Cambridge University Press & Assessment
978-1-009-20895-6 — Learner Language, Discourse and Interaction
A Corpus-Based Analysis of Spoken English
Tony McEnery , Isobelle Clarke , Gavin Brookes
Excerpt
More Information

CHAPTER 1

Exploring Learner Discourse
Context, Data and Methods

1.1 Learner Corpora

The idea of a learner corpus is a disarmingly simple one – it is a good idea, when exploring how a second language (L2) is learned, to study the language in the acquired L2 produced by a learner. Ideally this should be gathered in a range of contexts relevant to the intended use of the L2 by the learner. Through the examination of such data at any one point in time, we may be able, given appropriate data, to see a wide range of things. For example, we may be able to see systematic differences between different learners based on demographic variables (e.g. age or sex) or on the basis of variables that might be indicative of their aptitude for language learning (e.g. past educational experience and performance in exams directly related to their language learning). We may also have access to other variables that may allow us to explore the influence of social or other situational contexts which might impact on language learning, such as the learner's first language or languages (L1/L1s), their cultural background or the task they are engaged in. Through time, we may also be able to explore the process and experience of language learning. We may be able to look at individual performance across space (at one moment in time) and time (the same population, or what we believe to be equivalent populations, in the same space sampled across time). When looking across time, we may model learners as a group (we study the same features related to learners over time, not the same individuals) or as a cohort (we study those features in the same learner or learners across time).[1]

[1] Note our use of the term *cohort* here is deliberate. In the social sciences, cohort studies are longitudinal studies which look, at intervals, at individuals experiencing the same life events. The analogy here is clear – the testing of the same individuals over time would represent our sampling points; training in an L2 would be the experiences they share in common between the sampling points, for example. For a fuller introduction to cohort studies in the social sciences see Wadsworth and Bynner (2011).

Cambridge University Press & Assessment
978-1-009-20895-6 — Learner Language, Discourse and Interaction
A Corpus-Based Analysis of Spoken English
Tony McEnery , Isobelle Clarke , Gavin Brookes
Excerpt
More Information

The idea is appealing to the extent that, even before learner corpus research was developed, studies of learner language proceeded, typically in qualitative fashion, on the assumption that the output of learners was worth studying. Small-scale studies, such as those by Juvonen (1989) and Cornu and Delahaye (1987), used what were, effectively, small collections of learner language, produced in contexts in which an L2 was being produced for the purposes of communication, to draw conclusions about language learning. The development of corpus linguistics (see McEnery and Hardie, 2011, for an overview) opened up the prospect of increasing the scale and ambition of such studies. Rather than study small samples of L2 usage, one could instead look at such language on a scale which would allow insights that smaller studies could not provide, or which could provide evidence sufficient to corroborate or reject hypotheses based on such small studies. It was very much in this spirit that pioneering early work on learner corpora, notably that of Granger (1998), proceeded; that is, it used the emerging capacities of corpus linguistics to pursue what we have called this disarmingly simple idea, at scale.

Yet, as with all disarmingly simple ideas, actually realising the potential of the idea is much more complex than one might at first think. A major contribution of learner corpus research has been to demonstrate this clearly. A suggestion of that lies in the discussion so far – the enterprise outlined is actually complex, and that complexity becomes apparent when one tries to operationalise models of corpus construction that would facilitate a thorough investigation of hypotheses about second language acquisition (SLA), for example. The range of variables that are thought to impact on the process of language learning is large enough in the examples given, yet these are only examples – we may wish to look at many more such variables. As soon as we wish to examine a broad range of variables, then the demands made on our data increase rapidly, both in terms of the effort it takes to collect it and in terms of the sheer scale of data that we need. For example, let us imagine that we decide that we need to collect information on three variables for each person that we will include in our corpus. These variables are sex (which we decide will have one of three possible values: male, female and non-binary), age (which we decide can have one of ten possible values, as we group ages into ten equally sized bands to cover learners from age 7 to 106) and L1 background (for the sake of simplicity we record only what the learner believes their primary L1 to be, and we limit our study to one country in which we discover that fourteen L1s account for all language learners). Let us further imagine that we have determined that, for any combination of variables, we need

Cambridge University Press & Assessment
978-1-009-20895-6 — Learner Language, Discourse and Interaction
A Corpus-Based Analysis of Spoken English
Tony McEnery , Isobelle Clarke , Gavin Brookes
Excerpt
More Information

sub-corpora of 10 speakers, each producing approximately 10,000 words each, to support both the development of new hypotheses and the testing of established ones. We further decide that we would like to sample the learners across three years, with a sample being taken every year, over a two-week period at the beginning of the calendar year, for three years. This sounds like a very exciting project indeed, though note that even this ambitious project has limited itself – it is dealing with only three variables, it is dealing with one country, it is only sampling language learning across three years and age is being aggregated into decade-sized intervals. Yet even as it stands, a project of this scale would require a vast amount of data: it requires 10,000 (words) × 10 (speakers) × 3 (sexes) × 10 (age groups) × 14 (L1s spoken) × 3 (repeated samples). This means we need a corpus of 126 million words to meet what, at first glance, seems like a modest proposal for research.

Scale is an eye-catching problem, but the principal impediments to achieving scale are practical. Some variables are easier to collect from and balance out. For example, in our imaginary research context, we may discover it is easy to find plentiful L2 learners for each of the L1s. However, finding language learners in the age range 95–106 is likely to be impossible. While this may sound like good news – as there is less data for us to have to collect – the same is probably true further down the scale also. Yet the process of deciding to artificially limit the range of a particular variable, while it has appeal in terms of making corpus building easier, is achieved at a cost. For example, older language learners are relatively neglected by researchers, but, in a context where leisure learning and learning spurred by claims of cognitive benefit is causing this group to expand (Murray, 2011), the group may be important to study in research terms because, *inter alia*, it is under-researched and some of the claims made about language learning in the group (including those regarding its cognitive benefits) need to be critically explored. Therefore, the impulse to turn away from elements of corpus construction because the task is difficult should always be balanced against what is lost if we do so. Persistence is one response to such an opportunity swaddled in difficulty, yet a shift of method is another perfectly plausible response – where you have access to a small number of learners of a difficult-to-find age group, you may accept that pursuing the collection of a corpus from the group may not be possible, but more qualitative methods, or perhaps more psycholinguistically oriented methods (see Roberts, 2012), may be a better way of proceeding. Whatever the decision, the result is inevitably principled pragmatism in corpus building and this is especially true in the construction of learner corpora, as it is with other corpora.[2]

This principled pragmatism in learner corpus building is distinctly visible in another issue which we have not touched upon so far – the medium of communication. In our example scenario, we remained silent on the question of whether we intended to collect a corpus of speech or writing. Note that if we had decided that the answer is 'both', the size of the corpus we would need to collect would have ballooned to 252 million words. Of the two halves of this corpus, the written component would have appeared much easier to construct – with many texts, including those produced by learners, now 'born-digital' (Smith et al., 2014), it has become increasingly easy to construct corpora of written language (McEnery and Brookes, 2022). This in itself is well borne out by a look at existing learner corpora – these are principally written corpora and, moreover, they are typically composed of texts that are easy for researchers to gather in contexts they wish to research, with the result being that argumentative essays 'correspond to over half of the written corpora' in the list of written learner corpora maintained by the University of Louvain (Gilquin, 2015: 12). The dominance of essay data in learner corpora is likely to persist and the primacy of written, over spoken, language in available learner corpora is currently overwhelming – the source of evidence used by Granger now lists nearly 200 learner corpora.[3] Most are solely or partially composed of written material (128 and 24, respectively – 71 of which include student essays), set against 72 corpora which are either solely spoken (46) or contain a spoken component (26).[4] Likewise, the difficulty of studying language-learning cohorts means that there are few longitudinal corpora and those that exist are relatively limited in scope with regard to the range of variables which such a corpus might consider (see Meunier (2015) for a discussion). Importantly for this book, the difficulty of composing corpora of spoken language produced by language learners has meant that the number of such corpora is fewer, as noted. They are also smaller – some written learner corpora are very large indeed, with the Cambridge Learner Corpus, composed of essays, being 50 million words. Yet the largest publicly available conversational spoken learner

[2] See Biber, Egbert and Gray (2022) for a good account of how research design and principled pragmatism combine in corpus building.

[3] See https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpora-around-the-world.html. Website accessed 19 September 2023.

[4] Two corpora not included in this count are multimodal corpora, which could also count towards the spoken corpus count.

Cambridge University Press & Assessment
978-1-009-20895-6 — Learner Language, Discourse and Interaction
A Corpus-Based Analysis of Spoken English
Tony McEnery , Isobelle Clarke , Gavin Brookes
Excerpt
More Information

corpus of which we are aware is the one used in this book – the Trinity Lancaster Corpus (TLC). It is just over 4 million words in size. Typically, spoken learner data of this sort is available in corpora comprising tens of thousands of words (e.g. the 35,000-word Evaluation of English in Norwegian Schools corpus, Hasselgren, 1997) or hundreds of thousands of words (e.g. the approximately 500,000-word Corpus of Young Learner Interlanguage, Myles, 2005), but not millions of words. There are notable exceptions, such as the International Corpus Network of Asian Learners of English corpus, which, at the time of writing, includes 1.6 million words of conversational speech (Ishikawa, 2019, 2023). However, such exceptions are all the more remarkable given the usual size of learner corpora covering speech in general and conversational speech in particular.

The discussion so far has introduced the idea of the learner corpus in general terms, principally to show that what seems like a useful tool is difficult to construct. Readers interested in looking deeper into the advantages and disadvantages of using learner corpora should see McEnery et al. (2019) and Tracy-Ventura and Paquot (2021). While we have focused so far on noting the difficulties that working with learner corpora may bring, we have done so in part to highlight the effort that has been expended on, and the importance of, the corpora used in this book – all are spoken corpora, orthographically transcribed, and all come from contexts that are difficult to access. Thus, while this book is about the value of exploring large corpora of spoken interactions with language learners, we are well served to remember that the exploration of such data is relatively new and is, accordingly, likely to provide fresh insight. As well as providing insight into those interactions, we also hope that this book encourages researchers to invest more time in building a larger, more comprehensive set of corpus resources for exploring the spoken interactions of L2 learners, in particular spoken corpora. Constructing such corpora can be hard, but the results, as we hope this book will show, are worth the effort.

This context outlines the approach taken in this book – our approach is exploratory. We are investigating new corpus resources and, in doing so, we are taking new perspectives on that data. As this chapter proceeds, we will introduce these perspectives, but to begin with we focus on the new material on which all of the new perspectives are taken – our corpora.

1.2 The Trinity Lancaster Corpus

The learner corpus to be used in this book is the TLC. This corpus is solely a spoken language corpus. It is based upon completed examinations taken by students from a range of L1 backgrounds. The test that the students took, the General Examination of Spoken English (GESE), is a graded examination, for which the learners in the corpora had been preparing. In the exam, the learners interact with an examiner, who is an L1 speaker of British English. These examiners, drawn from a central pool of examiners based in the UK, administered the examinations as parts of tours that they did to different countries and regions across the world.

The corpus was constructed by transcribing the speech of the examiner and examinee to produce an orthographic transcription of their interaction. That orthographic transcription was quality checked and can be assumed to be of very high fidelity (see Gablasova et al., 2019). When transcribing the data, a number of non-linguistic features were also included in the corpus, such as pauses, so-called vocal events (e.g. laughter) and filled pauses. We will not introduce these features in full here, but we will provide relevant detail about them as and when they become important for our analysis.

Each file in the transcribed corpus is composed of interactions between one L1 British English-speaking examiner and one L2 British English-speaking examinee. The interactions in each file are split into distinct tasks. The number of tasks taken varies by level of exam. In this book, we will focus principally on the TLC data at grades 6–8. The data in the corpus as a whole corresponds, on the Common European Framework of Reference (CEFR), to grades B1 (GESE grade 6) and B2 (GESE grades 7 and 8). Hence our focus is learners at B1/B2, so-called Independent Users on the CEFR scale.

While the tasks may vary by grade, the grades looked at in this book are associated with a relatively stable set of tasks. At B1, speakers are expected to undertake three tasks – greeting, discussion and conversation. At B2, they perform the same tasks as at B1 but also complete an interactive task. At C1, in addition to the tasks performed at B2, they also have to undertake a presentation task.

The tasks vary the demands on the speaker and vary the roles that they take in conversation. The *Greeting task* need not detain us long – it is typically a short, formulaic sequence where the examiner and the student introduce themselves to one another. To the extent to which they interact, this may be called a jointly led task, as both contribute and either could, in principle, initiate the interchange. The *Discussion task* is a pre-prepared task for students at B1 and B2, where a student briefly introduces a topic they have chosen for a conversation and the discussion proceeds from there. At C1, the examiner chooses the topic for discussion. Hence, this is

Cambridge University Press & Assessment
978-1-009-20895-6 — Learner Language, Discourse and Interaction
A Corpus-Based Analysis of Spoken English
Tony McEnery , Isobelle Clarke , Gavin Brookes
Excerpt
More Information

a task which is broadly jointly led, though at the higher levels the candidate is required to start a discussion of a topic that may be unfamiliar to them. In the *Conversation task*, the focus, at all levels considered here, is a conversation, jointly led, about two topics chosen from a list of topics that the examinee is made aware of in advance of the exam. The topics are varied according to the age of the examinee and the cultural context in which the examinee is learning. Throughout these tasks, the goal is to test for specific features that the candidate is expected to produce in their English at that level and to maintain a sustained, relative to the level, conversation with the examiner. The *Interactive task* is examiner-led and starts with the examiner introducing a topic and then allowing the conversation to proceed from there. The examinee is expected to engage with the topic by asking questions about it, expressing opinions about it, and so on.

At the higher level, C1, the *Presentation task* requires the examinee to present a talk on a topic of their choosing. This leads to a conversation with the examiner about that topic. This task is clearly examinee-led, yet it also differs from the other task in that it is principally monologic. The dialogic nature of the lower level exams permits the examinee to show that they can interact with an L1 speaker (i.e. the examiner). However, the C1 level presentation task requires a sustained linguistic performance from the examinee, placing the student in a position where they clearly have to lead the interaction, largely in a context where they are typically the only speaker. While not relevant to the discussion of the TLC, another corpus used in the book, the TLC L1 (introduced in Section 1.4), does include this task.

It is important to understand the tasks in the TLC as they relate to function – the different tasks see the examinee and examiner performing different functions and, accordingly, we should expect that linguistic form will vary with those functions; that is, that the language usage will adapt to meet the situational demands of the different tasks. However, we also need to be mindful that in our data there is a range of other variables that may promote variation. A feature of Lancaster working with Trinity College London to collect the TLC was that Trinity could provide authentic language data, produced by learners for a purpose for which they were well prepared, and Trinity could, in advance of the exam, retrieve metadata from those learners that the Lancaster team thought might cause variation in performance. This metadata covers a wide range of variables, such as age, sex, L1s spoken by the examinee, years in education and proficiency as marked by the examiner. The corpus contains metadata relevant to the

Cambridge University Press & Assessment
978-1-009-20895-6 — Learner Language, Discourse and Interaction
A Corpus-Based Analysis of Spoken English
Tony McEnery , Isobelle Clarke , Gavin Brookes
Excerpt
More Information

Table 1.1 *Metadata available in the TLC.*

Learner	Examiner
Age band into which the student falls – young (8–15), adolescent (16–19), young adult (20–35), middle adult (36–50), older adult (51+)	Age band within which the examiner falls – middle adult (36–50) and older adult (51+ and over)
Sex	Sex
CEFR proficiency level (mnemonics are those used by CEFR)	Years of experience the examiner has arranged by band – 1–2 years, 3–5 years, 6–10 years or over 11 years
The grade of GESE exam taken (from 6 upwards)	
Country in which the exam took place	
L1 of examinee	
Overall exam mark (in ascending order, Fail, Pass, Merit, Distinction)	

examiner also. While not all of this metadata will be used in this book, for completeness Table 1.1 shows all metadata in the TLC.

Additionally, for those tasks which are allocated a score (that is, all tasks bar Greeting), the score by task is also encoded in the corpus. The marking scale is the same as for the exam overall.

The nature of the data in the TLC sets a series of methodological challenges for the studies in this book. Firstly, it makes little sense to study the interaction at the level of the whole exam. We may reasonably assume that the exam varies functionally, because of the differing nature of the tasks, and that this is likely to mean that the language use in the corpus varies across the tasks too. While we may look at the data as a whole, the task seems like an important, likely linguistically meaningful, level of organisation in the corpus. Accordingly, the task will be our default high-level unit of analysis in this book.

That decision leads to another which needs to be made. Within the task there are smaller units still – turns, for example – and we may assume that these combine together in distinct ways. The issue of making perfectly sensible and justified shifts to ever smaller units of analysis is that the possibility of using some important, functionally oriented, techniques of analysis begins to fade. For example, multi-dimensional analysis (MDA) (Biber, 1988) may appear perfect for our purpose. However, it is known to have problems dealing with small data samples, with texts shorter than 1,000 words being progressively more likely to present a

E: <unclear text='you'/>
S: I'm fine ma'am
E: good
S: how are you ma'am?
E: I'm good thank you very much okay so erm good morning my name is <anon type='name'/> what's your name?
S: my name is <anon type='name'/>
E: <anon type='name'/>
S: yes ma'am
E: and you're a grade seven
S: yes
E: and can you just turn your oh you don't need to take it off
S: mm
E: okay thanks thanks for showing me that that's your ID card thanks <anon type='name'/>

Figure 1.1 A Greeting task from the corpus.

challenge to the system (see Clarke, McEnery and Brookes, 2021, for a discussion). The reason for this is simple – if we use 1,000 words as a baseline for estimating linguistic distributions, following Biber (1993), we may then normalise frequency to occurrences per thousand words to provide a common baseline across texts of varying length. However, if analysing texts that are shorter than this, the normalisation may produce grossly distorted and unreliable frequencies. Imagine that we wish to run an MDA at the level of the task in the TLC. Consider the Greeting task shown in Figure 1.1, taken from file 2_7_IN_5 in the TLC, where an Indian student is starting a grade 7 exam. In the example, as in all other examples in this book, the speech of the student is introduced by S, and that of the examiner by E; features marked up within turns are shown as XML elements.

This task is sixty-eight words long. In it, BE as a main verb occurs 6 times, giving it a normalised frequency per 1,000 words of 88.2. This is a poor guide to the frequency of the features – if we look at the whole corpus, we discover that the frequency per 1,000 words of BE as a main verb is 37.64. This problem with frequency inflation in small-text sequences undermines MDAs of texts in general and presents a general problem for frequency-based approaches to small, functionally coherent sections of language in particular. This is a problem addressed in Clarke (2022), where it was proposed to shift to looking at patterns of presence and absence in short texts, rather than relative frequencies, instead. This is the approach taken in this book, as will be described in more detail in Section 1.6. We take this approach

Cambridge University Press & Assessment
978-1-009-20895-6 — Learner Language, Discourse and Interaction
A Corpus-Based Analysis of Spoken English
Tony McEnery , Isobelle Clarke , Gavin Brookes
Excerpt
More Information

because we require the discriminating power of a technique such as MDA. We have a host of variables, all of which may exert an influence on language, singly or in concert, in our corpus, as variables from the metadata and the tasks within the corpus interact. We need an approach to the TLC which will allow us to see, where we wish, how these variables combine and to what effect. Hence, the TLC, by presenting us with a complex dataset in which we have plentiful evidence of a range of variables and tasks in interaction, sets us the welcome task of rising to the methodological challenge that such data presents.

The issue that presents itself to MDA – data sparsity – is another that we have to engage with more generally in this book. As discussed earlier, apparently large corpora can become unbalanced and the data available for any given question vanishingly small when variables are combined. This problem amplifies when we consider larger units within a text. So, for example, while the TLC may contain millions of words, if we combine variables to look at a specific group we may find no data at all. For example, if we are interested in older learners, those sixty years of age and older, then the data is sparse. There is only one example of a learner in this age group in the TLC data used in this book. It follows that most of the possible combinations of other variables for this age group have no data associated with them. For example, our one older learner is a male who has Mexican Spanish as their L1. Many other L1s are represented in the corpus we use in this book, but none have older speakers. Likewise, we have no data at all for older female speakers. This is one manifestation of sparsity. Another manifests itself in a different way when we undertake linguistically meaningful aggregation in the data – for example, if we wish to look at turns in the examination of older learners and use those as our basic unit of analysis rather than words, then the evidence available to us for our older learner diminishes to 257 examples (the turns in the relevant corpus files) from 1,115 examples (the number of words in the relevant corpus files). We note these issues for now and will return to them to outline our response as this chapter proceeds.

Before leaving the introduction of the TLC, we can identify a further challenge from using it relating to the goal of the GESE examination. The examination is judged by an expert native speaker, and the target which it is assumed that the student is aiming at is conversational spoken British English, as produced by L1 speakers of that variety of English. This begs the question of how we know what conversational British English looks like. To consider that, we need to introduce another corpus to be used in this book – the Spoken BNC 2014.