Using corpus methods to identify subject specific uses of polysemous words in English secondary school science materials

Alice Deignan & Robbie Love
University of Leeds

Abstract

Many education professionals in Britain believe that school students have difficulty accessing academic texts because of inadequate vocabulary knowledge. Previous research has suggested that some high frequency words used in non-specialised contexts have academic meanings that can cause problems for school students. We take corpus techniques used in the study of higher education texts and apply them to a corpus of texts designed for school students aged 11-14, attempting to identify such words automatically. We use the Spoken BNC2014 as a reference corpus. We identify a list of semi-technical words (Baker, 1988), many of which are polysemous, having everyday meanings and related school subject meanings that may not be familiar to students. We investigate how semi-technical vocabulary can be identified and distinguished from both specialised and general vocabulary. Some supplementary qualitative analysis was needed, using collocation and concordance analysis. While time-consuming, the potential benefits for students struggling with school language make this a worthwhile exercise.

1. Introduction

Within the UK, teachers' increasing awareness of the importance of vocabulary for first language speakers of English (as well as EAL students) is evidenced by the success of a recent book on the subject ('Closing the Vocabulary Gap', Quigley, 2018), written by a former secondary school teacher and practitioner-researcher. There is general agreement among teachers who we have spoken to with the central theses of the book: that vocabulary knowledge is central to success in mainstream schooling, that there is a gap between different children's knowledge of vocabulary, which seems linked to social and economic background, and that the topic has largely been neglected in teacher education. Harley (2018) reports on a survey of over 1,000 British primary and secondary teachers, in which over half reported that 'at least 40% of their pupils lacked the vocabulary to access their learning' (p. 2). She adds, 'on average, secondary school teachers who took part in the survey reported that 43% of Year 7 pupils have a limited vocabulary to the extent that it affects their learning' (p. 4).

The issue may not be simply one of accumulating new words. In a project led by one of the authors, approximately 200 secondary school students were interviewed about their

understandings of climate change, resulting in a corpus of around 88,000 words of transcribed discussion (Deignan, Semino & Paul, 2019). A number of instances of language use appeared to indicate that they struggled with the language needed to describe the scientific processes involved. For example, the word *release* is used to refer to the emission of carbon dioxide into the earth's atmosphere as a result of burning fossil fuels. Analysis of a corpus of their teaching materials showed that students had been presented with this semi-technical use, yet they often used it inaccurately when interviewed as the following utterances show:

(1) If we're recycling stuff like the landfills, I don't know, it *releases* something like you know less landfills and less pollution and stuff like that.
(2) It's getting thicker because erm, there's more pollutants and they're like carbon dioxide, so cos it's getting thicker, less oxygen, over less gases, like bounce back off. So they're getting less *released* so there's holes in there, which makes it more warmer.

These and other extracts suggest that the speakers do not have a good understanding of the meaning and use of *release* in this register. A concordance analysis of the British National Corpus[1] confirms that *release* is widely used in non-academic language, but predominantly with two other meanings: allow someone out from prison or other confinement, and put on sale a piece of music, film or book. These are related semantically to the scientific meaning, but possibly not obviously so to a school student who has not encountered it before. Further, perceiving the semantic relationship is not helpful in understanding the very specific meaning of the term, nor its collocational constraints. We observed the same phenomenon for a number of other semi-technical words. Clearly, teachers have very limited time to talk in depth with individual students and explore understandings, and we hypothesise that they may not realise the difficulties that some students have with vocabulary that does not immediately appear to be technical. In this article, we bring techniques from corpus linguistics to explore the issue, attempting to develop automatic methods for identifying semi-technical words, including those that have multiple meanings in different registers.

2. Literature review

2.1. School academic language and levels of vocabulary

[1] The British National Corpus, version 3 (BNC XML Edition). 2007. Distributed by Bodleian Libraries, University of Oxford, on behalf of the BNC Consortium. URL: http://www.natcorp.ox.ac.uk/

A series of studies has argued for the existence of a cognitive academic language proficiency (CALP), which is not universally acquired, as a separate facility from basic interpersonal communicative skills (BICS), which are shared by all competent language users (Cummins, 2008, 2014). While the distinction has been problematised (Leung, 2014), it is nonetheless widely agreed that school language differs from everyday language (e.g. Barwell, 2013; Olin-Scheller & Tengberg, 2017). Mastery of school language is a central factor in academic success from a very early point (Snow & Uccelli, 2009; DiCerbo, Anstrom, Baker & Rivera, 2014).

A number of studies have described aspects of this language (e.g. Schleppegrell, 2001; Snow & Uccelli, 2009). At the level of detail, registers differ by discipline (Schleppegrell, 2007), and even within this 'each discipline… may have hundreds of sub-areas, each with its own "specialised language"' (Wilkinson 2018: 169). While each register and sub-register is characterised by different lexical choices and uses, a number of writers believe that there is a core of academic vocabulary that is shared by many disciplines (e.g. Coxhead, 2000, 2018; Wilkinson 2018, 2019). This idea is well-known among teachers in English secondary schools. In teacher interviews comprising part of the project mentioned above (Deignan et al., 2019), some teachers talked about three "tiers" of vocabulary. The classification can be articulated in full as follows:

- Tier 1 comprises general, everyday words;
- Tier 2 comprises words used in academic discourse but shared across different disciplines, e.g. *susceptible, grossly inadequate* (Quigley, 2018);
- Tier 3 comprises words specific to a particular academic discipline, e.g. *isosceles*, (Woolridge, 2018); *cyclone, storm surge, tsunami* (Quigley, 2018);

We focus on Tier 2, described in the original classification by Beck at al. (2002) as follows:

> The second tier contains words that are of high utility for mature language users and are found across a variety of domains. Examples include *contradict, circumstances, precede, auspicious, fervent* and *retrospect*... Because of the large role Tier Two words play in a language user's repertoire, rich knowledge of words in the second tier can have a powerful impact on verbal functioning... (p. 11).

In writing about higher education, these ideas have been more fully explored. Gardner and Davies (2014: 315) write of high frequency, academic core and academic technical vocabulary,

and Farrell describes "semi-technical" vocabulary. Fraser (2012: 124) discusses "cryptotechnical" words (e.g. *transmitter, dependence, relaxation*); words such as these are said to have 'a technical meaning which may be obscure to a non-specialist' (Fraser, 2012: 124). The most widely used term is "sub-technical vocabulary". An early definition is 'context independent words which occur with high frequency across disciplines' such as '*function, inference, isolate, relation, basis, presuppose, simulate, approximately*' (Cowan, 1974: 391). Baker (1988: 92) described sub-technical lexis as 'items which express notions general to all or several specialised disciplines (examples being *factor* and *method*)' (1988: 92). Baker designed corpus procedures to identify sub-technical lexis, which we return to in the Methods section.

2.2. Polysemy and academic lexis

Specialist terms, such as *point* and *frequency* in physics (Jacobs, 1989), are sometimes polysemous with everyday uses. Wignell, Martin and Eggins (1989) make the same point in their discussion of school geography, about the terms *environment, wind,* and *rain* (p. 370). Similar observations have been made regarding a number of other disciplines (Schleppergrell, 2007; Nagy & Townsend, 2012; Greene & Coxhead, 2015; Chan, 2015; Wilkinson, 2018, 2019; Yun & Park, 2018). Chan (2015) identifies a category of action verbs including *expand, find, give, convert, simplify, evaluate*, which are sub-technical in scientific discourse but are polysemous with everyday meanings (p. 308). He notes two further categories of polysemous words with specialist meanings: prepositions such as *ahead* and *over*; and conjunctions such as *assume*, *given* (2015: 308). Baker points out different kinds of polysemy, including that noted above, where a word has a meaning in everyday language, such as *bug* and *solution*, but different meanings in specialised disciplinary language. A second is where a word has an everyday meaning but its meaning in specialised language is more restricted, for example, 'in botany, *effective* simply means "take effect", it has no evaluative meaning' (1988: 92). With regards to "cryptotechnical" vocabulary, Fraser (2012) claims that 'learners…may erroneously think they know' these words, which is 'a source of concern' (p. 135), since the familiar everyday meaning may serve to mask another technical meaning.

Chan (2015) notes the problem of specialised collocations in mathematical problems, giving the example of *possible actual walking speed*, which consists of words that are probably known in isolation but may present problems as a long noun group. Wilkinson (2018: 169) finds 'specialised vocabulary that is exclusively mathematical' (e.g. *binomial*); 'everyday vocabulary that is repurposed as specialised' (e.g. *table, product, length, factor*); and 'dense

noun phrases that express specialised meaning' (e.g. *area under a curve*). She analyses a test item from the 2015 PISA, showing how the relatively simple and non-disciplinary specific vocabulary of the item is combined to form complex and precise noun groups, such as *seasonal large-scale movement of birds* (2018: 172). Furthermore, Fraser (2012: 134) discusses how semi-technical, polysemous words such as *cell* may combine with other words to form highly specialised technical multiword units, like *cell blood count* and *mast cell*.

There is also polysemy across different school subjects. Quigley (2018) notes that *variable* has different meanings in computer science, science and mathematics. Nagy and Townsend (2012) give the examples of *force* and *function* which have different meanings in different disciplines. Deignan et al. (2019) found that secondary school students in Year 7 (aged 11- 12) encountered two different meanings of the word *concentration*; *concentration camp*, in English, and *concentration of a liquid*, in science. Both of these differ from the most frequent everyday meaning of *concentration*, heard paraphrased for school students as "thinking hard".

There is evidence that this causes difficulties for students. Jacobs (1989) found that most first-year physics undergraduates were unable to demonstrate accurate discipline-specific understanding of many polysemous terms. Boyes and Stanisstreet (1990) observed that 'the same vocabulary (*energy, work, force, power, conservation*) is used in both situations, and this can result in pupils transferring ideas from everyday life to a formal scientific context' (p. 513). Meyerson, Ford, Jones and Ward (1991) found that students will have 'alternative conceptual possibilities' for some words; that is, non-scientific meanings for science vocabulary (p. 427). They found that some 3rd and 5th graders explained *mass* as 'something at a church' and *organ* as 'like a piano' (p. 425). The studies cited above have produced numerous examples of potentially problematic uses that have been noticed by teachers and researchers, but none has claimed an objective and comprehensive method for classifying vocabulary in school texts.

2.3. Corpus research into Tier 2

Corpora have been widely used to study subject-specific, technical lexis at university level (e.g. for biology, Conrad, 1996; for medicine, Wang, Liang & Ge, 2008) and at secondary school level (e.g. Green & Lambert, 2018, 2019; Coxhead & Boutorwick, 2018). An early list of semi-technical lexis was produced by Cowan (1974), by hand, but using frequency lists in procedures that are now performed automatically. Baker (1988) used automatic techniques to compile a list of 65 sub-technical items from a corpus of medical texts. The most widely used list of words generic to academic writing is Coxhead's (2000) Academic Word List, compiled from a corpus of 3.5 million words from texts from a wide range of academic disciplines. This has

been followed by a spoken word list (Dang, Coxhead & Webb, 2017). Less corpus research has been conducted with school level texts (Coxhead, 2011). Monaghan (1999) shows how corpora can be applied to the study of individual vocabulary items in school mathematics texts. Coxhead, Stevens and Tinkle (2010) built a corpus to analyse school science textbooks in the New Zealand context, identifying low frequency words. Here, we explore how corpus work can help with describing the issues of polysemy and Tier 2 words in school scientific texts used in England, using techniques and insights gained from previous work at university level, adapted for this context. Our research questions are:

RQ1. Can sub-technical (Tier 2) words in Key Stage 3[2] (KS3) science materials be identified using corpus methods, and distinguished from Tier 1 and Tier 3 words?
RQ2. Can polysemous Tier 2 school lexis be identified using corpus methods?

3. Method

3.1. Data

We compared two corpora, one specialised and one which we used as a reference. The first, referred to here as 'The Science Corpus', is a collection of KS3 science education materials on the topic of climate change, originally built as one of a set of corpora of climate change texts (Deignan et al., 2019). It contains 22,416 tokens of KS3 textbook material and 192,422 tokens of website material accessed by KS3 students according to their self-reports, a total of 214,838 tokens. The reference corpus is the Spoken British National Corpus 2014 (Spoken BNC2014) (Love, Dembry, Hardie, Brezina & McEnery, 2017), which consists of 11,396,292 tokens of transcribed recent, everyday British English informal conversation. In this research context, it could be argued that comparing our written data with a spoken reference corpus introduces an irrelevant variable, in that we did not intend to capture linguistic variation across the modes of speech and writing (cf. Biber, 1988; Coxhead, 2017). However, we considered this the best currently available proxy for the language that English school students encounter outside the classroom, while recognising that it contains no written material. In the absence of a recent, general corpus of the kind of written English that Key Stage 3 students might read outside school, we decided this was the best reference corpus available. Many students read little

[2] Key Stage 3 is the first part of secondary school education in England and Wales, consisting of either 2 or 3 years, when students are aged between 11 and 13 or 14. All school subjects are compulsory in this stage.

outside school in any case, and others read largely fiction, which presents another genre challenge.

3.2. Unit of analysis

Coxhead (2000) used the word family as her unit of analysis in the compilation of the AWL. Word families sometimes comprise word forms (i.e. types; cf. Chung & Nation 2003, 2004) with rather different meanings, such as *react* (V; to respond), *reactionary* (Adj; strongly opposed to political or social change) and *reactivation* (N; to make something happen again) (Gardner & Davies, 2014). A recently proposed replacement for word families in the production of academic wordlists is the lexeme (Dang et al., 2017: 12). Unlike word families, members of a lexeme belong to the same part of speech and so each member is expected to be closely related in meaning (although Wang & Nation 2004 claim that polysemy/homography is also rare in word families). However, corpus work on the close relationship between form and meaning has suggested that sometimes different inflections of a headword are associated with different meanings (e.g. Deignan, 2005), meaning that analysis at the level of the lexeme could mask some of the polysemy in our data. We therefore decided to use the word form as our unit of analysis.

3.3. Corpus analysis procedures

We explored four quantitative and qualitative procedures, rejecting the first but combining the remainder.

Procedure 1. Keyness analysis

We used the keyness analysis (Gabrielatos, 2018) tool in the corpus analysis software *AntConc* (Anthony, 2018) to identify which word forms are much more frequent in the Science Corpus than in the Spoken BNC2014. This helps identify which word forms in the Science Corpus might be new to school students. We noted that, though informative as a step towards answering our first research question, the analysis does not distinguish Tier 2 from Tier 3. Further, because it cannot help to identify word forms that occur in both corpora but have a different meaning in each, it will not pick up Tier 2 or 3 words that happen to also have an everyday meaning in the reference corpus. As discussed above, these may be some of the most challenging words for students. We did not therefore use keyness further.

Procedure 2. Ratio of Frequency Percentage

Baker (1988) aimed to identify sub-technical lexis in her small corpus of medical English by eliminating both specialised and general lexis. She does not use the term "tier", but we have assumed from her description of her categories that they are approximately aligned to the three tiers. She reasoned that general lexis (Tier 1) will be widely distributed across both academic texts and her general reference corpus,[3] while specialised lexis (Tier 3) are narrowly distributed in a very few corpora. She developed the measure Ratio of Frequency Percentage (RFP): the relative frequency of a word in the specialised corpus (expressed as a percentage) is divided by its relative frequency in a general reference corpus. The RFP can thus be described as a simple measure of effect size which indicates the size of difference between relative frequencies. Generally, Tier 1 words tend to have a low RFP, because they are distributed fairly evenly across all texts, and are not markedly more frequent in the specialised corpus. At the other extreme, words that are much more frequent in the specialised corpus will have a high RFP. A middle band will be candidates for sub-technical vocabulary (Tier 2). Table 1 shows the RFP ranges that Baker (1988) suggests correspond to each type.

Table 1. Relationship between RFP and vocabulary type (adapted from Baker, 1988: 95-96)

RFP range	**Vocabulary type (tier)**
0-5	General (Tier 1)
6-299	Sub-technical (Tier 2)
300 +	Specialised (Tier 3)

The procedure can be improved by repeating the comparison with a number of specialised corpora from different disciplines, and identifying which lexis recur in the middle, sub-technical band. This is because sub-technical lexis, unlike specialised lexis, should be fairly evenly distributed across academic corpora from different disciplines. We have not done this in the current study, since our express focus is placed upon language that appears to function sub-technically in the science discipline alone.

Procedure 3. Qualitative tier adjustment

Baker (1988) treats the classification developed using RFP as provisional and writes that qualitative analysis should be used to check meaning. This can lead to manual adjustments to

[3] The University of Birmingham COBUILD corpus, at that time comprising 7.3 million tokens.

the RFP-assigned tiers. We used concordance data from *AntConc* (Anthony, 2018) to conduct this stage. Findings that would lead us to overrule the automatic classification would include evidence of different uses. For example, *rising* has an RFP of 372 which puts it in Tier 3, but the two most frequent nouns which follow *rising* in the Science Corpus, *sea* and *temperatures*, evidence different senses, *rising seas* being literal and *rising temperatures* being metaphorical. This means that there are at least two competing senses of *rising*, each of which will have a lower RFP. Further, the word is used with a flexibility not associated with highly technical terms. *Rising* was therefore reclassified as Tier 2. Baker (1988) notes that a word's regular use in multiword units can be an indicator of degree of specialism. On this basis, we reassigned *effects* to Tier 3. It is frequent in semi-fixed collocations with a technical meaning, such as *effects of climate change* and *effects of global warming*.

Procedure 4. Collocation analysis of Tier 2 words

The previous stage, qualitative tier adjustment, was a starting point in identifying polysemy in Tier 2 words. We also conducted collocation analysis, using *AntConc* (Anthony, 2018) for the Science Corpus and *CQPweb* (Hardie, 2012) for the Spoken BNC2014. Because different collocates can be associated with different meanings of a word (Hunston & Francis, 1998), this can be an automatic way in to identifying polysemy. We identified significant collocates using log likelihood (Brezina, 2018), with a window of five tokens either side of the search terms. Collocates had to occur at least five times in the corpus, and at least three times as a collocate of the word under examination, to be used in analysis.

4. Findings

4.1. Tier 2 words in the Science Corpus

In this section, we address RQ1 (*Can Tier 2 words in Key Stage 3 (KS3) science materials be identified using corpus methods, and distinguished from Tier 1 and Tier 3 words?*) Using Baker's (1988) RFP to categorise the 100 most frequent content word forms in the Science Corpus gave us a candidate list of fifty-four Tier 2 words. We then analysed the full concordances of each of these 100 word forms in order to judge whether the tier they had been assigned to automatically using RFP was consistent with our observations of their use. We also analysed the concordances of the same words from the Spoken BNC2014. Based on this analysis, we adjusted the tier of fourteen of the 100 words.

Twelve words were adjusted downwards, that is, their use in context was less specialised on analysis than the classification suggested. We reassigned *world, UK, future,*

likely, since, during, century from Tier 2 to Tier 1 because they appear to be used in their general senses in the Science Corpus, and their higher relative frequency was probably due to subject matter rather than other genre features. A second group were also adjusted downwards, from Tier 3 to Tier 2: *global, scientists, rising, arctic, glaciers*, because qualitative analysis suggested that their use in the Science Corpus is not exclusive to the subject but rather sub-technical. We concluded this from the range of collocations and meaning which each was used with. The example of *rising* is given above.

Two words were adjusted upwards; that is, their use in context was found to be more specialised than their RFP had suggested: *found* was reassigned from Tier 1 to Tier 2, while, as discussed above, *effects* was reassigned from Tier 2 to Tier 3. *Found* tends to collocate with specialised lexis and is typically used in a different structure in the Science Corpus than the Spoken BNC2014, as discussed in more detail below. This process resulted in a list of fifty-two Tier 2 words, forty-six of which had been automatically identified, the remainder having been added following the qualitative adjustment. The final list of Tier 2 words in the Science Corpus is as follows:

Table 2. Tier 2 words in the Science Corpus.

rank	word	per million	rank	word	per million
1	change	8,485.46	27	level	991.44
2	global	4,687.25	28	plants	982.14
3	earth	3,020.88	29	planet	972.83
4	ice	2,639.20	30	warmer	940.24
5	atmosphere	2,560.07	31	average	926.28
6	energy	2,499.56	32	extreme	828.53
7	water	2,383.19	33	areas	823.88
8	sea	2,099.26	34	surface	823.88
9	scientists	1,973.58	35	rising	814.57
10	heat	1,768.77	36	land	805.26
11	temperature	1,666.37	37	cause	805.26
12	weather	1,657.06	38	found	800.60
13	gas	1,643.10	39	reduce	800.60
14	countries	1,512.77	40	research	791.29
15	changes	1,466.22	41	species	791.29

16	report	1,461.57	42	action	772.68
17	study	1,335.89	43	arctic	763.37
18	rise	1,280.03	44	coal	758.71
19	levels	1,280.03	45	increasing	758.71
20	human	1,261.42	46	evidence	749.40
21	increase	1,196.25	47	animals	744.75
22	air	1,159.01	48	health	716.82
23	ocean	1,140.39	49	science	707.51
24	natural	1,107.81	50	increased	707.51
25	power	1,051.96	51	amount	702.86
26	effect	996.10	52	glaciers	702.86

4.2. Polysemy

In this section, we address RQ2 (*Can polysemous Tier 2 school lexis be identified using corpus methods?*) Having identified Tier 2 words in the Science Corpus using the adjusted RFP scores, we conducted qualitative analysis. We examined concordances and collocational information for each word in both the Science Corpus and the Spoken BNC2014 and compared results. Our findings showed polysemy for all of these words. We illustrate this with examples from the following words: *ice, energy, land, health*, and *found*, while acknowledging that our choice of these words is necessarily subjective: we chose these words because we consider that teachers would be unlikely to think that any of these words present problems to secondary school students.

Ice

The top collocates, using log likelihood, for each corpus are given in Table 3.

Table 3. Ten most significant collocates of *ice* in the Science Corpus and the Spoken BNC2014.

	Science Corpus			Spoken BNC2014		
rank	collocate	co-occurrence	log likelihood	collocate	co-occurrence	log likelihood
1	the	490	1,704.47	cream	563	6,640.97
2	sea	135	1,054.64	ice	154	1,347.54
3	of	243	852.89	creams	28	332.24
4	arctic	81	735.01	cube	24	263.05

5	and	206	710.15	lollies	15	200.56
6	melting	65	602.07	skating	20	198.16
7	sheets	48	576.17	an	92	186.79
8	in	148	460.34	chocolate	36	172.92
9	caps	35	406.48	lolly	12	132.10
10	melt	37	344.25	cubes	13	117.77

The collocates indicate that the more frequent meaning of *ice* in non-specialised language, as represented in the Spoken BNC2014, is as a food, or means of cooling drinks, the exception being *ice skating*. All of these collocations refer to entities made and controlled by humans. In contrast, the collocates in the Science Corpus suggest that ice is a large-scale natural phenomenon, largely outside human control. While few school students would struggle with this meaning, we would argue that it is nonetheless an academic and semi-specialised term which happens to share the same form as a general, everyday term.

Energy

The top ten collocates of *energy* in the two corpora are given in Table 4.

Table 4. Ten most significant collocates of *energy* in the Science Corpus and the Spoken BNC2014.

	Science Corpus			Spoken BNC2014		
rank	collocate	co-occurrence	log likelihood	collocate	co-occurrence	log likelihood
1	the	339	981.98	energy	20	178.12
2	renewable	70	746.64	much	33	87.50
3	and	203	714.71	efficiency	6	79.12
4	to	171	549.96	levels	7	47.89
5	of	155	430.31	of	85	46.44
6	clean	41	393.06	more	22	34.81
7	efficiency	32	358.37	drinks	5	27.30
8	use	48	340.57	gives	5	25.51
9	sun	42	329.32	less	7	25.08
10	sources	35	300.41	got	37	22.25

In the Science Corpus, the collocates *the*, *and*, and *to* occur in a number of positions relative to the node, not indicating any meaningful patterns. *Renewable* and *clean* refer to the supply of power from non-carbon, environmentally-friendly sources. *Efficiency* occurs in the phrase *energy efficiency* in 25 of 32 citations, 9 of which are part of a longer noun phrase such as *energy efficiency improvements*. All refer to measures at a national or global scale. *Of* occurs in phrases that quantify energy, such as:

> The Earth's atmosphere traps some of the *energy* from the sun.

Of also occurs in the expressions *sources of [renewable] energy* and *uses of [renewable/ clean] energy*. The common factor to all collocational patterns is that *energy* refers to power on a large and abstract scale, either the power from the sun that reaches the Earth, or the power generated by humans to support their needs and lifestyles.

In the Spoken BNC2014, the frequency of the top collocate, *energy* itself, results from the tendency for speakers to rephrase themselves, repeating words. The collocates *much*, *levels*, *of*, *more*, *gives*, *less* and *got* all occur in phrases referring to an individual person's feelings of physical and mental capacity, such as:

> I wish I had that *much energy*.
> I'm thirty four but I wouldn't have the *energy levels*…
> He's got loads *of energy*.
> I've got *more energy* in the morning…

The collocate *drinks* mainly occurs in the expression *energy drinks*, related to the above sense. The collocation with *efficiency*, in the noun phrase *energy efficiency*, evidences the meaning found in the Science Corpus, but the uses touch on the personal and concrete, referring to speakers' houses in all citations. With the exception of this last use, the collocates of *energy* point to distinct senses in the two corpora, and for school students, the Science Corpus sense may be relatively unfamiliar.

Health

The top ten collocates in both corpora are shown in Table 5.

Table 5. Ten most significant collocates of *health* in the Science Corpus and the Spoken BNC2014

	Science Corpus			Spoken BNC2014		
rank	collocate	co-occurrence	log likelihood	collocate	co-occurrence	log likelihood
1	and	77	310.13	safety	87	1,077.48
2	the	102	303.07	mental	71	668.21
3	of	61	203.81	health	34	263.01
4	human	23	176.03	care	37	226.23
5	change	31	137.94	national	28	209.05
6	climate	31	117.78	occupational	13	174.51
7	to	38	103.94	issues	21	170.30
8	public	12	102.10	problems	24	157.31
9	a	31	98.36	and	293	143.82
10	benefits	8	88.95	champion	10	108.08

In the Science Corpus, the collocates of *health* point towards a discourse of health that is public and global as opposed to private and individual. *Human* occurs in the noun phrase *human health*, referring to the health of the human population in general:

> Climate change was already having an impact on every continent, affecting *human health*, agriculture and wildlife

Specifically, the negative effects of climate change on human health are discussed, using nouns such as *impacts, problems, threat* and *risks*. Even when *health* is not explicitly modified by *human*, it is clear that it is human health specifically that is being discussed.

In the Spoken BNC2014, it is also clear that *health* refers almost exclusively to human health; however, there are differences in usage. Collocates include words that contribute to a range of fixed noun phrases such as *health and safety*, *health and fitness*, *health and social care*, *occupational health*, *mental health* and *National Health Service*. Many of these refer to 'systems' of human health which are encountered in day to day life. *Health and safety* and *occupational health* for example, refer to sets of (workplace) rules and procedures.

and at Christmas if you got a sixpence in your pudding which *health and safety* probably wouldn't allow these days

Citations suggest that the individual and local aspects of health are more salient than the broad, global health which is discussed in the Science Corpus, also indicated by the collocates *you* and *your*. The noun phrase *your health* (optionally modified by an adjective) occurs 39 times (3.4 per million), where it is clear that the speakers are addressing individuals, and framing *health* as being possessed by the individual.

see you've got to take your own *health* seriously

Students might find it difficult to appreciate the scale which is used to describe global health in the context of climate change, as they are likely to have encountered discussion of *health* on a much more personal and local level.

Found

The top ten collocates in both corpora, ranked by log likelihood, are shown in Table 6.

Table 6. Ten most significant collocates of *found* in the Science Corpus and the Spoken BNC2014

	Science Corpus			Spoken BNC2014		
rank	collocate	Co-occurrence	Log likelihood	collocate	Co-occurrence	Log likelihood
1	that	72	363.32	out	475	1121.56
2	the	114	338.93	found	144	693.63
3	of	66	216.69	I	1,687	539.10
4	in	59	212.44	‘ve	316	227.66
5	study	25	187.20	and	854	114.64
6	they	29	162.71	they	456	114.23
7	and	42	113.42	guilty	17	100.00
8	a	33	101.82	he	351	85.73
9	researchers	11	85.94	we	323	73.33
10	were	13	79.55	that	657	54.57

In the Science Corpus, the phrase *found that* is used to present the results of research with the implications that these have the status of established fact, and is approximately three times as frequent as the literal use *found in*. The literal use tends to refer to the location of phenomena, for example:

Limestone *found* in Yorkshire would have been formed on the bottom of a seabed.

In the Spoken BNC2014, *found* has both metaphorical and literal senses. The grammatical collocates *the*, *a* and *in* occur where *found* is modified by concrete noun phrases and preposition phrases. The most frequent collocate is *I*, in *I found*. The pronoun *it* is used both as an object, where *found* is literal and as an object complement, and where *found* is metaphorical:

I *found* it and it was in a bookshop.
I *found* it difficult.

Found out is purely metaphorical, having a sense which is close to the metaphorical academic meaning observed in the Science Corpus, to learn something new:

I *found out* I was allergic.
I *found out* that it'd been stolen.

Found out occurs only once in the Science Corpus. In the Spoken BNC2014, *out* signals the metaphorical meaning 'learning'; this is a highly frequent meaning of *found* in the Science Corpus, and it is not signalled. The Science Corpus meaning of *found*, without the use of *out* to indicate metaphoricity, may be unfamiliar to school students.

Land

The top ten collocates of *land* in the two corpora are shown in Table 7.

Table 7. Ten most significant collocates of *land* in the Science Corpus and the Spoken BNC2014

	Science Corpus	Spoken BNC2014

rank	collocate	co-occurrence	log likelihood	collocate	co-occurrence	log likelihood
1	the	123	381.38	land	56	432.42
2	and	90	369.44	rover	32	395.14
3	of	65	210.88	the	403	191.02
4	to	49	146.27	of	204	125.65
5	on	30	132.55	on	125	102.60
6	for	30	129.67	registry	9	96.74
7	is	33	112.27	buy	26	76.30
8	use	16	112.20	owns	10	74.33
9	more	21	92.02	rovers	6	70.49
10	water	16	86.45	bought	21	65.56

Once the concordances are analysed in more detail, some distinct meanings emerge. In the Science Corpus, *land* occurs exclusively as a non-count noun, mostly to refer to all of the land on the planet or a subset of it such as agricultural land, as opposed to the sea. The collocates *the* and *and* occur with this use. *And* tends to coordinate *land* with other parts of the planet and the atmosphere, including *oceans*, *water* and *air*. *On* occurs before *land* 18 times, in citations such as the following:

Glaciers are large sheets of snow and ice that are found on *land* all year long.

Land also collocates with some topic-specific words such as *use*, in *land use planning*; and *clearing* in *land clearing*, all indicating a specialist, geographical meaning. Other collocates suggest land mass, its contrast with water, and its use in agriculture.

In the Spoken BNC2014, as in the Science Corpus, *land* is a non-count noun. However, it is not construed as one unspecified entity but rather can be divided into constituent parts. Collocations with *the* often refer to owning, buying or selling pieces of land. Collocations with *of* largely occur in expressions such as *bit/ piece/ amount of land*. In the Spoken BNC2014, *on* often occurs immediately to the right of *land*, where it is a verb, in citations such as:

… so we leave on the thirtieth and we *land* on the morning of the thirty-first.

The collocates *Rover(s)* indicate a brand of car, and other collocates indicate a concern with ownership, trading and dividing land.

The patterns demonstrated for these five words were found extensively in the Tier 2 words listed above, that is, polysemy between a specialised sense and a more everyday, familiar sense is widespread. There is a long tradition of research into the percentage of unknown words that a reader can cope with and still make sense of a text (e.g. Schmitt, Jiang & Grabe, 2011), with general agreement that readers need to know between 95% and 98% of words; in other words, comprehension is disrupted if something between 1 word in 20 and 1 in 50 is not known. One word not used in its most familiar sense is unlikely to be a problem, but we contend that the fairly frequent use of words in different contexts and with different meanings may render a text less accessible to students.

5. Conclusion

We set out to see whether corpus methods could assist in the identification of Tier 2 words, a category that teachers have identified using intuition. We conclude that they are of assistance, though need to be supplemented by manual adjustment. Further research will include following Baker's procedure of using a range of specialised corpora representing the subjects studied at school, which will enable us to produce a Tier 2 wordlist for general purposes. This will require the compilation of a range of corpora, a task which is ongoing.[4]

We then looked at the use of corpus data to identify polysemy within Tier 2 words. We found collocation analysis to be productive, both including grammatical collocates, and for some words, where only lexical collocates are examined. For the words that we studied, this always needed to be supplemented with concordance analysis.

Some of the Tier 2 words and examples of polysemy that we identified are known to teachers; *energy* for example was mentioned to us by science teachers in interviews. However, most have not been mentioned in the various discussions of Tier 2 and sub-technical meaning in school language. We would claim therefore that corpus analysis can contribute to the description of school language through, as is often the case for corpus studies, making evident uses that are hidden in plain sight. The procedures we have described are time-consuming and it may be possible to automate them further. Even if this is not possible, we would argue that the identification of features of school language is worth a good deal of painstaking corpus

[4] Corpora are being compiled as part of an ESRC-funded project, reference ES/R006687/1.

study, given the importance of supporting young people, especially those from the least advantaged educational backgrounds, to access the school curriculum.

The corpus techniques described here have identified a list of words that are central to Key Stage 3 science, particularly the study of climate change, the topic of the texts in our corpus. They have shown words with specific scientific meanings that may not be familiar to school students from everyday language. Further research will generate more such lists, both for specific school subjects, and topics within these, and for the language of school generally. From the earliest days of corpus research it was noted that introspection cannot produce accurate descriptions of the data, though corpus findings once described have a quality of obviousness. This applies to the discourse of school; while teachers are both fluent in academic language and aware of the problems it poses to students, they are not able to consistently identify specific problems in advance. The work described here will give them central information to support their students in developing the academic language needed to access the curriculum.

Acknowledgements

The research presented in this paper was supported by two grants: AHRC 'Translating science for young people' (grant reference AH/M003809/1) and ESRC 'The linguistic challenges of the transition from primary to secondary school' (grant reference ES/R006687/1).

We are grateful to the teachers and students at our participating schools, who directed us to textbooks and websites that they use, and allowed us to record and transcribe lessons and interviews.

References

Anthony, L. 2018. AntConc (Version 3.5.7) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/software

Baker, M. 1988. 'Subtechnical vocabulary and the ESP teacher: An analysis of some rhetorical items in medical journal articles', *Reading in a Foreign Language* 4 (3), pp 91-105.

Barwell, R. 2013. 'The academic and the everyday in mathematicians' talk: The case of the hyper-bagel'. *Language and Education* 27 (3), pp 207-222.

Beck, I., McKeown, M., and Kucan, L. 2002. *Bringing words to life: Robust vocabulary instruction.* New York: Guilford.

Biber, D. 1988. *Variation across Speech and Writing*. Cambridge: Cambridge University Press.

Boyes, E., and Stanisstreet, M. 1990. 'Pupils' ideas concerning energy sources', *International Journal of Science Education* 12 (5), pp 513-529.

Brezina, V. 2018. *Statistics in Corpus Linguistics: A Practical Guide*. Cambridge University Press.

Chan, S. 2015. 'Linguistic challenges in the mathematical register for EFL learners: linguistic and multimodal strategies to help learners tackle mathematics word problems', *International Journal of Bilingual Education and Bilingualism* 18 (3), pp 306-318.

Chung, T., and Nation, I. S. P. 2003. 'Technical vocabulary in specialised texts', *Reading in a Foreign Language* 15 (2), pp 103–116.

Chung, T., and Nation, I. S. P. 2004. 'Identifying technical vocabulary', *System* 32 (2), pp 251–263.

Conrad, S. 1996. 'Investigating academic texts with corpus-based techniques: An example from biology', *Linguistics and Education* 8, pp 299-326.

Cowan, J. R. 1974. 'Lexical and syntactic research for the design of TESOL materials', *TESOL Quarterly* 8 (4), pp 389-399.

Coxhead, A. 2000. 'A new academic word list', *TESOL Quarterly* 34 (2), pp 213-238.

Coxhead, A. 2011. 'The Academic Word List 10 years on: Research and teaching implications', *TESOL Quarterly* 45 (2), pp 355-362.

Coxhead, A. 2017. 'Academic Vocabulary in Teacher Talk: Challenges and Opportunities for Pedagogy', *Oslo Studies in Language* 9 (3), pp 29–44.

Coxhead, A., and Boutorwick, T. J. 2018. 'Longitudinal vocabulary development in an EMI international school context: Learners and texts in EAL, maths and science', *TESOL Quarterly* 52 (3), pp 588–610.

Coxhead, A., Stevens, L., and Tinkle, J. 2010. 'Why might secondary science textbooks be difficult to read?', *New Zealand Studies in Applied Linguistics* 16 (2), pp 37-52.

Coxhead, A. 2018. *Vocabulary and English for Specific Purposes Research: Quantitative and Qualitative Perspectives.* Abingdon: Oxford.

Cummins, J. 1980. 'The cross-linguistic dimensions of language proficiency: Implications for bilingual education and the optimal age issue', *TESOL Quarterly* 14 (2), pp 175-187.

Cummins, J. 2008. 'BICS and CALP: Empirical and theoretical status of the distinction' in B. Street & N. H. Hornberger (eds.) *Encyclopedia of language and education*, pp. 71-84. New York: Springer.

Cummins, J. 2014. 'Beyond language: Academic communication and student success', *Linguistics and Education* 26, pp 145-154.

Dang, T. N. Y., Coxhead, A. and Webb, S. 2017. 'The Academic Spoken Word List', *Language Learning* 67 (4), pp 959-997.

Deignan, A., 2005. *Metaphor and Corpus Linguistics*. Amsterdam: John Benjamins

Deignan, A., Semino, E. & Paul, S-A, 2019. Metaphors of climate science in three genres: Research articles, educational texts, and secondary school student talk, *Applied Linguistics* 40 (2), pp 479-403.

DiCerbo, P., Anstrom, K., Baker, L. and Rivera, C. 2014. 'A review of the literature on teaching academic English to English language learners', *Review of Educational Research* 84 (3), pp 446-482.

Farrell, P. 1990. *Vocabulary in ESP: A lexical analysis of the English of electronics and a study of semi-technical vocabulary.* CLCS Occasional Paper 25. Trinity College Dublin.

Fraser, S. 2012. 'Factors affecting the learnability of technical vocabulary: findings from a specialized corpus', *Hiroshima Studies in Language and Language Education* 15, pp 123-142.

Gabrielatos, C. 2018. 'Keyness analysis: nature, metrics, techniques' in C. Taylor and A. Marchi (eds.) *Corpus approaches to discourse: A critical review*, p. 225- 258. London: Routledge.

Gardner, D. and Davies, M 2014. 'A new academic vocabulary list', *Applied Linguistics* 35 (3), pp 305-327.

Gee, J. 2008. 'What is academic language?' in A. Roseberry and B. Warren (eds.) *Teaching science to English language learners: Building on strengths*, pp 57-70. NSTA Press.

Green, C. and Lambert, J. 2018. 'Advancing disciplinary literacy through English for academic purposes: Discipline-specific wordlists, collocations and word families for eight secondary subjects', *Journal of English for Academic Purp*oses 35, pp 105-115.

Green, C. and Lambert, J. 2019. 'Position vectors, homologous chromosomes and gamma rays: Promoting disciplinary literacy through Secondary Phrase Lists', *English for Specific Purposes* 53, pp 1-12.

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

Hardie, A. 2012. 'CQPweb - combining power, flexibility and usability in a corpus analysis tool', *International Journal of Corpus Linguistics* 17 (3), pp 380–409.

Harley, J. 2018. 'Foreword', in *Why closing the word gap matters: Oxford language report*, p. 2. Oxford: Oxford University Press.

Hunston, S. and Francis, G. 1998. 'Verbs observed: a corpus-driven pedagogic grammar', *Applied Linguistics 19* (1), pp 45-72.

Jacobs, G. 1989. 'Word usage misconceptions among first year university physics students', *International Journal of Science Education* 11 (4), pp 395-399.

Leung, C. 2014. 'Researching language and communication in schooling', *Linguistics and Education* 26, pp 136-144.

Love, R., Dembry, C., Hardie, A., Brezina, V., & McEnery, T. 2017. The Spoken BNC2014: Designing and building a spoken corpus of everyday conversations. *International Journal of Corpus Linguistics* 22 (3), pp 319-344.

Maher, P. 2016. 'The use of semi-technical vocabulary to understand the epistemology of a disciplinary field', *Journal of English for Academic Purposes* 22, pp 92-108.

Masrai, A. and Milton, J. 2018. 'Measuring the contribution of academic and general vocabulary knowledge to learners' academic achievement', *Journal of English for Academic Purposes* 31, pp 44-57.

Meyerson, M. J., Ford, M. S., Jones, W. P., and Ward, M. A. 1991. 'Science vocabulary knowledge of third and fifth grade students', *Science Education* 75 (4), pp 419-428.

Monaghan, F. 1999. 'Judging a word by the company it keeps: The use of concordancing software to explore aspects of the mathematics register', *Language and Education* 13 (1), pp 59-70.

Nagy, W. and Townsend, D. 2012. 'Words as tools: Learning academic vocabulary as language acquisition', *Reading Research Quarterly* 41 (1), pp 91-108.

Nation, I. S. P. (2013). *Learning vocabulary in another language*. Cambridge: Cambridge University Press.

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing.* Amsterdam: John Benjamins Publishing.

Olin-Scheller, C., and Tengberg, M. 2017. 'Teaching and learning critical literacy at secondary school: The importance of metacognition', *Language and Education* 31 (5), pp 418-431.

Quigley, A. 2018. *Closing the vocabulary gap*. London: David Fulton.

The Confident Teacher. 2018. 'The "reading gap" between primary and secondary school'. Available at: https://www.theconfidentteacher.com/2018/03/the-reading-gap-between-primary-and-secondary-school/ (last accessed August 2019).

Schleppegrell, M. 2001. 'Linguistic features of the language of schooling', *Linguistics and Education* 12 (4), pp 431-459.

Schleppergrell, M. 2007. 'The linguistic challenges of mathematics teaching and learning: A research review', *Reading and Writing Quarterly* 23 (2), pp 139-159.

Schmitt, N., Jiang, X. and Grabe, W. 2011. 'The percentage of words known in a text and reading comprehension', *The Modern Language Journal* 95 (1), pp 26-43.

Snow, C. and Uccelli, P. 2009. 'The challenge of academic language' in D. Olson and N. Torrance (eds.) *The Cambridge Handbook of Literacy*, pp 112-133. Cambridge: Cambridge University Press.

Wang, K. and Nation, I.S.P. 2004. 'Word meaning in academic English: Homography in the Academic Word List', *Applied Linguistics* 25 (3), pp 291–314.

Wang, J., Liang, S. and Ge, G. 2008. 'Establishment of a medical word list', *English for Specific Purposes* 27 (4), pp 442-458.

West, M. 1953. *A General Service List of English words*. London: Longman, Green and Co.

Wignell, P., Martin, J. R. and Eggins, S. 1993. 'The discourse of geography: Ordering and explaining the experiential world', *Linguistics and Education* 1, pp 359-391.

Wilkinson, L. 2018. 'Teaching the language of mathematics: What the research tells us teachers need to know and do', *Journal of Mathematical Behaviour* 51, pp 167-174.

Wilkinson, L. 2019. 'Learning language and mathematics: A perspective from Linguistics and Education', *Linguistics and Education* 49, pp 86-95.

Wooldridge, J. 2018. 'Case study: Ideas for developing vocabulary' in *Why closing the word gap matters: Oxford language report*, p. 13. Oxford: Oxford University Press.

Yun, E. and Park, Y. 2018. 'Extraction of scientific semantic networks from science textbooks and comparison with science teachers' spoken language by text network analysis', *International Journal of Science Education* 40 (17), pp 2118- 2136.