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?)

www.reading.ac.uk/centaur

CentAUR

Central Archive at the University of Reading

Reading's research outputs online

ARTICLE

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?)

Beverley Jennings ⬤ | Daisy Powell | Sylvia Jaworska | Holly Joseph

INTRODUCTION

Nationally set external exams have been a feature of the educational system in England for more than a century. These high-stakes exams inevitably have an impact on the curriculum that is taught in schools, as the grades achieved by students effect their education and work choices post-16. This study uses corpus linguistics to analyze the type of vocabulary that is found in a new format of one of the most important of these exams, the English language GCSE. The type of vocabulary that features in these exams is identified and the likely genres of reading that could help students build their knowledge of this vocabulary is also found through comparisons with reference corpora.

Background

Education in England is divided into four key stages: Key Stages 1 and 2 (ages 4–11) are taught in primary schools; Key Stages 3 and 4 (ages 12–16) are taught in secondary schools. At the end of Key Stage 4, there are national examinations in each different subject called the General Certificates of Secondary Education (GCSEs). These qualifications are administered by independent exam boards who are regulated by a government department called the Office of Qualifications and Examinations Regulation (Ofqual). Results in these GCSE exams are then used to gain admittance to post-16 education options, which include qualifications in traditional academic subjects at a school or college; vocational qualifications at a college; and apprenticeships or traineeships. Most of these options require students to have a minimum of a Grade 4 (previously grade C) in both English language and maths, as these are taken to indicate a competent level of literacy and numeracy. Any students without the minimum pass grades are required to retake the qualifications as part of their post-16 option. This makes the English language and maths qualifications very high stakes for students. GCSE results, in these two subjects especially, are also very high stakes for schools as pass rates are published by the government and are used to judge school performance and effectiveness.

Externally set exams have been a feature of the education system in England since the middle of the 19th century, with the first national qualification for 16-year-olds introduced in 1918. The content of the curriculum in England has therefore been influenced by externally set exams for over a century. Schools and teachers in England are therefore used to having a Key Stage 4 curriculum that is focused on high-stakes external exams and teaching a curriculum that is heavily influenced by content that is set by the Government DfE, regulated by Ofqual and administered by independent exam boards. While there may be a general belief that testing has the ability to raise standards (e.g., Hart & Teeter, 2001; Mitchell, 1997), high-stakes testing and exams have also been found to have a detrimental effect on teaching practices and curriculum decisions (Brown, 2015; Jennings & Bearak, 2014; Volante, 2004). The design and content of any test has the potential to skew classroom practices in favor of drills and practice testing (Sacks, 2000) and to eliminate any curriculum content that is not predicted to be on the test (Volante, 2004). The content of national exams, and their potential impact on teaching practices, is therefore an important area of research.

The current study focused on the reading part of a new specification of the English language GCSE. The format of this exam changed in 2017, as part of wider government reforms intended to raise standards, from being partly assessed through coursework, an oral assessment and exams that could be taken at several different points during the course, to being linear and solely assessed through two written exams at the end of the course. The unseen reading part of the exam is now worth 50% rather than 20% of the total, with the remaining 50% testing students' writing ability. The form and age of the reading texts also changed from the previous specification where "cultural diversity, multimodal study and connections to the real world and daily life" were more of a focus (Isaacs, 2014). Now the Government's Department for Education (DfE) specified that the texts must provide a high challenge and be in a traditional form such as an essay, review or print journalism and explicitly excluded forms of writing found online. They also specified that the exam texts must be literature or literary nonfiction and be drawn from each of the last three centuries (19th, 20th, and 21st) (Department for Education, 2013). An additional difference to the previous qualification format was that there are no longer two exams at different levels, one with more accessible texts for students working within the lower half of the grade range (C-G) and one with more challenging texts for the higher grades (A*-C). In the new specification, all students sit the same exams and read the same challenging texts (Grades awarded: 9-1).

Preparing students to successfully comprehend previously unseen literary texts that have been drawn from the previous two centuries has therefore become much more of a focus in Key Stage 4 English classrooms, and the potential impact of the new English language GCSE is relatively new. Text comprehension involves many different levels of processing, from decoding and understanding word meanings to working out the structure of the text and constructing a situation model (Kintsch et al., 2005). It also involves comprehension skills such as inference making and comprehension monitoring (Perfetti et al., 2005). While these are all worthy of study, at the heart of many of these processes is having access to a wide vocabulary which enables a reader to efficiently process texts, thus freeing up resources needed for high-level comprehension processes. Vocabulary is therefore the focus of the current study.

Vocabulary knowledge and reading experience

One of the strongest predictors of successful reading comprehension is vocabulary knowledge: at a very simple level, if you do not know what words mean (or have only basic knowledge of their meaning) in a text, then you cannot understand the text. This is especially the case for older students as reading materials increase in difficulty (Braze et al., 2007; Henderson et al., 2013; James et al., 2020; Lervåg et al., 2018; Nation & Snowling, 1998; Tilstra et al., 2009). Perfetti and Hart's lexical quality hypothesis (LQH) (Perfetti & Hart, 2002) describes high-quality vocabulary knowledge, needed for successful comprehension, as depending on repeated exposures to words through reading experience. Building on this, Nation's lexical legacy hypothesis (LLH) (Nation, 2017) suggests that exposures to words need to be multiple and diverse so that readers gradually build lexical quality as they encounter words in different contexts over time. For example, the word "crest" appears three times in the exam texts collected for the current study. One time it refers to a heraldic emblem on a tin; the other two times, it is referring to the top of a wave. These are two quite distinct meanings, as is a third possible meaning, part of the head of a bird or animal. Knowledge of these distinct meanings and more nuanced understandings within them, for example, that "crest" can also refer to the top of a mountain as well as a wave, would need to be built through repeated diverse experiences with the word.

Previous studies of the relationship between reading experience and reading ability have shown that it is fiction book reading that improves reading comprehension performance, rather than the reading of nonfiction, magazines, newspapers, or digital reading (McGeown et al., 2015; Pfost et al., 2013; Torppa et al., 2019). It is not clear from these studies why fiction book reading was a superior predictor, but it seems likely that it provides more diverse contexts within which to encounter and reencounter the kind of vocabulary that is found in the standardized and researcher developed reading measures that were used. What this study examines, using corpus linguistic methods, is the nature of the vocabulary in the reading tests themselves (in this case the English language GCSE exam) and then the genres in which that vocabulary is most likely to be found. If, as previously studies suggest, source genres for the vocabulary in the English language GCSE exams are predominantly fiction genres, then this could provide important information for practitioners. Whether or not practitioners should recommend particular genres of reading to students, on the basis that they may provide multiple exposures to the type of vocabulary that will be in the exam, or whether this could instead be regarded as "teaching to the test" is a matter for teachers, policy makers, and test developers to discuss.

Corpus linguistics

Corpus linguistics is the study of lexical and grammatical patterns in a body (corpus) or bodies (corpora) of texts using both quantitative and qualitative methods (Biber et al., 1998). It uses computer software, for example, corpus packages like *Sketch Engine*,

AntConc, *#Lancsbox*, and *Wmatrix*, to automatically retrieve and analyze language use (Anthony, 2013). Corpora can be used for a variety of research purposes, for example: by lexicographers as empirical frequency data for dictionary entries; by applied linguists to study language use in specific contexts and registers; and by language teachers and learners to explore language use with the view to inform pedagogical practice. Corpus software, like *Sketch Engine* (Kilgarriff et al., 2014), gives access to reference corpora, large collections of texts created to be representative of certain registers or genres, which can be compared to other purpose-built or smaller specialist corpora. New corpora can be created by uploading texts into the corpus software, which then allows the language in any new corpus to be interrogated using analytical tools such as frequencies, concordances, collocations, and keywords. These tools allow for an empirical, more systematic, and consistent analysis of words and their uses in larger datasets and for discovering patterns that might simply escape the attention of an analyst performing solely qualitative analysis based on "manual" reading of texts. They also reduce the possibility of human error when counting words and minimize certain biases such as primacy bias that might inadvertently influence qualitative research of texts and vocabulary therein.

This study

Previous research has highlighted the importance of vocabulary knowledge to reading comprehension and also the importance of reading experience to building this vocabulary knowledge. The key aims of the current study were to create a corpus of a sample of the texts used in the new exams to (1) identify vocabulary that is typical of the exam texts and (2) identify in which types of reading this vocabulary is most likely to be found. This could then suggest which genres of reading would provide the best reading experience for the types of vocabulary found in the exams.

The research questions for this study were as follows:

RQ1—What type of vocabulary is typical of the exam texts?

RQ2—In what types of reading material is the vocabulary that typifies the exam texts most likely to be found?

METHOD

Data

A small, specialized corpus was created in the corpus tool *Sketch Engine* (Kilgarriff et al., 2014), to be referred to here as the "Exam Text Corpus" (ETC), to represent the reading extracts from the English language GCSE exams. The extracts were sourced from the sample assessment materials and the three sets of past papers that were publicly available when the data were collected (June 2017, November 2017, and June 2018) from the four awarding exam boards in England: AQA, Edexcel, Educas, and OCR. In total, there were 59 extracts available from the exam board websites. The ETC contains 36585 words, of which 6854 are unique. The documents were categorized as fiction or nonfiction, and by their century of publication (see Tables 1 and 2).

The 59 exam texts were then divided into the David Lee Categories (Lee, 2001), which are genre categories devised by Lee from the contents of the British National Corpus (BNC; see Table 3). Lee created these categories, after the publication of the BNC, because he argued that the existing classification of texts within the corpus was too broad and that researchers would benefit from being able to identify specific genre categories. Lee took "genre" to mean a culturally constructed type, as compared to "register" which described linguistic patterns. Lee carried out the classification himself, which not only gives consistency but also means that it should be remembered that the categorizations are subjective. The final 70 categories, of which 24 are different genres of spoken language (e.g., broadcast news, conversations, courtroom speech, and meetings) and 46 are different genres of written texts (e.g., biography, prose fiction, letters, and newspapers), were decided upon to represent as far as possible the widest range of not only all the different types of texts in the BNC but also with reference to categories used in other

TABLE 1 Document distribution in the ETC: 19th, 20th, and 21st century and fiction and nonfiction subcorpora.

	19th century	20th century	21st century
Fiction	6	11	2
Nonfiction	14	10	16

TABLE 2 Words, unique words, and documents in the ETC and subcorpora.

	Words	Unique words	Documents
Whole corpus	36585	6854	59
Subcorpora			
Fiction	14946	3328	19
Nonfiction	21639	5154	40
19th century	11060	2856	20
20th century	15008	3455	21
21st century	10517	3264	18

TABLE 3 Exam texts by David Lee Genre Categories.

David Lee Genre Category	Number of exam texts in category
Scripted speech	1
Biography	17
Essay	1
Prose fiction	19
Personal letters	3
Miscellaneous	3
Newspaper articles	13
Popular magazines	2

corpora to enable comparisons (Lee, 2001). Using the David Lee Categories also allows for distinction, within BNC's wide genre categories of "imaginative" and "informative" written texts, of subgenres that may blur the boundaries of the genres. For example, biographical writing (informative writing) may have more in common, as a register, with fictional prose (imaginative writing), than it does with other informative writing (e.g., commercial writing).

Keywords

Keyword lists are generated in *Sketch Engine* by comparing the focus corpus, the ETC, to a reference corpus. The keywords are individual words that appear more frequently in the focus corpus than in the reference corpus. This is calculated by dividing the frequency per million (fpm) of each word in the focus corpus by the fpm of the same word in the reference corpus and by adding the simple maths parameter to account for the zero problem in divisions (Kilgarriff, 2005). The bigger the difference between the two fpm values, the higher the keyness score of the word. Keywords can therefore show what is specific or different about the language in the focus corpus compared to general language, as represented by the reference corpus (Evison, 2015; Kilgarriff et al., 2014). The keywords generated from the ETC were able to provide the answer to RQ1, what type of vocabulary is typical of the exam texts. The BNC was chosen as the reference corpus as it was designed to represent a cross section of both spoken and written British English and so best represents language that is generally used or experienced by students.

According to Koester (2015), there can be problems of local density in small corpora, like the ETC, where a word appears more frequently in the corpus due to just one document. Keywords were, therefore, selected from the 1000 generated by *Sketch Engine*, using the following criteria: (1) they appeared at least twice in the corpus, (2) they appeared in more than one exam and (3), they appeared as only one part of speech. *Louder*

was removed as a keyword as its selection was due to a tagging error in *Sketch Engine*.

After applying these criteria, 146 keywords remained from the initial list of 1000. Fpm was recorded for each keyword from the BNC to give an indication of the frequency in general language (see Appendix A).

Comparisons with other corpora

By selecting a range of register-specific corpora, it was possible to see which genres of texts were most like the ETC and were therefore most likely to contain the vocabulary in the exam texts (RQ2). The corpora were selected to represent the different types of texts that students may encounter and so came from a range of sources and dates of publication (see Table 5).

The corpus comparison tool in *SketchEngine* compares two corpora at a time by taking the 5000 most frequent words from each corpus and calculating keyword scores for words that are in both corpora. The mean of the highest 500 keyword scores becomes the overall score for the comparison. The lower the overall score is, then the closer the match between the reference corpus and the focus corpus. Comparing the ETC, which is the focus corpus in this study, to a range of reference corpora demonstrates the kind of registers and genres with which it aligns and from which it differs.

Genre sources for keywords

While the whole corpora comparison identified likely genre sources for the vocabulary in the ETC by comparing it as a whole to other whole corpora, this was explored further by searching the BNC for the 146 keywords from the ETC and identifying the specific genres sources for them. The BNC was selected as the reference corpus for these searches as it has the most diverse and specifically labeled range of genres. The raw frequencies and fpm for the ETC keywords were found for the general registers of spoken texts and written texts and then the raw frequencies and fpm were found for two subcategories within the written texts: written imaginative and written informative.

Sketch Engine uses relative frequencies to calculate how likely it is for a word to appear within one of the David Lee genre categories, compared to the whole corpus. This is calculated by taking the number of occurrences in a genre category divided by the total occurrences in the whole corpus and then dividing by the size of the genre category within the corpus. A score of 100 would mean there was an equal likelihood of finding the word in the David Lee genre category as in the corpus as a whole. As scores rise above 100, the relative frequency of the word in the genre increases; if the score is below 100, then

the relative frequency was lower in the genre than the corpus as a whole. For example, the noun *tea* has a score of 5 in the Hansard category (the record of debates in the UK Parliament), a score of 102 in the arts sections of regional newspapers and a score of 395 in spoken conversations. These relative frequencies show that *tea* is very infrequent in parliamentary debates, appears in the arts sections of regional newspapers with about the same frequency as the whole corpus, and is a more frequently occurring word in spoken conversations. Those genres with higher relative frequencies are likely to be the categories in which the vocabulary that typifies the exam texts is most likely to be found (RQ2).

FINDINGS

Keywords

The 146 keywords, which are the words that typify the ETC (RQ1), are displayed in Table 4 (see Appendix A for a more detailed list). It would be expected that

a corpus of written text would contain more low-frequency words than spoken language (Korochkina et al., 2023). Words with a fpm of less than five are considered to be low frequency in general language (Brysbaert et al., 2018) and 71% (104/146) of the ETC keywords have a fpm of less than five in the BNC, with 96% (48/50) of the top 50 keywords having a fpm of less than one and thus represent very unusual words. There was a small negative correlation between the two variables of keyness and fpm in the BNC, $r = -0.21$, $n = 146$, $p = 0.01$. This is shown in Figure 1, where the keywords cluster at the lower end of the frequency measure with the highest keyness scores also having low frequencies.

A small number of the keywords could be described as archaic, like the adverb "fro," the adjective "woolen," the verb "envelop," and the noun "tweed." Clothes and material feature, with "cravat," "nightdress," "tweed," "sock," "jersey," and "stocking" in the noun list and "woolen" as an adjective. Food appears too with "trifle," "tea," "soup," and "pizza" appearing as nouns as well as items to do with food such as "napkin," "crockery," "spoonful," and

TABLE 4 Keywords with scores (Ordered by keyword score).

Breaker (74.49)	Iceland (18.69)	Humiliate (12.58)	Vessel (9.77)
Nasally (69.44)	Wade (18.06)	Hasty (12.49)	Shore (9.74)
Thrill (42.59)	Stocky (17.82)	Horribly (12.46)	Boat (9.73)
Boulder (36.3)	Trifle (17.74)	Envelop (12.43)	Tow (9.63)
Uncontrollably (35.23)	Wistfully (17.7)	Scorch (11.96)	Consonant (9.61)
Clang (32.93)	Burnt (17.29)	Wardrobe (11.89)	Wrestle (9.61)
Napkin (32.63)	Nightdress (17.25)	Ooze (11.73)	Rotten (9.59)
Majestically (32.41)	Motionless (17.18)	Tweed (11.68)	Amusing (9.59)
Cravat (31.47)	Crumpled (16.72)	Cork (11.63)	Nelson (9.58)
Balloon (30.56)	Sickening (16.67)	Crest (11.59)	Ocean (9.5)
Slosh (29.58)	Horrid (16.57)	Fiercely (11.49)	Prison (9.43)
Dispirit (27.34)	Gust (15.8)	Prisoner (11.42)	Eyelid (9.34)
Incessantly (26.41)	Quicken (15.77)	Wail (11.36)	Sofa (9.29)
Swimmer (25.84)	Hoarse (15.75)	Soup (11.27)	Rejoin (9.28)
Lucy (25.37)	Bedside (15.69)	Float (11.27)	Pizza (9.28)
Handshake (25.14)	Pat (15.66)	Arrogant (11.17)	Moonlight (9.25)
Giddy (24.72)	Hearty (15.22)	Landing (11.0)	Dwindle (9.25)
Molten (23.74)	Agony (14.72)	Fro (10.9)	Packed (9.2)
Solitary (23.03)	Drip (14.49)	Sock (10.72)	Blanket (9.11)
Rut (22.93)	Fragrant (14.37)	Meaningless (10.72)	Tub (9.11)
Crockery (21.93)	Dangle (14.19)	Ghastly (10.69)	Dreadful (9.09)
Rekindle (21.72)	Neglected (14.0)	Kitten (10.67)	Woollen (9.06)
Divest (21.16)	Amiable (13.86)	Siren (10.61)	Stocking (9.05)
Idleness (20.99)	Rosy (13.86)	Frantically (10.61)	Dart (9.02)
Spoonful (20.91)	Upside (13.75)	Jersey (10.42)	Utmost (8.84)
Plank (20.75)	Housekeeping (13.72)	Hopelessly (10.34)	Nostril (8.8)
Nightfall (20.67)	Mantelpiece (13.72)	Shabby (10.3)	Expedition (8.8)
Homework (20.39)	Ledge (13.5)	Instant (10.21)	Defiance (8.8)
Endurance (20.28)	Sane (13.45)	Muffle (10.21)	Midday (8.77)
Gobble (19.91)	Hue (13.38)	Hideous (10.17)	Ice (8.66)
Shriek (19.68)	Thrilling (13.12)	Radiate (10.02)	Deck (8.64)
Ox (19.64)	Tea (13.09)	Mighty (9.99)	Creak (8.59)
Speck (19.21)	Throb (13.02)	Seep (9.98)	Foam (8.57)
Yank (19.21)	Weary (13.02)	Fury (9.87)	Vigour (8.5)
Savoury (19.14)	Miraculous (12.99)	Hillside (9.87)	Distressed (8.48)
Spiky (19.14)	Online (12.64)	Feeble (9.84)	
Smelt (18.74)	Spoon (12.59)	Ache (9.82)	

"spoon." Words to do with the sea, such as "breaker," "crest," "vessel," "shore," "ocean," and "deck," also feature in the nouns. These patterns suggest that extracts selected for the exams tend to describe people (and their clothes) and social gatherings (food), as well as travel or exciting events (represented by the sea). Verbs, adjectives, and adverbs add to this focus on description with the trend seeming to be toward extremes: the verbs indicating dramatic or negative events with "shriek," "yank," "wail," and "wrestle"; the adjectives either being pejorative, for example, "sickening," "horrid," "ghastly," "hideous," and "dreadful," or the more positive "fragrant," "rosy," "thrilling," and "miraculous"; and the adverbs cover a range from "frantically" and "fiercely" to "wistfully" and "majestically." Overall, these keywords seem to be centered on people or characters, be highly descriptive and tending toward either domestic affairs like meals or extreme or dramatic events.

Comparison with reference corpora

The corpus comparison tool in Sketch Engine takes the mean of the highest 500 keyword scores between two corpora as an overall score for similarity between them. The ETC was compared to a range of reference corpora (see Table 5). The corpus that had the lowest overall score when compared to the ETC, and therefore was the most like it, was Project Gutenberg English (2.16), a free digital library of mostly out of copyright literary texts). The corpus with the highest overall score and therefore the most different from the ETC was British National Corpus Spoken (3.69). The more modern corpora, English Web 2015 and English Broadsheets 1993–2013, also had high scores and so, along with the spoken corpus, were furthest from the ETC.

Genre sources for keywords

BNC genres

To explore the most likely source genres for the vocabulary from the exam, a search was run for each of

TABLE 5 Comparisons of Reference Corpora to ETC: the lower the score, the more alike the corpus is to the ETC.

Corpus compared to ETC	Score
Project Gutenberg English	2.16
Brown Family (written American and British English)	2.29
British National Corpus	2.41
English Broadsheets 1993–2013	2.65
English Web 2015	2.91
British National Corpus 2014, Spoken	3.69

the 146 keywords from the ETC in the BNC to identify the frequency counts in different genres of texts. Where the ETC keywords are most frequently found gives an indication of where the students might be most likely to encounter the words in their reading. Raw frequencies and normalized frequencies (fpm) are shown in Table 6 below. Fpm allows for comparison between different sizes of corpora. Frequencies are given for the two general registers of spoken and written texts and then for two sub-registers within the written text register: written imaginative and written informative.

Overall, the fpm scores, which were used for the comparison rather than the raw frequencies due to the different sizes of the subcorpora, were higher for written than spoken texts. There was also a much higher fpm score for written imaginative texts compared to written informative texts. This indicates that, as with the corpora comparison result where the ETC was most like a written corpus of older literary texts, the ETC keywords are most likely to be found in written imaginative texts.

Relative frequencies in the David Lee categories

For a more detailed breakdown of types of texts that contained the keywords, a relative frequency was calculated for the David Lee Categories (see Appendix B). Scores for the spoken genre categories in the classification were generally well below 100, meaning that the keywords were much less likely to be found in the spoken genres than in the BNC as a whole. There were higher scores in the written genre categories, with by far the highest averages in poetry (371.66) and prose fiction (227.17). Considering that only 19 (32%) of the exam texts were prose fiction, and none were poetry these are surprisingly high relative frequency scores. There were also some high relative frequencies amongst the nonfiction categories, for example, biography (134.22), the arts sections of broadsheet newspapers (145.21) and tabloid newspapers (125.12), mirroring the largest nonfiction David Lee Categories of the exam texts (biography and newspaper articles) and suggesting that the boundary between the general categories of imaginative and informational texts is not always distinct linguistically.

DISCUSSION

The background to this paper is the introduction in 2017 of a new specification of a high-stakes national exam in England, the English language GCSE. The reading part of the exam is now worth a higher percentage of the marks and also now has to include literary texts from the 19th, 20th, and 21st century, thereby introducing an increased focus on the types of vocabulary found in these texts.
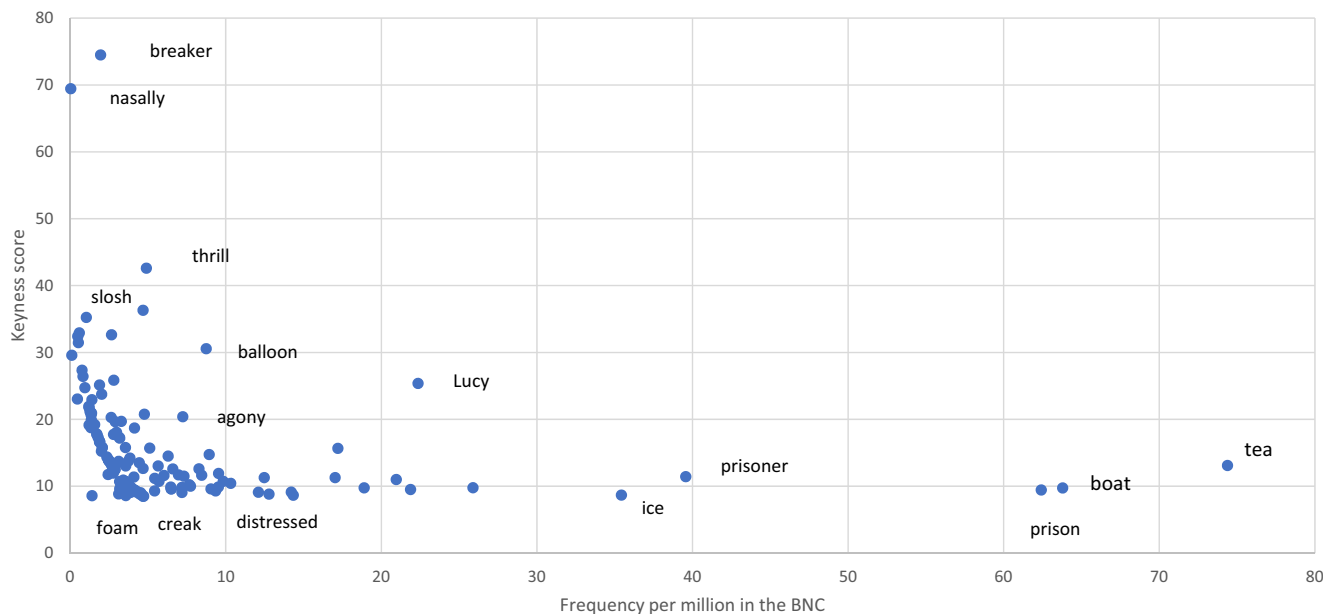
FIGURE 1 Scatterplot of keyword scores and frequency per million in the British national corpus.
Note. The color figure can be viewed in the online version of this article at http://ila.onlinelibrary.wiley.com.

TABLE 6 Frequencies and frequencies per million of the ETC keywords in the BNC.

Registers in BNC	Raw frequency of keywords	Fpm of keywords
Whole corpus	109841	977
Spoken texts	8133	689
Written texts	101771	1012
Written imaginative	36611	1851
Written informative	65159	806

The majority of the keywords from the ETC were low in frequency presenting a challenge to comprehension as it is more difficult for readers to experience them multiple times in diverse contexts to build high-quality representations. While the aim of the new English language GCSE was to create an improved qualification that was "more engaging and worthwhile to teach and study, as well as more resilient and respected" and "to prepare young people better for the next steps in their education or employment in years to come" (Ofqual, 2013, p.3), it is hard to see how the obscure nature of the vocabulary that typifies the exam texts is helping to fulfill these objectives.

Pressure to prepare students for these exams could lead teachers to feel they should include more older texts in the curriculum, to increase the exposure to archaic words, or design lessons that focus on the rote learning of low-frequency vocabulary, when there is little evidence that teaching word meanings directly improves comprehension (Wright & Cervetti, 2017). If

past exam papers are used for preparation lessons, with their focus noted above on characters, social and dramatic events, then students' curriculum reading experience could be narrowed. Such exam-focused activities, as Volante (2004) points out, are not always in the best interests of the students or necessarily effective activities for learning.

The genre in the BNC that contained the exam keywords most frequently was imaginative texts. This is in line with the literature that identifies fiction as the type of reading experience that best predicts reading ability (McGeown et al., 2015; Pfost et al., 2013; Torppa et al., 2019), supporting the hypothesis that fiction reading is superior because it provides the best source for vocabulary encounters due to its diversity. Fiction takes us to places we have never been, to times we could never travel back or forwards to, and puts us into action that we might never normally experience. This diversity of place, time, and action is described through a diversity of vocabulary that we might not otherwise encounter. But the closeness of the exam vocabulary to older literary fiction, through the match with Project Gutenberg, calls into question whether this is the type of fiction that adolescents are, or even should be, predominantly reading.

While fiction texts are obviously important to read, they are not the only type of genre that students will need in their future education and employment. Some nonfiction source genres featured in the findings, but these were limited to biography and newspaper articles. Familiarity with, for example, instructional texts, academic texts, and new media could also be considered essential or at least useful to students' future

literacy and employment. However, the reading of this range of genres is not the best preparation, according to the findings in this study, for the vocabulary in the current exam texts. This places schools and teachers in a difficult position when selecting classroom texts or activities. Gaining a good grade in this exam is essential for young people to access their next steps in education and employment. However, preparing students to obtain this grade is not necessarily going to prepare them for the literacy demands of their future. Given that curriculum time is not unlimited, teachers may have to choose between preparing students for the exam, by choosing reading that will exposure them to the types of vocabulary that is likely to be in the exam (using older, literary fiction), or choosing curriculum materials that they feel will prepare students for their further studies (e.g., academic texts), employment (e.g., instructional and commercial texts), and successful societal relations (e.g., new media and online texts). These external pressures on curriculum time could leave little, if any space, for reading which teachers might choose that is inspirational or enjoyable or thought provoking.

The English curriculum, before, during, and after the GCSE qualification, already receives criticism for its lack of diversity and representation (Elliott et al., 2021). Much good work has been done in schools to promote reading for pleasure that includes diverse voices, contemporary concerns, and spaces where students see themselves represented (Clark & Rumbold, 2006; Department for Education, 2012). If preparation for the English language exams is better served by reading traditional literary fiction, a canon of texts that is already covered by a separate English literature GCSE, then wider representation could be jeopardized and the dominance of the writings of dead White men could become further entrenched, as schools may feel that they should be recommending students read older literary texts instead of encouraging freedom of choice.

FURTHER RESEARCH AND LIMITATIONS

While the main finding of the current study is clear, there are some limitations which should be acknowledged. First, the focus of this study was solely on vocabulary as an important component of comprehension. Further research on other aspects of comprehension such as collocations, syntactic and morphological structures, and the need for high-level processing such as inferences and comprehension monitoring with regard to the exam texts would be advantageous, but were beyond the scope of this study.

A second limitation was the small size of the ETC. It would be useful to continue to grow the ETC, as more exam texts become publicly available, to monitor whether or not later extracts change the typical

vocabulary found, and to broaden the kinds of analysis that are possible.

This study makes some assumptions about students' reading experiences. Further research into the actual reading habits of students, for example, through a reading survey, would be valuable. The extent to which the exam text vocabulary is already found in curriculum materials and the choices that students are currently making about what to read for pleasure would be a valuable addition to the current literature.

CONCLUSION

Due to the long history of externally set exams, teachers in England have been used to teaching a curriculum, at Key Stage 4 especially, that is heavily influenced by the content in the exams. This study has found the vocabulary in the new English language GCSE to be typically low frequency and found predominantly in older, literary fiction. This calls into question whether the qualification is achieving its stated aim of preparing students for future study and work. It could also potentially skew what teachers feel they ought to choose for students to read in lessons and recommend that they read at home, as students' access to post-16 education and work opportunities depends on success in this exam. However, the promotion of this type of reading could limit students' reading experience to the literary canon of mostly dead White men, undermining efforts to increase diversity and representation in the curriculum and encourage freedom of choice in reading for pleasure. As the impact of this new qualification begins to become clear, teachers may feel that this is a step too far and move to question the influence of the test on the curriculum.

CONFLICT OF INTEREST STATEMENT
The authors declare that there are no conflicts of interest.

DATA AVAILABILITY STATEMENT
Data are not available until PhD of the first author is complete.

ORCID
Beverley Jennings https://orcid.org/0000-0002-3645-7411

REFERENCES
Anthony, L. (2013). A critical look at software tools in corpus linguistics. *Linguistic Research*, *30*(2), 141–161.
Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics: Investigating language structure and use*. Cambridge University Press.

Braze, D., Tabor, W., Shankweiler, D. P., & Mencl, W. E. (2007). Speaking up for vocabulary: Reading skill differences in young adults. *Journal of Learning Disabilities*, *40*(3), 226–243. https://doi.org/10.1177/00222194070400030401

Brown, C. P. (2015). Taking and teaching the test are not the same: A case study of first-year teachers experiences in high-stakes contexts. *Teachers and Teaching: Theory and Practice*, *21*(8), 1026–1044. https://doi.org/10.1080/13540602.2015.1005870

Brysbaert, M., Mandera, P., & Keuleers, E. (2018). The word frequency effect in word processing: An updated review. *Current Directions in Psychological Science*, *27*(1), 45–50. https://doi.org/10.1177/0963721417727521

Clark, C., & Rumbold, K. (2006). *Reading for pleasure: A research overview*.

Department for Education. (2012). *On Reading for Pleasure* (Issue May).

Department for Education, D. for. (2013). *English language: GCSE subject content and assessmentobjectives*. In *English language GCSE subject content and assessment objectives*. https://www.gov.uk/government/uploads/system/uploads/attachment_data/file/254497/GCSE_English_language.pdf

Elliott, V., Nelson-Addy, L., Chantiluke, R., & Courtney, M. (2021). *Diversity in literature in English schools*.

Evison, J. (2015). What are the basics of analysing a corpus? In *The Routledge Handbook of Corpus Linguistics*. Routledge. https://doi.org/10.4324/9780203856949.ch10

Hart, P. D., & Teeter, R. M. (2001). *A measured response: Americans speak on education reform*. Education Testing Service.

Henderson, L., Snowling, M., & Clarke, P. (2013). Accessing, integrating, and inhibiting word meaning in poor Comprehenders. *Scientific Studies of Reading*, *17*(3), 177–198. https://doi.org/10.1080/10888438.2011.652721

Isaacs, T. (2014). Curriculum and assessment reform gone wrong: The perfect storm of GCSE English. *Curriculum Journal*, *25*(1), 130–147. https://doi.org/10.1080/09585176.2013.876366

James, E., Currie, N. K., Tong, S. X., & Cain, K. (2020). The relations between morphological awareness and reading comprehension in beginner readers to young adolescents. *Journal of Research in Reading*, *44*(1), 110–130. https://doi.org/10.1111/1467-9817.12316

Jennings, J. L., & Bearak, J. M. (2014). "Teaching to the test" in the NCLB era: How test predictability affects our understanding of student performance. *Educational Researcher*, *43*(8), 381–389. https://doi.org/10.3102/0013189X14554449

Kilgarriff, A. (2005). Language is never, ever, ever, random. *Corpus Linguistics and Linguistic Theory*, *1*(2), 263–276. https://doi.org/10.1515/cllt.2005.1.2.263

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The sketch engine: Ten years on. *Lexicography*, *1*, 7–36. https://doi.org/10.1007/s40607-014-0009-9

Kintsch, W., Rawson, K. A., Nation, K., Perfetti, C. A., Landi, N., & Oakhill, J. (2005). In M. J. Snowling & C. Hulme (Eds.), *The science of Reading*. Blackwell Publishing Ltd.

Koester, A. (2015). Building small specialised corpora. In *The Routledge Handbook of Corpus Linguistics*. Routledge. https://doi.org/10.4324/9780203856949.ch6

Korochkina, M., Marelli, M., Brysbaert, M., & Rastle, K. (2023). The children and young People's books lexicon (CYP-LEX): A large-scale lexical database of books read by children and young people in the United Kingdom. https://doi.org/10.31234/osf.io/nha8t

Lee, D. Y. W. (2001). Genres, registers, text types, domains, and styles: Clarifying the concepts and navigating a path through the BNC jungle. *Language, Learning and Technology*, *5*(3), 37–72. https://doi.org/10.1163/9789004334236_021

Lervåg, A., Hulme, C., & Melby-Lervåg, M. (2018). Unpicking the developmental relationship between oral language skills and reading comprehension: It's simple, but complex. *Child Development*, *89*(5), 1821–1838. https://doi.org/10.1111/cdev.12861

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. *Reading and Writing*, *28*(4), 545–569. https://doi.org/10.1007/s11145-014-9537-9

Mitchell, K. J. (1997). What happens when school reform and accountability testing meet? *Theory into Practice*, *36*(4), 262–265. https://doi.org/10.1080/00405849709543777

Nation, K. (2017). Nurturing a lexical legacy: Reading experience is critical for the development of word reading skill. *Nnpj Science of Learning*, *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. *Journal of Memory and Language*, *39*(1), 85–101. https://doi.org/10.1006/jmla.1998.2564

Ofqual. (2013). *Reforms to GCSEs in England from 2015 (Issue November 2013)*.

Perfetti, C. A., & Hart, L. (2002). *The lexical quality hypothesis*. 189–213. https://doi.org/10.1075/swll.11.14per

Perfetti, C. A., Landi, N., & Oakhill, J. (2005). The acquisition of reading comprehension skill. In M. J. Snowling & C. Hulme (Eds.), *The science of reading* (pp. 227–247). Blackwell Publishing Ltd.

Pfost, M., Dörfler, T., & Artelt, C. (2013). Students' extracurricular reading behavior and the development of vocabulary and reading comprehension. *Learning and Individual Differences*, *26*, 89–102. https://doi.org/10.1016/j.lindif.2013.04.008

Sacks, P. (2000). Predictable losers in testing schemes. *School Adminstrator*, *57*(11), 6–9.

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. *Journal of Research in Reading*, *32*(4), 383–401. https://doi.org/10.1111/j.1467-9817.2009.01401.x

Torppa, M., Niemi, P., Vasalampi, K., Lerkkanen, M., Tolvanen, A., & Poikkeus, A. (2019). Leisure reading (but not any kind) and Reading comprehension support each other—A longitudinal study across grades 1 and 9. *Child Development*, *91*(3), 876–900. https://doi.org/10.1111/cdev.13241

Volante, L. (2004). *Teaching To the Test: What Every Educator and Policy-maker Should Know*. http://www.umanitoba.ca/publications/cjeap/articles/volante.html

Wright, T. S., & Cervetti, G. N. (2017). A systematic review of the research on vocabulary instruction that impacts text comprehension. *Reading Research Quarterly*, *52*(2), 203–226. https://doi.org/10.1002/rrq.163

AUTHOR BIOGRAPHIES

Beverley Jennings is a PhD candidate at the Institute of Education, University of Reading, Reading, UK; email: b.j.jennings@pgr.reading.ac.uk.

Daisy Powell is an Associate Professor at the Institute of Education, University of Reading, Reading, UK; email: d.a.powell@reading.ac.uk.

Sylvia Jaworska is a **PROFESSOR** of Language and Professional Communication at the Department of English Language and Applied Linguistics, University of Reading, Reading, UK; email: s.jaworska@reading.ac.uk.

Holly Joseph is a Professor of Language and Literacy Development at the Institute of Education, University of Reading, Reading, UK; email: h.joseph@reading.ac.uk.

APPENDIX A

A.1 | Keyword List

No.	Keyword score	Keyword	Part of speech	Fpm in BNC
1	74.49	Breaker	Noun	1.96
2	69.44	Nasally	Adverb	0.04
3	42.59	Thrill	Noun	4.9
4	36.3	Boulder	Noun	4.69
5	35.23	Uncontrollably	Adverb	1.04
6	32.93	Clang	Verb	0.59
7	32.63	Napkin	Noun	2.66
8	32.41	Majestically	Adverb	0.49
9	31.47	Cravat	Noun	0.53
10	30.56	Balloon	Noun	8.75
11	29.58	Slosh	Noun	0.11
12	27.34	Dispirit	Verb	0.77
13	26.41	Incessantly	Adverb	0.83
14	25.84	Swimmer	Noun	2.81
15	25.37	Lucy	Noun	22.36
16	25.14	Handshake	Noun	1.89
17	24.72	Giddy	Adjective	0.95
18	23.74	Molten	Adjective	2.03
19	23.03	Solitary	Adjective	0.47
20	22.93	Rut	Noun	1.4
21	21.93	Crockery	Noun	1.2
22	21.75	Rekindle	Verb	1.22
23	21.16	Divest	Verb	1.28
24	20.99	Idleness	Noun	1.34
25	20.91	Spoonful	Noun	1.38
26	20.75	Plank	Noun	4.77
27	20.67	Nightfall	Noun	1.35
28	20.39	Homework	Noun	7.24
29	20.28	Endurance	Noun	2.64
30	19.91	Gobble	Verb	1.36
31	19.68	Shriek	Verb	3.29
32	19.64	Ox	Noun	2.9

APPENDIX A (Continued)

No.	Keyword score	Keyword	Part of speech	Fpm in BNC
33	19.21	Speck	Noun	1.58
34	19.21	Yank	Verb	1.51
35	19.14	Spiky	Adjective	1.52
36	19.14	Savoury	Adjective	1.22
37	18.74	Smelt	Verb	1.34
38	18.69	Iceland	Noun	4.14
39	18.06	Wade	Verb	2.98
40	17.82	Stocky	Adjective	1.71
41	17.74	Trifle	Noun	2.79
42	17.7	Wistfully	Adverb	1.73
43	17.29	Burnt	Adjective	3.16
44	17.25	Nightdress	Noun	1.81
45	17.18	Motionless	Adjective	3.19
46	16.72	Crumpled	Adjective	1.89
47	16.67	Sickening	Adjective	1.9
48	16.57	Horrid	Adjective	1.9
49	15.8	Gust	Noun	2.07
50	15.77	Quicken	Verb	3.56
51	15.75	Hoarse	Adjective	2.07
52	15.69	Bedside	Noun	5.12
53	15.66	Pat	Noun	17.21
54	15.22	Hearty	Adjective	2.01
55	14.72	Agony	Noun	8.94
56	14.49	Drip	Verb	6.3
57	14.37	Fragrant	Adjective	2.36
58	14.19	Dangle	Verb	3.85
59	14	Neglected	Adjective	2.45
60	13.86	Rosy	Adjective	2.48
61	13.86	Amiable	Adjective	2.48
62	13.75	Upside	Adverb	3.74
63	13.72	Mantelpiece	Noun	2.52
64	13.72	Housekeeping	Noun	3.12
65	13.5	Ledge	Noun	4.44
66	13.45	Sane	Adjective	2.59
67	13.38	Hue	Noun	2.99
68	13.12	Thrilling	Adjective	2.68
69	13.09	Tea	Noun	74.39
70	13.02	Weary	Adjective	5.66
71	13.02	Throb	Verb	3.57
72	12.99	Miraculous	Adjective	2.71
73	12.64	Online	Adjective	4.69
74	12.59	Spoon	Noun	8.28
75	12.58	Humiliate	Verb	6.6
76	12.49	Hasty	Adjective	2.87
77	12.46	Horribly	Adverb	2.88

APPENDIX A (Continued)

No.	Keyword score	Keyword	Part of speech	Fpm in BNC
78	12.43	Envelop	Verb	2.88
79	11.96	Scorch	Verb	2.78
80	11.89	Wardrobe	Noun	9.54
81	11.73	Ooze	Verb	2.44
82	11.68	Tweed	Noun	6.97
83	11.63	Cork	Noun	8.46
84	11.59	Crest	Noun	6.03
85	11.49	Fiercely	Adverb	7.32
86	11.42	Prisoner	Noun	39.57
87	11.36	Wail	Verb	4.1
88	11.27	Soup	Noun	12.48
89	11.27	Float	Verb	17.03
90	11.17	Arrogant	Adjective	5.44
91	11.00	Landing	Noun	20.96
92	10.9	Fro	Adverb	3.42
93	10.72	Sock	Noun	9.84
94	10.72	Meaningless	Adjective	5.71
95	10.69	Ghastly	Adjective	3.18
96	10.67	Kitten	Noun	3.59
97	10.61	Frantically	Adverb	3.55
98	10.61	Siren	Noun	3.63
99	10.42	Jersey	Noun	10.32
100	10.34	Hopelessly	Adverb	3.67
101	10.3	Shabby	Adjective	3.69
102	10.21	Instant	Noun	7.66
103	10.21	Muffle	Verb	3.68
104	10.17	Hideous	Adjective	3.75
105	10.02	Radiate	Verb	3.82
106	9.99	Mighty	Adjective	7.74
107	9.98	Seep	Verb	3.4
108	9.87	Fury	Noun	9.53
109	9.87	Hillside	Noun	6.47
110	9.84	Feeble	Adjective	3.9
111	9.82	Ache	Verb	7.19
112	9.77	Vessel	Noun	25.89
113	9.74	Shore	Noun	18.9
114	9.73	Boat	Noun	63.79
115	9.63	Tow	Noun	3.19
116	9.61	Consonant	Noun	4.03
117	9.61	Wrestle	Verb	3.85
118	9.59	Rotten	Adjective	6.5
119	9.59	Amusing	Adjective	6.5
120	9.58	Nelson	Noun	9.05
121	9.5	Ocean	Noun	21.88
122	9.43	Prison	Noun	62.42
123	9.34	Eyelid	Noun	4.2

APPENDIX A (Continued)

No.	Keyword score	Keyword	Part of speech	Fpm in BNC
124	9.29	Sofa	Noun	9.35
125	9.28	Pizza	Noun	5.43
126	9.28	Rejoin	Verb	4.2
127	9.25	Moonlight	Noun	4.2
128	9.25	Dwindle	Verb	4.22
129	9.2	Packed	Adjective	4.25
130	9.11	Blanket	Noun	14.21
131	9.11	Tub	Noun	4.28
132	9.09	Dreadful	Adjective	12.1
133	9.06	Woollen	Adjective	3.88
134	9.05	Stocking	Noun	7.19
135	9.02	Dart	Verb	4.53
136	8.84	Utmost	Adjective	3.12
137	8.8	Nostril	Noun	4.49
138	8.8	Expedition	Noun	12.78
139	8.8	Defiance	Noun	4.58
140	8.77	Midday	Noun	4.56
141	8.66	Ice	Noun	35.43
142	8.64	Deck	Noun	14.34
143	8.59	Creak	Verb	3.58
144	8.57	Foam	Verb	1.41
145	8.5	Vigour	Noun	4.73
146	8.48	Distressed	Adjective	4.69

APPENDIX B

B.1 | Relative Frequencies of Keywords in the David Lee Categories

David Lee Category	Average relative frequency of keywords
S_parliament	15.22
S_pub_debate	6.79
S_sermon	37.43
S_speech_scripted	45.43
S_speech_unscripted	39.51
S_sportslive	35.48
S_tutorial	20.01
S_unclassified	32.17
W_ac_humanities_arts	51.33
W_ac_medicine	20.26
W_ac_nat_science	34.93
W_ac_polit_law_edu	17.65
W_ac_soc_science	46.94
W_ac_tech_engin	15.02

APPENDIX B (Continued)

David Lee Category	Average relative frequency of keywords
S_brdcast_discussn	46.78
S_brdcast_documentary	46.70
S_brdcast_news	36.51
S_classroom	61.94
S_consult	19.93
S_conv	52.61
S_courtroom	6.28
S_demonstratn	76.91
S_interview	22.95
S_interview_oral_history	43.29
S_lect_commerce	6.11
S_lect_humanities_arts	58.15
S_lect_nat_science	36.01
S_lect_polit_law_edu	21.83
S_lect_soc_science	25.52
S_meeting	21.87

APPENDIX B (Continued)

David Lee Category	Average relative frequency of keywords
W_admin	6.55
W_advert	69.08
W_biography	134.22
W_commerce	30.84
W_email	49.03
W_essay_school	108.73
W_essay_univ	36.30
W_fict_drama	113.11
W_fict_poetry	371.66
W_fict_prose	227.17
W_hansard	23.67
W_institut_doc	25.79
W_instructional	104.48
W_letters_personal	110.54
W_letters_prof	23.74
W_misc	92.96
W_news_script	62.25
W_newsp_brdsht_nat_arts	145.21
W_newsp_brdsht_nat_commerce	42.83
W_newsp_brdsht_nat_editorial	65.10
W_newsp_brdsht_nat_misc	113.61

APPENDIX B (Continued)

David Lee Category	Average relative frequency of keywords
W_newsp_brdsht_nat_report	55.52
W_newsp_brdsht_nat_science	66.51
W_newsp_brdsht_nat_social	79.17
W_newsp_brdsht_nat_sports	82.39
W_newsp_other_arts	92.48
W_newsp_other_commerce	40.82
W_newsp_other_report	68.36
W_newsp_other_reportage	56.76
W_newsp_other_science	53.74
W_newsp_other_social	92.87
W_newsp_other_sports	73.83
W_newsp_tabloid	125.12
W_non_ac_humanitites_arts	90.01
W_non_ac_medicine	76.70
W_non_ac_nat_science	74.06
W_non_ac_polit_law_edu	40.99
W_non_ac_soc_science	51.20
W_non-ac_tech_engin	29.42
W_pop_lore	110.12
W_religion	80.83