Copenhagen: Center for Sprogteknologi, 2002. v. 2, p. 471-781.

DE COCK, S. Preferred Patterns of Use of Positive and Negative Evaluative Adjectives in Native and Learner Speech: An ELT Perspective. *In*: FRANKENBERG-GARCIA, A.; FLOWERDEW, L.; ASTON, G. (ed.). *New Trends in Corpora and Language Learning*. London; New York: Continuum, 2011. p. 198-212.

DE COCK, S. *Speaking in Tongues*': EFL Learners' Use of 'Foreign Words' in Informal Interviews. Paper presented at the 4th Learner Corpus Research Conference (LCR 2017), Bolzano, Italy, 2017.

DUMONT, A. *Fluency and Disfluency*: A Corpus Study of Non-Native and Native Speaker (Dis)Fluency Profiles. 2018. Dissertation (PhD in Philosophy and Letters) – Université catholique de Louvain, Louvain-la-Neuve, 2018. Unpublished.

DUMONT, A. *The Effects of Speaking Tasks on L2 Fluency*. Paper presented at the 4th Learner Corpus Research Conference (LCR2017), Bolzano, Italy, 2017.

FIKSDAL, S. *The Right Time and Pace*: A Microanalysis of Cross-Cultural Gatekeeping Interviews. New Jersey: Ablex Norwood, 1990. p. 168.

FRIGINAL, E.; LEE, J. J., POLAT, B.; ROBERSON, A. *Exploring Spoken English Learner Language Using Corpora*: Learner Talk. [*S. l.*]: Palgrave Macmillan, 2017. p. 300.

GABLASOVA, D.; BREZINA, V.; MCENERY, T. The Trinity Lancaster Corpus: Development, Description and Application. *International Journal of Learner Corpus Research*, v. 5, n. 2, p. 126-158, 2019.

GABLASOVA, D.; BREZINA, V.; MCENERY, T.; BOYD, E. Epistemic Stance in Spoken L2 English: The Effect of Task and Speaker Style. *Applied Linguistics*, v. 38, n. 5, p. 1-26, 2017.

GABLASOVA, D.; HARDING, L.; BREZINA, V.; DUNLEA, J. Expressions of Epistemic Stance in Computer-Mediated L2 Speaking Assessment. A Corpus-Based Approach. *International Journal of Learner Corpus Research*, v. 10, n. 1, p. 183-215, 2024.

GHANEM, R. ESL Students' Use of Suprasegmental Features in Informative and Opinion-Based Tasks. *In*: CRAWFORD, W. J. (ed.). *Multiple Perspectives on Learner Interaction*: The Corpus of Collaborative Oral Tasks. Berlin; Boston: de Gruyter, 2021. p. 173-198.

GILQUIN, G. Learner Corpora. *In*: PAQUOT, M.; GRIES, S. (ed.). *A Practical Handbook of Corpus Linguistics*. [*S. l.*]: Springer, 2020. p. 283-303.

GILQUIN, G.; DE COCK, S.; GRANGER, S. (ed.). *The Louvain International Database of Spoken English Interlanguage*. Handbook and CD-ROM. Louvain-la-Neuve: Presses Universitaires de Louvain, 2010. p. 111.

GÖTZ, S.; WOLK, C.; JAESCHKE, K. Contextualizing Fluency in Advanced Spoken Learner Language. A Contrastive Interlanguage Analysis Across L1s, Task Types and Learning Context Variables. *In*: LENKO-SZYMANSKA, A.; GÖTZ, S. (ed.). *Complexity, Accuracy and Fluency in Learner Corpus Research*. Amsterdam; Philadelphia: John Benjamins, 2022. p. 273-297.

GRÁF, T. Speech Rate Revisited: The Effect of Task Design on Speech Rate. *In*: GÖTZ, S.; MUKHERJEE, J. (ed.). *Learner Corpora and Language Teaching*. Amsterdam; Philadelphia: John Benjamins, 2019. p. 175-190.

GRANGER, S. Contrastive Interlanguage Analysis. A Reappraisal. *International Journal of Learner Corpus Research*, v. 1, n. 1, p. 7-24, 2015.

GRANGER, S. Learner Corpora. *In*: LÜDELING, A.; KYTÖ, M. (ed.). *Corpus Linguistics*. An International Handbook. Berlin; New York: Walter de Gruyter, 2008. v. 1, p. 259-275.

GUT, U. The LeaP Corpus. A Multilingual Corpus of Spoken Learner German and Learner English. *In*: SCHMIDT, T.; WÖRNER, K. (ed.). *Multilingual Corpora and Multilingual Corpus Analysis*. Amsterdam; Philadelphia: John Benjamins, 2012. p. 3-33.

HANKS, E.; MCENERY, T.; EGBERT, J.; LARSSON, T.; BIBER, D.; REPPEN, R.; BAKER, P.; BREZINA, V.; BROOKES, G.; CLARKE, I; BOTTINI, R. Building LANA-CASE, a Spoken Corpus of American English Conversation: Challenges and Innovations in Corpus Compilation. *Research in Corpus Linguistics*, v. 12, n. 2, p. 24-44, 2024.

ISHIKAWA, S. I. *The ICNALE Guide*: An Introduction to a Learner Corpus Study on Asian Learners' L2 English. London; New York: Routledge, 2023. p. 229.

LAZARATON, A. The Structural Organization of a Language Interview: A Conversation Analytic Perspective. *System*, v. 20, n. 3, p. 373-386, 1992.

LOVE, R. *Overcoming Challenges in Corpus Construction*. The Spoken British National Corpus 2014. New York; London: Routledge, 2020. p. 202.

LOZANO, C. CEDEL2: Design, Compilation and Web Interface of an Online Corpus for L2 Spanish Acquisition Research. *Second Language Research*, v. 38, n. 4, p. 965-983, 2022.

LOZANO, C.; DÍAZ-NEGRILLO, A.; CALLIES, M. Designing and Compiling a Learner Corpus of Written and Spoken Narratives: COREFL. *In*: BONGART, C.; TORREGROSSA, J. (ed.). *What's in a Narrative?* Variation in Story-Telling at the Interface Between Language and Literacy. Berlin: Peter Lang, 2021. p. 21-45.

MCCARTHY, M. J. Fluency and Confluence: What Fluent Speakers Do. *In*: MCCARTHY, M. J. (ed.). *Explorations in Corpus Linguistics*. Cambridge: Cambridge University Press, 2006. p. 1-6.

MCCARTHY, M.; CARTER, R. *Language as Discourse*: Perspectives for Language Teaching. London; New York: Longman, 1994. p. 230.

MÜLLER, S. *Discourse Markers in Native and Non-Native English Discourse*. Amsterdam; Philadelphia: John Benjamins, 2005. p. 290.

MYLES, F. Commentary: An SLA Perspective on Learner Corpus Research. *In*: LE BRUYN, B.; PAQUOT, M. (ed.). *Learner Corpus Research Meets Second Language Acquisition*. Cambridge; New York: Cambridge University Press, 2021. p. 258-273.

NACEY, S.; GRAEDLER, A.-L. Communication Strategies Used by Norwegian Students of English. *In*: GRANGER, S.; GILQUIN, G.; MEUNIER, F. (ed.). *Twenty Years of Learner Corpus Research*: Looking Back, Moving Ahead. Louvain-la-Neuve: Presses Universitaires de Louvain, 2013. p. 345-356.

NESSELHAUF, N. Learner Corpora and Their Potential in Language Teaching. *In*: SINCLAIR, J. (ed.). *How to Use Corpora in Language Teaching*. Amsterdam; Philadelphia: Benjamins, 2004. p. 125-152.

OSBORNE, J. Fluency, Complexity and Informativeness in Native and Non-Native Speech. *In*: GILQUIN, G.; DE COCK, S. (ed.). *Errors and Disfluencies in Spoken Corpora*. Amsterdam: John Benjamins, 2013. p. 139-161.

PAQUOT, M.; KÖNIG, A.; STEMLE, E. W.; FREY, J.-C. The Core Metadata Schema for Learner Corpora: Collaborative Efforts to Advance Data Discoverability, Metadata Quality and Study Comparability in L2 research. *International Journal of Learner Corpus Research*, v. 10, n. 2, p. 280-300, 2024.

PÉREZ-PAREDES, P.; MARK, G. Rethinking Interviews as Representations of Spoken Language in Learner Corpora. *Research in Corpus Linguistics*, v. 12, n. 2, p. 111-145, 2024.

REDER, S.; HARRIS, K.; SETZLER, K. The Multimedia Adult ESL Learner Corpus. *TESOL Quarterly*, v. 37, n. 3, p. 546-557, 2003.

TRINITY COLLEGE LONDON. *Exam Information*: Graded Examinations in Spoken English (GESE). London: Trinity College London, 2009. Available at: https://www.trinitycollege.com/resource?id=5755. Accessed on: 2 Sept. 2024.

TONO, Y. The Roles of Oral L2 Learner Corpora in Language Teaching: The Case of the NICT JLE Corpus. *In*: CAMPOY, M. C.; LUZÓN, M. J. (ed.). *Spoken Corpora in Applied Linguistics*. Bern: Peter Lang, 2007. p. 163-179.

TROIANI, G.; DU BOIS, J. W.; FILCHENKO, A. Corpus as a Slice of Life: Representing Naturally Occurring Language and Its Speakers. *Research in Corpus Linguistics*, v. 12, n. 2, p. 174-202, 2024.

VAN DE MIEROOP, D.; SCHNURR, S. Candidates' Humour and the Construction of Co-Membership in Job Interviews. *Language & Communication*, v. 61, p. 35-45, 2018.

History

Date of submission: 27/11/2024.
Date of approval: 01/08/2025.

Responsible Editor

Andréa Machado de Almeida Mattos, Universidade Federal de Minas Gerais (UFMG), Belo Horizonte, Minas Gerais/MG, Brasil. Lattes: http://lattes.cnpq.br/7749222257907067, ORCID: https://orcid.org/0000-0003-3190-7329, e-mail: andreamattos@ufmg.br.

Declaration of Conflict of Interest

The author declares that there is no conflict of interest.
Data Availability Statement
Information on how to access the LINDSEI data can be found here: https://www.uclouvain.be/en/research-institutes/ilc/cecl/lindsei-cd-rom-and-handbook. The LOCNEC data will be made available in the next release of the LINDSEI data.

Use of AI

No GenAI was used in the writing of this article.

Reviews

As part of the commitment made by the Brazilian Journal of Applied Linguistics to Open Science, the journal publishes the reviews issued regarding its published works, when authorized by all parties involved.

Review 1

The text provides a valuable contribution to the field by introducing a meaningful discussion about the role of spoken interlanguage corpora and, as such, learner interview corpora. The focus on features of spoken learner interview corpora highlights aspects that can certainly contribute to the compilation of future spoken corpora.

Reviewed on: 25/01/2025.
Recommendation: Accept Submission.

Reviewer: Ana Luiza Pires de Freitas.