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).

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

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.

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

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

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

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.

1.3 The Spoken BNC 2014

One of the key questions to ask when using a learner corpus relates to what 'target' the learners' linguistic performance should be measured against. If, for example, one is seeking to rate a learner's proficiency, then there is an implication that there must be a norm against which the performance is being judged.[5] What is the norm and why should it be something that all learner corpus studies should consider? In this case, the nature of the exam itself points towards some of the answers. Students taking the exam on which the TLC is based are learning British English. That is the variety of English in which they are examined and to which the materials they studied prior to the examinations oriented them. Thus, a nominal target of British English may be assumed for our learners as comparison to another variety of English is likely to misrepresent the learners' performance. The exam also dictates other features of our target corpora. They should be collections of *spoken* British English – GESE is an exam in which oral production and reception are assessed in an interactive, conversational setting. We know that spoken English varies systematically from written English (McCarthy and O'Keeffe, 2014), so we also know that using a written British English corpus, such as the written BNC 2014 (Brezina, Hawtin and McEnery, 2021), as a target would be inappropriate. In addition to this, we also know that the students are being prepared for an examination where the accent is on conversational competence (Gablasova, Brezina and McEnery, 2019). While it has long been relatively easy to gather spoken corpora based on broadcast speech (see, e.g. Graff, 2002) or to collect and analyse scripts of television programmes or films (e.g. Bednarek, 2018), they are imperfect for our purpose. The (mostly scripted, or at the very least produced and edited) interactions taking place in TV programmes and films are not necessarily at all like authentic spoken conversation. Meanwhile, although written-to-be-spoken materials such as film scripts might represent what a writer *believes* conversation to be like, such texts cannot, *prima facie*, reasonably be held to actually represent spoken conversation, and the degree to which different representations of conversation are actually similar to it may vary (see Al-Surmi, 2012). We see this tacitly acknowledged by the creators of such datasets. For example, in a description of the TV and Movie corpora available at mark-davies.org, the creator of the corpus states that 'the corpora contain extremely informal

[5] See McEnery and Brezina (2022: 89) for a more general discussion of the role of normative epistemology in corpus linguistics.

language … in many cases it is more informal than the language in actual spoken corpora, like the spoken portion of the BNC'.[6] A series of phrases which are more frequent in the script corpus are then noted; these are much more frequent in the scripts than in the spoken portion of the BNC.

Of course, this assertion shows that what is in the script corpus is not, in fact, representative of spoken English; the features that are more frequent in the TV and Movie corpora than in the spoken portion of the BNC are, by dint of their higher frequency, over-represented relative to authentic spoken interaction. Relevant here is Cameron and Kulick's (2005: 118) reference to the creation of a 'pleasurable illusion'; that is, the creation of a world and events in that world which in some way transcend the naturalness of the real world. This 'unordinariness' gives the events depicted in (particularly fictional) TV shows their tellability and/or watchability. The overall effect is something like saying that a version of a medieval castle at a Disney resort is even more of a medieval castle than real medieval castles because it has so many features of the original present in more exaggerated and frequent forms – more battlements, more gargoyles, the biggest keep ever seen, a wise king and noble queen waving from each tower and so on. While each battlement and gargoyle generates a pleasurable illusion, allowing you to see more examples of them, their overall form and frequency just underlines that what we are seeing is a distortion of the original, not the original. It is a fantasy construct, projected from reality, but not to be confused with reality. Indeed, what we can expect of television dialogue, 'as a result of a commitment to intelligibility, is a "tidying up" of dialogue' that results in a reduction of features that abound in naturally produced language, such as interruptions, false starts and hesitations (Brookes and Collins, 2023: 130).

This leads us to our first choice of target corpus – the Spoken BNC 2014 (Love et al., 2017). The Spoken BNC 2014 is one of two components of the updated British National Corpus 2014 (see also Brezina et al., 2021). This is an approximately 100 million-word corpus representing contemporary British English language use (written and spoken), assembled by a team based at Lancaster University. The Spoken BNC 2014 consists of approximately 11.5 million words (tokens) of informal British English conversation, spoken by 672 speakers. The corpus is available via Lancaster University's *CQPweb* server (Hardie, 2012) as well as a file download.[7] The texts comprising the Spoken BNC 2014 each correspond to a recorded informal

[6] See www.english-corpora.org/files/tv_movie_corpora.pdf.
[7] See http://corpora.lancs.ac.uk/bnc2014/. Website accessed 19 June 2024.

conversation, and the data is contemporaneous with the recordings used to construct the TLC.

The corpus is, of course, not exactly the same as the TLC – the speakers in the TLC, while engaged in conversational interaction, are limited to a set number and type of tasks. The speakers in the Spoken BNC 2014 data are not. However, by comparing the Spoken BNC 2014 data with the TLC, we can begin to address the question of how conversation-like the language use in the TLC is. Should the TLC vary from the spoken BNC 2014, we can develop an appreciation of what the similarities and differences between the two datasets are. This book builds towards that comparison, which is presented in Chapter 7. The process which allows us to undertake the comparison unfolds across the earlier chapters. For now, let us simply note that a comparison of the TLC and Spoken BNC 2014 is well motivated in terms of target variety of English for the learners (British English), mode of communication (speech) and type of interaction (spontaneously occurring conversational English). At the same time, we also acknowledge that the fit between the two, being imperfect, will allow us to observe similarities and differences between the two corpora that may tell us about differences driven by the nature of the exam or the students taking it. However, to better discriminate between the types of differences, we need to compare the TLC and the Spoken BNC 2014 with a third corpus – the Trinity Lancaster L1 Corpus.

1.4 The Trinity Lancaster L1 Corpus

The Trinity Lancaster L1 Corpus (TLC L1) is a match for the TLC – the only differences between the two are that in the TLC L1 the GESE exam is being taken by L1 British English speakers and, accordingly, they are given some tasks to do in addition to those in the TLC because their level of proficiency matches a much higher grade of exam than that of the L2 learners. However, this does not mean that the TLC and TLC L1 are not comparable – the point of comparison is generated by the shared tasks between the corpora (principally, Conversation and Discussion).[8]

The corpus was constructed in collaboration between Lancaster University and Trinity College London. The Lancaster team recruited the L1 speakers to take the tests and Trinity provided experienced examiners to administer and score the tests. The corpus is still in development, led

[8] The additional tasks are a listening task, a presentation task and an interactive task.

by a team that includes Vaclav Brezina, Lorrae Fox and Dana Gablasova. The version used in this book is composed of the exams of 191 L1 British English speakers. It amounts to just over 1 million words of data. The people recruited to take the test were a broad range of L1 British English speakers producing a diverse dataset. For example, the speakers are male (32.6 per cent) and female (67.4 per cent), have a highest level of education stretching from secondary school (40.24 per cent) to doctoral level (7.69 per cent) and a range of ages (from twelve to seventy-seven), though, as with the TLC corpus, many speakers are twenty years of age or younger (32.78 per cent of speakers are in this group).[9]

Taking the TLC and the Spoken BNC 2014 together, we can consider the similarity – or otherwise – of the spoken interactions in the TLC with conversational British English. Using the TLC L1, we can explore whether any differences observed are a product of the tasks in the exam, or something relating to the language learners themselves. For example, if we find that interrogatives are more marked in frequency in the TLC than in the Spoken BNC 2014, we may be unsure as to whether this is a product of the tasks represented in the TLC, or related to some feature of language learning. However, if we discover that L1 speakers in the exam also have a high frequency of interrogative use relative to conversational data, then we have evidence that task, and not proficiency, is the probable source of this difference.

Having introduced the three corpora to be used in this book, we can now turn to the question of annotation and analysis. In the next section, the main annotation used in the book – that of discourse units – is introduced. Following from that, we return to the question of how to analyse our data and explore, in more depth, the approach to MDA taken in this book. We conclude with a brief word on narrative as, later in the book, this is a form of annotation that becomes important. For now, let us consider the annotation that provides the basis for most of the analyses in this book – discourse units.

1.5 Discourse Units

One response to the issue of data sparsity that we discussed in Section 1.2 is to maximise the volume of data in a study by focusing on small units – lexis, for example. The history of learner corpus research is strongly oriented towards studies which are lexically driven, in part because identifying

[9] For more details of the corpus, see Fox (2024).

words is relatively tractable (though not trivial, see McEnery and Brezina, 2022, for example) but also because the type of problem brought about by linguistically meaningful aggregation is eased by this decision – putting it simply, we can count lots of small things (e.g. words) rather than fewer large things composed of small things (e.g. turns or larger units). This tendency has consequences as it limits the horizons of ambition in such studies. Discourse is an interesting case in point. In this book we will treat discourse as the study of organisation of language above the level of the sentence/utterance. In that process of organisation, meaning is both created and sequenced. The later view aligns closely with van Dijk's (1977) view of discourse, in which structures above the level of the sentence are formed in chains of propositions which help provide coherence in discourse. A more general view, of the organisation of sentences into functionally coherent groups that we will term macro-structures (after van Dijk), is of 'a particular unit of language (above the sentence), and a particular focus (on language use)' (Schiffrin, 1994: 20). Some of these macro-structures we may conceive of easily – a narrative, for example. Others, as this book will show, are discernible with appropriate analysis, even if their immediate function is not one which is readily apparent. Discourse is also, of course, organised at the turn level and below – what we will call the micro-structural level. A good example of this is discourse markers-lexical items that often signal macro-structural effects, such as a change of topic, including, for example, in learner speech (see Buysse, 2020). Yet the focus in research, especially learner corpus research, is decidedly upon discourse micro-structures. If we look at the helpful learner corpus bibliography maintained at Université catholique de Louvain this point becomes clearer.[10] At time of writing, the bibliography contains 2,095 entries. Of these, few (seventy-three) focus on discourse, be it micro- or macro-structure. Of those studies, fifty-eight of them clearly focus on micro-structures, half (twenty-nine) focus on discourse markers. Work at the macro-structural level is rare and often limited to lexically realised coherence (e.g. Zinsmeister and Breckle, 2010) rather than macro-structures *per se*. This preference, in the few studies of discourse in learner corpus studies, for micro-structural analyses is almost certainly because, as noted, frequency data is more easily gathered at the lexical level but also because, beyond the turn, there is no consensus, or available data, on which to base an approach to macro-structural discourse analysis. Such is the dominance of written corpus analysis in

[10] See https://uclouvain.be/en/research-institutes/ilc/cecl/learner-corpus-bibliography.html. Website accessed 19 June 2024.

learner corpus studies that, even where discourse markers are studied, the focus may be on their use in writing rather than speech.[11] While we may have differing views on what words constitute discourse markers, for any given word or multi-word unit that one believes to be a discourse marker, corpus search packages will very readily show one exactly where that word occurs in a corpus and how many times it occurs. No such facility exists for macro-structures. This book is a response to the lack of work on macro-structures in learner corpus research. In this book we will view micro-structures as occurring at or below the turn level – turns are clearly marked in our corpora and hence exploring this level of organisation is facilitated. But how do we identify and search macro-structures? While we will look at micro-structures in this book, we wish to mainly explore the unexplored – macro-structures. This level, as noted in Egbert et al. (2021), is a level which has been acknowledged as important by many linguists, but it is also one where a robust, widely applicable, annotation scheme has not really emerged. Egbert et al. (2021) developed such a scheme and it is that scheme which informs how we identify the stretches of discourse that represent a discourse in this book.

Egbert et al. identify discourse units as the basic level of macro-structure. These discourse units are recognizably self-contained segments which have a coherent communicative purpose. In the process of annotation, they discovered that each unit typically had to be a minimum of 5 utterances or 100 words. To help illustrate how discourse units work, Figure 1.2 shows a section from the TLC corpus in which two discourse units are identified. The example is taken from the Discussion task in the exam of an Argentinian student taking the grade 6 GESE exam. The student was awarded a mark of B for the task and a B for the exam overall. While it is marked up with XML in the corpus, the break between the two discourse units is shown in the figure by a short dashed line.

The macro-structures here are driven by questions. The first discourse unit begins with a question that elicits an attitude towards rugby. This then forms a focus of coherence that binds micro-structures together within the macro-structure. The second discourse unit starts with another question and it represents a break – the focus is no longer upon eliciting an attitude, it is now about finding out about the examiner's experience of watching rugby. In this context, errors at the lexical level do not necessarily

[11] For example, it is telling that in the *Cambridge Handbook of Learner Corpus Research*, the chapter on discourse (Neff-van Aertselear, 2015) defines its focus as being upon writing by the end of its first paragraph, perhaps understandably given the limited availability of spoken learner corpora.

S: do you think rugby is <pause length='short'/> is offensing? Or
E: I'm not sure I'm not sure erm sometimes when I watch rugby I think oh that's really aggressive
S: yes
E: and really dangerous erm but I know there are rules and I know that the rules are quite strict
S: uhu
E: er so and and do you think the rules are strict enough or do you think they they need to more strict?
S: no they are strict enough
E: you think they're strict enough
S: yes
E: yeah okay

S: have you ever seen a foot a rugby a rugby match?
E: er only on TV
S: ah
E: on TV I have never been to to watch one live
S: yes
E: er and you
S: er yes too here in Argentina
E: <unclear/>
S: to <pause length='short'/> er er Los Pumas that
E: oh
S: are <pause length='short'/> that just are
E: okay thank you well thank you for talking about er rugby in my life
S: yes

Figure 1.2 An example of a division of a stretch of text into two discourse units.

have an impact at the level of coherence – the first turn of the first discourse unit contains a clear error, but it has no impact at the level of macro-structure. The conversation proceeds entirely felicitously in spite of the error. However, had the examiner replied 'I like eating pasta', this non-sequitur would have clearly been an error within the context of the macro-structure and would have disrupted it.

It is this structural annotation that is the foundation of the work undertaken in Chapters 3–7. Our approach to identifying the function of the discourse units differs from that of Egbert et al., however. In their scheme functional codes were applied to the discourse units identified by the analysts.[12] In our case we are interested in deriving the function from the bottom up, using form-to-function relations to discover the functions

[12] In the Egbert et al. coding, the discourse units in Figure 1.2 are coded as (time-neutral) feelings and descriptions respectively.

relating to the macro-structures in the texts, for reasons that will be explained shortly. Nonetheless, we undertook the coding based on Egbert et al. (2021) of our corpora to ensure that our analysts could divide the corpora into discourse units, based on an understanding of the nature of the annotation they were undertaking.

To prepare our texts for analysis, using the coding framework a team of coders read and annotated transcripts in terms of the discourse units for all three corpora used in this book.[13] The coders were trained, undertook sample annotations in pairs in which they critiqued one another's work and re-edited transcripts as later decisions arising from double coding arose. Throughout, one of the authors of this book (McEnery) oversaw the annotation process, including training, and acted as arbiter where coders required guidance in reaching a decision.[14] While we used the functional coding of Egbert et al. as a framework to guide analysts' decisions, we will not discuss that framework here because, as noted, our overall approach will be bottom up. We made that decision because we wanted one method that would allow us to functionally categorise the macro-structures and the uppermost level of micro-structure, the turn, so that we could look at interplay between the two. As the Egbert et al. scheme is designed for the macro-structural level only, we set that coding aside for our purposes, helpful though it was for guiding the process of annotation.[15]

The coding process was time-consuming and, as a consequence, while we annotated the whole of the TLC L1, for the TLC and the BNC we only annotated a subset of the corpus, albeit a substantial one in each case. This means that for the discourse unit analyses in this book we use 901,085 words of the TLC L1, 3,265,194 words of the Spoken BNC 2014 and 1,737,822 words of the TLC.

The decisions made so far have consequences for our analysis, of course. We have identified our goal, an exploration of discourse at the uppermost macro-structural and micro-structural levels in three corpora selected to

[13] The coding was led by Tony McEnery, who worked with coders employed to undertake the coding. The coders were trained and each coder had 10 per cent of their analyses checked for plausibility (see Egbert et al., 2021: 728). If this led to revisions, the analyst then went back over previous files they had annotated to try to ensure consistency of analyses across the corpora. The authors would like to thank the analysts who worked on this task and whose work is used in this book: Alexandra Antonova, James Balfour, Kevin Gerigk, Josephine Gwizdala, Abi Hawtin, Beth Malory, Poppy Plumb, Gillian Smith, Vera Stepanyan, James Wright and Liying Zhou. Additional help was provided by Isobelle Clarke, Marius Henius Dreyer and Sam Hollands.

[14] The process followed was that further described by Egbert et al. (2021) and McEnery et al. (2023).

[15] Where the underlying Egbert et al. (2021) coding is of some relevance for the analysis presented, we will include that, for the interested reader, in a footnote.

allow for a meaningful comparison. Our goal in doing this is to cast light upon what we think is an important, but neglected, research area. The neglect arises in part from the lack of suitable resources to explore macro-structures – we have remedied this by applying a system designed to identify macro-structures in the form of discourse units in all three corpora. Rather than use the functional analysis introduced in that coding, however, we wish to develop, by a focus on form–function relations, a way of deciding the functions both of the micro-structures and macro-structures studied. However, in making these choices we have also effectively eliminated the possibility of a meaningful quantitative exploitation of much of the metadata in our corpus – the aggregation of micro-structures into macro-structures, while linguistically well motivated, also reduces the number of data points, i.e., the objects that we may observe and count in our data. How we will deal with this will be returned to towards the end of the chapter. Before considering that, we need to address a substantial problem which our decisions so far have caused – what technique to use to derive the functions of both micro- and macro-structures in our data. That is the focus of our next section.

1.6 Language, Form and Function: The MDA Approach

The technique that we will use to attempt to characterise both discourse micro-structure (turns) and macro-structure (discourse units) will be based on MDA. This sets a stern test for the technique. Both the micro-structures and macro-structures we want to account for are small, certainly small enough to engage with the problems of MDA outlined earlier. In this section we briefly investigate MDA to show its overall suitability for the task and then consider how we address the issues with text length already discussed.

MDA is an approach rooted in the analysis of register. The perspective of register combines analysis of the linguistic characteristics in a text variety with analysis of the real-world situations in which that variety is used. The linguistic analysis of register is thus underpinned by the assumption that linguistic features are functional, and that the linguistic features which characterise a particular register are associated with its communicative purposes and situational contexts. Registers can therefore be considered groupings of texts that are defined by factors that are external to the texts themselves, such as the social or situational conditions of their medium, their contexts of production or their purpose – making the perspective ideal, in principle, for our discourse-oriented analysis. According

to Biber and Conrad (2009: 6), the description of a register covers three major components: 'the situational context, the linguistic features, and the functional relationships between the first two components'. They elaborate:

> Registers are described for their typical lexical and grammatical characteristics: their linguistic features. But registers are also described for their situational contexts, for example whether they are produced in speech or writing, whether they are interactive, and what their primary communicative purposes are. […] [L]inguistic features are always functional when considered from a register perspective. That is, linguistic features tend to occur in a register because they are particularly well suited to the purposes and situational context of the register. Thus, the third component of any register description is the functional analysis. (Biber and Conrad, 2009: 6)

MDA is a corpus-based text linguistic approach for identifying the major patterns of linguistic variation across a corpus of texts. The approach was pioneered by Biber (1984, 1985, 1986, 1988), who set out to examine variation across spoken and written English registers. He was influenced by theoretical discussions (e.g. Ervin-Tripp, 1972; Hymes, 1974; Brown and Fraser, 1979) which emphasised the importance of considering linguistic co-occurrence for studying registers and understanding the functional differences between them (Biber, 2019). MDA is thus grounded in the notion of linguistic co-occurrence; that is, that frequent patterns of co-occurring linguistic features tend to reveal at least one shared underlying communicative function (Biber, 1988). The assumption that underpins the analysis of linguistic co-occurrence therefore dictates that if two texts, for example, exhibit similar patterns of co-occurring lexico-grammatical features, this similarity is not random. Rather, the two texts are viewed as sharing at least one underlying communicative function which influences the shared selection of linguistic features. In this book we extend that reasoning to discourse units – those that are functionally equivalent should have a shared selection of linguistic features.

The aim of MDA, then, is to identify the major patterns of linguistic co-occurrence across a corpus of texts. MDA is driven by a lexico-grammatical account of register variation, since Biber's argument is that registers are formed by distinct combinations of lexico-grammatical categories. Biber (1988) uses sixty-seven of these, which are grouped into sixteen broader categories including, *inter alia*, tense and aspect markers, place and time adverbials and pronouns and pro-verbs (see also Conrad and Biber, 2001: 18–19). To give an example of how this approach helps to identify groups of linguistic features relevant to identifying a register, Biber (1988) observed

a pattern whereby texts with a high frequency of, among other things, private verbs (e.g. *believe*, *think*) are also likely to exhibit a high frequency of *that*-deletion and contractions, as well as the lower frequency of such features as nouns, prepositions and attributive adjectives. Texts displaying these patterns were functionally interactive and involved and prototypically came from the casual conversations register. The lexico-grammatical features that are analysed as part of the MDA approach are thus functionally related and combine in different ways to form distinct 'dimensions' – that is, 'sets of syntactic and lexical features that co-occur frequently in texts' (Biber, 1989: 5).

Dimensions in MDA are interpreted and labelled in terms of their perceived functions. MDA-based analyses are thus concerned with identifying the functions of texts and registers – or in our work, discourse units and turns – by exploring the dimensions along which they are placed, based on the frequency or infrequency of particular lexico-grammatical co-occurrence patterns. In other words, based on their co-occurrence in the texts of the corpus, the linguistic features bundle to create a number of dimensions reflective of these co-occurrence patterns, and the registers, or discourse units, place themselves in distinct configurations along those dimensions based on their exhibition of these patterns.

In practical terms, MDA first involves tagging each text in a corpus for a variety of lexico-grammatical features. The relative frequencies of these lexico-grammatical features in each text are then measured and subjected to the multivariate statistical technique known as factor analysis. Factor analysis is used to identify underlying or latent variables in a dataset by finding variation in observed and correlated variables. On this basis, factor analysis returns a smaller number of dimensions which reflect the most common patterns of co-occurring variables, with each dimension representing another pattern of covariation. Based on the notion of linguistic co-occurrence, in MDA these dimensions of co-occurring linguistic features are analysed and interpreted in terms of their underlying communicative function. Each factor has a weighted combination of all the linguistic (i.e. lexico-grammatical) features, where each linguistic feature has some weight for each factor. Each linguistic feature's weight is its loading, where loadings range from −1 to +1 for each factor. The loading indicates the amount of shared variance with the total pool of variance. The strength of the loading represents how associated it is to the factor. Loadings that are closer to 0 tend to be ignored, as these do not really influence the factor. Loadings that are assigned scores closer to −1 or +1 are given prominence, as the variables assigned high weights are relevant

for the factor. In other words, the loadings show which linguistic features tend to co-occur with each other most frequently. Most features load strongly on the early factors, and features can also load strongly on more than one factor. Note that, for our purposes, this technique, applied to discourse units, will forge a direct link between macro-structures and low-level micro-structural features, in this case lexis and morphosyntax. Likewise, if applied at the turn level, we can use this technique to tie that uppermost level of micro-structure to such lower level features. So, from the perspective of the goals of our analysis, MDA is, in principle, very promising.

Following this quantitative statistical analysis, the next step in MDA is interpretative and functional, as the dimensions of aggregated linguistic co-occurrence patterns are interpreted by a human analyst. Specifically, the analyst must decide which linguistic features to take into account for each dimension, based on the strength of the loadings. Most MDA studies consider factor loadings above 0.3 as strong. The analyst also computes factor scores for each text. These indicate how associated each text in the corpus is to the particular patterns of linguistic co-occurrence captured by a dimension. Based on the notion of linguistic co-occurrence, the linguistic features with strong loadings for each dimension returned by the factor analysis are subsequently interpreted along with the texts displaying these patterns (i.e. those with high factor scores) for the underlying communicative function. The interpretations of the dimensions thus need to capture the function that is shared by the co-occurring features.

Each dimension consists of a positive pole and a negative pole. Each pole is associated with a set of co-occurring linguistic features that are in complementary distribution with the set on the opposite pole. Each dimension is conceptualised as a continuum of variation, where the set of co-occurring linguistic features on one side is more associated with a given function and the set of co-occurring features on the other end of the pole is less associated with that function. Thus, when it comes to interpreting dimensions, the label the analyst assigns to a dimension must capture the function that explains the difference between the two sets of co-occurring features. Using this method, in one of the first large-scale studies examining variation across spoken and written English, Biber (1988) proposed six dimensions of linguistic variation to which he assigned the following functional labels: Dimension 1: Informational versus Involved Production; Dimension 2: Narrative versus Non-Narrative Concerns; Dimension 3: Explicit versus Situation-Dependent Reference; Dimension 4: Overt

Expression of Persuasion; Dimension 5: Abstract versus Non-Abstract Information; and Dimension 6: Online Informational Elaboration. These are introduced in the following to illustrate the reasoning underpinning the analysis and to show how the micro and macro mesh in the functions that the dimensions represent.

Dimension 1: Involved versus Informational Production: in this dimension, the linguistic features with negative loadings mark high informational density and specific informational content (nouns, prepositions, attributive adjectives, type/token ratio, word length etc.), whereas the linguistic features with positive loadings are less specific and mark generalised content. Features with positive loadings are also more affective and interactive (e.g. first- and second-person pronouns, private verbs, demonstrative pronouns, contractions and WH-questions (*what*, *why*, *where*, *who* etc.)). This dimension not only reflects the primary purpose of the author/speaker, but also the circumstances in which the discourse is produced. For example, informational texts tend to have complex structures and dense noun phrase modification which are particularly likely to occur in texts where discourse producers have time to edit and select precise lexis. On the other hand, interactive texts, which are influenced by real-time constraints, tend to exhibit less precise lexis and a higher density of pronouns and contracted forms.

Dimension 2: Narrative versus Non-Narrative Concerns: on the one hand, linguistic features with positive loadings in this dimension mark narrative concerns in that they function to mark past time, third-person animate referents, reported speech and depictive discourse (e.g. third-person pronouns, past tense verbs, perfect aspect and public verbs). On the other hand, the linguistic features with negative loadings in this dimension are non-narrative as these function to mark immediate time and a more frequent elaboration of nominal referents (e.g. present tense verbs, attributive adjectives, past participle whiz-deletions).

Dimension 3: Explicit versus Situation-Dependent Reference: linguistic features with positive loadings in Dimension 3 mark exophoric references that are highly explicit and context-independent (e.g. WH-relative clauses on object and subject positions, pied-piping constructions, nominalisation and phrasal coordination). The features with negative loadings mark endophoric, non-specific and situation-dependent references (e.g. time and place adverbials, adverbs).

Dimension 4: Overt Expression of Persuasion: this dimension only has linguistic features with positive loadings, and these features function to persuade the addressee. Features which characterise this dimension include

those which indicate the speaker's point of view or which are used to assess the advantages or likelihood of an event (e.g. infinitives, prediction modals, suasive verbs, conditional subordination, necessity modals, split auxiliaries, possibility modals).

Dimension 5: Abstract versus Non-Abstract Information: the linguistic features with positive loadings on this dimension are used to indicate informational discourse that is abstract, formal and technical. These include, for example, conjuncts, passives, adverbial subordinators, past participle clauses and WHIZ deletions, and predicative adjectives. The negative loadings are used to mark other kinds of discourse.

Dimension 6: Online Informational Elaboration: in this dimension, linguistic features with positive loadings function to mark fragmented informational elaboration that is relatively spontaneously produced, especially under strict real-time constraints (e.g. *that* clauses as verb and adjective complements, *that* relative clauses and WH-relative clauses on object position, final preposition, existential *there*, demonstrative pronouns). Linguistic features with negative loadings, meanwhile, are associated with informational integration (e.g. phrasal coordination).

A key to understanding MDA is to realise that texts are not allocated to only one dimension. When carrying out MDA, each text in a corpus is simultaneously scored for each dimension using standardised counts of the relevant features (see Biber, 1988). This means that each text will be assigned a score for each dimension. For example, Table 1.2 (adapted from Biber, 1988: 125) provides the mean factor scores for all texts in the face-to-face conversations register that were included in Biber's (1988) study of spoken and written English registers. This table shows that a typical face-to-face conversation has a high positive Dimension 1 factor score, a low negative Dimension 2 and 4 score, a moderate negative Dimension 3 and 5 score, and a low positive Dimension 6 score.

Dimension 1, 'Involved versus Informational Production', comprises twenty-five features with high positive scores, including, as noted, the use of private verbs, that-deletion, contractions, present tense verbs and second-person pronouns; meanwhile, features with high negative loadings include nouns, word length, prepositions and attributive adjectives. In this case, if texts comprise, on the one hand, the frequent co-occurrence of the features along the positive pole, then they will have a high positive Dimension 1 factor score and will be interpreted as having an 'involved production' communicative function. On the other hand, the co-occurrence of features along the negative pole indicates a shared communicative function of 'informational production'. Because

Table 1.2 *Mean factors for all texts in conversation in Biber (1988).*

Dimension	Mean	Minimum	Maximum	Range	Standard deviation
1	35.3	17.7	54.1	36.4	9.1
2	−0.6	−4.4	4.0	8.4	2.0
3	−3.9	−10.5	1.6	12.1	2.1
4	−0.3	−5.2	6.5	11.7	2.4
5	−3.2	−4.5	0.1	4.6	1.1
6	0.3	−3.6	6.5	10.1	2.2

the features along the positive and negative poles tend not to occur with similar frequency within the same texts, the presence of the features of one – in this case, 'involved production' – usually indicates that the features of the other – that is, 'informational production' – are largely absent. Dimension 4 has features that only load strongly on the positive pole, meaning this dimension is characterised by the frequency or absence of a single set of linguistic features. This means that, in Dimension 4, texts can be characterised either as being more or less associated with overt persuasion. Analysts typically proceed by calculating the mean dimension scores for each of the registers represented in their data (as in Table 1.2), leading to the characterisation of registers in terms of the aforementioned dimensions. The characteristics of individual registers become more salient, and are rendered more apparent, when their mean dimension scores are compared against each other, essentially illuminating the most salient linguistic characteristics of each one.

Whilst MDA began as an approach to investigating variation across spoken and written language in English, it has since been applied to the investigation of an ever-widening range of languages and specialised discourse domains, where the patterns of register variation originally put forward by Biber (1988) have proven to be a useful starting point for analyses. For example, MDA has been used to examine variation across registers in languages such as, *inter alia*, Nukulaelae Tuvaluan (Besnier, 1988), Somali (Biber and Hared, 1992, 1994), Korean (Kim and Biber, 1994), Taiwanese (Jang, 1998), Dutch (Grieve et al., 2017), Brazilian Portuguese (Berber Sardinha et al., 2014), Gaelic (Lamb, 2008), Spanish (Asención-Delaney, 2014) and World Englishes (Xiao, 2009; Bohmann, 2017).

Although language-wide studies have been carried out using MDA, the majority of research utilising this approach focuses on language use in specific contexts, producing accounts of the registers that are characteristic

of domains as diverse as schools (Reppen, 1994) and universities (Biber, 2006), academic research articles (Gray, 2013), televised dialogue (Quaglio, 2009), call centre interactions (Friginal, 2009), job interviews (White, 1994), pop song lyrics (Bértoli-Dutra, 2014), medical encounters (Staples, 2015), legal texts (Goźdź-Roszkowski, 2011), newspaper editorials (Westin and Geisler, 2002), extremist texts (McEnery and Brookes, 2022) and, of particular relevance to this book, casual conversation (Biber, 2004), as well as spoken and written exam responses by learners of English as an L2 (Biber and Gray, 2013). As well as addressing a variety of domains, MDA has also been utilised in the study of register across specific time periods, for example Biber's (2001) study of written and speech-based registers in the eighteenth century and Egbert's (2012) study of fictional novels written in the nineteenth century.

MDA is thus a powerful method for investigating large corpora of language in use. It has enabled research on discourse in a range of languages and taken place in a variety of contexts to provide rich descriptions of language in use, documenting how language users frequently make certain lexical and grammatical selections in particular contexts and situations in order to achieve particular purposes. An important finding issuing from such studies is that Biber's (1988) first (Informational versus Involved Production) dimension, and often the second (Narrative versus Non-Narrative) dimension, have been found consistently across various languages and discourse domains, supporting the notion that these may be universal dimensions in language use (Biber, 2014, 2019). Another feature of this body of research is that many studies compare English registers to the dimensions of spoken and written English proposed by Biber (1988) (discussed earlier). These studies enable rich, comparative descriptions of registers that were not included in Biber's (1988) study with the ones that were (see Berber Sardinha et al., 2019).

Applying the reasoning behind MDA to the level of the discourse macro-structures in a way that integrates a micro- and macro-structural view is the crux of the work presented in this book. In taking the original insight of Biber and applying it to a slightly different domain and text unit we follow a growing tradition in corpus-based research where MDA has been applied to texts representing an ever-widening range of languages and contexts.

Despite its popularity, like any methodological approach MDA has shortcomings. One of these, discussed briefly earlier in the chapter, concerns text length. This is a fundamental problem for our work. For MDA to be effective, the texts subjected to it must be sufficiently long to

allow for the relative frequencies of the relevant grammatical forms to be accurately estimated. In other words, the shorter the texts being analysed are, the likelier it is that certain grammatical forms will not be sufficiently represented to permit accurate estimation of their relative frequencies. MDA is underpinned by measurements of relative frequencies (as opposed to raw frequencies) so that texts of differing lengths can be compared and analysed together without the generally higher frequencies of words in the longer texts confounding the analysis. On the one hand, Biber (1993) suggests that the relative frequencies of most forms can be inferred accurately on the basis of text samples that are at least 1,000 words in length. On the other hand, Passonneau et al. (2014) found text samples of 500 words to be sufficient in length to allow the accurate estimation of the relative frequencies of the features in their feature set. Where texts are shorter than this, however, a problem arises with normalised frequencies, as discussed already. The macro-structures we will look at are shorter than 500 words typically. The turns we will look at when considering micro-structures are shorter still. While the specific length of the text required to perform MDA varies from study to study, the general consensus is that the rarer the features included in the feature set, the longer the texts need to be (Passonneau et al., 2014). Passonneau et al. also noted that, relying on frequent features, the texts could be, as reported, around 400–500 words and still be analysed by MDA. However, Biber (1993) was using a feature set much more similar to ours. This includes low-frequency grammatical features and thus requires texts above 1,000 words in length. Our macro-structures are typically never of this length and if we push the analysis down to the micro-structure level and analyse turns, the size of the units to be analysed is well below the thresholds suggested by either Biber or Passonneau and colleagues.

For these reasons, MDA studies have generally restricted their focus to texts that are at least 500 words in length (although this can also depend on the features under examination). For our study, in which we seek to explore micro- and macro-structures, this represents a significant problem.

An approach to this problem – concatenation – is one we reject. Concatenation was used by Passonneau et al. (2014). They concatenated short texts to achieve texts sizes of 500 words. If we took that approach in this book we would, in effect, be concatenating shorter units (micro- and macro-structures of discourse) into relatively arbitrary chunks that had only one property of interest – that is, they would be the right size for the analysis we wished to undertake. Their capacity to reveal the functions of individual discourse units, for example, would be lost. One approach we

could consider is to look at the tasks (e.g. Connor-Linton and Shohamy, 2001), for example, in the TLC, and analyse them. Some of them exceed 500 words. However, our discourse unit annotation shows that beneath the level of the task there is meaningful variation – functions can vary *within* a single task, shifting from one discourse unit, or even one turn, to the next. It is that variation that we wish to capture in our analysis. While we would like to gain a perspective on the relationship of discourse units to tasks, we want to understand tasks through their relationship to their constituent discourse units. Thus, focusing on the task, while a legitimate approach, would not yield the kinds of insights we wish to gain. Therefore, while concatenation may solve a mathematical problem, it would do so at the expense of a more refined – and, we would argue, more meaningful – view of the data.

Another approach to the problem is to remove short texts from the data (e.g. Liimatta, 2020, 2022). This approach has been advocated in cases where short texts constitute only a small fraction of the dataset overall, as this is assumed to mean that their removal will not impact on the results unduly. This is not an approach we can take, as all of the units we want to look at are short in length.

Given that existing approaches to dealing with short texts when carrying out MDA will not work for us, we need to use another method – one which permits the study of small text sequences rather than one which aggregates them blindly or involves simply deleting them. It is for this reason that we turn to short-text MDA.

1.7 Short-Text MDA

In response to this issue of text length, in this study we retain the ability to focus on individual micro- and macro-structures by adopting an approach which has been devised to overcome both the short-text limitation of MDA as well as the limitations of the afore-discussed approaches that have been put forward to work around this. This approach is known as 'short-text MDA'. It is the approach taken for the study of spoken interaction in L1 and L2 speech in this book. Firstly, each text is tagged for a variety of lexico-grammatical features using tagging and annotation programmes. These features are selected to represent the language variety or varieties under investigation and include, amongst others, tense and aspect markers, general and specialised verb categories, and different kinds of adjectives and adverbials. The features used in this study vary slightly from Biber's and are based on the work of Clarke (2018; 2020).

These are shown in Appendix A. Secondly, using another programme, the occurrence of each linguistic feature in each text is recorded in a data matrix. Following this, using a statistical software package, this data matrix is subjected to a multivariate statistical technique to uncover the relations amongst linguistic features and texts. We will introduce the technique used shortly. With short-text MDA, as is the case with MDA, in the statistical analysis the measured variables are the linguistic features and, based on the notion of linguistic co-occurrence, the latent variables are the communicative functions that are influencing the linguistic co-occurrence patterns. Thus, the analysis is used to reveal a series of dimensions comprising the most common patterns of co-occurring linguistic features across the texts of the corpus. Based on the notion of linguistic co-occurrence, these dimensions of co-occurring linguistic features are then interpreted to determine the underlying communicative functions associated with the distributional patterns.

The procedure described so far is very similar to MDA itself. However, what is processed is different and how it is analysed statistically is different. The system does not record in the data matrix relative frequencies of lexico-grammatical features as standard MDA does. Rather, short-text MDA measures the *occurrence* of features (i.e. whether they are present or absent). This is what populates the matrix analysed. Working with presence and absence in turn influences our choices with regards to factor analysis, as factor analysis does not work with categorical data, but instead works with continuous data. Thus, in short-text MDA, rather than using factor analysis which is used for MDA, information pertaining to the presence or absence of features across the texts of a corpus is subjected to multiple correspondence analysis (MCA), which is used in a way that is similar to how factor analysis is used in traditional MDA; that is, to return dimensions comprising the most common patterns of co-occurring linguistic features across the texts of a corpus. MCA is good for our purpose as it is a geometric data-analytic method which allows the identification and visualisation of the most dominant relationships between three or more categorical variables in a low-dimensional space. The method was popularised by Jean-Paul Benzécri, who used it to analyse sociological data from questionnaires (Benzécri, 1979). This is because MCA can be used to observe relationships between individuals in a sample, as well as the relationships between variables. For example, for the analysis of questionnaire data, Benzécri used MCA to understand the relationships between individuals, in terms of respondents who answered questions similarly or dissimilarly, and relationships between variables, in terms of

understanding which answers tended to be selected together and which answers were rarely selected together. MCA visualises the relationships between individuals and variables in terms of distance and produces two clouds of points, where the points on one cloud represent the individuals and the points on the other cloud represent the categorical variables. The distance between each point is based on how similar they are with respect to their distribution. For example, for Benzécri's questionnaire data, points representing people are situated more closely together in the space if people give the same responses to the questions. Meanwhile, points representing responses are placed closer together if they distribute similarly across the respondents. Therefore, if many respondents select the same responses, those responses will be placed closer together in the space. In addition to analysing data from surveys or questionnaires (Greenacre and Pardo, 2006), MCA has been used in a range of exploratory studies, including those concerned with the identification of factors contributing to motorcycle crashes (Jalayer and Zhou, 2017), different tastes (Le Roux et al., 2008; Le Roux and Rouanet, 2010), different patterns of cultural consumption (Kahma and Toikka, 2012), patterns of ageing (Costa et al., 2013) and for linking crimes (Yokota et al., 2016). MCA has also been used in a small number of linguistic studies, mainly to identify confounding variables (Tummers et al., 2012) and to identify patterns of usage of polysemic words (Glynn, 2009).

MCA is used in short-text MDA much like factor analysis is used in standard MDA – that is, to identify the major patterns of linguistic co-occurrence across the texts of a corpus. However, MCA is better suited to the short-text MDA approach than factor analysis is, as the former deals with categorical data (e.g. the presence or absence of a feature), whereas the latter requires continuous data (e.g. the relative frequency of a feature). In practical terms, within short-text MDA, MCA is used to identify the major sets of co-occurring linguistic features in the texts of a corpus, as well as to identify the texts that are associated most strongly with these patterns of co-occurrence. As in standard MDA, the patterns of co-occurrence are then analysed within their wider textual contexts (i.e. in the contexts of the texts that are most strongly associated with them) by the human analyst. The objective of this analysis is to identify the underlying communicative functions that are fulfilled by the co-occurring features in question.

To date, short-text MDA has been applied mainly to the analysis of tweets (Clarke, 2019; Clarke and Grieve, 2019). These studies have demonstrated the utility of the short-text MDA approach for analysing the major

communicative functions in texts or text units which, due to their short size, would have otherwise evaded analytical focus. In this book, we apply the approach in our analyses of L1 and L2 spoken English. Chapters 2–7 of this book, as well as representing an exploration of discourse micro- and macro-structures in L1 and L2 speech, are also an exploration of the extent to which short-text MDA can facilitate such an analysis. Before we commence the process of investigating the use of short-text MDA to investigate our corpora, however, we will briefly introduce a macro-structure which is one of those which is readily identifiable and well studied by linguists – narrative.

1.8 Narrative

As the analyses in this book develop, a focus will be formed around narrative – what might informally be called 'storytelling' but might more accurately be called sequences, rooted typically in the past and exhibiting a flow of time across the narrative, where some focus of the narrative is introduced, actions or events being relayed within the narrative proceed, and a resolution of the narrative is reached. Given the central role of narrative to the analyses in the latter part of this book, we will introduce it only briefly here – a much fuller introduction to narrative is needed and that is provided in Chapter 8. However, what should be noted here is that our focus on narrative in this book does not arise from a pre-determined wish to focus on narrative. Rather, it is brought about by the analysis of the language in the TLC, TLC L1 and Spoken BNC 2014. In this book, studies flow from one to the other, with the nature of the study in one chapter leading on, to some extent, from the findings arising from our exploration in the previous chapter. Hence, our focus on narrative in Chapter 8 emerges from the analyses presented over the preceding six chapters. While it may be tempting to call this a 'corpus-driven' approach (see Tognini-Bonelli, 2001), we view it rather in the context of the searchlight paradigm of research (McEnery and Brezina, 2022: 93). By starting our focus on the data in one study with our notional searchlight, we see things in our data and produce findings on the basis of those observations which lead us to move the searchlight to other areas and even other datasets. It is the rational process of discovery that determines the movement of the searchlight, though we admit that at the end of Chapter 7 at least, we could in principle have shone our searchlight in a number of places. Yet the scale and variety of narrative use across the corpora forces our focus. For now,

we simply note that narrative will become a key focus of the book and that we will return to it later.

We will proceed now with a test of the approach to short-text MDA outlined here. In the next chapter we test the approach by focusing on micro-structures – providing a robust test of the technique by providing it with very little data on which to work.