Adolescent reading experience, independent choices and curriculum materials

Applied Corpus Linguistics 5 (2025) 100124

Contents lists available at ScienceDirect

Applied Corpus Linguistics

journal homepage: www.elsevier.com/locate/acorp

Articles

Adolescent reading experience, independent choices and curriculum materials

Beverley Jennings [a,*], Daisy Powell [a], Sylvia Jaworska [b], Holly Joseph [a]

[a] Institute of Education, University of Reading, London Road Campus, 4 Redlands Road, Reading, RG1 5EX
[b] Department of English Language and Applied Linguistics, Whiteknights, PO Box 218, Reading, RG6 6AA

ARTICLE INFO

Keywords:
Reading comprehension
Corpus linguistics
Reading experience
Vocabulary

ABSTRACT

Reading comprehension ability is assessed in England within the English language GCSE exam. This is a high stakes exam, taken by all 16-year-olds, and a pass grade is needed to progress onto the next stage of education and employment. Since reading experience is an important predictor of reading comprehension ability, two different types of reading materials were explored to see how well they matched the reading required in the exam: 1) curriculum reading; and 2) independent reading. Two corpora of texts representing the two types of reading were created and explored using the methods of Corpus Linguistics. The curriculum reading corpus (CRC) had lower linguistic diversity, and higher frequency of nouns but lower frequency of adverbs, than the independent reading corpus (IRC). Exploratory analysis of the most frequent parts of speech revealed that the CRC had words that were more abstract and conceptual, whereas the IRC featured words about the concrete and the everyday, suggesting that curriculum reading presents a different type of vocabulary challenge. The CRC was not as close a match to the exam texts as the IRC. As the English language GCSE exam is used as a measure of literacy competency for both future study and future employment, this suggests that the types of texts chosen for the exam are not a good match for this purpose. The choice of texts in assessments therefore needs careful consideration.

1. Introduction

In England, as part of a suite of General Certificate of Secondary Education (GCSE) exams, taken at the end of their full-time compulsory education, students (age 16) sit an English language GCSE that is taken to indicate their literacy competency and suitability for future study and employment. There is a separate English literature qualification to assess the critical analysis of literary fiction texts. The English language GCSE exam was reformed by the Government in 2015, with the new exam introduced in 2017. This new specification changed the form and age of the texts that have to be read in the exam and, instead of mostly modern and accessible texts in the old version (Isaacs, 2014), now texts are literary fiction and literary non-fiction and have to be from all three of the 19th, 20th and 21st centuries. Students' ability to read and understand these types of texts is therefore an important area of research.

An important predictor of comprehension ability is reading experience (Acheson et al., 2008; Chateau and Jared, 2000; Davidse et al., 2011; Mol and Bus, 2011). This is explained by the lexical quality hypothesis (LQH) (Perfetti and Hart, 2002) as the gradual building of an increasingly secure and coherent, but also nuanced, understanding of words each time they are encountered. The lexical legacy hypothesis (LLH) (Nation, 2017) builds on this by specifying that encounters with words need to be in diverse contexts for greater quality to be built (Joseph and Nation, 2018; Pagán and Nation, 2019; Rosa et al., 2017, 2022). It is therefore important to examine students' actual reading experience, to understand how far it is providing exposures to words in order to build good vocabulary knowledge and comprehension ability, in preparation for the final exam.

2. Literature review

2.1. Reading experience

The Simple View of Reading (SVR) (Gough and Tunmer, 1986; Hoover and Gough, 1990) describes reading as being the product of two parts, the ability to decode written words (either by sounding them out

* Corresponding author.
E-mail addresses: b.j.jennings@pgr.reading.ac.uk (B. Jennings), d.a.powell@reading.ac.uk (D. Powell), s.jaworska@reading.ac.uk (S. Jaworska), h.joseph@reading.ac.uk (H. Joseph).

https://doi.org/10.1016/j.acorp.2025.100124
Received 16 November 2024; Received in revised form 17 February 2025; Accepted 23 February 2025
Available online 26 February 2025

or by recognising them immediately) and the linguistic comprehension of the words. The importance of the two components of the SVR does not, however, remain consistent for readers across all ages. As students become more skilled, and their reading more proficient, the decoding element of the SVR (in which proficiency has been reached) declines in importance and the linguistic element becomes more important (Braze et al., 2007; Francis et al., 2005; Gough et al., 1996; Henderson et al., 2013; Nation and Snowling, 1998; Ouellette, 2006; Tilstra et al., 2009). This linguistic comprehension includes vocabulary knowledge and, as the reading materials of secondary or high school education increase in difficultly, the vocabulary in this kind of written register becomes more and more different from the vocabulary used in spoken registers (Biber and Conrad, 2009; Braze et al., 2007; Cunningham, 2005; Landauer and Dumais, 1997; Tilstra et al., 2009). Written registers, especially informational registers such as those prominent in curriculum reading materials, tend to include more complex lexico-grammatical features especially noun phrases (Biber and Conrad, 2009), more low frequency words and more exception words (words that do not follow usual spelling rules) (Nation and Snowling, 1998). Perfetti and Hart's LQH (2002) defines the ability to read a word efficiently as when the reader is able to access high quality representations of three components of a word: its written form (orthography); its sound (phonology); and its meaning (semantic information). As specified by the LQH, building high-quality representations of words depends on experiences with them, each encounter enabling the components of high lexical quality to become more secure and coherent (Perfetti, 2007).

Reading experience is therefore an important predictor for reading ability (Acheson et al., 2008; Chateau and Jared, 2000; Davidse et al., 2011; Mol and Bus, 2011), and several studies show that it is fiction reading specifically that is a superior predictor of that ability (Mar and Rain, 2015; Martin-Chang et al., 2020; McGeown et al., 2015; Pfost et al., 2013). That reading experience predicts reading skill can be explained by the LLH (Nation, 2017) as it creates a bank of previous experiences with words that each reader has built up. If these experiences are diverse, then lexical quality is gradually increased through each new context or nuance of meaning encountered (Pagán and Nation, 2019; Rosa et al., 2017, 2022).

Whilst explicit teaching of vocabulary is, of course, an essential part of good classroom practice, using the theoretical background of the LQH and the LLH, it is clear that it is not enough to be taught words from lists. Instead, in order to build lexical quality (Perfetti and Hart, 2002) words must be experienced in diverse contexts (Nation, 2017). It has also been estimated that the number of words taught in classrooms each year is approximately 200–300 (Nagy and Herman, 1984), whereas the estimate of the number of words learnt by children each year is approximately 3000 (Nagy et al., 1987). The gap between words learnt overall and words taught is filled, according to Nagy et al. (1987), by learning from context, that is through listening and reading. For older children most new words will be acquired through reading as they will have already encountered, by age 12, words that are found in spoken language (Landauer and Dumais, 1997). It is reading experience that is needed at this age, therefore, for a vocabulary growth to occur (Nagy et al., 1987).

Being able to read well and access the learning materials of the curriculum is crucial for students at the secondary levels of education (Shanahan and Shanahan, 2017). Analysis of the Programme for International Student Assessment research showed that 20 % of 15-year-old students in England were below the reading level considered the minimum required to be able to participate in society (Ingram et al., 2023). The 2023 national Statutory Assessment Tests, taken in England by students at the end of their primary (elementary) education (age 11), showed even lower levels of proficiency, with only 73 % of students meeting the expected standard in reading (Department for Education, 2023b). As adolescents progress through school, subjects are taught more discretely and reading materials become more complex, use increasingly specialised and more academic language (Schleppegrell, 2001, 2007). Many different types of words, for example technical or subject-specific vocabulary and also words that are used for cohesion like connectives, are found more frequently in written language than in spoken. It is reading experience therefore, that will provide encounters with this type of vocabulary (Tilstra et al., 2009).

2.2. Curriculum reading

Corpus studies of vocabulary in education have tended to focus more on higher education than on schools (Coxhead, 2000, 2011; Gardner and Davies, 2014). The main focus of these studies has been on creating lists of academic words (Coxhead, 2000; Gardner and Davies, 2014) and of disciplinary language (Hyland, 2008, 2017; Hyland and Tse, 2007). Some similar work has been done in secondary schools with the creation of lists of school vocabulary and phrases (Green and Lambert, 2018, 2019). There have also been corpus studies of the language used in maths resources (Monaghan, 1999), science textbooks (Coxhead et al., 2010; Deignan and Love, 2019) and of reading materials from a range of KS3 lessons (ages 11–13) (Deignan et al., 2022). These studies and lists have provided teachers and students with valuable teaching and learning resources. However, as shown above (Nagy and Herman, 1984), being taught or learning words from lists in class is not sufficient for vocabulary to grow adequately. Diversity and meaningful context are lacking in lists of words or phrases, compared to the reading experience required to build lexical quality (Nation, 2017; Perfetti and Hart, 2002). Texts that students read, as part of their classes, form part of each students' bank of prior reading experience (Nation, 2017). Studying samples of class reading, at GCSE level, can therefore provide useful data about the vocabulary that students are (and are not) exposed to through the curriculum. Most of the school studies outlined above have relied on collecting text from curriculum textbooks, to represent what is read in the classroom. However, the increasing use of technology, both from teachers' use of slides, worksheets and online quizzes in the classroom, and students' increasing use of their own devices and electronic resources, mean that textbooks can no longer be taken as a good example of the kind of reading that students are expected to do and are exposed to through the curriculum (Deignan et al., 2022).

2.3. Independent reading

Corpus studies of children's non-curriculum reading (reading for pleasure) have compared book language to spoken language and found that book language is more complex (Cameron-Faulkner and Noble, 2013; Dawson et al., 2021; Hsiao et al., 2022; Montag, 2019; Montag et al., 2015; Montag and MacDonald, 2015). These studies have generally used existing collections of texts written for children, like the Oxford Children's Corpus (Wild et al., 2013) or a children's reading subset of the Corpus of Contemporary American English (University of Arizona Libraries, 2021).

A report into the results of a large national (UK) survey by the National Literacy Trust (Clark et al., 2023) about children's reading practices shows that only 43.4 % of children (aged 8–18) said they enjoyed reading, the lowest level recorded since the survey started in 2005. The number of young people enjoying reading drops as age increases. In the same survey only 28 % of respondents said that they read daily in their free time, which followed the trend of gradually decreasing numbers since 2005. Fiction is still the most popular choice for free time reading (73.5 %), but there was no further detail on the types of fiction that were being read. In an earlier report Clark and Rumbold (2006) showed that children's choices, when reading for pleasure, were diverse but that fiction dominated. A report based on data from the school reading programme, Accelerated Reader (Topping et al., 2023), for readers from the UK and the Republic of Ireland attending secondary school years 9–11 (age 13–16), showed that the most read titles were either fiction books that were likely to have been studied in class (e.g. *Of Mice and Men*), or titles by children's and YA authors (e.g. J.K. Rowling). It should

be noted that this data will be affected by the books that are stocked by school libraries and the books that are listed on the Accelerated Reader platform itself.

2.4. The English language GCSE

The new specification of the English language GCSE was first taken by students in England in 2017. A grade 4 (equivalent to a C) in this qualification is needed by students to access most post-16 options, including further study, apprenticeships and employment. It is a government funding requirement for post-16 courses that any students who did not gain a grade 4 or above, must continue to study English and ideally retake the qualification. GCSE results, with special focus on maths and English language, are published each year and are used as a measure by which to judge the quality of education being provided by each school. This means preparing for this exam is important for students, teachers, and schools. A corpus study of a selection of exams texts from the new English language GCSE identified 146 keywords, that appeared more frequently in the exam text corpus (ETC), created for the study, than a reference corpus and therefore were taken to typify the vocabulary in the exam texts (Jennings et al., 2024). These keywords were low in frequency in general language and were typically found in fictional texts, especially older classic fiction. The LQH and LLH show that if these words are to be understood, then students must have experienced them in their prior reading. Identifying the vocabulary content of students' reading experience therefore becomes key.

As the ability to proficiently comprehend the vocabulary in a text depends on previous reading experience having provided enough diverse exposure to that vocabulary (Acheson et al., 2008; Chateau and Jared, 2000; Davidse et al., 2011; Mol and Bus, 2011; Nation, 2017; Perfetti and Hart, 2002), adolescents preparing for their English language GCSE exam will be relying on their previous reading experience to enable their comprehension of the exam texts. That reading experience may have been gained inside and/or outside school. Previous corpus studies of academic language have focused on producing lists of vocabulary that are either common across disciplines (Coxhead, 2000; Gardner and Davies, 2014) or needed within disciplines (Green and Lambert, 2018, 2019; Hyland, 2008, 2017; Hyland and Tse, 2007). For independent reading or reading for pleasure outside school, there is evidence that fewer students, especially in this adolescent age group, choose to read in their free time (Clark et al., 2023). When children and young people do choose to read independently, fiction seems to remain the most popular choice. This is key for reading proficiency as previous research shows that fiction is a superior predictor of reading skill (Mar and Rain, 2015; Martin-Chang et al., 2020; McGeown et al., 2015; Pfost et al., 2013). It is also important to note that corpus studies of children's reading for pleasure usually depend on using collections of texts that are based on the target age for readers of the texts, rather than from any data about what children or adolescents are actually choosing to read. Although the data that we do have would suggest that this is likely to be children's or YA fiction ((Topping et al., 2023), it is crucial to find out what young people are actually choosing to read so that we can have a more accurate picture of the vocabulary they are encountering, rather than just the vocabulary that they would encounter if they read the books and genres targeted at their age group.

2.5. This study

The focus of this study is the vocabulary content of the reading experience of adolescents, both across the curriculum at school and in any independent reading. The intention was to collect a small but manageable number of texts and to carry out an exploratory analysis. One aim was to look at a sample of text drawn from a range of curriculum reading at school, rather than just have word lists or collections that only represent single or limited numbers of subjects, as previous corpus studies have done. A second aim was to add to the primary data on adolescent reading by creating a sample of students' actual independent reading, rather than looking at a collection that is defined by suggested age ranges or specific genres. It would then be possible to compare this new collection of adolescent reading materials to the corpus of exam texts created in a previous study (Jennings et al., 2024). This study therefore created two corpora of texts to explore students' actual reading materials: 1) from lesson materials, to explore curriculum reading in school; and 2) from students' independent reading, outside of school.

In order to explore the different reading experiences offered by the two different genres of reading (curriculum and independent), this study examined the linguistic content of the two new corpora created. The occurrences of different parts of speech were compared, as these can be an indication of linguistic register and could therefore suggest the types of registers present in each corpus (Biber et al., 1999). The lexical diversity of the two corpora were compared as a measure of linguistic richness, a high lexical diversity score indicates that there are more unique words in the text. This is important for reading experience as a higher lexical diversity will provide more encounters with different words and therefore have the potential to build greater lexical quality (Nation, 2017; Perfetti and Hart, 2002) with a greater range of words. The most frequent words in the two corpora were then compared, these most frequent words lists were separated into the four main parts of speech to enable a close comparison. Again, this was a useful method to use to consider the reading experiences and potential vocabulary encounters offered by the two different types of texts.

The level of difficulty presented by the words on the most frequent words lists from the two corpora was analysed by comparing the average number of letters in the words. Longer words have been shown, by eye tracking studies to have longer reading times (e.g. Joseph et al., 2009), and this can impact comprehension due to the increase in processing time (Martin-Chang et al., 2020). The level of difficulty for nouns was measured using concreteness and imageability scores. Words with higher scores for these two measures are easier to comprehend as the reader can draw on perceptual memory (Brysbaert et al., 2014; Cortese and Fugett, 2004; Khanna and Cortese, 2021; Sadoski and Paivio, 2013). Using these three measures (word length, concreteness and imageability) allowed for a comparison of the difficulty of the words in two focus corpora, an important indication of the kind of reading experience being offered by them.

2.6. Research questions

1. What is the linguistic make-up of the corpora of students' curriculum and independent reading?
2. What types of words typify the student reading material that were collected?
3. How far do the student reading materials match the vocabulary in the English language GCSE exam?

3. Methodology

3.1. Ethical approval

This study was granted ethical approval by the University of Reading's Institute of Education.

3.2. Curriculum reading

Curriculum materials were collected for year 10 classes (age 14–15) from an online platform used by teachers to share resources with their classes (Google Classroom). A week in June was chosen for expediency and resources were downloaded from each class. Resources were accessed from the following 16 subjects: Art, Computing, Media, Technology, English, Geography, History, Maths, Music, Physical Education, Religious Studies, Biology, Chemistry, Physics, Childcare, Graphics.

Collecting data from a range of subjects across the curriculum is important because students' bank of experience with words (Nation, 2017), is formed by all their experiences with text, not just from subjects like English where reading is explicitly being taught. With recent reports suggesting that only 28 % of children read every day in their free time (Clark et al., 2023), reading within the curriculum may represent the only reading that some children do, so the full range of subjects is essential to study. The types of resources downloaded included: worksheets; slides; pages from textbooks; quizzes; exam questions and answers; and coursework tasks. In subjects where there were more than 1000 words (11 subjects), the first 1000 words were taken as representative (Biber, 1990). Five subjects had less than 1000 words (see Table 1).

There were some challenges in converting the documents that were shared on the online platform into text files that were suitable for uploading to the corpus tool, *Sketch Engine* (Kilgarriff et al., 2014). Slides often used pictures and graphics with the text presented in separate boxes, so the process of extracting the text was difficult to automate. There were similar challenges with PDF files, pages from textbooks, exam papers, and worksheets; where the design and presentation of text meant that many manual adjustments were needed when converting the format. Considerable time was therefore needed to create a relatively small corpus.

3.3. Independent reading

Retrospective opt-out permission was used to access a list of reading materials submitted by students in a year 10 (age 14–15) mixed ability English class for a free choice reading homework task over a half-term holiday. Of the twenty-five students in the class: twenty-three submitted what they had read for homework (two students did not complete the original homework task); and twenty-one did not opt-out. One book was submitted twice (*One of Us is Lying* by Karen McManus), this left a list of twenty different source texts. These twenty texts consisted of: nine young adult (YA) fiction books; three newspaper articles, two autobiographies; two classic children's books; two modern literary fiction books; one crime/thriller fiction book; and one classic literary fiction book (see Table 2). One thousand words from the beginning of each text were collected, as Biber (1990) showed that 1000-word sub samples from texts, when compared, had high level of linguistic stability.

3.4. Creation of the corpora

Two corpora were created from the texts collected: 1) the CRC, using the 16 curriculum documents; 2) the IRC, using the 20 independent reading documents collected from the homework task. Details of the two corpora are given in Table 3. Documents were uploaded to the corpus tool *Sketch Engine* (Kilgarriff et al., 2014).

Table 1
Curriculum Subjects in the Curriculum Reading Corpus with Word Counts.

Subject	Word Count
Art	206
Computing	1000
Media	909
Technology	1000
English	1000
Geography	1000
History	1000
Maths	1000
Music	603
Physical Education	1000
Religious Studies	1000
Biology	437
Chemistry	1000
Physics	1000
Childcare	1000
Graphics	276

Table 2
Texts used to create Independent Reading Corpus.

No.	Text	Genre
1	*Harry Potter and the Chamber of Secrets* by J. K. Rowling	Young Adult Fiction
2	*Checkmate* by Malorie Blackman	Young Adult Fiction
3	*Harry Potter and the Deathly Hallows* by J. K. Rowling	Young Adult Fiction
4	*One Of Us Is Lying* by Karen McManus	Young Adult Fiction
5	*Rule of Wolves* by Leigh Bardugo	Young Adult Fiction
6	*Divergent* by Veronica Roth	Young Adult Fiction
7	*Twilight* by Stephenie Meyer	Young Adult Fiction
8	*Harry Potter and the Philosopher's Stone* by J. K. Rowling	Young Adult Fiction
9	*The Maze Runner* by James Dashner	Young Adult Fiction
10	News article from online daily newspaper for young people	Non-fiction (news)
11	News article from online daily newspaper for young people	Non-fiction (news)
12	Sports article from an online newspaper	Non-fiction (news)
13	*The Storyteller: Tales of Life and Music* by Dave Grohl	Autobiography
14	*I am Malala* by Malala Yousafzai	Autobiography
15	*Biggles of the Camel Squadron* by W. E. Johns	Classic Children's Fiction
16	*The BFG* by Roald Dahl	Classic Children's Fiction
17	*Everything I Never Told You* by Celeste Ng	Modern Literary Fiction
18	*Woman in Black* by Susan Hill	Modern Literary Fiction
19	*Body Language* by A. K. Turner	Crime/thriller
20	*The Great Gatsby* by F. Scott Fitzgerald	Classic Fiction

3.5. Frequencies of parts of speech and lexical diversity

To answer RQ1, what is the linguistic make up of students' curriculum and independent reading, total occurrences for nouns, verbs, adjectives, adverbs and other parts of speech were calculated for the CRC and IRC. A comparison of frequencies of parts of speech showed how the linguistic make up of these two corpora differed. In order to compare totals between corpora that are not the same size, frequencies need to be normalized. This was calculated by converting raw scores to frequency per million (fpm) (raw occurrences of part of speech/total words in corpus x 1,000,000). Chi squared tests of independence were used to compare whether differences between the frequencies of the parts of speech in each corpus were significant. Lexical diversity, which is measure of how many different (unique) words are used in a corpus was calculated using a type to token ratio (TTR) (Jarvis, 2013; Richards, 1987). This measure showed which of the two corpora contained the most unique words and therefore could potentially be a richer source of reading experience. A simple TTR can be calculated by dividing the number of types (unique words) by the number of tokens (total words) within a text or corpus, with higher scores representing higher diversity. However, this calculation does not account for the impact that the length of a text will have on this ratio (Covington and McFall, 2010; Kyle et al., 2021). To account for the sizes of the corpora, a moving-average

Table 3
Corpora Contents.

Corpus	Documents	Tokens	Words	Types (Unique Words)
Curriculum Reading Corpus	16	15, 574	13,210	3356
Independent Reading Corpus	20	25,467	21,553	5169

type-token ratio (MATTR) (Covington and McFall, 2010) was calculated for both corpora using the MATTR computer program (Covington and McFall, 2008) which averages the TTR for every rolling 500 words.

3.6. Most frequent words

Frequent words are important to study as these are the words that students are most likely to encounter in these different types of reading experiences and therefore gave an indication of how the vocabulary in the reading texts might be different. Word lists, which rank words by their frequency in the corpus, were produced from *Sketch Engine* for four parts of speech (nouns, verbs, adverbs and adjectives) from the CRC and IRC corpora. The parts of speech labels were allocated to the words in the corpora through the automatic tagger in *Sketch Engine*. The 100 most frequent words of each of the four parts of speech from the two corpora were compared using the MRC psycholinguistic database (Coltheart, 1981). The first measure used was word length to identify if there were significant differences in word lengths between the word lists from the two corpora. Word length is compared as longer words are an indication of greater difficulty (Carver, 1976). Longer words can also lead to longer processing times which can have a negative impact on comprehension (Martin-Chang et al., 2020). The two lists of the 100 most frequently occurring nouns were then compared to see if there were significant differences in concreteness and imageability. Concreteness is measure of the closeness of what the word denotes to a "perceptual entity" (Brysbaert et al., 2014, p. 904). A word that has a high concreteness score is understood to be easier to process because perceptual memory can be used, as compared to abstract words where it cannot (Brysbaert et al., 2014; Khanna and Cortese, 2021). Imageability scores give a measure of the extent to which the word is related to the senses and the formation of a mental image (Sadoski and Paivio, 2013). High scores for imageability indicate that the word is easier to process (Cortese and Fugett, 2004; Khanna and Cortese, 2021).

The word lists from the two corpora were then compared to identify which occurrences, in the 100 most frequent words in each of the four parts of speech, were common to both corpora and which occurrences were only in one of the corpora. Qualitative analysis was then conducted to further describe and compare the words on these eight lists.

3.7. Corpora comparisons

In order to see how far students' prior reading, represented by the CRC and IRC, matched the texts that they would need to comprehend in their English language GCSE (RQ3), a comparison was run in *Sketch Engine* (Kilgarriff et al., 2014) between the two corpora created for this study, the ETC created for a previous study (Jennings et al., 2024) and a range of reference corpora. The *Sketch Engine* comparison tool compares the keyword scores (frequency per million in the focus corpus divided by the frequency per million in the reference corpus) of the 5000 most frequent words in each corpus and then creates an overall comparison score from the mean of the highest 500.

4. Findings

4.1. Parts of speech in the corpora

Raw numbers and fpm are reported for occurrences for each part of speech in both corpora created for this study and for the ETC created in a previous study (Jennings et al., 2024) (see Table 4).

A chi-square test of independence was performed to examine the relationship between the frequency of different parts of speech in the CRC and IRC. For adjectives, the difference was not significant, χ^2 (1, N = 34,763) = 0.56, p = 45. For verbs, the difference was also not significant, χ^2 (1, N = 34,763) = 0.01, p = 91. However, for adverbs, the difference was significant, χ^2 (1, N = 34,763) = 174.21, p = < 0.01: adverbs were significantly more frequent in the IRC compared to the CRC. For nouns, the difference was also significant, χ^2 (1, N = 34,763) = 251.60, p = < 0.01, with nouns significantly more frequent in the CRC compared to the IRC.

Table 4
Raw Occurrences and Frequency per Million (in brackets) for Parts of Speech in the Corpora.

Corpus	Nouns (fpm)	Verbs (fpm)	Adjectives (fpm)	Adverbs (fpm)	Other[a] (fpm)
Curriculum Reading Corpus	4467 (338,153)	2424 (183,497)	1025 (77,593)	406 (30,734)	4888 (370,023)
Independent Reading Corpus	5576 (258,711)	3945 (183,037)	1625 (75,396)	1351 (62,683)	9056 (420,174)
Exam Text Corpus	8191 (223,890)	7085 (193,659)	2535 (69,291)	2396 (65,491)	16,378 (447,670)

Note.
[a] other includes: conjunctions, prepositions, pronouns and numerals.

Reference data on the frequencies of parts of speech in different registers (Biber et al., 1999) identifies verbs and adverbs being most common in conversation and fiction, nouns as being most common in newspaper language and then academic prose, and adjectives being most common in academic prose and then newspaper language. Whilst none of the four registers used in Biber et al. (1999) (conversation, fiction, newspaper language and academic prose) are a complete match for the make-up of the CRC and IRC, the frequencies of the parts of speech in them generally follow the same pattern. The CRC had significantly more nouns, as is found in newspaper language and academic prose. In contract, the IRC, which contains mostly fiction, some narrative non-fiction and three newspaper articles, had a significantly higher frequency of adverbs, which fits with adverbs being most common in fiction. The frequencies were closer for verbs and adjectives, perhaps due to the mix of registers contained in the two corpora.

4.2. Lexical diversity

Lexical diversity, measured by MATTR was slightly higher in the IRC (0.55) than the CRC (0.46), suggesting that the independent reading (mostly fiction) had a higher lexical diversity than the curriculum reading.

4.3. Comparing the most frequent words in the different parts of speech in the two corpora

Independent-samples *t*-tests were conducted to compare the number of letters, as an indication of difficulty, in the 100 most frequent words for each part of speech from the two corpora. Nouns in the CRC contained on average a higher number of letters than nouns in the IRC, and this was also the case for verbs and adjectives. For nouns there was a significant difference between the CRC (M = 5.48, SD = 1.99) and the IRC (M = 4.87, SD = 1.35), t (158) = 2.45, p = .02, two-sided. The effect size was small, with a Cohen's d of 0.36. Verbs had significantly more letters in the CRC (M = 5.44, SD = 1.84) than in the IRC (M = 4.37, SD = 1.20), t (171) = 4.87, p = < 0.001, two-sided. The effect size was medium, with a Cohen's d of 0.69. Adjectives also had significantly more letters in the CRC (M = 6.22, SD = 2.27) than in the IRC (M = 5.18, SD = 1.85), t (185) = 3.49, p = < 0.001, two-sided. The effect size was medium, with a Cohen's d of 0.50. For the length of adverbs there was no significant difference between the CRC (M = 6.04, SD = 2.59) and the IRC (M = 5.63, SD = 2.00), t (193) = 1.25, p = .211, two-sided. The effect size was small, with a Cohen's d of 0.18.

An independent samples *t*-test was also conducted to compare the concreteness and imageability scores of the 100 most frequent nouns in both corpora. Higher scores for both these attributes suggest lower difficulty. For concreteness the score was significantly lower for the CRC

($M = 450.54$, $SD = 100.18$) than the IRC ($M = 507.57$, $SD = 99.99$), $t(131) = -3.28$, $p = .001$, two-sided. The effect size was medium, with a Cohen's d of 0.57. For imageability the scores were also significantly lower for the CRC ($M = 469.48$, $SD = 91.15$) compared to the IRC ($M = 532.16$, $SD = 82.79$), $t(132) = -4.17$, $p = < 0.001$, two-sided. The effect size was medium, with a Cohen's d of 0.72. The lower scores in the CRC indicate that the words in the curriculum texts would be more difficult to comprehend.

Qualitative exploratory analysis, on the 100 most frequent words for each part of speech in the two corpora, was then conducted to identify any similarities and differences between them.

4.3.1. Nouns

Nouns are the most frequent word class (Biber et al., 1999) so it was perhaps to be expected that there was a high diversity of occurrences in the two noun frequency lists. Only 13 of the same nouns occurred in both corpora's top 100 for frequency. The nouns that appeared on both 100 most frequent word lists were all high frequency nouns and were concrete entities (e.g. *school, queen*) and qualities and states (e.g. *time, year, word, day*). The CRC top 100 nouns by frequency (Appendix B) had a small number of proper nouns (5) (e.g. *London, Essex, Elizabeth*). Common nouns were materials (e.g. *metal, copper, carbon*), were about space (e.g. *galaxy, earth, universe, sun*) or were to do with the classroom (e.g. *paper, line, mark*). There were a wide range of nouns that were about qualities or states (e.g. *probability, aggression, personality, spectrum*). In the IRC however, the top 100 nouns by frequency (Appendix A) there were 21 proper nouns, 20 of which were for people (e.g. *Harry, Voldemort, Lydia, Drogba*) and one for a place (*Chelsea*), a far higher number than in the CRC (5). Instead of common nouns that were topic based, the IRC featured 18 common nouns for domestic or everyday objects (e.g. *house, room, table, car*), eight common nouns for the body or parts of it (e.g. *eye, hand, hair, head*), and 12 common nouns for people (e.g. *man, mother, queen, brother*). Of the 22 nouns that denoted qualities or states, just over half related to time (e.g. *year, day, moment, night*), whereas the CRC only had 2 (*day* and *year*).

4.3.2. Verbs

There were far more shared verbs in the top 100 frequency lists of the CRC (Appendix D) and the IRC (Appendix C) than there were for nouns, with 43 verbs appearing on both lists. These are mostly simple, high frequency actions and states verbs (e.g. *be, do, have, create, learn*). *Be* and *do* are also always likely to be very frequent due to their grammatical use in tense building. The verbs that appear in the CRC top 100 most frequent, that are not shared in the IRC top 100, were different to those on the shared list and included: verbs that are parts of instructions for class tasks (e.g. *explain, describe, write, extract, identify*); verbs that are part of a mark scheme or answer sheet (e.g. *accept, demonstrate*) and verbs that describe causation or relationships (e.g. *help, develop, involve, increase, produce*). However, the verbs that only appear in the IRC top 100 are very similar to the shared ones and were mostly simple actions or states (e.g. *turn, feel, want, call, tell*).

4.3.3. Adjectives

Just under half (42) of the adjectives were on both the CRC (Appendix F) and IRC (Appendix E) top 100 most frequent, and these were largely physical qualities (e.g. *red, long, small, green, big*) or simple qualitative attributes (e.g. *good, different, important, major*). The adjectives that only appeared in the CRC top 100 featured abstract qualities (e.g. *relative, reactive, random, holistic*). Whereas the 58 adjectives in the IRC top 100, that did not appear in the CRC, were either similar to the those on the shared list that expressed physical qualities (e.g. *little, tall, hard, black, pale*) or had more complex qualitative attributes (e.g. *strange, magnificent, extraordinary, prominent*).

4.3.4. Adverbs

Adverbs were the part of speech that had the most crossover between the CRC (Appendix H) and the IRC (Appendix G) top 100 frequency lists. This would be expected as it is the smallest word class (Biber et al., 1999). Those that appeared on both lists were simple adverbs, including of time and place (e.g. *now, then, back, down*), and of manner (e.g. *quickly, especially, directly, exactly*). The adverbs that were only on the CRC list only contained technical examples of manner (e.g. *randomly, artificially, extrinsically, functionally, aesthetically*). Some adverbs that only appeared in the IRC were of time and place like the shared list (e.g. *finally, soon, forever, upwards, behind*), but most of manner but less technical than the CRC list (e.g. *obviously, purely, completely, barely, excitedly*).

4.4. Corpora comparisons

The comparisons between corpora are presented in Table 5. The comparison score represents the mean of the highest 500 scoring keywords created by calculating the frequency per million in one corpus divided by the frequency per million in the other, the closer the score to 1 the more alike the corpora are.

The IRC is a closer match to all the other corpora than the CRC, suggesting that the CRC is very particular in its register. The IRC is a closer match to the exam text corpus than the CRC.

5. Discussion

This exploratory analysis and comparison of two collections of adolescent reading material has revealed that the curriculum texts in the CRC contain a very particular set of vocabulary, that did not match either the IRC, the ETC or any of the other reference corpora. The CRC is obviously a very small corpus that only represents one week of curriculum materials, so its particular nature could be due to the high density of the specific topics covered in lessons that week. For example, in chemistry the topic was metals, in physics it was red shift and in history the 1601 rebellion by the Earl of Essex – all these topics featured in the top 100 most frequent word lists. However, there is no reason to think that these topics are not representative of the subjects from which they were taken. Moreover, what this shows is that the vocabulary and specifically the nouns being used in the curriculum reading are more challenging nouns that point to scientific terminology typical of informational written registers.

The significant differences between the number of letters in the 100 most frequent nouns, verbs and adjectives in the CRC and IRC, with the CRC nouns, verbs and adjectives having significantly more letters, suggested that, on the simple measure of word length, that the vocabulary challenge was higher in the CRC. Eye tracking studies have shown that, for adults and children, longer words have longer reading times (e.g. Joseph et al., 2009) and longer processing times for words can impact comprehension (Martin-Chang et al., 2020). The large overlap in the 100 most frequent adverbs in the two corpora probably accounts for the lack

Table 5
Comparison of Curriculum Reading Corpus and Independent Reading Corpus with Exam Text Corpus and Other Reference Corpora.

	Curriculum Reading Corpus	Independent Reading Corpus
Curriculum Reading Corpus	~~1.0~~	5.56
Independent Reading Corpus	5.56	~~1.0~~
Exam Text Corpus	5.44	2.83
British National Corpus (spoken part)	7.34	4.66
British National Corpus	4.33	2.72
Brown Family	4.35	2.62
Project Gutenberg	5.02	2.73
English Web 2015	4.04	3.09
English Broadsheet Newspapers	4.51	2.88
Cambridge Academic English	4.30	3.89

of significant difference in the number of letters between the two corpora for this part of speech.

The significant differences between the concreteness and imageability scores for the 100 most frequent nouns in the two corpora, with the scores being lower in the CRC, again suggests that the challenge of the vocabulary is greater in the CRC. Concreteness and imageability can indicate the closeness of the meaning of a word to perceptual experience, the idea being that the closer the meaning of a word is to perceptual experience, the easier the word is to process (Brysbaert et al., 2014; Cortese and Fugett, 2004; Khanna and Cortese, 2021). Therefore, since concreteness and imageability scores were lower for the most frequent nouns in the CRC, it suggests that these are harder words to process as the words are further from perceptual experience.

The qualitative exploratory analysis, of the word lists of the 100 most frequent words in each part of speech (nouns, verbs, adjectives and adverbs), revealed interesting differences between the two corpora. In the CRC there was a particular vocabulary group that was specific to the classroom and learning tasks, both in nouns (e.g. *paper, line, mark*) and verbs (e.g. *explain, describe, write, extract, identify, accept, demonstrate)*. Not surprisingly, the CRC also had subject-specific tier three vocabulary (Beck et al., 2002) that was specific to topics being studied in the week the curriculum texts were collected (e.g. *metal, copper, carbon, galaxy, earth, universe, sun*). There was also more abstract vocabulary in the CRC, across nouns (e.g. *probability, aggression, personality, spectrum*), adjectives (e.g. *relative, reactive, random, holistic*) and adverbs (e.g. *randomly, artificially, extrinsically, functionally, aesthetically*), demonstrating the more theoretical and scientific content of the curriculum materials. Despite the IRC having higher lexical diversity, it could be argued that the challenge of the words in the top 100 most frequent word lists from the CRC was much higher. Not only was there a set of words that were specific to the classroom and learning tasks but also words that require conceptual understanding, none of which were present in the top 100 frequency word lists of the IRC.

Whilst the majority of the independent reading, chosen by the class of year 10 students, was, as expected, by children's and YA authors, there were exceptions with autobiographies, fiction written for adults, and newspaper articles included in the choices. From this small sample at least, this suggests that analysis of children's and adolescent's reading materials should not focus solely on texts that are targeted at their age group. This is especially important with the age group in this study, mid-adolescents, as they transition from reading books by children's and YA authors to more mainstream and general genres (e.g. crime/thrillers) or to non-fiction genres (e.g. autobiography). Whilst acknowledging that the concept of genres is contested (Bawarshi and Reiff, 2010; Biber, 1990; Chandler, 1997; Sabao, 2014), and using genres to describe reading materials can only give an imperfect indication of the type of language that might be found within them, the different range of genres represented by the reading materials chosen does warrant attention.

The whole corpora comparisons supported the findings, from comparing the parts of speech, that the two corpora were different linguistically. The CRC seemed particularly unlike any of the other corpora, even a corpus of academic English. This suggests that there might be a real particularity to curriculum resources in schools. The high scores, and therefore large difference, between both the CRC and the IRC and the reference corpus of spoken language, supports the literature that reading is providing experience with different vocabulary to that which is experienced through listening (Braze et al., 2007; Cunningham, 2005; Landauer and Dumais, 1997; Tilstra et al., 2009). The IRC was a closer match to the ETC, that represents the vocabulary found in the English language GCSE, which is surprising as this qualification is meant to demonstrate proficiency for work and future study, rather than fiction reading ability, which is measured separately in the English literature GCSE. The closer relationship between independent (mostly fiction) reading and the corpus created from exam texts, suggests that it is the independent reading of fiction that is going to provide the best preparation for comprehending the reading texts in these high stakes exams. This is concerning as large numbers of students say that they do not read outside of school (Clark et al., 2023).

The policy ambition behind the construction of the new GCSE exams was 'to prepare young people better for the next steps in their education or employment' (Ofqual, 2013, p. 4). However, with the exam texts having so little in common with the curriculum materials collected for this study, it is hard to see how far the English language GCSE tests the comprehension abilities that will be needed for the curriculum materials in further or higher education – especially when it comes to the more abstract vocabulary found on the CRC but not the IRC most frequent word lists. It is also hard to see how the close match to the vocabulary found in fiction links to the literacy needs of most employers.

The differences in the frequencies of nouns and adverbs in the two corpora, suggests that both types of reading, curriculum and independent, are important in a students' reading experience, as they contain different proportions of parts of speech. This suggests that the types of texts read will impact the number of encounters readers could have with different types of words. For example, if students only read curriculum texts, then they are less likely to have experienced a wide range of adverbs. The slightly higher lexical diversity of the IRC, as measured by the MATTR could also suggest that independent reading offers experience with a wider range of vocabulary than curriculum reading. As the numbers of students who read independently outside of school regularly is decreasing (Clark et al., 2023) this will mean that students who do not read fiction independently could potentially miss out on the most lexically diverse texts. It is important to note however that these results are from the comparison of two very small corpora and further research would be needed with larger collections of text to support these exploratory findings.

We also want to be careful not to create a deficit narrative with these findings. The word 'gap' has been an influential concept in education in England in recent years (e.g. Department for Education, 2023a; Ofsted, 2022; Quigley, 2018, 2020). This concept of groups of children or students having a deficit or 'gap', compared to other groups, dates back to Hart and Risley's (1995) influential study in which they claimed that there was a thirty-million-word gap between the lowest socioeconomic group they studied (the 'welfare' group) and the highest (the 'professional' group). There has been further research on this perceived 'gap' (e.g.Duff and Brydon, 2020; Fernald et al., 2013; Sullivan et al., 2021) and Cushing (2023) has shown that it has been a very influential concept in the English educational context from 2010 to the present day. However, there has also been extensive critique of the deficit narrative (Baugh, 2017; Cushing, 2023; García and Otheguy, 2017; Johnson, 2015), where the concept of a 'gap' is seen as positioning the linguistic practices of traditionally powerful and dominant groups above those used by more marginalised groups and defining the difference between their practices as the marginalised group's deficit.

In order to avoid creating a simplistic deficit narrative in our findings, that of students' lack of independent reading being judged as deficient, as far as preparation for the language in the exam is concerned, what should also be questioned or critiqued instead is the rationale behind the choices of what is included in the exam. As this qualification operates as a gatekeeper to future study, training and work opportunities, it is important to question any assumptions or value judgements about what have been deemed to be appropriate texts to include in the exam. Older, literary texts are now required, instead of the multi-modal and more deliberately accessible texts used in the past (Isaacs, 2014), non-fiction choices must be 'extended literary' and 'transient' (online) texts are specifically listed as not to be included (Department for Education, 2013a, p. 4). These choices reveal an inherent valuing of literary and traditional genre forms, to the exclusion of new and non-literary forms and genres, that are explicitly devalued. This narrow focus could be considered as much of a 'deficit' as any so called 'gap' in students' reading. If experience of the vocabulary, found in the types of texts that have been specified for the exam, depends on the independent reading of fiction, then this could exclude students for

all sorts of reasons. There can be financial, social or time barriers to adolescents accessing the kinds of reading materials that will most likely prepare them for the vocabulary in their exams and there are also huge swathes of alternative types of texts and vocabulary that could be being read but are not currently being included in the texts in the exams.

6. Limitations and further study

The limited scope of this study meant that the corpora created, and the findings generated, were only ever intended to be exploratory rather than representative. The very small sample of independent reading was collected from a just one class of students and a wider range of participants would be desirable in the future as different students may make very different reading choices. The curriculum materials accessed were also only from a very small number of lessons, that took place in just one week. A greater number of texts from a greater number of lessons would create a larger corpus with which to test some of the initial findings from this paper. More sophisticated methods of extracting the text from highly designed formats like slides and PDFs could also help further study of classroom materials, as online resource formats continue to replace textbooks.

The difficulty with preparing the curriculum texts for uploading to the corpus tool highlights a research challenge now that classroom resources come in a wider variety of formats. With the growing use of slides and other formats that use sophisticated design features, collating and formatting classroom materials for corpus studies will be much more difficult than it was when there were standard textbooks that could be taken as a representation of what was being read in classrooms.

7. Conclusion

The English language GCSE is seen, in England, as an indication of a student's literacy ability and serves as a gatekeeping qualification for access to further and higher education and to employment and training. This study sought to create and explore two different types of reading that students are most likely to be exposed to: curriculum reading and independent reading. The curriculum reading was not as close a match for the vocabulary found in the exam texts as the independent reading. This suggests that unless students are reading independently outside of school, something that has been shown to be in decline, they will not have experience with, and therefore have had the chance to build sufficient knowledge of, the type of vocabulary that will be found in the exam.

However, instead of creating a simple deficit narrative, that some students are not reading enough independently or reading enough fiction, the choice of exam text should be critiqued too. The specification that the new exam should only have texts that are literary fiction and literary non-fiction, prioritises and values one genre of reading over any others. Students' ability with a range of fictional texts is already assessed in the English literature GCSE, instead of duplicating this valuing of fiction, maybe the English language GCSE should be filling in the 'gap' and including texts that are more like the curriculum texts that will be read in any future studies and also including texts that are common in the workplace and society. The exploratory analysis of the curriculum texts suggested that there may be a higher frequency of more abstract vocabulary, as well as a set of vocabulary that was exclusive to the classroom and learning activities. If the exam is used as an indication of having the reading skills needed for further study, then perhaps more vocabulary representative of curriculum materials should feature in the reading texts. There could also be an argument to consider other language practices, that will be useful in adult life, not just more formal and privileged language practices.

This exploratory study has shown that the collection and analysis of actual reading materials is possible, if challenging. Continued development in the methods and techniques of studying the content of reading experience, especially as it moves outside traditional formats, will help to improve our understanding of reading and reading content.

Funding

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Ethics statement

This project is part of the first author's PhD research which has been granted ethical approval by the Institute of Education at the University of Reading.

Data Availability Statement: Data is not available until PhD of 1st author is complete.

CRediT authorship contribution statement

Beverley Jennings: Writing – original draft, Software, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Daisy Powell:** Writing – review & editing, Supervision. **Sylvia Jaworska:** Writing – review & editing, Methodology. **Holly Joseph:** Writing – review & editing, Supervision.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

We thank all the participating school staff and students for their contributions to this study.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.acorp.2025.100124.

References

Acheson, D.J., Wells, J.B., MacDonald, M.C., 2008. New and updated tests of print exposure and reading abilities in college students. Behav. Res. Methods. https://doi.org/10.3758/BRM.40.1.278.

Baugh, J., 2017. Meaning-less differences: exposing fallacies and flaws in "the word gap" hypothesis that conceal a dangerous "language trap" for low-income American families and their children. Int. multiling. res. j. 11 (1), 39–51. https://doi.org/10.1080/19313152.2016.1258189.

Bawarshi, A.S., Reiff, M.J., 2010. Genre: An Introduction to History, Theory, Research, and Pedagogy. Parlor Press.

Beck, I.L., McKeown, M.G., Kucan, L., 2002. Bringing Words to Life. The Guildford Press.

Biber, D., 1990. Methodological issues regarding corpus-based analyses of linguistic variation. Liter. Linguis. Comput. 5 (4), 257–269. https://doi.org/10.1093/llc/5.4.257.

Biber, D., Johansson, S., Leech, G., Conrad, S., Finegan, E., 1999. Longman Grammar of Spoken and Written English. Pearson Education.

Biber, D., Conrad, S., 2009. Register, Genre, and Style. Cambridge University Press.

Braze, D., Tabor, W., Shankweiler, D.P., Mencl, W.E., 2007. Speaking up for vocabulary: reading skill differences in young adults. J. Learn Disabil. 40 (3), 226–243. https://doi.org/10.1177/00222194070400030401.

Brysbaert, M., Warriner, A.B., Kuperman, V., 2014. Concreteness ratings for 40 thousand generally known English word lemmas. Behav. Res. Methods 46 (3), 904–911. https://doi.org/10.3758/s13428-013-0403-5.

Cameron-Faulkner, T., Noble, C., 2013. A comparison of book text and Child Directed Speech. First Lang. 33 (3), 268–279. https://doi.org/10.1177/0142723713487613.

Carver, R.P., 1976. Word length, prose difficulty, and reading rate. J. Liter. Res. 8 (2). https://doi.org/10.1080/10862967609547176.

Chandler, D., 1997. An Introduction to Genre Theory. http://visual-memory.co.uk/daniel//Documents/intgenre/chandler_genre_theory.pdf.

Chateau, D., Jared, D., 2000. Exposure to print and word recognition processes. Memory Cognit. https://doi.org/10.3758/BF03211582.

Clark, C., Picton, I., Galway, M., 2023. Children and Young People's Reading in 2023.

Clark, C., Rumbold, K., 2006. Reading For Pleasure: A research Overview.

Coltheart, M., 1981. The mrc psycholinguistic database. Quarter. J.Experiment. Psyc. Sec. A 33 (4). https://doi.org/10.1080/14640748108400805.
Cortese, M.J., Fugett, A., 2004. Imageability ratings for 3,000 monosyllabic words. In: Behavior Research Methods, Instruments, and Computers, 36. https://doi.org/10.3758/BF03195585.
Covington, M.A., McFall, J., 2008. *MATTR 2.0.3018.28419 2008/04/07* (2.0). CASPR Project.
Covington, M.A., McFall, J.D., 2010. Cutting the gordian knot: the moving-average type-token ratio (MATTR). J. Quant. Linguist. 17 (2), 94–100. https://doi.org/10.1080/09296171003643098.
Coxhead, A., 2000. A New Academic Word List. TESOL Quarter. 34 (2), 213. https://doi.org/10.2307/3587951.
Coxhead, A., 2011. The academic word list 10 years on: research and teaching implications. TESOL Quarter. 45 (2), 355–362. https://doi.org/10.5054/tq.2011.254528.
Coxhead, A., Stevens, L., Tinkle, J., 2010. Why might secondary science textbooks be difficult to read. New Zealand Studies Appl. Linguist. 16 (2), 1.
Cunningham, Anne.E., 2005. Vocabulary growth through independent reading and reading aloud to children. In: Kamil, M., Hiebert, E. (Eds.), Teaching and Learning New vocabulary: Bringing research to Practice. Erlbaum, pp. 45–68.
Cushing, I., 2023. Word rich or word poor? Deficit discourses, raciolinguistic ideologies and the resurgence of the 'word gap' in England's education policy. Crit. Inq. Lang. Stud. 20 (4), 305–331. https://doi.org/10.1080/15427587.2022.2102014.
Davidse, N.J., de Jong, M.T., Bus, A.G., Huijbregts, S.C.J., Swaab, H., 2011. Cognitive and environmental predictors of early literacy skills. Read. Writ. 24 (4). https://doi.org/10.1007/s11145-010-9233-3.
Dawson, N., Hsiao, Y., Wei, A., Tan, M., Banerji, N., 2021. Features of lexical richness in children's books: comparisons with child-directed speech. Lang. Dev. Res. https://doi.org/10.34842/5we1-yk94.
Deignan, A., Candarli, D., Oxley, F., 2022. The Linguistic Challenge of the Transition to Secondary School. Routledge. https://doi.org/10.4324/9781003081890.
Deignan, A., Love, R., 2019. Using corpus methods to identify subject specific uses of polysemous words in English secondary school science materials. Corpora. https://eprints.whiterose.ac.uk/154115/.
Department for Education, 2013. English language: GCSE subject content and assessment objectives. English Language GCSE Subject Content and Assessment Objectives. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/254497/GCSE_English_language.pdf.
Department for Education, 2023a. The Reading Framework.
Department for Education, 2023b. Key Stage 2 Attainment. December 14. National Statistics. https://explore-education-statistics.service.gov.uk/find-statistics/key-stage-2-attainment/2022-23.
Duff, D., Brydon, M., 2020. Estimates of individual differences in vocabulary size in English: how many words are needed to 'close the vocabulary gap'? J. Res. Read. 43 (4), 454–481. https://doi.org/10.1111/1467-9817.12322.
Fernald, A., Marchman, V.A., Weisleder, A., 2013. SES differences in language processing skill and vocabulary are evident at 18 months. Dev. Sci. 16 (2), 234–248. https://doi.org/10.1111/desc.12019.
Francis, D.J., Fletcher, J.M., Catts, H.W., Tomblin, J.B., 2005. Dimensions affecting the assessment of reading comprehension. Children's Reading Comprehension and Assessment. https://doi.org/10.4324/9781410612762.
García, O., Otheguy, R., 2017. Interrogating the Language Gap of Young Bilingual and Bidialectal Students. Int. multiling. res. j. 11 (1), 52–65. https://doi.org/10.1080/19313152.2016.1258190.
Gardner, D., Davies, M., 2014. A new academic vocabulary list. Appl. Linguist. 35 (3), 305–327. https://doi.org/10.1093/applin/amt015.
Gough, P.B., Hoover, W., Peterson, C., 1996. Some observations on a simple view of reading. Reading Comprehension difficulties: Processes and Intervention.
Gough, P.B., Tunmer, W.E., 1986. Decoding, Reading, and Reading Disability. Remedi. Special Educ. 7 (1), 6–10. https://doi.org/10.1177/074193258600700104.
Green, C., Lambert, J., 2018. Advancing disciplinary literacy through English for academic purposes: discipline-specific wordlists, collocations and word families for eight secondary subjects. J. Engl. Acad. Purp. 35, 105–115. https://doi.org/10.1016/j.jeap.2018.07.004.
Green, C., Lambert, J., 2019. Position vectors, homologous chromosomes and gamma rays: promoting disciplinary literacy through secondary phrase lists. Eng. Specific Purposes 53, 1–12. https://doi.org/10.1016/j.esp.2018.08.004.
Hart, B., Risley, T.R., 1995. Meaningful Differences in the Everyday Experience of Young American children. Paul H Brookes Publishing.
Henderson, L., Snowling, M., Clarke, P., 2013. Accessing, integrating, and inhibiting word meaning in poor comprehenders. Sci. Studies Read. 17 (3), 177–198. https://doi.org/10.1080/10888438.2011.652721.
Hoover, W.A., Gough, P.B., 1990. The simple view of reading. Read. Writ. 2 (2), 127–160. https://doi.org/10.1007/BF00401799.
Hsiao, Y., Dawson, N.J., Banerji, N., Nation, K., 2022. The nature and frequency of relative clauses in the language children hear and the language children read: a developmental cross-corpus analysis of English complex grammar. J. Child Lang. https://doi.org/10.1017/S0305000921000957.
Hyland, K., 2008. As can be seen: lexical bundles and disciplinary variation. Engl. Specific Purposes 27 (1), 4–21. https://doi.org/10.1016/j.esp.2007.06.001.
Hyland, K., 2017. English in the disciplines: arguments for specificity. ESP Today 5 (1), 5–23. https://doi.org/10.18485/esptoday.2017.5.1.1.
Hyland, K., Tse, P., 2007. Is There an "Academic Vocabulary"?. In: Quarterly, 41.
Ingram, J., Stiff, J., Cadwallader, S., Lee, G., Kayton, H., 2023. PISA 2022: National Report For England December 2023.
Isaacs, T., 2014. Curriculum and assessment reform gone wrong: the perfect storm of GCSE English. Curriculum J. 25 (1), 130–147. https://doi.org/10.1080/09585176.2013.876366.
Jarvis, S., 2013. Capturing the Diversity in lexical diversity. Lang. Learn. 63 (SUPPL. 1), 87–106. https://doi.org/10.1111/j.1467-9922.2012.00739.x.
Johnson, E.J., 2015. Debunking the "language gap. J. Multicult. Educ. 9 (1), 42–50. https://doi.org/10.1108/JME-12-2014-0044.
Joseph, H., Liversedge, S.P., Blythe, H.I., White, S.J., Rayner, K., 2009. Word length and landing position effects during reading in children and adults. Vision Res. 49 (16), 2078–2086. https://doi.org/10.1016/j.visres.2009.05.015.
Joseph, H., Nation, K., 2018. Examining incidental word learning during reading in children: the role of context. J. Exp. Child Psychol. 166, 190–211. https://doi.org/10.1016/j.jecp.2017.08.010.
Khanna, M.M., Cortese, M.J., 2021. How well imageability, concreteness, perceptual strength, and action strength predict recognition memory, lexical decision, and reading aloud performance. Memory 29 (5), 622–636. https://doi.org/10.1080/09658211.2021.1924789.
Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., Suchomel, V., 2014. The Sketch Engine: ten years on. Lexicography. https://doi.org/10.1007/s40607-014-0009-9.
Kyle, K., Crossley, S.A., Jarvis, S., 2021. Assessing the Validity of Lexical Diversity Indices Using Direct Judgements. Lang. Assess. Q 18 (2), 154–170. https://doi.org/10.1080/15434303.2020.1844205.
Landauer, T.K., Dumais, S.T., 1997. A solution to Plato's problem: the latent semantic analysis theory of acquisition, induction, and representation of knowledge. Psychol. Rev. 104 (2), 211–240. https://doi.org/10.1037/0033-295X.104.2.211.
Mar, R.A., Rain, M., 2015. Narrative fiction and expository nonfiction differentially predict verbal ability. Sci. Studies Read. 19 (6), 419–433. https://doi.org/10.1080/10888438.2015.1069296.
Martin-Chang, S., Kozak, S., Rossi, M., 2020. Time to read Young Adult fiction: print exposure and linguistic correlates in adolescents. Read. Writ. 33 (3), 741–760. https://doi.org/10.1007/s11145-019-09987-y.
McGeown, S.P., Duncan, L.G., Griffiths, Y.M., Stothard, S.E., 2015. Exploring the relationship between adolescent's reading skills, reading motivation and reading habits. Read. Writ. 28 (4), 545–569. https://doi.org/10.1007/s11145-014-9537-9.
Mol, S.E., Bus, A.G., 2011. To read or not to read: a meta-analysis of print exposure from infancy to early adulthood. Psychol. Bull. https://doi.org/10.1037/a0021890.
Monaghan, F., 1999. Judging a word by the company it keeps: the use of concordancing software to explore aspects of the mathematics register. Lang.Educ. 13 (1), 59. https://doi.org/10.1080/09500789908666759.
Montag, J.L., 2019. Differences in sentence complexity in the text of children's picture books and child-directed speech. First Lang. 39 (5), 527–546. https://doi.org/10.1177/0142723719849996.
Montag, J.L., Jones, M.N., Smith, L.B., 2015. The words children hear: picture books and the statistics for language learning. Psychol. Sci. 26 (9), 1489–1496. https://doi.org/10.1177/0956797615594361.
Montag, J.L., MacDonald, M.C., 2015. Text exposure predicts spoken production of complex sentences in 8-and 12-year-old children and adults. J. Experiment. Psychol. General 144 (2), 447–468. https://doi.org/10.1037/xge0000054.
Nagy, W.E., Anderson, R.C., Herman, P.A., 1987. Learning Word Meanings From Context During Normal Reading. Am. Educ. Res. J. 27 (2), 237–270. https://doi.org/10.3102/00028312024002237.
Nagy, W.E., Herman, P.A., 1984. Limitations of vocabulary instruction. Centre For the Study of Reading.
Nation, K., 2017. Nurturing a lexical legacy: reading experience is critical for the development of word reading skill. NPJ. Sci. Learn. 2 (1), 3. https://doi.org/10.1038/s41539-017-0004-7.
Nation, K., Snowling, M.J., 1998. Semantic processing and the development of word-recognition skills: evidence from children with reading comprehension difficulties. J. Mem. Lang. 39 (1), 85–101. https://doi.org/10.1006/jmla.1998.2564.
Ofqual, 2013. Reforms to GCSEs in England from 2015. Issue November 2013.
Ofsted, 2022. Research Review Series: English. https://www.gov.uk/government/publications/curriculum-research-review-series-english/curriculum-research-review-series-english.
Ouellette, G.P., 2006. What's meaning got to do with it: the role of vocabulary in word reading and reading comprehension. J. Educ. Psychol. https://doi.org/10.1037/0022-0663.98.3.554.
Pagán, A., Nation, K., 2019. Learning words via reading: contextual diversity, spacing, and retrieval effects in adults. Cogn. Sci. 43 (1). https://doi.org/10.1111/cogs.12705.
Perfetti, C.A., 2007. Reading ability: lexical quality to comprehension. Sci. Studies Read. 11 (4), 357–383. https://doi.org/10.1080/10888430701530730.
Perfetti, C.A., Hart, L., 2002. The Lexical Quality Hypothesis, pp. 189–213. https://doi.org/10.1075/swll.11.14per.
Pfost, M., Dörfler, T., Artelt, C., 2013. Students' extracurricular reading behavior and the development of vocabulary and reading comprehension. Learn. Individ. Differ. 26, 89–102. https://doi.org/10.1016/j.lindif.2013.04.008.
Quigley, A., 2018. Closing the [Vocabulary] Gap. Routledge.
Quigley, A., 2020. Closing the Reading Gap. Routledge.
Richards, B., 1987. Type/token ratios: what do they really tell us? J. Child Lang. 14 (2), 201–209. https://doi.org/10.1017/S0305000900012885.
Rosa, E., Salom, R., Perea, M., 2022. Contextual diversity favors the learning of new words in children regardless of their comprehension skills. J. Exp. Child Psychol. 214. https://doi.org/10.1016/j.jecp.2021.105312.
Rosa, E., Tapia, J.L., Perea, M., 2017. Contextual diversity facilitates learning new words in the classroom. PLoS One 12 (6). https://doi.org/10.1371/journal.pone.0179004.

Sabao, C., 2014. Towards a theory of genre ? Reflections on the problems and debates on theorising ‘ genre’. The Dyke 8 (2).
Sadoski, M., Paivio, A., 2013. Imagery and text: a dual coding theory of reading and writing: second edition. Imagery and Text: A Dual Coding Theory of Reading and Writing: Second Edition. https://doi.org/10.4324/9780203801932.
Schleppegrell, M.J., 2001. Linguistic Features of the Language of Schooling.
Schleppegrell, M.J., 2007. The linguistic challenges of mathematics teaching and learning: a research review. In: Reading and Writing Quarterly, 23, pp. 139–159. https://doi.org/10.1080/10573560601158461.
Shanahan, T., Shanahan, C., 2017. Disciplinary literacy: just the FAQs. Educ. Leadersh. 74, 18–22.
Sullivan, A., Moulton, V., Fitzsimons, E., 2021. The intergenerational transmission of language skill. British J. Sociol. 72 (2), 207–232. https://doi.org/10.1111/1468-4446.12780.
Tilstra, J., McMaster, K., Van den Broek, P., Kendeou, P., Rapp, D., 2009. Simple but complex: components of the simple view of reading across grade levels. J. Res. Read. 32 (4), 383–401. https://doi.org/10.1111/j.1467-9817.2009.01401.x.
Topping, K., Clark, C., Picton, I., Cole, A., 2023. What and How Kids Are Reading: The Book-Reading Behaviours of Pupils. https://renaissance.widen.net/view/pdf/priipdurrj/UK-What-Kids-Are-Reading-report-2023.pdf?t.download=true&u=zceria.
University of Arizona Libraries, 2021. Corpus of Contemporary American English (COCA) 1990 to 2012 Dataset Version 2). University of Arizona Research Data Repository.
Wild, K., Kilgarriff, A., Tugwell, D., 2013. The oxford children’s corpus: using a children’s corpus in lexicography. Int. J. Lexicogr. 26 (2), 190–218. https://doi.org/10.1093/ijl/ecs017.
Jennings, B., Powell, D., Jaworska, S., Joseph, H., 2024. A Corpus Study of English Language Exam Texts: Vocabulary Difficulty and the Impact on Students’ Wider Reading (or Should Students be Reading More Texts by Dead White Men?). J. Adolesc. Adult Lit. 67, 303–316. https://doi.org/10.1002/jaal.1331.