The Trinity Lancaster Corpus

Development, description and application

Dana Gablasova, Vaclav Brezina and Tony McEnery
Lancaster University, UK

This paper introduces a new corpus resource for language learning research, the Trinity Lancaster Corpus (TLC), which contains 4.2 million words of interaction between L1 and L2 speakers of English. The corpus includes spoken production from over 2,000 L2 speakers from different linguistic and cultural backgrounds at different levels of proficiency engaged in two to four tasks. The paper provides a description of the TLC and places it in the context of current learner corpus development and research. The discussion of practical decisions taken in the construction of the TLC also enables a critical reflection on current methodological issues in corpus construction.

Keywords: Trinity Lancaster Corpus, spoken language, transcription, corpus design, corpus compilation

1. Introduction

In order to move forward, a discipline needs to continually reflect on its practices and draw on this reflection when developing new resources, methods and theories. Corpus linguistics as a field has evolved dramatically over the last twenty years, extending and refining its methods and resources as well as increasing its interdisciplinary reach (see for example Semino et al. 2017; Dayrell & Urry 2015). This development has been, on the one hand, enabled by technological advances from which the field has benefited greatly; on the other hand, it has been driven by active reflection from inside the field on the available resources, methods and their applications (e.g. Brezina 2018; Baker & Egbert 2016; Gries 2015). Therefore, ideally, a new corpus resource should react not only to an immediate research need or to the availability of data; it should also be used as an opportunity to advance the field and to engage with broader issues in the discipline (Leech 2007). This paper seeks to contribute to this ongoing reflection in the area of corpus building by providing a description of a new corpus resource for investigating spoken L2

https://doi.org/10.1075/ijlcr.19001.gab
International Journal of Learner Corpus Research 5:2 (2019), pp. 126–158. issn 2215-1478 | e-issn 2215-1486

English interaction, the *Trinity Lancaster Corpus* (TLC), and by placing it in the context of current approaches to (learner) corpus development.

The TLC was developed in a cooperation between the Centre for Corpus Approaches to Social Science (CASS) at Lancaster University and Trinity College London, a major international examination board, which conducts tests of spoken English in over sixty countries worldwide. The data used in the corpus were collected from 2012 to 2018 as part of the Graded Examinations in Spoken English (GESE), an exam developed and administered by Trinity College London (Trinity College London 2016). Overall, the corpus contains 4.2 million words (tokens) of transcribed spoken interaction between exam candidates (L2 speakers of English) and examiners (L1 speakers of English). The L2 data come from over 2,000 L2 speakers from different cultural and linguistic backgrounds and with a range of sociolinguistic characteristics. In addition to the language samples, the corpus also contains further metadata about the L2 speakers collected through background questionnaires.

The properties of a corpus, such as representativeness, structure and amount of evidence, directly affect the ability of researchers to interpret findings and generalise to contexts outside of the corpus (Leech 2007; Gablasova, Brezina & McEnery 2017). Decisions made at the corpus-building stage can thus have far-reaching consequences for the quality of research studies based on them; this is especially true of large-scale corpus-building projects, with their products expected to be used in a large number of research studies (e.g. such as the recently developed Spoken BNC 2014; see Love et al. 2017). Despite the importance of corpora and the process of their development, in the last decade this area has not been given as much attention as methodological issues related to the analysis and interpretation of the data (e.g. Brezina 2018).

This article thus has a dual aim. First, it seeks to provide a transparent and documented account of key decisions made in a large project focused on building the TLC, which represents, to date, the largest corpus resource for the investigation of spoken (interactive) L2 English production. Second, the discussion of the methodological decisions involved in building the corpus provides a unique opportunity to reflect on current trends in corpus creation, acknowledging both advances and limitations, and considering some of the key conceptual and methodological questions faced during corpus construction. It also demonstrates how, when creating corpora, it is important to reflect on the expected users of the resource and the research questions that they may wish to address. In doing so, the paper moves questions of corpus-building to the forefront of attention in corpus-based L2 research.

1.1 The Trinity Lancaster Corpus in the context of learner corpus research on spoken L2

The TLC contains transcribed oral production from different types of interaction between L1 and L2 speakers of English. Producing this dataset took over five years, included more than 3,500 hours of transcription and countless hours of planning, checking the accuracy of the transcripts and further processing of the data. The creation of corpora of spoken language remains a time-consuming process. We therefore need to ask why, given the existence of several corpora of spoken L2 English, we believed an additional resource was needed. There are several reasons that acted as a motivation for this project.

First of all, corpora representing speech are an invaluable resource in investigations of (second) language acquisition and use (Myles 2015). Speaking, both in first and second language, is an essential channel of social communication. Nevertheless, to successfully engage in oral, face-to-face, interactive communication is a cognitively and linguistically demanding task. Unlike writing, which usually offers greater control, speaking does not provide interlocutors with the possibility to revisit and edit their output at a later time; the production has to be monitored and, where necessary, repaired while the interaction is ongoing. In addition, the speaker also needs to engage with the other interlocutor and react flexibly to the development of the conversation (Kormos 2014). As a result, the observation of speech can provide invaluable insight into how users acquire and produce language. As du Bois (1991:73) puts it,

> it is in spoken discourse that the process of the production of language is most accessible to the observer. Hesitations, pauses, glottal constrictions, false starts, and numerous other subtle evidences observable in speech but not in writing provide clues to how participants mobilize resources to plan and produce their utterances, and to how they negotiate with each other the ongoing social interaction.

Corpora of spoken language can thus provide descriptions of distinctive features of spoken language and of the ways in which it may differ from writing (e.g. Leech 2000). Speech also holds a special place in psycholinguistic research into L1 and L2 acquisition and production, providing unique insights into the links between language use and cognitive processes (Tomasello 2003). Furthermore, machine-searchable records of spoken language enable the systematic investigation of frequency-related patterns in language use, a major area of research relevant both for psycholinguistic accounts of language production (e.g. Ellis 2002) as well as for language pedagogy (McEnery, Xiao & Tono 2006). Spoken interactive communication also provides rich data for corpus-based pragmatic research into interpersonal factors in language use (e.g. Aijmer 2014). These are but a few of the

many areas which can benefit directly from empirical evidence based on corpora of spoken language.

Despite the prominent role of speaking in language use in general, large-scale corpora of spoken language are still relatively limited in number and corpora of spoken interlanguage are rarer still. Although there are both larger and smaller corpora representing different kinds of spoken L2 English such as, for example, the Louvain International Database of Spoken English Interlanguage (LINDSEI; Gilquin, de Cock & Granger 2010), the Giessen – Long Beach Chaplin Corpus (GLBCC; Jucker, Smith & Lüdge 2003) and the Barcelona English Language Corpus (BELC; Muñoz 2006), given the sheer diversity of language settings in which spoken language is used as well as the broad range of L2 speaker characteristics, the current corpora can capture only a fraction of these contexts. More corpora are needed 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 between social, linguistic and psycholinguistic variables involved in language communication. New learner corpora of speech can increase the robustness of corpus-based models of L2 use in two major ways: first, by providing accounts of spoken language use in new or previously under-described linguistic settings and, second, by allowing for (at least a partial) replication of previous studies.

The TLC contains language from a well-defined environment (semi-formal institutional discourse) and includes a range of metadata about the speakers (see Section 3). These characteristics make it a valuable resource for both addressing novel research questions as well as for re-visiting the effect of variables studied previously. For example, contributing to a novel area of inquiry, the TLC contains a substantial component consisting of L2 English speakers from India with different L1 backgrounds (e.g. Hindi and Gujarati), so far a rather under-researched group of English users in corpus-based L2 studies. It also contains L2 English speakers from more commonly investigated L1 backgrounds (e.g. Chinese and Spanish), allowing for replication of studies on learners from these backgrounds. The scientific value of replication studies should not be underestimated (Porte 2012); they provide information about whether findings based on one corpus are just typical of the linguistic settings/tasks and learners represented in that corpus, or whether they capture more general traits of a particular learner group (e.g. advanced L1 Chinese speakers) or a setting (e.g. academic discourse).

2. The development of the Trinity Lancaster Corpus: Key issues

In this section, we will discuss some of the key issues in the development of learner corpora of spoken language and relate them to the building of the TLC. In particular, the focus will be on (i) corpus representativeness, (ii) the amount of evidence in the corpus and (iii) transcription as a process of transformation of spoken language into a written, machine-readable format.

2.1 Representativeness

Representativeness of language samples included in a corpus is one of the major concepts in corpus linguistics; the information about what population of language users and type of language the corpus represents is directly related to the ability to interpret findings from a corpus and to generalise from them to a context beyond the corpus (McEnery & Hardie 2011; Leech 2007). A corpus, no matter how large, is usually only a small sample of language used in the 'real world' and obtaining this sample involves the selection of data, hence it is vital to understand how the sample relates to the population from which it has been drawn, i.e. to the language used by speakers outside of the corpus. For example, in a corpus-based study of English produced by speakers from a Spanish L1 background, we should consider whether the findings can be generalised to all English speakers from this background or whether the generalisation needs to be narrowed in some way (e.g. is it likely to apply only to speakers with university education or of a certain age?) (Gilquin 2015).

In order to establish corpus representativeness and the validity of the conclusions based on it, two approaches are recommended. First, sufficient information about the language samples and collection methods should be provided to "allow the reader to determine the extent to which the results of [the] study are indeed generalizable to a new context" (Mackey & Gass 2015: 215). As further stressed by McEnery, Xiao and Tono (2006: 18), corpus creators must also "document corpus design criteria explicitly and make the documentation available to corpus users so that the latter may make appropriate claims on the basis of such corpora" (see also Gilquin 2015; Sinclair 2005). While it is relatively common to provide a description of the learners in the corpus, considerably less information seems to be available about how the data were collected and how reliability of data collection methods was ensured. In a (learner) corpus of speech, corpus background information should include all the steps involved in obtaining the data and converting it into a machine-readable corpus; the description should thus minimally provide details about any data collection methods involved and steps taken to ensure their reliability (e.g. the training available to the interviewers and any train-

ing or instructions received by the participants) as well as information about the transcription process.

The second approach to establishing the representativeness of data involves replication, that is, comparing the corpus data and findings to those from other linguistic settings and speakers, thus empirically evaluating the expectations of representativeness (e.g. Mackey & Gass 2015:176). As Leech noted (2007), although desirable and recognised as important by corpus linguists, this is still not a common approach in corpus research; the representativeness of corpora remains predominately based on assumptions about the language sample, rather than being established empirically (but see Gablasova, Brezina & McEnery 2017 for a study comparing several corpora of spoken British English). In order to determine what language is represented in the TLC, in this paper we follow the first approach by providing a description of the main characteristics of the corpus in terms of information about the speakers, linguistic setting and data collection methods (including transcription), see Sections 2.3 and 3. The second approach is to be pursued through a series of empirical studies investigating different aspects of language use in the TLC and employing comparison with relevant corpora.

2.2 Amount of evidence in a corpus: Corpus size, number of speakers and size of samples

The amount of evidence on which any piece of research is based is a crucial factor in evaluating the strength of its findings. The question of 'how much evidence is enough' or 'how much evidence do I need' to justify a claim is a common one in most quantitative disciplines. For much of quantitative social science research this question concerns the number of participants needed in the sample (e.g. Mackey & Gass 2015:176); in corpus-based research, the question is often answered in terms of the number of words in the corpus which are then taken as an indicator of the amount of evidence available in a corpus. However, there are also different ways in which the amount of corpus evidence can be quantified and evaluated with respect to its suitability for a research study (Sinclair 2005). In addition to the word count, the amount of corpus evidence can be considered from the perspective of the number of speakers or text samples, the size of each individual sample and the amount of evidence about the target linguistic feature(s).

First, we should consider the number of language users (speakers) who contributed samples to the corpus. This information is important in establishing the strength of any pattern identified in the corpus; if a relatively homogeneous pattern is observed across a number of independent language users, it is more likely to reflect a regularity in use than if a similar pattern is found in the same amount

of data from a smaller number of speakers. In the latter case, the pattern could be largely idiosyncratic (Brezina & Meyerhoff 2014).

Second, the size of each individual contribution and its effect on the nature of linguistic evidence (i.e. frequency of linguistics features) in the corpus should be taken into consideration. It is generally known (and documented by Zipf's law) that words and linguistic structures are not distributed equally in language and many of them can be relatively rare and dependent on the communicative setting. As Leech (2007: 141) put it, "with frequent grammatical characteristics, small text samples of 1,000 words are sufficient"; however, he notes that for linguistic structures such as collocations much larger samples may be required. As a result of these language patterns, the occurrence and range of some linguistic features may be limited and we should bear this in mind when interpreting the data, and especially when dealing with absence of evidence. In this case, when there is no occurrence of a particular linguistic feature and thus no evidence of its use, this could be equally due to the fact that the feature is not in the speaker's repertoire, or that the speaker chooses not to use it or did not get a chance to use it (Gablasova, Brezina & McEnery 2017). When sampling language we are not likely to capture the whole range of an individual's linguistic knowledge; as a result, there may not be a reasonable chance for rarer words or structures to appear without direct elicitation.

Third, we should consider the nature of the task (linguistic setting) and how controlled the language production in this linguistic context is. There are different degrees of restrictions applied when collecting language samples for a corpus; the more controlled the context, the more homogeneous and comparable the language samples are likely to be. The linguistic setting can be controlled for attributes such as the topic, the nature of a (speaking) task and the time allowed to produce language. These controls directly influence the type of linguistic evidence obtained and can encourage or supress specific linguistic patterns (e.g. specific vocabulary or grammatical structures). This will in turn have an effect on the nature of research questions that can be addressed with this linguistic evidence. For example, if the speech is elicited in response to a limited set of topics, this can likely result in the so-called topic bias when certain lexical items, related to the set topics, recur frequently in the corpus (Brezina 2018: 17). In such situations, caution is needed when interpreting findings related to lexical frequency or diversity. Another important factor related to the nature of the data from the tasks is the reliability in the data collection procedure; ideally, if we want to observe the effect of an independent variable (such as L1 background), the same data collection procedures should be followed so as not to introduce additional variables into the data and to ensure that the findings can be attributed to the effect of independent variables rather than idiosyncrasies in data collection methods (Gablasova, Brezina & McEnery 2017).

When planning the TLC design criteria, one of the main tasks was to reflect on the different types of linguistic evidence available with the goal of ensuring that the corpus offers sufficient evidence for a range of research questions (i.e. questions that would allow investigating L2 use with respect to a number of grammatical, lexical and pragmatic features in relation to different task and speaker-related variables). Some of the corpus properties were determined by the exam framework from which the speech samples came (such as the amount of evidence from individual speaker in terms of, for example, the length of individual contributions); others were decided by the corpus creators with a view to potential uses of the corpus. As the individual samples are relatively short (around six minutes of interaction per one task, with each L2 speaker taking part in two to four tasks), they would not provide an opportunity for each speaker to demonstrate the full repertoire of their linguistic knowledge. In designing the corpus, we therefore included a large number of speakers with the same characteristics (e.g. age, L2 proficiency) doing the same tasks in order to provide a sufficient amount of evidence about regular patterns in L2 speech elicited through these tasks and to increase the range of language features provided collectively.

Next, it may appear that, compared to other currently available learner corpora, e.g. the Vienna-Oxford International Corpus of English (VOICE), the TLC offers a narrower range of L1 backgrounds and includes a large number of speakers from one background at the expense of larger diversity in this area. However, the focus on fewer backgrounds enabled us to provide a sufficient amount of evidence for studying variables that so far have been underrepresented in learner corpus research, such as age. For example, the sub-corpus of the 349 Italian L2 speakers in the TLC includes sub-samples balanced for different age bands that provide a robust basis for studying the effect of age on language production: young (11–15), adolescent (16–19), young adult (20–30), adult (31–45) and mature (46–70) speakers of English.

In terms of the linguistic context in which the language was produced, the speaking tasks in the TLC represent a balanced combination of more and less controlled tasks. With respect to the structure of the interaction, each task is clearly defined by its speaker roles and communicative aim, with an explicit task description available to the exam candidates; however, the nature of interaction develops dynamically between the L1 and L2 speaker (see Section 3.3 for a detailed description of the tasks). As regards the topics, 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. In two of the tasks, therefore, the topics represent a very broad range, reflecting the candidate's interests and expertise, with no topic bias assumed. In the remaining two tasks, the range of topics used by the examiners is relatively broad, especially as the topic

lists are changed yearly and differ for each proficiency level. Moreover, while the lists contain general topics such as "education", the examiners personalise and narrow the topics further as illustrated by the following conversation openers used for the topic of education: "What do you feel are the advantages of a good education?" and "If you could change something about the Spanish school system would you change anything?". However, the possibility of topic bias should be taken into consideration when planning a study based on these particular tasks.

Across all tasks, the reliability in data collection can be considered to be high. All tasks were conducted by examiners who are rigorously trained in administering the GESE exam with their performance regularly standardised and monitored. In addition, each year, a certain proportion of exams is recorded and examiners receive feedback. It should be noted that some examiners took part in more examinations, thus potentially increasing the presence of their communicative styles in the corpus (the individual examiners have been tagged and their effect can thus be controlled for and/or investigated in the corpus).

Finally, bearing in mind the considerations above, let us reflect on the balance of the material in the TLC. Balance relates to the overall property of a corpus to include a "range of text categories" (McEnery et al. 2006:16). The TLC was designed specifically to include data across different speaker (e.g. gender, age, L1) and situational (e.g. tasks, speaker performance) categories. At the same time, the corpus reflects the population of test takers of the GESE exam with certain categories being more represented than others (see Appendix A). With the shift from working with aggregate data sets (Brezina 2018: 21–22) to a stronger focus on subcorpora and individual texts and speakers (e.g. Brezina & Meyerhoff 2014), we would argue that the overall balance is secondary to the amount of evidence, the structure of the corpus and availability of metadata (McEnery et al. 2006), especially since the balance of a dataset is often defined individually in different studies by selecting an appropriate subset of the corpus that can answer a particular research question.

2.3 Transcription

Often relatively little information is available on the methodological and theoretical considerations involved in the transcription of spoken corpus data beyond plain transcription guidelines (however, see e.g. Breiteneder et al. 2006 for an example of a detailed discussion of spoken data transcription). Transcription quality has direct implications for the reliability and usability of the corpus as a research resource and should thus be given more attention in the field. As Cameron (2001: 43) points out, "transcribing is not just a tedious, mechanical process that has to be got out of the way before the more 'interesting' part –

analysing and interpreting – can begin […] transcribing is effectively the first stage of analysis and interpretation". In this section, we discuss several issues in the transcription of spoken learner language and how they were addressed in a systematic manner in the development of the TLC (a more detailed account of the methodological decisions in the transcription of the TLC can be found in Gablasova & Brezina in prep). In particular, the focus will be on two issues in the TLC transcription process – (i) the type of transcription and the role of transcription conventions (Section 2.3.1) and (ii) the consistency and reliability of transcription (Section 2.3.2).

2.3.1 *The type of transcription and the role of transcription conventions*

The production of corpora of spoken language involves a number of important decisions and a considerable amount of skill (Cook 1995; Thompson 2005; Adolphs & Knight 2010; Gilquin et al. 2010). The main decisions relate to the amount of information to be captured about the spoken language as reflected in more or less detailed transcription (e.g. du Bois 1991; Cameron 2001; Breiteneder et al. 2006). The transcription also reflects our understanding of the way language functions and what features, therefore, are worth capturing and analysing (Ochs 1979). The type of transcription used in corpus creation directly affects (and is often guided by) the research questions that can be answered using the corpus (Gilquin 2015; Adolphs & Knight 2010; Thompson 2005).

With respect to the type of transcription, having considered existing transcription systems and their principles (e.g. O'Connell & Kowal 1994; Crowdy 1994; Arche 2008; Gilquin et al. 2010; MacWhinney 2000), the TLC opted for simple orthographic transcription, with attention to several non-linguistic and paralinguistic features such as hesitations, filled and unfilled pauses, laughs etc. (see Figure 1). Following Crowdy (1994), standard written forms of words were used with a set of non-standard forms (e.g. contracted forms), all of which are listed in the transcription conventions (see Appendix C). This approach was adopted to maximise the amount of data while providing sufficient empirical evidence for research on a variety of lexical, grammatical and discourse-related features in spoken language. The transcription system follows the "principle of adaptability" recommended by du Bois (1991: 94–95) which allows additional layers of transcription to be added systematically later if the dataset is to be adopted for new research questions (see also Adolphs & Knight 2010). The transcription process was guided by a set of explicit guidelines which can be found in full in Appendix C and are further discussed in Gablasova & Brezina (in prep.). Due to ethical considerations (principally the need for anonymisation), the sound files on which the transcripts are based are currently not available as part of the corpus; we are, however, exploring the possibility of the automatic anonymization of these files.

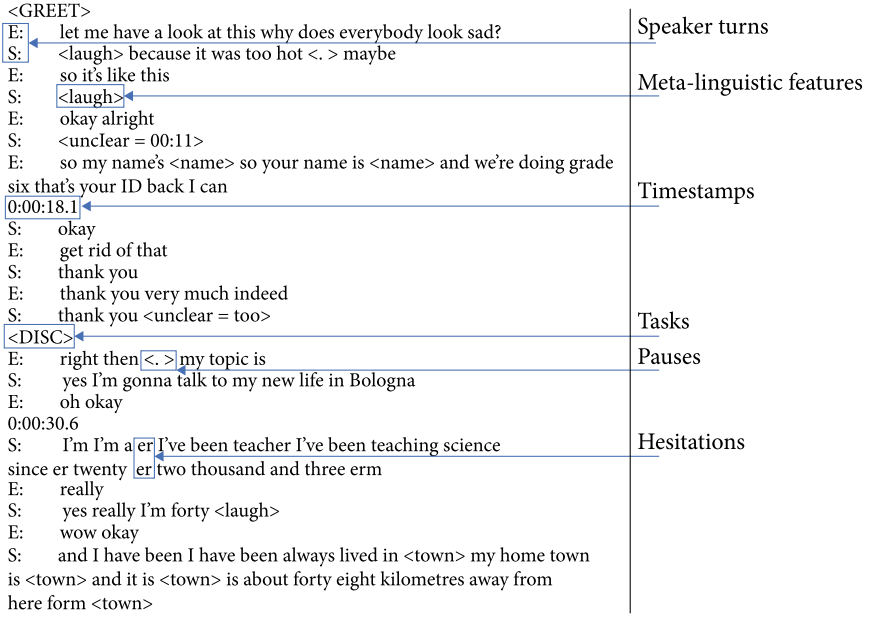

Figure 1. An example of a transcript

To demonstrate the role of the transcription conventions in the reliability of transcription, we will use an example of how filled pauses/backchannels were dealt with in a systematic manner in the TLC. Filled pauses are a very frequent marker of spoken language (e.g. Wong & Kruger 2018); at the same time, they can include a high number of expressions which may be difficult to distinguish from each other reliably (Thompson 2005). Many corpora allow transcribers to capture these expressions as they hear them (e.g. BNC1994, Aston & Burnard 1998). However, a transcription of the same passages in the TLC by two trained transcribers showed that this resulted in very variable outcomes (e.g. the same expression was transcribed as *mhm, hm, mm* and *mhmm*). As a result, we decided to use a closed set of back-channelling expressions that would allow corpus users to distinguish between major categories of these expressions (e.g. *erm* and *hm*). The standardisation practice (see Section 2.3.2) proved that this considerably increased the consistency of the transcription of these spoken features. Other systematic decisions had to be reached regarding features of (spoken) learner language such as the transcription of non-standard words (e.g. *discoverments*) or features related to learner L1 backgrounds (e.g. adding an additional syllable at the end of words by many L1 Italian speakers of English such as in the pronunciation of the word *laptop* as /lʌptʊpɛ/); these decisions are recorded in the transcription guidelines in Appendix C.

2.3.2 *Consistency and reliability of transcription*

The transcription of speech always involves an interpretation of what is heard (e.g. Cameron 2001). In order to monitor and maximise reliability of the transcription, as recommended by Thompson (2005), several practices were used in the transcription of the TLC, targeting different stages of the transcription process. These steps are briefly described below.

Before the transcription began, a thorough training programme was put in place for two transcribers. The transcribers were given the transcription guidelines, which were explained to them. They were then closely monitored throughout the first two months of transcribing. The objective of the first transcriptions during this period was for the transcribers to reach an in-depth understanding of the guidelines and achieve consistency in applying them. The transcribers therefore transcribed the same recordings and compared the resulting transcripts, identifying any discrepancies and resolving them according to the guidelines. Members of the project team listened to the recordings while reading the transcripts, noting any places of interest (e.g. where guidelines were not applied appropriately). This approach proved to be very fruitful in identifying unclear or ambiguous parts in the guidelines which could then be resolved. Only when the transcripts produced by the two transcribers were fully consistent and the researchers identified no issues did the individual transcriptions begin.

The process of monitoring transcription reliability (i.e. the consistency of transcribing and adherence to the guidelines) continued during the transcription using the following measures. First, ten percent of the transcripts were double-coded, i.e. the transcribers exchanged randomly selected transcripts and listened to the recording while reading the transcript, noting any differences in perceived spoken production. The transcribers then met to discuss each transcript, listening to the instances where a discrepancy was identified and resolving it if possible. All decisions arising from this process were recorded and the outcomes analysed by members of the project team. The rate of disagreement between the transcribers was found to be three percent, a very small minority of which was related to the meaning of the utterance (e.g. *do a lot* instead of *develop*; omission of *the*) with the vast majority related to the transcription of contractions (e.g. *they've* vs *they have*) or the spelling of a particular word (e.g. *saree* instead of *sari*). In the phase of the project when only one transcriber continued to work on the transcription, the monitoring process was modified so that the transcriber, after a substantial period of time (e.g. five months), relistened to ten percent of the recordings she had transcribed and noted any differences. To further monitor the quality of transcription, after two years of transcribing, ten per cent of transcripts from L2 speakers from Spanish and Chinese L1 backgrounds were selected to be double-coded by native speakers of the language with a background in linguistics, trained in the transcription guidelines.

The exercise showed that the trained native speakers were able to improve the transcripts further; the improvements were largely related to words marked as 'unclear' in the original transcripts rather than to correcting transcription errors.

A second practice established to ensure consistency in the transcription concerned keeping transcription logs (Adolphs & Knight 2010). Throughout the transcription, the transcribers kept a "decisions log", a shared folder in which all decisions related to the transcription were noted. For example, a note on the transcription of *Blu-ray* was made with respect to hyphenation. The log allowed checking whether each decision was applied consistently throughout the whole corpus. In addition to the shared decisions log, each transcriber kept an individual transcription log in which they noted information about each recording (e.g. the length of the recording) and transcript (e.g. the number of words in the transcript) as well as any issues with the transcription. These logs were further used to discuss any need for changes to the transcription guidelines (e.g. one transcript contained an L2 speaker singing for a few seconds to demonstrate a musical concept; following this a tag <sing> was introduced). These two logs allowed the researchers to regularly monitor the transcription process, notice any difficulties that emerged and address them with the transcribers.

The aim of this section was to raise awareness of the role of transcription in the final product (corpus) and to highlight the types of decisions related to day-to-day transcribing. While it may not be feasible to capture the full complexity of spoken (learner) language in a corpus transcript, it is important to create a procedure that is transparent and maximises the reliability (and replicability) of the transcription process.

3. TLC: Structure and variables

This section offers a more detailed description of the TLC in terms of major speaker-related and language-related variables that play a role in the representativeness and generalisability of the corpus as discussed above (Sections 2.1 and 2.2). The speaker-related variables discussed in this section (and listed in Appendix A) have been included in the corpus metadata and are searchable in the corpus. The corpus can also be searched according to each of the four speaking tasks described in Section 3.2. Appendix B offers an overview of the structure of the corpus according to the main variables and provides the information about the size of the relevant subcorpora.

3.1 Speaker-related characteristics

3.1.1 *Proficiency in English*

Despite L2 proficiency being a key variable in L2 studies, there are several issues with reliably establishing L2 proficiency levels of language users in a learner corpus which make the interpretation of research findings and comparison of groups of learners difficult (e.g. Gablasova, Brezina & McEnery 2017; Callies 2015). A major concern is that, as Carlsen (2012:162) argues, "the levels of proficiency are not always carefully defined, and the claims about proficiency levels are seldom supported by empirical evidence" which may lead to incorrect conclusions about learner language development and use. A common way of establishing proficiency in learner corpora has been through external criteria such as the educational level of L2 users or years spent studying the L2. However, these proxy measures can rarely ensure a reliable estimate of L2 proficiency (Callies 2015), especially if L2 users come from different educational backgrounds and from countries with different levels of exposure to English.

In order to respond to calls for greater reliability and transparency in establishing L2 proficiency in learner corpora, special care was taken during the development of the TLC to provide information about this variable. The TLC bases the information about L2 speakers' proficiency on their performance in the speaking tasks. This provides a direct rating (assessment) of the performance included in the corpus, rather than a proxy measure of the speaker's proficiency established on the basis of, for example, overall school performance or performance on a different test (e.g. the TOEFL). The rating is awarded by trained raters (the examiners), who take part in regular standardisation sessions. The GESE exam operates on a twelve-grade system, with the grades ranging from emerging knowledge to advanced mastery of English. To ensure that candidates are examined at an appropriate level for their linguistic knowledge and skills (to avoid them failing the exam or exceeding the requirements of the grade), guidance is offered about what is required at each grade and level of proficiency (see Trinity College London 2016 to read more about the exam). In order to help with the interpretation of the proficiency information and making it comparable outside of the corpus, the marks given in the GESE exam have been formally validated to the bands of the Common European Framework of Reference for Languages (CEFR) (Papageorgiou, 2007). The TLC includes L2 speakers from the B1 to the C2 levels of the CEFR.

Using the proficiency ratings, L2 speakers representing different levels of English language proficiency were selected based on a combination of their overall performance in the exam as well as in individual speaking tasks. In addition to the CEFR level, each performance in the exam is given an overall rating ("Distinction" referring to a high, "Merit" to a middle-level and "Pass" to a low performance

which still, however, meets the criteria of the CEFR band). Each speaking task is awarded a separate mark: *A, B, C* or *D* (*A* referring to a high performance and *D* to a fail). The vast majority of language samples selected for the corpus came from L2 speakers in the middle range of the grade/proficiency band as determined by the marks awarded for the tasks.

Reflecting the population of GESE test-takers, certain levels of English proficiency (especially at the most advanced levels) contain a relatively small number of L2 speakers from some L1 backgrounds (see Appendix B) and this fact may have implications for the research design of studies based on the TLC. Researchers wishing to study L2 speakers across a range of proficiency levels can, for example, focus on the B1 to C1 range of the data or they can opt to combine data from certain categories, e.g. to combine C1 and C2 level speakers in an 'Advanced' category, or to combine speakers from different L1 backgrounds at each proficiency level.

3.1.2 *Linguistic and cultural background*

The TLC contains L2 speakers from a number of linguistic backgrounds: the main L1 backgrounds are Chinese, Italian, Russian and Spanish with another large group of speakers representing some of the major languages spoken in India, e.g. Hindi, Gujarati and Marathi (see Appendix A for a full list of L1 backgrounds). L1 background can be a relatively complex concept in some local contexts. For example, the L1 of most speaker groups in the corpus can be easily determined and often matches the community language (e.g. Spanish in Mexico). However, in contexts such as Catalonia or India the linguistic situation is more complex and many L2 speakers identify two or more languages as their L1s (e.g. Hindi, English and Punjabi were commonly identified by speakers from New Delhi as the languages they speak at home). This information is included in the corpus metadata and can be used by researchers to control the characteristics of linguistic backgrounds of the speakers in the corpus.

In addition to linguistic background, the potential influence of cultural background on L2 use should be also recognised (e.g. Roever 2010). While there are different ways to define culture and cultural background (e.g. through social, occupational or ethnic affiliation), the definition through national affiliation, which is also available in the TLC, has been used for the purposes of linguistic research before (Spencer-Oatey 2008). Cultural background could be of particular interest in studies on L2 communicative strategies and could further contribute to the understanding of cross-linguistic effects in language learning. The TLC composition offers a great opportunity to gain insights into the interaction of linguistic and cultural background; for example, it contains Spanish L1 speakers from different countries, e.g. Spain, Mexico and Argentina, allowing the investigation of L1 background effects other than language typology on L2 use.

3.1.3 *Sociolinguistic characteristics: Age, gender and education*

The influence of sociolinguistic characteristics, with the exception of gender, has not yet been given much attention in learner corpus research to date (Gablasova, Brezina & McEnery 2017). While information about variables such as age and education is available in a number of learner corpora (e.g. the LINDSEI), the sample in these corpora is often relatively homogenous, not lending itself to a fuller examination of the effect of these variables. To contribute to a more systematic exploration of sociolinguistic influences in L2 use, the TLC contains L2 speakers across different age bands, ranging from schoolchildren to people of retirement age, with a diversity of educational backgrounds; for the examiners, the information about their age, gender and examining experience is available. For example, information about the age of speakers at the time of recording can be used to compare young L2 speakers (11–13 year olds) with more mature language users (speakers in their 20s or 40s) and to identify linguistic patterns that may be affected by cognitive and linguistic maturity such as lexical choice, grammatical complexity and pragmatic ability.

3.1.4 *Learning experience: Age of exposure, learning history and patterns of L2 use*

The different dimensions of the language learning experience play a significant role in learning outcomes, allowing a better understanding of variation in L2 acquisition and use (Mackey & Gass 2015). While these variables represent a major research area in SLA research, they have not been routinely used in corpus-based studies (for an example see Fuchs, Götz & Werner 2016). The TLC contains information about different aspects of the L2 learning experience such as the age of first exposure to English and the type of exposure (e.g. through schooling or time abroad). For a full list of variables see Appendix A.

3.2 The nature of interaction: Linguistic setting and the speaking tasks

The linguistic setting in which interaction occurs, defined by factors such as the aim of communication, speaker roles and the register, is one of crucial determiners of the nature of the language represented in a corpus (Biber & Conrad 2009). Factors related to the nature of interaction such as topic familiarity and planning time also affect linguistic variables of interest to L2 researchers such as complexity and fluency of production (e.g. Alexopoulou et al. 2017; Tracy-Ventura & Myles 2015). The information about the nature of interaction determines whether the corpus can be meaningfully compared with other corpora, but also what type of conclusions and generalisations can be reached about the language in the corpus.

3.2.1 *The linguistic setting*

The TLC represents interaction in an institutional setting with the speaker roles to some extent predefined by the framework of the examination. Overall, the language contained in this setting is semi-formal in nature and is close to academic interaction. In this regard, research on the TLC can be used to complement findings from corpora that represent more informal interaction such as that contained in the LINDSEI and the VOICE corpora. However, although the language provided in the corpus can be situated in a particular genre (that of institutional interaction), each of the different speaking tasks represents a specific linguistic setting and different conditions of language use (e.g. Gablasova, Brezina, McEnery & Boyd 2017; Gablasova & Brezina 2015; Wall & Taylor 2014). The four speaking tasks are described below (for more details and for guidelines for the exam candidates see Trinity College London 2016). In all of the tasks, one L2 speaker interacts with one L1 speaker of English.

3.2.2 *Speaking tasks*

There are four different speaking tasks represented in the TLC. Each L2 speaker participated in at least two tasks with the number of tasks increasing with the grade of examination (proficiency level) as shown in Table 1. The tasks consist of one largely monologic (presentation) and three dialogic and highly interactive communicative settings (discussion, conversation and the interactive task). For the number of tokens produced in each task per proficiency band see Appendix B, Tables 2 to 4. There is a short warm-up chat between the candidate and the examiner before they proceed to the first task; no additional preparation time is given before any of the tasks during the exam.

Table 1. Speaking tasks[1] according to proficiency band

	CEFR proficiency band		
Speaking task	**B1: Threshold**	**B2: Intermediate**	**C1 & C2: Advanced**
Presentation			✓
Interactive task		✓	✓
Discussion	✓	✓	✓
Conversation	✓	✓	✓

The presentation task is based on a topic freely chosen and prepared by the L2 speakers in advance of the exam, allowing them to select a topic they are familiar with and interested in. The task represents formal to semi-formal monologic

1. The labels of the tasks are directly adapted from the GESE exam.

speech, with the examiner providing back-channelling throughout the task. The topics vary widely across social and political issues and both historical and current affairs. Example (1) below shows an excerpt from a transcript of a presentation (Speaker 1, S1 – examiner; Speaker 2, S2 – exam candidate).

(1) S2: erm my topic for presentation today is the social conditioning I will first of all speak a bit about definition of what social conditioning means and secondly I'm going to speak about the pros and the cons of social conditioning and what I believe erm are the risks and benefits
S1: mm
S2: and finally I will conclude with my opinion and observation of how we can tackle social conditioning and how <unclear> relate erm social conditioning is natural behaviour or phenomenon which has been handed down through centuries of tradition and erm certain progressive changes through generations er now social conditioning infor= er reflects in human behaviour it's said that a child up to the age of three er i-in that child right brain is dominant so the child has natural adjustment to the world around him or her but after the age of three the left brain er takes […]

The presentation is followed by the discussion task, in which the examiner begins by asking about ideas and information expressed in the presentation and a conversation develops from there. In the lower grades (B1 and B2), which do not include the presentation, the candidates introduce their topic briefly and the discussion proceeds from there. Example (2) offers an excerpt from the discussion task.

(2) S1: okay well thank you very much that was interesting yes interesting I'm a bit surprised that you chose Roosevelt as an example of a planned economy <.> erm because I associate planned economies with er Russia and China in the period
S2: yeah
S1: nineteen forty nine to nineteen seventy five you know
S2: yeah I I also think so however I think that what China's done with planned economy is not really good it is not good enough to be a positive example
S1: mm
S2: and Roosevelt he used planned economy he made a great success with planned economy
S1: mm
S2: and China is not as great but it's er unique and special as well
S1: sure but er Roosevelt never nationalised any industries did he he didn't take over Ford and General Electric and er […]

As illustrated by the transcript, the task is highly interactive with both speakers actively contributing to the conversation and with the L2 speaker usually being in the position of an expert on the topic.

The third task included in the corpus is the interactive task. This task starts with a prompt from the examiner, often presented as an issue of personal relevance or interest. Examples (3) and (4) illustrate typical prompts used in this task.

(3) "I haven't heard from a friend of mine who lives in Australia and I'm beginning to get a bit worried." (Grade 7, B2 level)

(4) "Some people think it's really important to like the people you work with. I'm not sure how necessary this is." (Grade 11, C1 level)

This task requires the L2 speaker to show initiative in maintaining the conversation by asking questions, commenting on what was said and by providing opinions and advice. An extract from the interactive task is provided in Example (5).

(5) S2: but do you think that technology is good for the society for for us?
S1: erm well I think it's a tool I'm not sure that it's good for us but it's like any tool it's the people isn't it?
S2: right I I think so er I find the technology is really useful for me er I think that that's my thinking because of my work
S1: mm
S2: er I teach and I like to teach with technology
S1: mm mm
S2: er do you use your t= the technology in your work?
S1: in my teaching yes absolutely all the time
S2: what kind of er techniques do you apply with technology?
S1: erm well I I I use I use it in in teaching all aspects of English […]

As can be seen from the task transcript in which the speakers talk about the value of technology, the L2 speaker is 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 (e.g. discourse markers such as *I think*).

The last task is the conversation task in which the L2 speaker is invited to engage in a conversation about two topics of general interest. The topics for candidates at the B1, B2 and C1 levels are selected from a list of topics (e.g. "Society and living standards", "Personal values and ideal" and "National environmental concerns") known to them before the exam; candidates at the C2 level can be given any topic considered suitable by the examiner. Two lists of topics are available for levels B1-C1. One is more appropriate for younger and the other one for more mature L2 speakers. The examiner introduces the topic by asking a more specific

question, such as "Let us talk about some issues in education. Do you believe education is important for young people nowadays?". An extract from the conversation task can be seen in Example (6).

(6) S2: although the consumers are <.> upset sometimes with so many adverts on tv
S1: uhu okay
S2: er what do you think about it?
S1: erm I find advertising annoying <.> I find it annoying
S2: d= I not d= I'm not for for the same opinion sometimes the advertising are t= good are funny or entertaining
S1: mm
S2: they they have a a history behind the so <unclear>
S1: yeah but don't you think it's it's overkill the the amount of advertising that they have
S2: yeah I
S2: mm for example on er if if you're trying to watch a tel= if you're trying to watch a film on tv <.> if you're trying to watch a film on tv
S1: you waste a lot of time because of the adverts […]

As demonstrated by the conversation above, dealing with issues of advertising, both speakers co-manage the flow of discourse and contribute their opinions as well as seek the opinion of the other interlocutor. The meaning is co-developed through their contributions, which show a range of discursive patterns such backchannelling ("uhu okay"), interruptions (S2: "yeah I" interrupted by "mm for example…") as well as a co-constructed sentence (S2: "if you are trying to watch a film on tv" completed by S1: "you waste a lot of time because of the adverts").

In terms of communicative contexts, the four tasks represent different situations, both monologic and dialogic, which allow contrasting language use by the same L2 speakers (see Table 2 for a summary of the discourse characteristics of the tasks). As illustrated by the examples, the dialogic tasks contain a range of features typical of spoken interaction and discourse management such as interruption, use of floor-holding devices, joint management of the conversation (e.g. switching between asking and responding) as well as completing the other speaker's sentence (e.g. the last adjacency pair in Example (6)). With regard to cross-task comparison, the greatest range is available for the most advanced speakers (C1 & C2 levels); in terms of the progress across proficiency levels, two tasks can be studied at each level (discussion and conversation).

When planning a comparison of tasks across proficiency levels, it should be noted that the same tasks at different levels of proficiency (e.g. discussion at B1 and C1 levels) seek to elicit both general features related to spoken communication

Table 2. Speaking tasks: Discourse characteristics (adapted from Gablasova et al. 2017)

Task	Topic familiarity	Interlocutor roles	Type of interaction
Presentation	pre-selected topic	candidate-led	monologic
Discussion	pre-selected topic	jointly-led	dialogic
Interactive task	general topic	candidate-led	dialogic
Conversation	general topic	jointly-led	dialogic

that are fully comparable across proficiency levels as well as a set of language functions defined in task specifications for each level (e.g. past progressive is highlighted as a grammatical language function at the B1 level). Researchers are thus advised to refer to the task specification for different levels of the GESE exam (Trinity College London 2016) in order to ensure that they avoid circularity in their analysis of the data (e.g. focusing on a linguistic feature that may be highlighted at a particular level of the GESE exam and interpreting it as characteristic of that proficiency level).

In terms of communicative settings, the tasks allow for contrasting variables such as topic familiarity and interlocutor roles and observing their effect on language production. Also, the multiple tasks offer an opportunity to observe the same speakers in different speaker roles (e.g. as an expert in the discussion task and a knowledge-seeker in the interactive task) (Gablasova & Brezina 2015). These different settings enable us to gain deeper insight into speakers' language mastery and their ability to adjust language choice according to the requirements of a specific task – a dimension of L2 proficiency which is still largely understudied in SLA and learner corpus research (Gablasova, Brezina, McEnery & Boyd 2017). The large number of L2 speakers in the TLC offers rich data for investigating systematic variation with respect to these contextual variables.

4. Conclusion: Applications and future directions

4.1 Applications in research and teaching

As stated in the introduction, the TLC was planned with a broad range of research applications in mind. It is evident from the research conducted on the corpus to date that the potential of the TLC is indeed large. The corpus has already been successfully used to investigate a broad range of lexical, grammatical and pragmatic features. For example, the studies have so far focused on light verb constructions (Gilquin this issue), verb-argument constructions (Römer & Garner this issue), phrasal verbs (Marin Cervantes & Gablasova 2017), lexical back-channelling

(Castello & Gesuato this issue), filled pauses (Götz this issue), disagreements (Gablasova & Brezina 2017), epistemic stance (Gablasova et al. 2017) and the effect of speaker roles (Gablasova & Brezina 2015). A range of speaker-related variables (e.g. L2 proficiency and L1 background) as well as variables related to the linguistic setting (e.g. speaker role and task effects) have been explored in these studies.

In addition to the research-oriented applications, the corpus has been used for pedagogical purposes and the development of teaching materials (Gablasova, Brezina & McEnery 2019; Gablasova & Brezina 2017). The materials have so far addressed topics in pragmatics, communicative strategies, spoken language features and collocations demonstrating the value of using learner corpora in language teaching.

When considering the use of the corpus for specific studies, researchers should carefully evaluate the suitability of the TLC for their purposes in terms of the amount and nature of evidence available (discussed in Sections 2 and 3). It is also recommended that researchers consult the information about the GESE exam (Trinity College London 2016) and view the videos available of the exams to gain further insight about the context in which the data were collected.

4.2 Future directions: The Trinity Lancaster Corpus of L1 English interaction (TLC-L1)

A contrastive research design, in which L1 and L2 speakers are compared, is used frequently in corpus-based studies of L2 use (Granger 2015). Although this is a very productive approach, it has proved considerably difficult to find corpora of L2 and L1 use that match each other in terms of language that they represent and can thus be meaningfully compared (Gablasova et al. 2017; Callies 2015; Leech 2007, 1998). In order to provide a suitable L1 corpus that would permit such a comparison with the TLC, the Trinity Lancaster Corpus of L1 English interaction (TLC-L1) is being developed in the context of the ongoing collaboration between Lancaster University and Trinity College London. So far, the corpus contains 145 recordings and over 450,000 words from L1 speakers of British English and is planned, when completed, to consist of 250 recordings, amounting to approximately one million words. In order to ensure comparability in terms of linguistic setting and speaking tasks, the data collection follows the GESE interview guidelines for Grade 12 (as it includes all four speaking tasks) used for the TLC. This approach offers high reliability with respect to the data collection procedure, as all interviews are conducted by trained examiners from Trinity College London following the same principles as in the L2 interviews.

Another consideration related to the comparability and representativeness of the TLC-L1 concerns the diversity of the L1 speaker group. Particular effort has been taken to collect a balanced sample of L1 British English speakers in terms of socio-economic background, age and professional experience for the TLC-L1 thus reflecting the diversity of the data in the TLC. This makes the TLC-L1 a very valuable resource for studying variation in L1 English in its own right. Together, the two corpora will thus offer datasets well suited to a sophisticated, multi-variable analyses of L1 and L2 use which can take into account the effect of a range of linguistic and sociolinguistic factors, making them a major resource in the study of SLA (e.g. Plonsky 2016; Gablasova et al. 2017).

Acknowledgements

We would like to thank three anonymous reviewers and the general editors of the journal for their valuable comments on the manuscript. The work on the corpus was supported by the ESRC grants no. EP/P001559/1, ES/K002155/1 and ES/R008906/1.

References

Adolphs, S. & Knight, D. 2010. "Building a spoken corpus". In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge Handbook of Corpus Linguistics*. London: Routledge, 38–52. https://doi.org/10.4324/9780203856949.ch4

Aijmer, K. 2014. "Pragmatic markers". In K. Aijmer & C. Rühlemann (Eds.), *Corpus Pragmatics: A Handbook*. Cambridge: Cambridge University Press, 195–218.

Alexopoulou, T., Michel, M., Murakami, A., & Meurers, D. 2017. "Task effects on linguistic complexity and accuracy: A large-scale learner corpus analysis employing natural language processing techniques". *Language Learning* 67(S1), 180–208. https://doi.org/10.1111/lang.12232

Arche, M. J. 2008. SPLLOC Transcription Conventions. <http://www.splloc.soton.ac.uk/trancon.html> (accessed August 2019).

Aston, G., & Burnard, L. 1998. *The BNC Handbook: Exploring the British National Corpus with SARA*. Capstone.

Baker, P. & Egbert, J. (Eds.). 2016. *Triangulating Methodological Approaches in Corpus Linguistic Research*. London: Routledge. https://doi.org/10.4324/9781315724812

Biber, D. & Conrad, S. 2009. *Register, Genre, and Style*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511814358

Breiteneder, A., Pitzl, M. L., Majewski, S. & Klimpfinger, T. 2006. "VOICE recording-Methodological challenges in the compilation of a corpus of spoken ELF". *Nordic Journal of English Studies* 5(2), 161–187.

Brezina, V. & Meyerhoff, M. 2014. "Significant or random. A critical review of sociolinguistic generalisations based on large corpora". *International Journal of Corpus Linguistics* 19(1), 1–28. https://doi.org/10.1075/ijcl.19.1.01bre

Brezina, V. 2018. *Statistics in Corpus Linguistics. A practical guide*. Cambridge: Cambridge University Press.

Callies, M. 2015. "Using learner corpora in language testing and assessment: Current practice and future challenges". In E. Castello, K. Ackerley & F. Coccetta (Eds.), *Studies in Learner Corpus Linguistics: Research and Applications for Foreign Language Teaching and Assessment*. Frankfurt: Peter Lang, 21–35.

Cameron, D. 2001. *Working with Spoken Discourse*. London: Sage.

Carlsen, C. 2012. "Proficiency level – A fuzzy variable in computer learner corpora". *Applied Linguistics* 33(2), 161–183. https://doi.org/10.1093/applin/amr047

Cervantes, I. M. & Gablasova, D. 2017. "Phrasal verbs in spoken L2 English: The effect of L2 proficiency and L1 background". In V. Brezina & L. Flowerdew (Eds.), *Learner Corpus Research: New perspectives and applications*. London: Bloomsbury, 28–46.

O'Connell, D. C. & Kowal, S. 1994. "Some current transcription systems for spoken discourse: A critical analysis". *Pragmatics* 4(1), 81–107. https://doi.org/10.1075/prag.4.1.04con

Cook, G. 1995. "Theoretical issues: transcribing the untranscribable". In G. Leech, G. Myers & J. Thomas (Eds.), *Spoken English on Computer: Transcription, Mark-up and Application*. Harlow: Longman, 35–53.

Crowdy, S. 1994. "Spoken corpus transcription". *Literary and Linguistic Computing* 9(1), 25–28. https://doi.org/10.1093/llc/9.1.25

Dayrell, C. & Urry, J. 2015. "Mediating climate politics: The surprising case of Brazil". *European Journal of Social Theory* 18(3), 257–273. https://doi.org/10.1177/1368431015579962

Du Bois, J. W. 1991. "Transcription design principles for spoken discourse research". *Pragmatics* 1(1), 71–106. https://doi.org/10.1075/prag.1.1.04boi

Ellis, N. C. 2002. "Frequency effects in language processing. A review with implications for theories of implicit and explicit language acquisition". *Studies in Second Language Acquisition* 24(2), 143–188.

Fuchs, R., Götz, S. & Werner, V. 2016. "The present perfect in learner Englishes: A corpus-based case study on L1 German intermediate and advanced speech and writing". In V. Werner, E. Seoane & C. Suárez-Gómez (Eds.), *Re-Assessing the Present Perfect*. Berlin: Mouton de Gruyter, 297–338. https://doi.org/10.1515/9783110443530-013

Gablasova, D. & Brezina, V. In preparation. Challenges in transcribing spoken learner language: Lessons from the Trinity Lancaster Corpus.

Gablasova, D., Brezina, V. & McEnery, T. 2019. "The Trinity Lancaster Corpus: Applications in language teaching and materials development". In S. Götz & J. Mukherjee (Eds.), *Learner Corpora and Language Teaching*. Amsterdam: John Benjamins, 8–28.

Gablasova, D., Brezina, V. & McEnery, T. 2017. "Exploring learner language through corpora: Comparing and interpreting corpus frequency information". *Language Learning* 67(S1), 130–154. https://doi.org/10.1111/lang.12226

Gablasova, D. & Brezina, V. 2017. "Disagreement in L2 spoken English: From learner corpus research to corpus-based teaching materials". In V. Brezina & L. Flowerdew (Eds.), *Learner Corpus Research: New perspectives and applications*. London: Bloomsbury, 69–89.

Gablasova, D., Brezina, V., McEnery, T. & Boyd, E. 2017. "Epistemic stance in spoken L2 English: The effect of task type and speaker style". *Applied Linguistics* 38(5), 613–637. https://doi.org/10.1093/applin/amv055

Gablasova, D. & Brezina, V. 2015. "Does speaker role affect the choice of epistemic adverbials in L2 speech? Evidence from the Trinity Lancaster Corpus". In J. Romero-Trillo (Ed.), *Yearbook of Corpus Linguistics and Pragmatics 2015*. Dordrecht: Springer, 117–136. https://doi.org/10.1007/978-3-319-17948-3_6

Gilquin, G., De Cock, S. & Granger, S. 2010. *The Louvain International Database of Spoken English Interlanguage*. Handbook and CD-ROM. Louvain-la-Neuve: Presses universitaires de Louvain.

Gilquin, G. 2015. "From design to collection of learner corpora". In S. Granger, G. Gilquin & F. Meunier (Eds.), *Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, 9–34. https://doi.org/10.1017/CBO9781139649414.002

Granger, S. 2015. "Contrastive interlanguage analysis: A reappraisal". *International Journal of Learner Corpus Research* 1(1), 7–24. https://doi.org/10.1075/ijlcr.1.1.01gra

Gries, S. Th. 2015. "Some current quantitative problems in corpus linguistics and a sketch of some solutions". *Language and Linguistics* 16(1), 93–117. https://doi.org/10.1177/1606822X14556606

Jucker, A. H., Smith, S. W. & Lüdge, T. 2003. "Interactive aspects of vagueness in conversation". *Journal of Pragmatics* 35(12), 1737–1769. https://doi.org/10.1016/S0378-2166(02)00188-1

Kormos, J. 2014. *Speech production and second language acquisition*. London: Routledge. https://doi.org/10.4324/9780203763964

Leech, G. 1998. "Preface. Learner corpora: what they are and what can be done with them". In S. Granger (Ed.), *Learner English on Computer*. London: Longman, xiv–xx.

Leech, G. 2000. "Grammars of spoken English: New outcomes of corpus-oriented research". *Language Learning* 50(4), 675–724. https://doi.org/10.1111/0023-8333.00143

Leech, G. 2007. "New resources, or just better old ones? The Holy Grail of representativeness". In M. Hundt, N. Nesselhauf & C. Biewer (Eds.), *Corpus Linguistics and the Web*. Amsterdam: Rodopi, 134–149. https://doi.org/10.1163/9789401203791_009

Love, R., Dembry, C., Hardie, A., Brezina, V. & McEnery, T. 2017. "The spoken BNC2014". *International Journal of Corpus Linguistics* 22(3), 319–344. https://doi.org/10.1075/ijcl.22.3.02lov

Mackey, A., & Gass, S. M. 2005. *Second Language Research: Methodology and Design*. New York NY: Routledge.

MacWhinney, B. 2000. *The CHILDES Database: Tools for analyzing talk*, 3rd edn. Mahwah NY: Lawrence Erlbaum Associates.

McEnery, T., Xiao, R. & Tono, Y. 2006. *Corpus-based Language Studies: An Advanced Resource Book*. London: Taylor & Francis.

McEnery, T. & Hardie, A. 2011. *Corpus Linguistics: Method, Theory and Practice*. Cambridge: Cambridge University Press. https://doi.org/10.1017/CBO9780511981395

Muñoz, C. (Ed.) 2006. *Age and the Rate of Foreign Language Learning*. Clevedon: Multilingual Matters. https://doi.org/10.21832/9781853598937

Myles, F. 2015. "Second language acquisition theory and learner corpus research". In S. Granger, G. Gilquin & F. Meunier (Eds.), *Cambridge Handbook of Learner Corpus Research*. Cambridge: Cambridge University Press, 309–332. https://doi.org/10.1017/CBO9781139649414.014

Ochs, E. 1979. "Transcription as theory". *Developmental Pragmatics* 10(1), 43–72.

Papageorgiou, S. 2007. *Relating the Trinity College London GESE and ISE Examinations to the Common European Framework of Reference*. Final project report, February 2007. London: Trinity College London.

Plonsky, L. 2016, February. *The N crowd: Sampling practices, internal validity, and generalizability in L2 research*. Presentation given at University College London, London, UK.

Porte, G. (Ed.). 2012. *Replication Research in Applied Linguistics*. Cambridge: Cambridge University Press.

Roever, C. 2010. "Effects of cultural background in a test of ESL pragmalinguistics: A DIF approach". In G. Kasper, H.t. Nguyen, D. R. Yoshimi & J. K. Yoshioka (Eds.), *Pragmatics and Language Learning*, Vol. 12. Honolulu: National Foreign Language Resource Center, University of Hawai'i at Mānoa, 187–212.

Semino, E., Demjén, Z., Demmen, J., Koller, V., Payne, S., Hardie, A., & Rayson, P. 2017. "The online use of violence and journey metaphors by patients with cancer, as compared with health professionals: a mixed methods study". *BMJ Supportive & Palliative Care* 7(1), 60–66. https://doi.org/10.1136/bmjspcare-2014-000785

Sinclair, J. 2005. "Corpus and text – basic principles". In M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books, 1–16.

Spencer-Oatey, H. 2008. "Introduction". In H. Spencer-Oatey (Ed.), *Culturally Speaking. Culture, Communication and Politeness Theory*, 2nd edn. London: Bloomsbury, 1–8.

Thompson, P. 2005. "Spoken language corpora". In M. Wynne (Ed.), *Developing Linguistic Corpora: A Guide to Good Practice*. Oxford: Oxbow Books, 59–70.

Tomasello, M. 2003. "Introduction: Some surprises for psychologist's". In M. Tomasello (Ed.), *The New Psychology of Language*. London: Taylor and Francis, 7–20.

Tracy-Ventura, N. & Myles, F. 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. https://doi.org/10.1075/ijlcr.1.1.03tra

Trinity College London. 2016. *Exam Information: Graded Examinations in Spoken English (GESE)*. Available at <http://www.trinitycollege.com/site/?id=368>

Wall, D., & C. Taylor. 2014. 'Communicative Language Testing (CLT): Reflections on the "Issues Revisited" from the perspective of an examinations board.' *Language Assessment Quarterly* 11(2): 170–185.

Wong, D. & Kruger, H. 2018. "*Yeah, yeah yeah* or *yeah no that's right*: A multifactorial analysis of the selection of backchannel structures in British English". In V. Brezina, R. Love & K. Aijmer (Eds.), *Corpus Approaches to Contemporary British Speech*. London: Routledge, 120–156.

Appendix A. Composition of the Trinity Lancaster Corpus

1. *L1 and cultural backgrounds in the TLC*

The TLC contains speakers from a variety of linguistic and cultural backgrounds. The major subcorpora contain speakers from the following countries (L1 backgrounds): Argentina (Spanish), China (Mandarin, Cantonese), Italy (Italian), India (Hindi, Bengali, Gujarati, Marathi, Tamil), Mexico (Spanish), Russia (Russian), Spain (Spanish), Sri Lanka (Sinhala, Tamil). In addition, the TLC contains data from the following language backgrounds: Arabic, Bulgarian, Czech, Danish, French, German, Japanese, Kannada, Konkani, Korean, Lithuanian, Malayalam, Marwari, Persian, Polish, Portuguese, Romanian, Sindhi, Slovak, Telugu, Turkish and Ukrainian.

2. *Speaker-related variables in the TLC*

There are several pieces of information available about the speakers in the TLC. For the L2 speakers who participated in the first phase of the data collection (2012), information on variables (a) to (g) is available; for the L2 speakers who participated in the second phase of the data collection (2013–2018), additional data on L2 learning and use have been collected as well (variables h-l). Overall, one quarter of the data in the corpus come from the first phase of the project and the rest have been collected in the second phase.

I. *Information about L2 speakers (exam candidates)*

a. GESE band (6, 7, 8, 10, 11 and 12)
b. CEFR band (B1.2, B2.1, B2.2, C1.1, C1.2, C2.2)
c. Overall mark (Fail, Pass, Merit and Distinction)
d. Mark for each speaking task (A, B, C and D)
e. Linguistic and cultural background
f. Age at the time of the interview
g. Gender
h. Highest completed education level (primary, secondary, tertiary – bachelor, masters, PhD)
i. Age of exposure (when the L2 speaker started learning English)
j. Uses of English (at school, at work, with family, with friends, on the Internet, other – can specify)
k. The context of L2 acquisition (at school in a foreign language classroom, at school where English was the medium of instruction, in an English speaking country, from TV or the Internet, self-study, other – can specify)
l. How long was spent in each kind of learning context (in years and months)

II. *Information about L1 speakers (examiners)*

a. Age at the time of interview
b. Gender
c. Experience – length of time examining for Trinity College London

Appendix B. Structure of the TLC according to key variables

This section shows the distribution of some of the key variables in the TLC and provides the information about the size of the relevant subcorpora in terms of number of speakers. Table 1 provides information about the number of speakers at each proficiency level per main L1 and cultural backgrounds in the corpus.

Table 1. L2 proficiency: The number of L2 speakers across CEFR bands

	B1 [grade 6]	B2 [gr. 7 & 8]	C1 [gr. 10 & 11]	C2 [grade 12]
Argentina	77	96	3	0
China	156	99	35	5
Italy	153	131	58	10
India – Hindi	55	59	1	4
India – Gujarati	30	20	0	2
India – Marathi	15	13	2	4
Mexico	174	78	52	9
Russia	26	28	1	1
Spain	131	159	62	17
Sri Lanka	42	45	12	4
Other	74	77	21	12
Total	**933**	**805**	**247**	**68**

Tables 2 to 4 provide information about the size of evidence available in the TLC according to the individual tasks at each proficiency level. The tables show information about the mean number of word tokens produced by L2 speakers in each task followed by the total number of word tokens available for L2 speakers (exam candidates) and L1 speakers (examiners) in each of the tasks per proficiency level.

Table 2. Word tokens per speaking task by proficiency: Advanced L2 users (C1 & C2 levels)

Task	Candidates (mean per task)	Candidates (total)	Examiners (total)	Sub-corpus (total)
Presentation	637.0	200,643	25,644	226,287
Discussion	496.3	156,331	110,241	266,572
Interactive task	399.2	125,736	107,989	233,725
Conversation	701.4	220,956	137,578	358,534
Sub-corpus Total	**2,233.9**	**703,666**	**381,452**	**1,085,118**

Table 3. Word tokens per speaking task by proficiency: Intermediate L2 users (B2 level)

Task	Candidates (mean per task)	Candidates (total)	Examiners (total)	Sub-corpus (total)
Discussion	519.7	418,380	213,036	631,416
Interactive task	309.0	248,745	256,049	504,794
Conversation	494.0	397,692	253,681	651,373
Sub-Corpus total	**1,322.8**	**1,064,817**	**722,766**	**1,787,583**

Table 4. Word tokens per speaking task by proficiency: Threshold users (B1 level)

Task	Candidates (mean per task)	Candidates (total)	Examiners (total)	Sub-corpus (total)
Discussion	436.8	407,534	263,447	670,981
Conversation	360.4	336,219	299,259	635,478
Sub-Corpus total	**797.2**	**743,753**	**562,706**	**1,306,459**

Table 5 reports the number of word tokens and L2 speakers (shown in brackets) for different age groups per the main proficiency levels.

Table 5. Word tokens and number of speakers (in brackets) per age bands by proficiency

Age band	Threshold	Intermediate	Advanced	Total
8–15	399,761 **(499)**	500,414 **(374)**	52,292 (23)	952,467 (896)
16–19	117,484 **(147)**	254,040 **(194)**	288,369 **(128)**	659,893 (469)
20–35	168,108 (200)	201,307 (149)	247,832 (106)	617,247 (455)
36–50	53,944 **(63)**	107,271 **(76)**	122,031 **(53)**	283,246 (192)
51+	15,612 (19)	13,494 (11)	8,221 (4)	37,327 (34)

Appendix C. Lancaster Spoken Language Transcription Guidelines

Feature	Transcription guideline	Example
ID info	Do not put ID info at the top of the transcript.	
Speaker identification	Speaker labels are followed by a colon, then a tab, then continuous text until the end of the utterance (indicated by a line break). Use 'E' for the examiner and 'S' for the candidate.	E: what's your name? S: it is <name>
Time stamp	Every 10th line on a separate line enter a timestamp using a keyboard shortcut	0:03:43.4
Segmentation	Indicate beginnings of the individual parts of the exam	<GREET> <PRESENT> <DISC> <TASK> <LISTEN> <CONV>

Feature	Transcription guideline	Example
Emphasis	Do not mark stressed words	
Reading	Mark the beginning and the end of each reading passage.	<reading> </reading>
Capitalisation	Use word-initial-capital for proper nouns and "I". If a proper noun includes a number, it is spelled out (see numbers). Proper nouns include:	
	1. NAMES OF PEOPLE (BUT SEE ANONYMISATION)	Roger, Shakespeare
	2. PLACE NAMES AND DERIVATIVES (ALL WORDS ARE CAPITALISED)	England, English, North Sea, American English, Mars, Earth (planet), Statue Of Liberty
	3. NAMES OF PRODUCTS AND INSTITUTIONS (THE INITIAL LETTER OF EACH WORD IS CAPITALISED REGARDLESS OF THE OFFICIAL SPELLING)	Google, Facebook, Iphone, Microsoft, American Broadcasting Company, University Of Vienna, Kendrick School, Atari CX Twenty Two
	4. RELIGIONS, RELIGIOUS INSTITUTIONS AND DERIVATIVES	Christianity, Buddhism, Catholicism, Catholic, Buddhist
	5. NAMES OF DAYS, MONTHS AND FESTIVALS	Monday, February, Christmas, Chinese New Year, Hanukkah
	6. BOOK AND FILM TITLES (ALL WORDS ARE CAPITALISED INCLUDING PREPOSITIONS)	Harry Potter And The Philosopher's Stone, Twelve Years A Slave [numbers spelled out], I Frankenstein [no comma], The Wolf Of Wall Street [capital The]
	7. PHENOMENON / DISEASES NAMED AFTER INVENTORS / RESEARCHERS	Parkinson's disease; Alzheimer's disease
	MULTI-WORD EXPRESSIONS	
	In multi-word proper nouns all words are capitalised including prepositions and articles. Initial articles are capitalised only in book/film titles.	The Australian [newspaper] The Wolf Of Wall Street A Companion To The History Of The Book
		x the American Broadcasting Company the European Union the University Of Vienna the Sistine Chapel
	ABBREVIATIONS/ACRONYMS AND INDIVIDUAL LETTERS ARE CAPITALISED	BBC, Mr, Ms, Mrs, Miss, PhD, PPT, AIDS, PDF H O R S E [spaces used when spelling out a word letter by letter]

Feature	Transcription guideline	Example
	NOT CAPITALISED	
	– No capitalisation is used for titles or 'honorific' uses.	archbishop, pope, king, duke, god, doctor, reverend, her majesty, his highness
	– No capitalisation is used when originally proper nouns are employed as common nouns or verbs.	I googled this he was facebooking she tweeted that she had no time
Overlapping speech	Do not mark overlaps	
Punctuation	Question mark indicates a surface form of a question:	?
	1. YES/NO QUESTION This type of question is indicated by inversion of the verb and the subject (V-S).	Can you help me? Are you ready to go?
	2. WH-QUESTION This type of question is indicated by a question word (why, when, what, who, how etc.) and inversion of the verb and the subject (V-S).	When will you help me? How did you do this?
	3. TAG QUESTION A statement with an interrogative element (tag) at the end.	This book is great isn't it? She's not French is she? Let's have a beer shall we? I haven't heard from him have you?
	N.B. If the question is interrupted or incomplete, do not use a question mark.	S: is this <.> E: I don't know S: important
	Otherwise do not use any other punctuation marks to indicate sentence or clause boundaries	
Pauses	Clear pauses should be marked with a period inside angled brackets. (1) Pauses up to 3 s. <.> (2) Measured pauses for pauses longer than 3 seconds.	<.> <pause=20>
Meta-linguistic behaviour and comments	Mark within angled brackets	<laugh> <cough> <clears throat>
Fillers	These need to be standardised to a small subset as shown in the examples. No other fillers should be used.	ah er erm huh mm oh uhu

Feature	Transcription guideline	Example
Unclear speech	Mark as unclear with a guess if possible or a timestamp	<unclear=lift> <unclear = 38.2>
Foreign word	Mark if possible (i.e. if the foreign word can be recognised)	<foreign=empresario>
Grammar error	Do not correct learner errors	
Anonymisation	Anonymised name of person (do not anonymise place names or names of authors etc.)	<name> Thames Shakespeare
Pronunciation, spelling and contracted forms	Use normal British spelling if word is clear, otherwise use unclear feature. Non-standard forms that appear in the dictionary are transcribed orthographically in their dictionary accepted way: *cos, dunno, gonna, gotta, kinda, lotta, sorta, wanna* and *yeah*.	
	Where the intended word is mispronounced resulting in a different word form being produced transcribe in brackets: <misp=original, target>	<misp=liquorice, lyrics>
Numbers	All numbers should be spelt out.	zero three nineteen ninety two twenty thirteen nine thirty twenty pounds
Okay	Always spell out *okay*	okay
False starts and repairs	Mark these using hyphenation (no space)	Au-Au-August any-anything cro-control
Truncated (abandoned/ unfinished) words	Mark these using equals sign	resem= I don't under= I don't totally understand
Learner language features	PRONUNCIATION: Do not attempt to transcribe different accents or non-standard pronunciation. Use standard (dictionary) forms of words.	
	WRONG WORD: If an incorrect word is produced, record it.	this people; mine husband; this was very loud er low
	NON-WORD: Transcribe as produced, if possible; otherwise mark as unclear.	discoverments

Sources consulted

LINDSEI corpus guidelines: <http://cecl.fltr.ucl.ac.be/Cecl-Projects/Lindsei/lindsei.htm>
CHILDES CHAT guidelines: <http://childes.psy.cmu.edu/>
BNC transcription guidelines: <http://www.natcorp.ox.ac.uk/docs/Burnage93a.htm>

Crowdy, S. 1995. "The BNC spoken corpus". In G. Leech, G. Myers and J. Thomas (Eds.), *Spoken English on Computer: Transcription, Mark-up and Application.* London: Longman, 224–234.

Address for correspondence

Dana Gablasova
Lancaster University
Department of Linguistics and English Language
County South
Lancaster LA1 4YL
United Kingdom
d.gablasova@lancaster.ac.uk

Co-author information

Vaclav Brezina
Lancaster University
Department of Linguistics and English Language
v.brezina@lancaster.ac.uk

Tony McEnery
Lancaster University
Department of Linguistics and English Language
a.mcenery@lancaster.ac.uk