Corpus data and methods

Candarli, Duygu

Published in:
The Linguistic Challenge of the Transition to Secondary School

DOI:
10.4324/9781003081890-3

Publication date:
2022

3 Corpus data and methods

Duygu Candarli

Introduction

This chapter overviews the corpus linguistic data and methods used in the studies described in the remaining chapters of this book. Corpus linguistics is a methodology that analyses collections of language data known as corpora, which are normally too large to read in full and search by hand using traditional text analysis procedures. They are compiled in a principled way in an attempt to represent a language or variety of language (Baker, 2006). Corpus linguists use a range of computational techniques to examine recurrent linguistic patterns in these corpora. The advantage of corpus linguistics is that it allows the analyst to access reliable information on the frequency and nature of linguistic patterns in the corpora, which would be not possible through intuition or the use of a small number of text extracts (McEnery & Hardie, 2012). There are many corpora in existence, some very large ones attempting general coverage of a language and, increasingly, more specialised ones, enabling study of a variety or register.

In the field of education, a number of corpora of spoken and written university registers have been built (e.g., Biber et al., 2002; Biber, 2006; Nesi & Gardner, 2012; Römer & O'Donnell, 2011; Thompson & Nesi, 2001). These have been used to describe both the spoken and written language used in higher education, especially at English-medium universities. A growing body of the literature that used such corpora has made aspects of the language of university visible for instructors of English for academic purposes, content lecturers and materials writers. This can support first-year students to develop the skills required to understand a range of university registers and succeed in transitioning from school to university, and master's students from international backgrounds, among others.

At the school level, fewer corpora have been built, and none that we are aware of that cover the transition from primary to secondary school, but there is a small number representing other aspects of schooling. Durrant and Brenchley (2019) created a corpus of writing produced by students in Years 2, 6, 9 and 11 – that is, the ends of Key Stages 1–4 – in the subjects of English, science, history, geography and religious studies, which they used

DOI: 10.4324/9781003081890-3

to study the development of children's vocabulary use. Their corpus comprises a large number of texts, 2898, and contributors, 983 children, and it remains unique in representing pre-university students' writing at schools in England, but nonetheless, it is not large. The median token count of Year 2 texts is approximately 63, increasing across the years of data collection points (Durrant & Brenchley, 2019).

School textbook corpora are easier to compile in volume. Coxhead and White (2012) compiled corpora of textbooks used for English, science and social studies (a subject incorporating socially relevant themes from subjects such as history, geography and economics), at secondary schools in New Zealand, to create a relatively large corpus of 1,211,373 tokens. This is perhaps slightly unbalanced however, as more than half the tokens, 751,638, are from fiction registers in the sub-corpus of English textbooks. Green and Lambert (2018) built a corpus of 16,253,350 tokens from secondary school textbooks from the Singapore national syllabi. They used this to develop subject-specific word lists for eight secondary school subjects. Greene and Coxhead (2015) built a corpus of 18,202,382 tokens of middle school textbooks used by state school students in the United States. Middle school covers students aged approximately 10–14 years. They used their corpus as the basis for subject word lists, following Coxhead's methodology for the New Academic Word List (Coxhead, 2000). This focus on textbooks in school corpora is at odds with corpora of the language of university, which as well as student writing (Römer & O'Donnell, 2011; Nesi & Gardner, 2012), cover talk in lectures and other university speech registers (Thompson & Nesi, 2001; Simpson et al., 1999). The TOEFL 2000 Spoken and Written Language (T2K-SWAL) corpus (Biber, 2006) covers an extensive range of spoken and written university genres including course packs, course management and institutional writing texts as well as textbooks.

There are several possible reasons for the relatively small number of corpora and corpus studies of school language, and the limited number of registers that have been collected. Researchers tend to be university-based, often within language support centres, or with close links to them, so corpus research into the language of university study directly supports their teaching and students. They may also have ready access to texts from their own or co-researchers' institutions and may be able to discuss materials with discipline experts. By contrast, collecting school data requires making contacts across different educational cultures, often with more complex ethical considerations, as school students are not adults. Once identified, texts are less easy to prepare for corpus work. Although online copies of school textbooks are often available, they may not be straightforward to convert into data that can be accessed using corpus software, due to their presentation, with numerous boxed charts and figures, and embedded graphics. Collecting other kinds of data in schools can be even more challenging: it is time-consuming and resource-intensive to collect

and prepare teacher worksheets and PowerPoint presentations for corpus analysis. Spoken classroom data are very difficult to obtain, for practical and ethical reasons, and time-consuming to transcribe. A further issue, seen above in Durrant and Brenchley's (2019) study, is the small token counts of many school language texts and registers. For example, a mean token count (running words) of mathematics worksheets at Key Stage 2 in our corpus is only 240 tokens. This is symptomatic of one of the difficulties of compiling corpora of school-level texts, particularly at the younger end of the age range.

In the following sections, we describe how we constructed our corpus and the content and size of the final database. We then discuss our written and spoken corpora and the various techniques we used to analyse them.

Constructing our corpus

Characteristics of our partner schools

The data for our corpora were provided by our 13 partner schools, which we mentioned briefly in Chapter 1. We approached Huntington School in York, which at the time, in 2016, was one of five UK schools in the newly formed Research Schools network (https://researchschool.org.uk/). (The network has since developed considerably, at the time of writing comprising 28 Research Schools and ten Associate Research Schools (EEF, 2022)). Research Schools are state schools which have applied for and gained Research School status through a competitive process. The aim of Research Schools is 'to lead the way in the use of evidence-based practice and bring research closer to schools' (EEF). Their brief does not include being involved in primary research such as our project, but as we had hoped, the staff were enthusiastic to participate, and the Literacy Lead teacher agreed to be a consultant for the project (Jones & Deignan, 2021). His collaboration was invaluable at all stages, from recruiting additional schools to the project, through data collection and discussion, to disseminating findings through networks of education professionals. We approached a number of other schools known to us through our teacher training and university networks and visited schools that expressed an interest; 13 schools eventually participated and were paid an honorarium for their participation. Of the 13, eight were primary schools and five were secondary. Five of the participating primary schools directly 'feed' three of the secondary schools. That is, most or all of the students from the primary school move together to the same secondary school for Year 7. Secondary schools are considerably larger than primaries and may have around six or eight feeder primaries, with some other students coming from other primary schools in addition to the feeder schools. The relationships between our partner schools are represented in Figure 3.1.

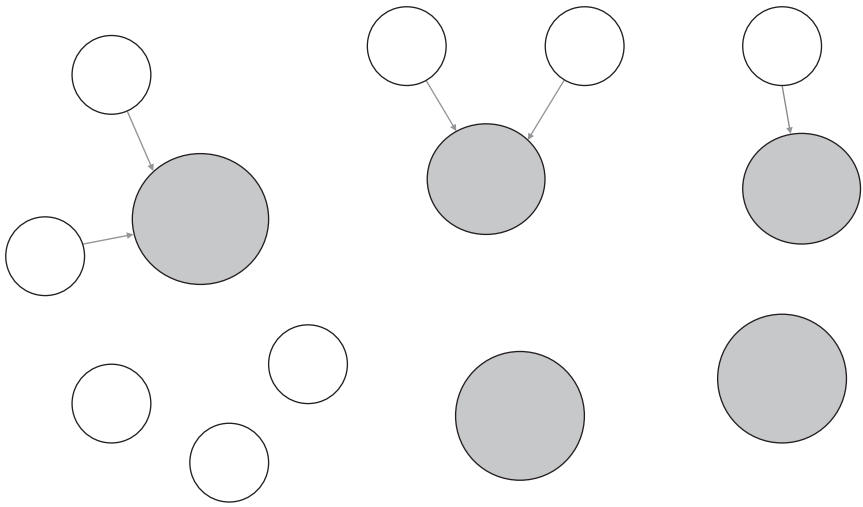

Figure 3.1 Relationships between partner schools.

The large, grey circles represent secondary schools, and the smaller clear circles, primary schools. Arrows show where primary schools feed secondary schools. It can be seen that there are three small clusters of primary and secondary schools and five schools which have no connections to other partner schools. They are geographically dispersed across Yorkshire and the North East and include inner city, suburban and rural schools. All are state funded and non-selective.

All of the schools provided written and spoken data towards the corpus. As mentioned in Chapter 1, we also interviewed groups of students and teachers. These were from the five primary and three secondary schools that are part of clusters. We spoke to the students, six from each of the five primary schools, when they were in Year 6, and then after they had moved to Year 7, secondary school. The interview data is discussed very briefly in Chapter 1 and in more detail elsewhere (e.g., Chambers, 2020).

The characteristics of the schools are outlined here in terms of external measures as follows:

- The most recent ratings from the Office for Standards in Education, Children's Services and Skills (Ofsted) at the time of data collection;
- Eligibility for free school meals, a characteristic of the student population;
- Academic scores: for primary schools, we used the pupil progress score in reading, writing and mathematics, and for secondary schools, the Progress 8 score.

For state-funded schools in England and Wales, Ofsted inspectors make judgements on the following four areas:

(1) effectiveness of school management;
(2) quality of education provided;
(3) personal development of pupils;
(4) outcomes for pupils.

Inspectors use a four-tier rating scale: 'outstanding', 'good', 'requires improvement' and 'inadequate' (Ofsted, 2018).

Free school meals (FSM) eligibility is based on a low family income. FSM eligibility is widely correlated with potential disadvantage as well as attainment levels of pupils and schools (Gorard, 2012). We used FSM data from 2017–2018, the most recent data available at the time of the corpus compilation.

Pupil progress scores are concerned with the progress that pupils make between the end of Key Stage 1 and the end of Key Stage 2 and are used in assessing and comparing the performance of primary schools. They are calculated by comparing pupils' KS2 assessment and test results at one school with those of other schools' pupils at the national level. A score of 0 means that the students in the school perform at the same level at the end of KS2 as students with the same KS1 attainment nationally. Positive and negative scores indicate that students in the school make above average or below average progress respectively, relative to students nationally. In secondary schools, the Progress 8 score refers to the progress made between the end of Key Stage 2 and the end of Key Stage 4. It is based on GCSE results in up to eight qualifications, which include the core subjects of English, mathematics, sciences, history and geography. As for the KS2 pupil progress score, a score of 0 means that students have progressed in line with others with the same prior attainment nationally, and positive and negative scores indicate progress that is better or worse than comparable students nationally.

We gave the schools codes (school_a, school_b, school_c, etc.) to ensure their anonymity. Tables 3.1 and 3.2 show their characteristics.

As seen in Table 3.1, all the primary schools in our sample were rated 'good' at their most recent inspection at the time of data collection. Although there was no variation between the primary schools' ratings in our sample, this closely reflected the rating of the majority of the schools at the national level, since 69% of all primary schools were rated 'good' (Ofsted, 2018). At Key Stage 3, three different categories are represented in our sample, and the mean of the Ofsted ratings corresponded to 'good'; 53% of the secondary schools are rated 'good' nationally (Ofsted, 2018). It should be noted that Ofsted ratings remain highly controversial (see Perryman et al., 2018 for a discussion) and that they can only provide crude information about the overall effectiveness of schools.

Table 3.1 Characteristics of our partner primary schools.

School code	*Ofsted category*	*FSM*	*Pupil progress reading*	*Pupil progress writing*	*Pupil progress mathematics*
school_a	Good (2)	8.9%	–0.9 (average)	–1.2 (average)	–0.8 (average)
school_b	Good (2)	28.4%	–2.5 (below average)	–1.4 (average)	–2.6 below (average)
school_c	Good (2)	9.4%	0 (average)	1 (average)	–0.4 (average)
school_d	Good (2)	48.5%	–0.1 (average)	2.2 (average)	–0.9 (average)
school_e	Good (2)	8.3%	–1.8 (average)	1.9 (above average)	–0.6 (average)
school_f	Good (2)	9.5%	3.4 (well above average)	–1.4 (average)	1.1 (average)
school_g	Good (2)	13.6%	–1.2 (average)	0.5 (average)	2.4 (average)
school_h	Good (2)	17.3%	–3.7 (well below average)	–2 (average)	–2.8 (below average)
Mean	Good (2)	17.99%	–0.85 (average)	–0.05 (average)	–0.58 (average)
Standard deviation	0	14.05	2.11	1.65	1.73

Table 3.2 Characteristics of our partner secondary schools.

School code	*Ofsted category*	*FSM*	*Progress 8 score*
school_i	Outstanding (1)	15.2%	–0.13 (average)
school_j	Outstanding (1)	30.4%	0.7 (well above average)
school_k	Outstanding (1)	12%	0.28 (above average)
school_l	Good (2)	14.7%	0.12 (average)
school_m	Inadequate (4)	17.7%	–0.14 (average)
Mean	Good (1.8)	18%	0.17 (above average)
Standard deviation	1.3	7.22	0.31

The mean percentage of the pupils who had been eligible for FSM at any time during the past six years in our primary school sample was 17.99%. This was below the national average, 24.3% at the time of data collection. The same figure in our secondary school sample was 18%, below the national average of 28.6%. The congruence between the primary and secondary schools' mean Ofsted ratings and mean percentage of pupils eligible for FSM made the profile of the schools similar at KS2 and KS3 levels.

As shown in Table 3.1, most of the scores in reading, writing and mathematics were average scores in our sample. The mean scores approximately correspond to the national average, equivalent to 64% of all schools in reading, 67% of all schools in writing and 57% of all schools in mathematics (DfE, 2016). Table 3.2 shows that the mean Progress 8 score in our secondary sample was 0.17, which corresponds to the above-average score that only 17% of all schools received nationally (DfE, 2016).

Taken together, the mean Ofsted ratings for both our primary and secondary school sample and pupil progress score for our primary school sample are very similar to the average school at the national level. The mean percentage of FSM was below the national average for both our primary and secondary school samples, meaning that the student population of our sample schools is probably slightly more advantaged than the national average. The mean Progress 8 score was slightly above the national average in our secondary school sample. It was not possible to recruit a more diverse sample of secondary schools even though we made multiple attempts to do so. Schools in the categories 'requires improvement' and 'inadequate' are subject to reinspection monitoring by inspectors (Ofsted, 2022), which arguably leaves little time for teachers and head teachers to collaborate with universities for research. We were told several times by leaders in such schools that while they were interested in our research, it could not be a priority for them.

Corpus design and representativeness

As noted earlier, we built a corpus using data supplied by our 13 partner schools. The corpus can be split in two ways: into written and spoken texts, and into Key Stage 2 and Key Stage 3 texts. It consists of texts from the subjects of English, mathematics, science, history and geography, on the basis of the subjects used for Progress 8, which we took as a proxy for valued subjects. Each subject can be analysed separately. As we intended the corpus to represent the language that students encounter during the academic part of their schooling from teachers and other educationalists, it contains no student-produced texts. With our focus on the transition, we collected data from Years 5 and 6 for the Key Stage 2 corpus, and Years 7 and 8 for the Key Stage 3 corpus, although the complete Key Stages are comprised of additional years (see Chapter 1). We collected the data in the school year of 2018–2019.

Biber describes representativeness in corpus design as 'the extent to which a sample includes the full range of variability in population' (1993, p. 243). He notes that preconditions for achieving representativeness are that the population from which the corpus is sampled is clearly defined and that the range of text types that the population comprises is fully known. Taking the first of these, our population is the academic language encountered by students at English state schools in Years 5 to 8 in Progress 8 subjects, and the sample is texts sourced from the 13 schools that had agreed to be project partners. This leads to a restriction on the situational parameters, the geographical representation, as all our partner schools are located in northern England. We have reasonable confidence that our sample is representative of the population in terms of academic content because all state-funded schools in England are obliged to follow the detailed specifications of the National Curriculum, and

textbooks have national reach. The previous section outlines to what extent the schools are representative in other ways.

The second of Biber's preconditions, knowing the range of text types, is not straightforward, as students study multiple subjects, and within these, encounter many registers. We sought to ensure ecological validity (Stangor, 1998) as far as possible, that is, to ensure that the resources that were collected are similar to the everyday life experience of language users – school students. Some school subjects are taught almost daily and some less frequently. To ensure that the weighting of subject materials in the corpus approximately reflected the time students spent on each, we obtained sample timetables from the schools. Table 3.3 shows the timetable of a class of Year 7 students at one of our partner secondary schools.

We also discussed the composition of the corpus with teachers at our partner schools, and in particular with the project consultant from Huntington School. A sampling frame was designed to include both the written and spoken registers of the subjects of English, mathematics, science, history and geography that would reflect their class times to create a representative and balanced corpus as much as possible; however, no target was set for the number of texts or text length and resources were collected in 'an opportunistic mode' (McEnery & Hardie, 2012, p. 64). As McEnery and Brookes (2022, p. 37) note, 'balanced, representative corpora are best viewed as a theoretical ideal rather than being necessarily achievable in practice'.

In addition to the corpus design that involves representativeness and balance, ethics and copyright are the other important considerations in building a corpus (McEnery & Hardie, 2012; McEnery & Brookes, 2022). In our written corpus, textbooks and some of the commercial presentations and worksheets are subject to copyright restrictions, and they cannot be redistributed publicly. In teacher-created resources, such as assessments and worksheets, we anonymised the names of the schools when the school name was present. As we describe below, our spoken corpus only includes the anonymised transcriptions of teachers who provided written informed

Table 3.3 A weekly timetable for Year 7 students.

Monday	**Tuesday**	**Wednesday**	**Thursday**	**Friday**
Registration	Registration	Registration	Registration	Registration
Geography	ICT and Computing	Religion, Philosophy and Ethics	Mathematics	Science
Physical Education	Art	History	Geography	Physical Education
Tutor Report	French	Mathematics	Science	Drama
French	Science	Food & Textiles Technology	English	Music
Technology	English	English	History	Mathematics

consent to participate in our project. We discuss the registers within the corpus in the following sections on the written and spoken corpora.

The written corpus

Representativeness and data gathering

We have discussed our use of the student timetables in our decisions about the overall balance of the corpus. In order to increase the degree of representativeness of the written corpus at the level of individual subject, we reviewed the Department for Education's (DfE) national curriculum documents for each subject for KS2 and KS3 in England (DfE, 2013, 2014). This gave us an understanding of attainment targets and topics as well as notes and guidance aimed at teachers and led to the inclusion of additional materials. For example, the programme of study of English has a word list for Years 5 and 6, and this was included in our corpus.

A particular issue for the written corpus was the wide range of written registers. We consulted the teachers in each school in order to determine registers that were used in each subject and identify the approximate extent of their use during lessons. For instance, we found that presentations and worksheets are central registers of academic language in lessons. Textbooks are used only occasionally, around 10% of the class time, or in some subjects, not used at all. These distributions differed from one subject to another. Naturally, teachers could only give us rough estimates, but they were nonetheless a useful guide to informing decisions about what proportions of each register to include for each subject. The practice of consulting informants is a crucial step in developing corpora for English for specific purposes and validating registers and their representation in accurate proportions in the corpus (Gray, 2015).

Where possible, a soft copy version of the written resources was collected. When no soft copy version of the resources was available, a hard copy of these resources was collected and scanned. Then, we used the software package *ABBYY PDF Transformer+* for optical character recognition (OCR). We manually checked all the scanned resources and corrected any OCR errors. All the written resources were converted to plain text files with UTF-8 encoding for corpus analysis, though it should be noted that some corpus software, including #LancsBox (Brezina et al., 2020) and AntConc v.4 (Anthony, 2022) can read PDF and Word files. We used #Lancsbox v.6.0 (Brezina et al., 2020) to calculate token counts.

Composition of the written corpus

Tables 3.4 and 3.5 show the composition of the written corpora, divided into the five subject areas that we collected.

As we noted above, we sought to make the corpus ecologically valid through consulting with teachers about the balance of subjects and

Table 3.4 KS2 written corpus.

Subject	*Texts*	*Tokens*	*Mean length*	*SD text length*
English	600	303,257	505	1381
Mathematics	614	174,337	284	904
Science	177	160,355	906	3069
History	140	83,998	600	683
Geography	152	62,300	410	541
Total	1683	784,247		

Table 3.5 KS3 written corpus.

Subject	*Texts*	*Tokens*	*Mean length*	*SD text length*
English	334	260,806	781	3552
Mathematics	872	257,459	295	353
Science	675	356,319	528	3046
History	156	233,600	1497	9141
Geography	170	70,503	415	346
Total	2207	1,178,687		

consulting timetables. We did not therefore attempt to gather additional materials to increase the size of the small sub-corpora, as this might have distorted the importance of that subject in the corpus as a whole, threatening representativeness. In particular, the KS2 history and geography written sub-corpora are very small because these subjects are not taught explicitly in primary schools. Instead, students have a timetable slot for 'topics', which covers content related to science, geography and history. We classified these texts into the subjects of science, geography and history, consulting with the primary school teachers who used the materials, and who had sometimes designed them. The heavy weighting of English and mathematics in KS2 is almost certainly partly due to the amount of time that is spent in Year 6 on preparing for the national SATs (Standard Attainment Tests, see Chapter 1), which cover English and mathematics. It can be seen that there is a big increase in the relative size of the science sub-corpus at KS3, which may constitute one aspect of the linguistic challenge for students. In addition to the subject categorisation of the written school language registers, we also categorised them into sub-registers to explore lexico-grammatical variation in the written school language resources in terms of both subjects and sub-registers across the Key Stages.

Sub-registers

Register studies use a number of situational characteristics to describe texts, including participants, relations among participants, channel/mode, setting,

communicative purposes and topic (Biber & Conrad, 2019). The participants of school registers were teachers and students in a classroom setting, and at the top level of analysis, the registers are the written and spoken resources of English, mathematics, science, history and geography. We conducted a systematic data-driven categorisation of the school sub-registers in our written corpus and focused on mode and communicative purposes of the texts in order to categorise them into sub-registers. Our findings are shown in Table 3.6. As can be seen, mode refers to the channel of the school sub-registers that were presented to students. Two resources were used for the identification of the school sub-registers and description of their situational characteristics, as recommended by Biber and Conrad (2019): (1) the insights that we gained from the teachers, expert informants in this context, into the school registers and the purposes of texts; (2) our examination of the texts within the registers that we conducted to identify their communicative purposes. With the exception of textbooks and fiction, all the texts in our written corpus were read and analysed inductively in order to describe their primary and other communicative purposes and identify their sub-registers.

In addition to their primary communicative purposes shown in Table 3.6, the school sub-registers served other purposes. For example, worksheets, which were presented both electronically and in written mode to students, contained exercises and questions that students were expected to complete in order to practise subject topics and strengthen their learning, and they also included short reading extracts, accompanied by questions related to them, to convey information. Presentations, which were electronic resources, primarily included informational subject-specific content on the topics but also contained warm-up questions and practice exercises to enable students to practise content. Like presentations, textbooks also served a multifunctional purpose at schools. The primary function of textbooks was to provide students with information on subject topics, and they included assessment tasks that assessed students' knowledge as well as exercises and

Table 3.6 Written school language sub-registers and their situational characteristics.

Sub-registers	*Mode*	*Primary communicative purpose*
Worksheets	Written/electronic written	Practising subject content and reinforcing learning
Presentations	Electronic written	Presenting subject content
Textbooks	Written	Presenting subject content
Assessment tasks	Written/electronic written	Assessing students' knowledge
Reading extract	Written	Presenting exposition
Glossary	Written	Presenting vocabulary and its definitions

questions that were aimed to reinforce students' learning. Assessment tasks involved exams, quizzes, peer assessment tasks and self-assessment criteria that evaluated students' summative or formative progress. We describe reading extracts, unaccompanied by any exercises or questions, as non-fiction expository texts on subject-specific topics that introduced information to students. Similarly, the glossary included vocabulary and its definitions without any exercises or questions. It should be noted that there was also the register of fiction that students encountered in their English classes. This fiction register was in the form of novels and stories that students read as part of their English classes. Although we collected these written resources, we have excluded the register of fiction in this book, since it does not meet our definition of the academic language of school, introduced in Chapter 2.

The spoken corpus

We also constructed a spoken corpus that comprised transcribed teacher talk in Years 5–8. As for the written corpus, this is divided into KS2 and KS3, and can be further sub-divided by year group and subject. We aimed to represent the teacher talk encountered by students in the subjects of English, mathematics, science, history and geography. In this book, we report our analysis of teacher talk in English, mathematics and science.

Audio recordings were collected from our partner schools. A number of teachers at our partner schools gave written informed consent to be recorded. We had anticipated some reluctance but did not find any. The teachers were provided with audio recorders and microphones worn on a lanyard, and they were asked to record their lessons themselves, without an observer. We did not set out to record student talk, in line with the project aims, so a lanyard microphone was ideal.

The teacher talk was transcribed by a professional agency. Any student contributions that happened to be audible were ignored and not transcribed. We did not have informed consent from students for their utterances to be recorded and transcribed and are not analysing these data. Occasionally, this makes interpreting the teacher utterances difficult, when they are responding to student questions, for example. Obtaining informed consent from all the students, usually around 30 per class, and from their parents or caregivers would have been unmanageable.

The transcribers used an orthographic transcription scheme adapted from the spoken British National Corpus (BNC) 2014 transcription scheme (see Love et al., 2017). The teachers were allocated codes to ensure anonymity. In order to ensure accuracy and consistency of the transcription, a research assistant manually checked all the transcribed texts and corrected any errors. Below is an example extract of an English lesson in Year 5.

line	*speaker*	*utterance*
1	T061	with your partner (.) there's a few tricky ones on here (.) today (.) okay and don't worry if you don't know it this is why we practise it isn't it so that we can go through it again and again (.) who's got is there anyone's that's got ten on one of our SPaG tests ye=yet? (.) okay look only a few very few people in the class (.) so well done if you have (.) but it's obviously very tricky (.) okay anybody now hasn't got their partner's test in front of them ready to mark it? (.) you might have two to mark somebody's gone (.) gone off somewhere (.) okay this one you should have been able to do I hope which sentence uses a relative clause the map that I brought with me is out of date or the map I bought yesterday is out of date (.) so which one is the relative clause? go on <name M>?#

Extract 3.1, Year 5 English lesson recording, Teacher 061.

While transcribing the teacher talk, no punctuation marks were used, except for question marks. A short pause was marked by a tag (.). The sampling frame for the collection of audio recordings was designed to represent all five subjects in proportion to their distribution within the timetable for one week at school (see Figure 3.2). For instance, we collected three lesson recordings of English, mathematics and science separately in Years 7 and 8 (KS3) to represent teacher talk. A similar procedure was followed for the lesson recordings of Years 5 and 6 (KS2), taking into account the timetable of our partner primary schools. In total, we collected 218 audio recordings. Due to the time-consuming and resource-intensive nature of high-quality transcription procedures, to date, we have only 108 fully transcribed audio recordings of English, mathematics, science, geography and history subjects, as shown in Table 3.7. This means that the spoken corpus of teacher talk in this book is not a balanced corpus of teacher talk at the transition. To our knowledge, however, it is still the largest corpus of teacher talk at the transition from primary to secondary school.

The corpus size, at 506,517 tokens, was calculated using #Lancsbox v.6 (Brezina et al., 2020). The mean text length of teacher talk showed an increase in all the subjects from KS2 to KS3, suggesting that the volume of teacher talk that the students encountered in one lesson on average increased at KS3. The larger standard deviations in text lengths at KS2 than KS3 indicated that the length of the teacher talk varied to a greater extent at KS2 than KS3, except for the history subject.

Our written and spoken corpus for both KS2 and KS3 totals 2,469,451 tokens. We divide this up in a number of different ways for the various analyses presented in the following chapters. In Chapter 4, we analyse the written data only, and take out the texts that include fewer than 100 tokens; in Chapters 5, 6 and 7, we focus on specific subjects and treat written and spoken data together.

Table 3.7 The spoken corpus of teacher talk.

	Number of texts		*Number of tokens*		*Mean text length (tokens)*		*Standard deviation text length (tokens)*	
	KS2	*KS3*	*KS2*	*KS3*	*KS2*	*KS3*	*KS2*	*KS3*
English	15	8	72,475	47,595	4832	5949	1284	992
Subtotal (English)	23		120,070					
Mathematics	18	11	82,031	51,171	4557	4652	2225	1467
Subtotal (Mathematics)	29		133,202					
Science	18	10	62,375	47,772	3465	4777	2923	1410
Subtotal (Science)	28		110,147					
History	6	7	22,114	39,319	3686	5617	1621	1354
Subtotal (History)	13		61,433					
Geography	6	9	24,472	57,193	4079	6355	1345	2440
Subtotal (Geography)	15		81,665					
Grand total	108		506,517					

Corpus analytical methods used

Thompson and Hunston give an elegant description of their use of corpus methods, as follows: 'We apply to the data that we have collected […] corpus investigation methods that rearrange and process that data. Our challenge then is to make sense of the rearranged data' (2019, p. 6). We have described above how we collected a relatively large quantity of school language data; we now describe how we rearranged and processed it, and in subsequent chapters, how we made sense of it. To explore our corpus, we have used a range of methods: quantitative, qualitative and mixed methods. Thompson and Hunston (2019, p. 6) place their corpus methods on a cline from qualitative to quantitative. Figure 3.2 is taken from their discussion.

We also use a range of corpus methods. As our study centrally concerns comparing two corpora, KS2 and KS3, in different ways, it is to be expected

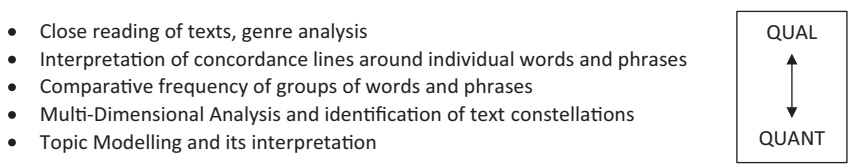

Figure 3.2 Thompson and Hunston's representation of methods used in their studies of interdisciplinary genres (2019, p. 6).

that the central method, comparing the frequency of words and phrases, is at its heart. We also use multi-dimensional (MD) analysis and undertake detailed interpretation of concordance lines, thus covering the central three methods from Figure 3.2. We now overview these methods, starting from the quantitative end of the cline, with MD analysis, as this reflects the order of our chapters. Further methodological detail is given in individual chapters.

Quantitative data analysis procedures

Multi-dimensional analysis

MD analysis was originally developed by Biber (1988). It is a quantitatively driven analytical approach to corpus analysis, which aims to provide a comprehensive linguistic description of registers. MD analysis is 'derived from factor analysis […] which observes the sequential, partial, and observed correlations of a wide range of variables in order to produce groups of co-occurring factors' (Friginal, 2013, p. 138). It is based on the premise that texts in the same register will exhibit clusters of co-occurring linguistic features, which reflect the underlying communicative functions of the register (Biber, 1988; Friginal, 2013; Biber & Conrad, 2019). For example, 'private verbs', such as *assume*, *believe*, *doubt* and *know* are found through factor analysis to co-occur with a group of other linguistic characteristics that includes present tense, second person pronoun and use of *DO* as a pro-verb (Biber, 1988, p. 75). Taken together, qualitative analysis shows that these features are associated with 'involved' discourse. Their relative absence and the presence of other features such as agentless passives and attributive adjectives are associated with 'informational' discourse. This analysis leads to the construction of a 'dimension', 'involved vs informational'.

There are several steps to conducting MD analysis. First, frequencies of lexico-grammatical features are counted across the registers in the corpus. Linguistic co-occurrence patterns that constitute an underlying dimension of variation are identified quantitatively using factor analysis. Then, each dimension of variation, statistically determined, is analysed qualitatively to construct the underlying communicative functions associated with each dimension in different registers. In his seminal work, Biber (1988) found six main dimensions of variation in a general corpus of written and spoken registers, and a seventh, which was not matched to a functional interpretation. The first five dimensions that we focus on in this study are as follows:

Dimension 1: Involved versus informational discourse

A positive score on Dimension 1 indicates involved discourse (e.g., conversational registers), while a negative score indicates informational discourse

(written registers, such as academic prose). The positively loaded (involved) linguistic features are as follows (Biber, 1988, p. 102):

> private verbs, that-deletions, contractions, present tense verbs, second pronouns, *do* as pro-verb, analytic negations, demonstrative pronouns, first person pronouns, pronoun *it*, *be* as main verb, causative subordination, discourse particles, indefinite pronouns, general hedges, amplifiers, sentence relatives, wh-questions, possibility modals, non-phrasal coordination, wh-clauses, final prepositions.

The negatively loaded (informational) linguistic features include 'nouns, word length, prepositions, type/token ratio, attributive adjectives' (Biber, 1988, p. 102).

Dimension 2: Narrative versus non-narrative discourse

A positive Dimension 2 score represents narrative discourse marked by past events (e.g., fiction) whereas a negative Dimension 2 score represents non-narrative discourse (e.g., academic prose). The positively loaded (narrative) linguistic features are 'past tense verbs, third person pronouns, perfects aspect verbs, public verbs, synthetic negation, present participial clauses' (Biber, 1988, p. 102). The negatively loaded linguistic features are 'present tense verbs, attributive adjectives, past participial WHIZ deletions (past participial forms of verbs as post-nominal modifiers – the solution proposed by the team), and word length' (Biber, 1988, p. 102).

Dimension 3: Situation-dependent versus elaborated reference

A positive Dimension 3 score characterises discourse dependent on the situation (e.g., a sports broadcast) while a negative Dimension 3 score exhibits elaborated reference and independence of the context (e.g., academic prose). The positively loaded linguistic features are 'wh-relative clauses on object positions, pied piping constructions, wh-relative clauses on subject positions, phrasal coordination, nominalisations' (Biber, 1988, p. 102). The negatively loaded features are 'time adverbials, place adverbials, and adverbs' (Biber, 1988, p. 102).

Dimension 4: Overt expression of persuasion

A positive Dimension 4 score is characteristic of persuasive discourse (e.g., editorials). The positive linguistic features of this dimension are 'infinitives, prediction modals, suasive verbs, conditional subordination, necessity modals, split auxiliaries' (Biber, 1988, p. 103). There are no negatively loaded linguistic features of this dimension.

Dimension 5: Abstract versus non-abstract information

A positive Dimension 5 score denotes abstract discourse (e.g., scientific discourse) whereas a negative Dimension 5 score denotes non-abstract discourse. The positive linguistic features are 'conjuncts, agentless passives, past participial clauses, by-passives, past participial WHIZ deletions, other adverbial subordinators' (Biber, 1988, p. 103). A relatively low type/token ratio, that is, lack of lexical variation, is the only linguistic feature negatively loaded to this dimension. (Biber notes that although this may seem surprising, abstract discourse is often technical, and tends to repeat key terms rather than seeking stylistic variation.)

MD analysis is ideally suited to investigating register changes between KS2 and KS3, with its potential for finding subtle, measurable distinctions along a large number of linguistic features and dimensions. The method has been used with many different corpora to date (see Berber Sardinha & Veirano Pinto, 2019). In Chapter 4, we discuss MD analysis studies relevant to our own. We then report an MD analysis of our written corpus of English, mathematics and science subjects at KS2 and KS3.

Mixed and qualitative data analysis

Towards the qualitative end of Thompson and Hunston's cline shown in Figure 3.2 is the method: 'Interpretation of concordance lines around specific words and phrases' (2019). This method was an integral part of our studies; in order to help teachers and students with the linguistic challenges of secondary school, we need to be able to provide details of usage and meaning. However, we needed a way into our corpus before examining concordances. Using concordance examination on its own is indicated when the central research questions entail the detailed analysis of pre-determined words and expressions. For example, Auge (2021) sought to identify the associations of the expression *greenhouse effect* across a range of registers. In other cases, studying one set of words and expressions can lead to further concordance analysis. For example, Islentyeva and Kafi (2021) studied attitudes towards the EU in the British press from 2016–2018. They began with the words *Britain*, *European* and the *EU*, and used corpus software to identify the most significant collocates immediately before and after each of the three words, finding words such as *voters*, *people*, *migrants* and *culture*. These were then classified semantically and analysed in detail using concordances. Another approach, if the corpus is fairly homogenous, has been to manually analyse a sample and identify candidates of interest. Charteris-Black (2004) has taken this approach in his study of the ideological use of metaphorical meanings of words in corpora.

With our corpora and for the research questions that we look at in Chapters 5, 6 and 7, none of these approaches would be sufficient as a starting point. We know from teachers' reports, and from examples that have

come to our notice, that KS2 and KS3 language use is likely to be different at the level of detail. However, we begin from the position of not knowing in advance which words would be significant, and our corpus is far from homogenous, so sampling would not be effective. We therefore began by using tools that showed us what was frequent in our corpus and sub-corpora, and what was more frequent in each sub-corpus relative to the others. The results of these analyses are valuable in themselves, and also give us the starting points for more detailed, qualitative studies. In discourse studies, corpus techniques increase the rigour of the analysis and minimise the researcher's subjective selection of texts and linguistic features for analysis since corpus techniques, such as keyness analysis that we use, point to frequent linguistic features that are important in the corpus and provide quantitative information on their frequency of occurrence that would underpin qualitative interpretations (e.g., Baker, 2006; Mautner, 2022). Hence, cherry-picking of texts and linguistic features is avoided in corpus-informed qualitative studies of the meaning and function of words and phrases.

A number of previous studies have taken a frequency-based approach, using wordlists and keyness analysis to compare different corpora. Deignan et al. (2019), comparing metaphorical uses in different corpora of texts on the topic of climate change, used word lists to identify the most frequent lexical words, and then studied concordances of these words in detail. Baker et al. (2013) conducted a detailed Critical Discourse Analysis of a corpus of British newspapers, following a number of steps. They began with word lists to get an overall sense of 'aboutness', and to look for expected and unexpected semantic domains. They then compared different sections of their corpus against each other, highlighting frequent words, followed by detailed concordance examination. We use variations of Baker et al.'s approach in Chapters 5, 6 and 7 when we study KS2 and KS3 English, science and mathematics sub-corpora.

Frequency was measured using two related tools. First, to produce a list of the most frequent words in each sub-corpus, we used the Words tool in #LancsBox 6.0 (Brezina et al., 2020). Second, to compare word frequencies across different sub-corpora, we used the keywords technique (Baker, 2006; Rayson, 2019), also available within #LancsBox 6.0. The keywords tool allows the researcher to compare the lexical make-up of two corpora, by showing us which words are significantly more frequent in one than the other. There are a number of different statistical options within the tool, and when we describe the studies in the following chapters, we explain the choices that we made. Using the keywords tool, we generated lists of words that were, for example, significantly more frequent in KS3 English than KS2 English, or than a general or reference corpus. We discuss the use of reference corpora, and how we developed a reference corpus for this project, in Chapter 5. The resulting list has to be carefully checked manually, as it will contain items that are not of interest, such as proper names from literature texts and text codes. Once irrelevant items have been deleted, the lists are of interest in themselves and also as the starting point for concordance analysis.

This follows the approach used by researchers, including Gabrielatos (2018) and Partington and Duiguid (2021), of using keywords as a way into a specialised corpus.

> [The researcher derives] a list of key items ranked according to the value of the keyness metric used in the study. At this point, the researcher may switch to a targeted approach and select particular types of items for concordance analysis according to explicit criteria, such as their normalised or raw frequency, part of speech, core sense, or relation to a particular topic.
>
> (Gabrielatos 2018, p. 3)

We studied concordance data for all words identified through frequency lists and the keywords tool, reflecting our aim to describe as fully as possible the language challenge of secondary school. Our concordance analyses followed well-established procedures such as those described by Sinclair (e.g., 1991, 2003). That is, we identified the different meanings of the words and phrases under study, and considered their function, using expanded context to support this. We examined the syntactic patterns and collocations that these words and phrases occur. As the following chapters show, this qualitative analysis often showed subtle but important differences in meaning and use between the different registers in our corpus, and sometimes between school registers and non-school language, as represented by a reference corpus.

Conclusion

In this chapter, we have explained how we attempted to tackle our questions about the challenge of the language of secondary school faced by KS3 students. To our knowledge, this is the first corpus to represent both written and spoken school language at the transition from primary to secondary school. This is also the first study to explore to what extent, if any, school language changes from primary to secondary school in both spoken and written modes by using corpus methods and quantitative and qualitative corpus techniques in order to investigate lexico-grammatical variation across the subjects and Key Stages. In the following four chapters, we describe various studies that we have conducted using these data and methods.

References

Anthony, L. (2022). *AntConc* (Version 4.0.3) [Computer Software]. Waseda University. https://www.laurenceanthony.net/software

Auge, A. (2021). From scientific arguments to scepticism: Humans' place in the greenhouse. *Public Understanding of Science*, *31*(2), 179–194. https://doi.org/10.1177/09636625211035624.

Baker, P. (2006). *Using corpora in discourse analysis*. Continuum.

Baker, P., Gabrielatos, C., & McEnery, T. (2013). *Discourse analysis and media attitudes: The representation of Islam in the British press*. Cambridge University Press.

Berber Sardinha, T., & Veirano Pinto, M. (Eds.). (2019). *Multi-dimensional analysis: Research methods and current issues*. Bloomsbury Academic. https://doi.org/10.5040/9781350023857

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Biber, D. (1993). Representativeness in corpus design. *Literary and Linguistic Computing*, *8*(4), 243–257. https://doi.org/10.1093/llc/8.4.243

Biber, D. (2006). *University language: A corpus-based study of spoken and written registers* (Vol. 23). John Benjamins Publishing.

Biber, D., & Conrad, S. (2019). *Register, genre, and style* (2nd ed.). Cambridge University Press.

Biber, D., Conrad, S., Reppen, R., Byrd, P., & Helt, M. (2002). Speaking and writing in the university: A multidimensional comparison. *TESOL Quarterly*, *36*(1), 9–48. https://doi.org/10.2307/3588359

Brezina, V., Weill-Tessier, P., & McEnery, A. (2020). #LancsBox v. 6. [software]. http://corpora.lancs.ac.uk/lancsbox.

Chambers, G. (2020). *Linguistic encounters: What year 7 students say about 'difficult words'*. https://linguistictransition.leeds.ac.uk/linguistic-encounters-what-year-7-students-say-about-difficult-words/

Charteris-Black, J. (2004). *Corpus approaches to critical metaphor analysis*. Palgrave Macmillan.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238. https://doi.org/10.2307/3587951

Coxhead, A., & White, R. (2012). Building a corpus of secondary school texts: First you have to catch the rabbit. *New Zealand Studies in Applied Linguistics*, *18*(2), 67–73. https://search.informit.org/doi/abs/10.3316/informit.108520980622633

Deignan, A., Semino, E., & Paul, S.-A. (2019). Metaphors of climate science in three genres: Research articles, educational texts, and secondary school student talk. *Applied Linguistics*, *40*(2), 379–403. https://doi.org/10.1093/applin/amx035

Department for Education (DfE). (2013). *The national curriculum in England: Key stages 1 and 2 framework document*. https://www.gov.uk/national-curriculum

Department for Education (DfE). (2014). *The national curriculum in England: Key stages 3 and 4 framework document*. https://www.gov.uk/national-curriculum

Department for Education (DfE). (2016). *Understanding school and college performance measures*. https://www.gov.uk/government/publications/understanding-school-and-college-performance-measures

Durrant, P., & Brenchley, M. (2019). Development of vocabulary sophistication across genres in English children's writing. *Reading and Writing*, *32*(8), 1927–1953. https://doi.org/10.1007/s11145-018-9932-8

Education Endowment Foundation (EEF). (2022). https://educationendowmentfoundation.org.uk/support-for-schools/research-schools-network

Friginal, E. (2013). Twenty-five years of Biber's multi-dimensional analysis: Introduction to the special issue and an interview with Douglas Biber. *Corpora*, *8*(2), 137–152. https://doi.org/10.3366/cor.2013

Gabrielatos, C. (2018). Keyness analysis: Nature, metrics and techniques. In Taylor, C., & Marchi, A. (Eds.), *Corpus approaches to discourse: A critical review* (pp. 225–258). Routledge.

Gorard, S. (2012). Who is eligible for free school meals? Characterising free school meals as a measure of disadvantage in England. *British Educational Research Journal*, *38*(6), 1003–1017. https://doi.org/10.1080/01411926.2011.608118

Gray, B. (2015). *Linguistic variation in research articles: When discipline tells only part of the story*. John Benjamins.

Green, C., & Lambert, J. (2018). Advancing disciplinary literacy through English for academic purposes: Discipline-specific wordlists, collocations and word families for eight secondary subjects. *Journal of English for Academic Purposes*, *35*, 105–115. https://doi.org/10.1016/j.jeap.2018.07.004

Greene Wells, J., & Coxhead, A. (2015). *Academic vocabulary for middle school students: Research-based lists and strategies for key content areas*. Brookes Publishing.

Islentyeva, A., & Abdel Kafi, M. (2021). Constructing national identity in the British Press: The Britain vs. Europe dichotomy. *Journal of Corpora and Discourse Studies*, *4*(0), 68. https://doi.org/10.18573/jcads.64

Jones, M., & Deignan, A. (2021, September 4). *The linguistic challenges of the transition from primary to secondary school*. [Paper presentation] ResearchED National Conference 2021, London, United Kingdom.

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

Mautner, G. (2022). What can a corpus tell us about discourse? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 250–262). Routledge.

McEnery, T., & Brookes, G. (2022). Building a written corpus: What are the basics? In A. O'Keeffe & M. McCarthy (Eds.), *The Routledge handbook of corpus linguistics* (pp. 35–47). Routledge.

McEnery, T., & Hardie, A. (2012). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.

Nesi, H., & Gardner, S. (2012). *Genres across the disciplines: Student writing in higher education*. Cambridge University Press.

Ofsted. (2018). *State-funded schools inspections and outcomes as at 31 August 2018*. https://www.gov.uk/government/statistics/state-funded-schools-inspections-and-outcomes-as-at-31-august-2018

Ofsted. (2022). *School inspections: A guide for parents*. https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1044285/School_inspections_-_a_guide_for_parents_January_2022.pdf

Partington, A., & Duiguid, A. (2021). Political media discourses. In Friginal, E., & Hardy, J. (Eds.), *The Routledge handbook of corpus approaches to discourse analysis* (pp. 116–135). Routledge.

Perryman, J., Maguire, M., Braun, A., & Ball, S. (2018). Surveillance, governmentality and moving the goalposts: The influence of Ofsted on the work of schools in a post-panoptic era. *British Journal of Educational Studies*, *66*(2), 145–163.

Rayson, P. (2019). Corpus analysis of key words. In Chapelle, C. A. (Ed.), *The concise encyclopaedia of applied linguistics* (pp. 320–326). Wiley.

Römer, U., & O'Donnell, B. M. (2011). From student hard drive to web corpus (part 1): The design, compilation and genre classification of the Michigan corpus of upper-level student papers (MICUSP). *Corpora*, *6*(2), 159–177. https://doi.org/10.3366/cor.2011.0011

Simpson, R. C., Briggs, S. L. Ovens, J., & Swales, J. M. (1999). *The Michigan corpus of academic spoken English*. The Regents of the University of Michigan.

Sinclair, J. M. (1991). *Corpus, concordance, collocation*. Oxford University Press.

Sinclair, J. M. (2003). *Reading concordances: An introduction*. Longman.

Stangor, C. (1998). *Research methods for the behavioral sciences*. Houghton Mifflin Company.

Thompson, P., & Hunston, S. (2019). *Interdisciplinary research discourse: Corpus investigations into environment journals*. Routledge.

Thompson, P., & Nesi, H. (2001). The British academic spoken English (BASE) corpus project. *Language Teaching Research*, *5*(3), 263–264. https://doi.org/10.1177/136216880100500305