Advancing disciplinary literacy through English for academic purposes

Discipline-specific wordlists, collocations and word families for eight secondary subjects

Green, Clarence; Lambert, James

Published in:
Journal of English for Academic Purposes

DOI:
10.1016/j.jeap.2018.07.004

Published: 01/09/2018

Document Version
Early version, also known as pre-print

Advancing disciplinary literacy through English for academic purposes: Discipline-specific wordlists, collocations and word families for eight secondary subjects

Article *in* Journal of English for Academic Purposes · July 2018

DOI: 10.1016/j.jeap.2018.07.004

CITATIONS
22

READS
726

2 authors:

Clarence Green
Federation University Australia
43 PUBLICATIONS **90** CITATIONS
SEE PROFILE

James Lambert
National Institute of Education (NIE), Singapore
46 PUBLICATIONS **114** CITATIONS
SEE PROFILE

Contents lists available at ScienceDirect

Journal of English for Academic Purposes

journal homepage: www.elsevier.com

Advancing disciplinary literacy through English for academic purposes: Discipline-specific wordlists, collocations and word families for eight secondary subjects

Clarence Green*, James Lambert

Department of English Language and Literature, National Institute of Education, Nanyang Technological University, NIE3-03-118, 1 Nanyang Walk, 637616, Singapore

ARTICLE INFO

Article history:
Received 26 September 2017
Received in revised form 17 April 2018
Accepted 2 July 2018
Available online xxx

Keywords:
Corpus linguistics
Disciplinary literacy
Academic vocabulary
Wordlists
Secondary education

ABSTRACT

The knowledge of academic vocabulary is crucial for educational success, and recently there has been a push amongst teachers and researchers to assist students at the pre-tertiary level to develop their disciplinary literacy and understanding of how academic English varies across disciplines. EAP research has developed advanced methods for producing corpus-informed vocabulary resources, but these have yet to be fully leveraged to promote disciplinary literacy within the secondary school context. For example, the focus of most previous wordlists has been on general academic vocabulary or the discipline-specific vocabulary needed in tertiary education. The current research contributes a series of discipline-specific wordlists for secondary school education, the *Secondary School Vocabulary Lists* (SVL), covering eight core subjects: Biology, Chemistry, Economics, English, Geology, History, Mathematics, and Physics. Further, the SVL goes beyond wordlists alone in developing accompanying word family and word association (i.e. collocation) lists for the disciplinary lexis. The SVL thus provides secondary education teachers with an unprecedented set of resources covering key vocabulary for the eight core disciplines informed by innovative EAP corpus methods.

1. Introduction: EAP affordances to advance disciplinary literacy in secondary education

With an increasing number of school students continuing to tertiary education, developing students' academic literacy during secondary school is essential. As Wingate (2011) notes, "in today's mass higher education system, many students are not fully prepared for the demands of academic study" (p. 66). Thus, research-based resources with the potential to assist teachers prepare secondary students for tertiary education are essential. Vocabulary is a vital component of educational success in both first and second language contexts (Webb & Nation, 2017), and in English for Academic Purposes (EAP) pedagogical materials to facilitate instructed vocabulary acquisition have been the focus of much research; in particular, advanced methods have developed for producing corpus-informed wordlists (Gardner & Davies, 2014; Lei & Liu, 2016). However, such wordlists have largely been developed in the context of tertiary level ESL education. The current advances in EAP for developing such educational resources have yet to be fully leveraged in other educational contexts, but there are recent trends in this direction as exemplified by the *Middle School Vocabulary Lists* (Greene & Coxhead, 2015). This paper aims to further advance the trend by reporting on a large corpus project that releases to the research and teaching community the *Secondary School Vocabulary Lists* (SVL), a series

* Corresponding author.
Email address: clarence.green@nie.edu.sg (C. Green)

https://doi.org/10.1016/j.jeap.2018.07.004
1475-1585/

of lemma-based discipline-specific academic vocabulary lists. These lists cover important vocabulary from eight core subjects that students need to master during secondary education: Biology, Chemistry, Economics, English, Geology, History, Mathematics, and Physics. Furthermore, the current study goes beyond previous wordlist development by contributing accompanying pedagogical lists containing word associations (i.e. collocations) and word families.

The SVL is designed for secondary school as a resource to assist teachers in the development of disciplinary literacy. Disciplinary literacy is a pedagogical approach and research area of increasing significance in secondary education internationally (Airey, Lauridsen, Räsänen, Salö, & Schwach, 2017). It emphasizes the connection between language and the disciplines, e.g. rather than English teachers focussing on a general academic vocabulary and subject area teachers on content, this approach calls for more explicit teaching of disciplinary language by all teachers. Shanahan and Shanahan (2017) suggest that students benefit the most from the disciplinary literacy approach during middle and secondary school as language becomes increasingly specialised. A resource that profiles the important language of secondary disciplines by adapting the methods of EAP research could therefore be very useful for such pedagogy. It would provide a resource from which, for example, teachers might select study words, knowing that the vocabulary chosen is important to the disciplines of secondary education (Ogle, Blachowicz, Fisher, & Lang, 2016). As Römer (2011, p. 209, p. 209) notes, even if a content area teacher does not focus on certain vocabulary until they reach a particular topic, language teachers can still introduce such discipline-specific vocabulary beforehand with the confidence that students will be better prepared for content classes. In the broader context, corpus-informed wordlists developed using the advanced methods of EAP within ESL research are increasingly relevant to schooling in predominantly English speaking countries. In the past, secondary classrooms have perhaps been seen as largely native speaker environments, but this is certainly no longer the case and secondary teachers are fully aware they need pedagogies and materials that support the different language backgrounds and proficiencies of contemporary classrooms. Further, the SVL can assist teachers in the growing number of international schools and EFL classes at the secondary level world-wide.

2. The pedagogical issues addressed by corpus-informed wordlists

The value of pedagogical material informed by objective methodological procedures developed in corpus linguistics is widely recognized (Brezina & Gablasova, 2017). The *General Service List* (GSL) of West (1953) has had no less than two recent updates (Brezina & Gablasova, 2013; Browne, Culligan, & Phillips, 2013), the *Academic Wordlist* (Coxhead, 2000) has recently been supplemented by the *Academic Spoken Wordlist* (Dang, Coxhead, & Webb, 2017), and there is an increasing trend toward the development of discipline-specific wordlists (Lei & Liu, 2016). The insight underlying such wordlists is that frequency, combined with metrics such as range and dispersion, profiles for teachers and students the relative usefulness of words (Brezina & Gablasova, 2017). Nation (2006, p. 79), for example, calculated that for any general text, comprehension requires about 98% understanding of its vocabulary, and that remarkably a teacher can provide this coverage by targeting the most frequent 8000–9000 word families of English. The general idea has a long tradition in educational research. Thorndike and Lorge (1944) compiled a corpus of primary and secondary school texts to produce a series of wordbooks to "enable a teacher to know not only the general importance of each word so far as frequency of occurrence measures that" (p. 1).

Vocabulary provides a foundation from which grammar, phonology, and morphology emerge, and in a subject area it provides access to conceptual knowledge (Coxhead, 2018). Vocabulary selection for pedagogical purposes is therefore crucial. Nation (2016) suggests a well-planned curriculum might progress from teaching general high frequency words, to general academic words, to discipline-specific and technical words, guided at each stage by corpus-informed resources that assist teachers in selecting vocabulary. Ogle et al. (2016), discussing how secondary teachers might intuitively select vocabulary, offer the example *pinnae* and recommend that although a word in secondary science textbooks that students would not understand, it should be treated as low-utility by teachers planning a vocabulary curriculum given its low frequency and topic-specificity. A corpus-informed wordlist could clearly facilitate such decisions. There are limitations to teacher intuition; as Schmitt (2010, p. 67) notes, given the idiolectal nature of language, teachers vary on which vocabulary they think important. This is also true of individual textbook writers (Harwood, 2014).

3. Corpus-informed wordlists: academic and discipline-specific vocabulary

To succeed in school, students need to develop an academic vocabulary (Shanahan & Shanahan, 2017). Perhaps the seminal wordlist for academic vocabulary is Coxhead's (2000) *Academic Wordlist* (AWL), designed to assist ESL speakers prepare for university. The AWL's impact on tertiary-level language education has been significant (Schmitt, Schmitt, & Mann, 2011) and there has been a growing awareness of corpus-based wordlists in the K-12 context. DiCerbo, Anstrom, Baker, and Rivera (2014) in a review of pedagogical approaches to academic vocabulary in US schools recommend the AWL as containing much vocabulary relevant to secondary school, even though the AWL is based on a corpus of university texts, largely sourced in New Zealand. Methodologically, Coxhead (2000) developed the AWL using the definition of word families in Bauer and Nation (1993), with a word family being a headword and its inflectional and derivational morphological forms, e.g. *react, reaction, reactor*. Coxhead (2000, p. 221) was particularly concerned with establishing corpus metrics that could profile what constituted general academic vocabulary and developed the following criteria: 1. the word was not in the GSL (West, 1953), considered to contain general vocabulary (i.e. the 2000 most frequent English words); 2. the word occurred in 15 of 28 subjects in the corpus,

thus ensuring the word was in general academic use; 3. the word occurred with a minimum frequency of 100 in the 3.5 million word corpus, since a useful academic word should occur frequently; 4. the word should have a minimum of 10 occurrences in arts, commerce, law, and science, since general academic words should occur across domains. These methods produced 570 word families, organized in the AWL by descending frequency. As a demonstration of its utility, Coxhead (2000) reported that the AWL covers around 10% of all academic texts. A supplement to the AWL, the Academic Spoken Wordlist (Dang et al., 2017), has recently been developed following similar methods but built from a spoken corpus.

Despite the impact of the AWL, the *Academic Vocabulary List* (AVL) has recently been produced by Gardner and Davies (2014), who critiqued the AWL on several grounds. First, the AWL lacks part-of-speech information so does not allow users to see which vocabulary can be used for multiple grammatical categories nor which grammatical category is more frequent. Second, Gardner and Davies (2014, p. 3) argue that the word families approach groups together words with significant meaning differences, e.g. *react* (a verb meaning respond) is in the same word family *reactor* (a noun usually associated with nuclear power). They cite other problems such as the size and sampling of the AWL corpus, and the use of the 50-year-old GSL. To address these issues, the AVL is lemma-based, part-of-speech tagged, and derived from approximately 120 million words of the academic subsection of the Corpus of Contemporary American English (Davies, 2010). The AVL contains 3015 lemmas, and in its development, Gardner and Davies (2014) introduced an advanced set of technical procedures into EAP, expanded further by Lei and Liu (2016) in the context of discipline–specific vocabulary. The current research draws on several of these procedures, so they are further explained in the methodology section of this paper.

3.1. The trend from general to discipline-specific wordlists

General academic wordlists such as the AWL and AVL have been challenged recently, with scholars questioning the extent and even existence of a general academic vocabulary. Hyland and Tse (2007, p. 114, p. 114) in a study of the AWL across eight disciplines found its 10% coverage held when the disciplines were considered a single academic corpus, yet there was so much variation by discipline that they conclude as many as ¼ of AWL words are not relevant to at least 1 discipline, thus calling into question whether these can be called general academic vocabulary. Durrant (2016) demonstrated a similar finding for the AVL. He explored coverage of the AVL in university papers and found good general coverage but fluctuation by discipline ranging from 10.25% of tokens in Classics to 21.54% in Economics. Durrant (2016) concludes "these findings support previous arguments that vocabulary is largely discipline-specific" (p. 60). Such findings have resulted in the current trend in EAP toward discipline-specific wordlists, most of which so far have been designed for tertiary education as reflected in the stated goals of the papers or their corpora, e.g. journal articles, dissertations, university textbooks etc. Within the past decade or so there have been corpus-derived lists developed for engineers (Todd, 2017), nurses (Yang, 2015), environmental studies (Liu & Han, 2015), medical professionals (Lei & Liu, 2016), and more besides.

Though limited work has been done thus far, a good example that recently expanded this research into new educational contexts are the *Middle School Vocabulary Lists* (Greene & Coxhead, 2015). These authors argue that lists based on university-level reading material cannot be directly imported into contexts such as middle school since the lexical profiles of subjects and their vocabulary demands are different. Furthermore, existing academic and discipline-specific wordlists have been produced in individual studies utilising a range of differing methodologies which raises a challenge for teachers, who would benefit from having wordlists for different school subjects consistent in their underlying principles. Greene and Coxhead's (2015) resources cover five content areas: English, Health, Mathematics, Science, and Social Science/History, derived from a corpus of 109 textbooks for grades 6–8 (approx. 11–14 years old). The methodology is similar to the AWL, taking a word families approach and excluding GSL words. The lists contain between 600 and 800 headwords for each of the five subjects, and the vocabulary covers from between 5.83% of texts for social studies/history up to 10.17% of texts for science.

3.2. Operationalising vocabulary constructs: general, academic, technical and discipline-specific

It is worth discussing further the constructs of general, academic and discipline-specific vocabulary. DiCerbo et al. (2014) in a review of vocabulary instruction in US secondary school suggest academic vocabulary "consists of words students must comprehend in order to access the concepts associated with a particular discipline and also use in order to display their acquisition of these concepts'' (p. 452). In EAP, one common classification has been between general vocabulary, academic vocabulary and technical vocabulary. General vocabulary is often defined as a common core of English words and operationalized as the most frequent words in a balanced and representative corpus of English, while academic vocabulary is defined as words that are relatively high frequency across many academic subjects, yet not part of this general vocabulary. A technical word is a word with a specific meaning or function in a discipline (Ha & Hyland, 2017). To isolate academic vocabulary from general vocabulary, the AWL and Middle School Vocabulary Lists decided to exclude any words in the GSL (West, 1953) on the basis that it was a list of general vocabulary. While reasonable, an issue that arises from this methodology is that a technical word, for example *set* in mathematics, is excluded from these lists of academic vocabulary because the same form occurs in the GSL. Further, Gardner and Davies (2014, p. 309) argued that the GSL used in the AWL was problematic as it is an outdated representation of general vocabulary, and indeed, when the authors of this current paper computed matches with the New GSL (Brezina & Gablasova, 2013), we found 372 New GSL words amongst the AWL word families. Another point made by Gardner and Davies (2014) was

that the AWL can be still be considered general vocabulary despite the GSL exclusion since the AWL is largely captured by the 4000 most frequent English words. Therefore, rather than using the GSL to isolate academic vocabulary for the AVL, they relied on methods such as the relative frequency of a word in an academic corpus compared to general corpus. However, despite their different methods for operationalising the construct of academic vocabulary and not using the original GSL, we also computed the matches between the AVL and New GSL and found 493 shared lemmas, i.e. approximately 16% of this academic wordlist overlaps with a wordlist representing general vocabulary.

Overlap between academic and general wordlists, however, does not mean that such words are not part of an academic vocabulary. After all, the research discussed in the previous section shows words in the AWL and AVL, even though they might constitute a general academic vocabulary, are also statistically associated with specific disciplines, and there is pedagogical value in such information (Durrant, 2016). The point is rather that general and academic vocabulary constructs are difficult to operationalize and probably overlap. Capturing the construct of technical vocabulary is also challenging, as such words exist along a technicality gradient (Ha & Hyland, 2017). Chung and Nation (2004) proposed that a technical word might be identified in a corpus if its frequency is 50 times greater than a general corpus; however, a wordlist that only contained words at this ratio would not capture the findings of the above research, i.e. that words in the AVL or AWL may be particularly associated with a discipline and constitute discipline-specific vocabulary. No wordlist has yet been developed dealing perfectly with such issues, and methodological decisions in different wordlists often balance gains and losses. The same is true for the SVL. The current approach, similar to Lei and Liu (2016), is that a discipline-specific wordlist should capture both the technical vocabulary of a discipline as well as vocabulary statistically prominent in the subject. Thus, the SVL does not exclude words if they were on general lists such as the GSL, AWL, or AVL.

4. Methodology

4.1. Corpus construction

To connect the SVL to current secondary school needs, the corpus was based on secondary school textbooks; the majority (82%) published within the past five years. Texts were taken from the textbook lists of the Ministry of Education (Singapore), and United Kingdom and Singapore A-level/O-level syllabi documents (the UK and Singapore exams being aligned). All textbooks were published in these two countries. When recommended texts were exhausted, texts specifically marketed at O-Level or A-level were sampled to reach the desired word count. Drawing on texts recommended in curricula and controlling for two countries with benchmarked curricula improves the external representativeness of the corpus (Miller & Biber, 2015). In total, the corpus consisted of 206 textbooks. Representative textbooks are given in Table 1.

Texts were scanned using *Omnipage 18* and converted via optical character recognition to plain text. Indexes, references, front matter, and contents pages were removed. A few issues encountered are worth reporting. The nature of secondary school textbooks, which have less running linear text than university material (e.g. visual boxes highlighting key ideas, activities, flow charts, etc.), introduced slightly more noise into the corpus than the researchers have experienced with other corpus-building projects. An example is given in (1):

(1) Derivatives of x", for any rational n, sin x, cos x, tan x, ex and In x, together with constant multiples, sums and differences.

This extract from an OCR of a Mathematics' textbook demonstrates missing punctuation and scanning errors. Other examples of noise include OCR errors such as 'I' sometimes converted to '1', occasional word spaces not recognized, etc. Despite extensive data cleaning and preprocessing, it was not possible to completely eliminate all noise in such a large corpus. Although this restricts potentially interesting measures such as mean sentence length, paragraph length etc. for the disciplines, the effect on vocabulary items (e.g. *derivatives, rational, multiples*) was minimal.

The corpus was approximately 16.25 million words. Corpus size estimates by discipline according to *Wordsmith* (Scott, 2016) are given in Table 2.

Table 1
Representative textbooks included in the corpus.

Author(s)	Date	Title	Publisher
Sang & Jones	2012	*Cambridge O Level Physics*	Cambridge University Press
Jones et al.	2013	*Cambridge International AS and A Biology Coursebook 3rd Edition*	Cambridge University Press
Sharma & Menon	2016	*All About English Comprehension*	Hodder Education
Christian	2014	*Essential Geography Skills*	Oxford University Press

Table 2
Word counts by discipline.

Biology	Chemistry	Physics	Geography
2,011,083	1,908,228	1,911,574	2,221,239
English	Mathematics	Economics	History
2,110,857	1,404,280	2,297,055	2,389,034
Total:	16,253,350		

4.2. Development of the SVL

Following Gardner and Davies (2014) and Lei and Liu (2016), the SVL is based on lemmas rather than word families. A lemma had to meet 6 criteria within each discipline to be included in the SVL. Table 3 presents each criterion with its justification.

To include part-of-speech (POS) information, the corpus was tagged using the CLAWS tagger. The CLAWS error rate is reported to be approximately 3% on standard text, but given the noise discussed above, we computed our own CLAWS error rate based on the procedures outlined for the BNC (Garside, Leech, & McEnery, 1997). 200 random samples of every tag were extracted and the resulting 26,489 tags checked by hand for accuracy. The estimated error rate based on this procedure is 6.2%. Every lemma in the final SVL was checked to remove POS errors. Since the objective approach to wordlist development was methodologically adopted, the advantages (and limitations) of which have recently been outlined in Brezina and Gablasova (2017), Dang et al. (2017) and Gardner and Davies (2016, p. 62), SVL words were not vetted on the basis of subjective judgements about teaching worth (cf. Martinez & Schmitt, 2012); for example, content area teachers were not asked whether a word was worth keeping. The SVL research team only vetted from the final lists problematic tags, scanning noise, proper nouns/adjectives (Nation, 2016), and agreed that items such as DNA should be retained but not chemical formulas such as NaCL. The corpus was lemmatized using *Wordsmith* (Scott, 2016), using a combination of several lemma conversion lists to extend coverage to over 70,000 lemmas, namely the Someya list (1998), a list from data mining software Prosuite (Provalis 2016), and a lemma list by Měchura (2016).

The fifty most frequent lemmas for each discipline are given in rank order in Table 4, with the complete SVL provided in the supplementary materials of this journal.

Table 4 reflects that the methods produced what one would intuitively consider representative vocabulary of the eight disciplines. The procedures produced 880 lemmas for Biology, 519 for Chemistry, 477 for Economics, 686 for English, 702 for Geography, 717 for History, 546 for Physics, and 253 for Mathematics. Together, there are 4781 lemmas that can support teachers' vocabulary selection. The lists capture technical vocabulary such as *photosynthesis* and *enzyme* in Biology, *molecule* and *hydrogen* in Chemistry, concepts such as *government* and *revolution* in History, *supply* and *demand* in Economics, as well as discipline-specific words such as *reaction,* which although in the AWL is here profiled as important to Chemistry.

As noted by Gardner and Davies (2016) in regard to the AVL (containing words such as *low, both, male*), not every objectively included word need be considered of equal complexity. For example, in Mathematics, the SVL includes *find* and *set*, which are not as orthographically complex as other entries such as *quadrilateral* or *trigonometric,* and furthermore commonly occur outside mathematics. However, it must be noted that even a word like *set*, which is in the GSL, has in mathematics a technical meaning, evidenced by collocations such as *empty*, *universal*, and *ordered*. Similarly, the research team discussed the mathematics list with a secondary textbook writer who informed us that secondary students need to learn that *find* within a mathemat-

Table 3
SVL: Lemma inclusion criteria.

1. ***Minimum Frequency***: > 28.57 occurrences per million words in the discipline.
   This number derives from Lei and Liu (2016) and Coxhead (2000) and excluded low frequency lemmas that a student would not commonly encounter.
2. ***Range***: > 50% of texts in a discipline.
   Following Gardner and Davies (2014), this ensured that over and above frequency, a lemma occurred in the majority of a discipline's texts.
3. ***Dispersion***: > 0.5 dispersion in the discipline.
   The Oakes Dispersion test divided the disciplinary corpora into 8 equal parts and computed homogeneity of occurrence for every lemma. Lei and Liu (2016) suggest below 0.5 dispersion indicates occurrences are clustered in a few texts.
4. ***Range Ratio***: > 20% of its minimum frequency in more than 50% of texts.
   As pointed out by Lei and Liu (2016) and Gardner and Davies (2014), this overcomes a potential issue that a word might have >50% range but still be infrequent in many of these texts.
5. ***Frequency Ratio/Keyness***: > 3 times more frequent in the discipline than the rest of the corpus.
   Gardner and Davies (2014) excluded AVL words that were 3 times higher in academic texts than a general corpus in order to exclude overly discipline-specific lemmas. We therefore used this ratio to target them, computing ratio against all other disciplines.
6. ***Major Part of Speech***: = noun, verb, adjective, adverb.
   Other word classes were excluded even if they met the statistical benchmarks as they were deemed to have low teachability.

Table 4
SVL: 50 most frequent lemmas for eight core subjects with part-of-speech.

rank	Biology	freq. p/m	Chemistry	freq. p/m	Economics	freq. p/m	Geography	freq. p/m
	lemma		lemma		lemma		lemma	
1	cell.n	11982	reaction.n	9319	price.n	8204	area.n	4746
2	blood.n	6125	acid.n	7469	cost.n	4906	country.n	3613
3	plant.n	3809	ion.n	5692	demand.n	4273	population.n	2465
4	enzyme.n	2971	atom.n	5533	rate.n	4147	food.n	2302
5	molecule.n	2844	form.v	5268	firm.n	4004	river.n	1985
6	gene.n	2762	solution.n	5107	income.n	3688	land.n	1900
7	dna.n	2700	electron.n	4942	good.n	3546	development.n	1658
8	protein.n	2515	gas.n	4479	market.n	3514	city.n	1645
9	body.n	2410	bond.n	4140	government.n	3166	place.n	1627
10	glucose.n	2332	metal.n	4051	supply.n	3014	rock.n	1390
11	concentration.n	2279	carbon.n	3922	business.n	2934	soil.n	1380
12	organism.n	2070	molecule.n	3884	tax.n	2875	global.adj	1336
13	contain.v	2062	compound.n	3696	bank.n	2861	sea.n	1336
14	chromosome.n	2059	hydrogen.n	3690	profit.n	2846	impact.n	1229
15	oxygen.n	2057	element.n	3492	curve.n	2781	tourist.n	1220
16	membrane.n	2039	group.n	3418	economy.n	2733	local.adj	1183
17	structure.n	2020	sodium.n	2889	money.n	2728	map.n	1173
18	carbon.n	1984	mass.n	2813	output.n	2636	million.n	1172
19	muscle.n	1629	react.v	2812	product.n	2627	wind.n	1166
20	allele.n	1577	oxide.n	2614	total.adj	2364	climate.n	1160
21	leaf.n	1523	formula.n	2487	worker.n	2363	activity.n	1159
22	dioxide.n	1516	chloride.n	2394	service.n	2292	tourism.n	1139
23	tissue.n	1448	oxygen.n	2184	trade.n	2247	urban.adj	1137
24	substance.n	1433	mole.n	2183	quantity.n	2215	environment.n	1135
25	amino.n	1364	table.n	2090	account.n	2156	earthquake.n	1093
26	wall.n	1342	structure.n	2060	consumer.n	2104	natural.adj	1018
27	carry.v	1312	contain.v	2031	pay.v	2028	erosion.n	942
28	photosynthesis.n	1278	substance.n	1979	economic.adj	1978	south.n	937
29	animal.n	1193	copper.n	1854	interest.n	1956	crop.n	920
30	tube.n	1183	volume.n	1818	increase.n	1929	live.v	908
31	species.n	1146	concentration.n	1795	production.n	1876	farm.n	887
32	potential.n	1143	dioxide.n	1793	capital.n	1838	plate.n	885
33	bacterium.n	1131	chemical.adj	1698	balance.n	1794	weather.n	880
34	genetic.adj	1105	iron.n	1698	sale.n	1734	beach.n	849
35	respiration.n	1046	oxidation.n	1668	revenue.n	1731	rainfall.n	846
36	heart.n	1043	salt.n	1571	wage.n	1697	study.n	842
37	light.n	1037	particle.n	1565	resource.n	1609	international.adj	817
38	nucleus.n	1030	hydroxide.n	1529	buy.v	1597	location.n	803
39	human.adj	995	aqueous.adj	1499	inflation.n	1585	tropical.adj	794
40	neurone.n	975	mixture.n	1440	sell.v	1519	over.adv	794
41	root.n	960	molecular.adj	1407	policy.n	1515	health.n	787
42	disease.n	914	nitrogen.n	1381	company.n	1504	coastal.adj	786
43	chain.n	912	property.n	1380	gdp.n	1369	north.n	784
44	hormone.n	908	ionic.adj	1368	marginal.adj	1343	building.n	779
45	active.adj	905	magnesium.n	1267	asset.n	1326	grow.v	757

Table 4 (*Continued*)

rank	Biology	freq. p/m	Chemistry	freq. p/m	Economics	freq. p/m	Geography	freq. p/m
	lemma		lemma		lemma		lemma	
46	transport.n	902	ammonia.n	1243	unemployment.n	1300	disease.n	755
47	sugar.n	880	chlorine.n	1240	real.adj	1289	flow.n	755
48	vessel.n	878	proton.n	1229	industry.n	1257	desert.n	743
49	release.v	857	strong.adj	1200	growth.n	1256	develop.v	741
50	gamete.n	838	calcium.n	1174	investment.n	1225	ice.n	736
	Mathematics		**History**		**English**		**Physics**	
1	find.v	13421	war.n	5894	word.n	3968	energy.n	7273
2	value.n	7406	source.n	3790	language.n	2408	force.n	6680
3	equation.n	7327	state.n	3430	say.v	2117	wave.n	4168
4	point.n	6818	world.n	2629	text.n	2010	object.n	3753
5	line.n	5323	government.n	2556	write.v	1977	field.n	3686
6	graph.n	4672	power.n	2360	writer.n	1915	current.n	3646
7	give.v	4307	party.n	1659	think.v	1867	speed.n	3516
8	solution.n	3956	military.adj	1624	read.adv	1665	direction.n	2964
9	example.n	3808	policy.n	1578	go.v	1643	mass.n	2939
10	area.n	3795	communist.adj	1539	life.n	1436	temperature.n	2866
11	curve.n	3587	political.adj	1479	child.n	1397	magnetic.adj	2558
12	term.n	3497	nation.n	1446	man.n	1354	resistance.n	2557
13	function.n	2999	leader.n	1251	paragraph.n	1299	move.v	2543
14	diagram.n	2830	treaty.n	1158	idea.n	1264	charge.n	2540
15	angle.n	2827	support.v	1053	look.v	1245	distance.n	2468
16	axis.n	2553	army.n	1018	student.n	1244	circuit.n	2367
17	let.v	2448	president.n	1015	come.v	1234	wire.n	2365
18	hence.adv	2267	west.n	971	sentence.n	1210	diagram.n	2314
19	circle.n	2121	control.n	937	essay.n	1074	particle.n	2305
20	solve.v	2093	east.n	930	just.adv	1017	air.n	2091
21	gradient.n	2089	cold.adj	927	feel.v	977	electric.adj	2028
22	length.n	2051	nuclear.adj	923	thing.n	959	calculate.v	2026
23	form.n	1963	foreign.adj	923	school.n	953	unit.n	1944
24	coordinate.n	1899	missile.n	875	passage.n	929	ray.n	1939
25	side.n	1895	crisis.n	874	story.n	870	velocity.n	1923
26	probability.n	1784	agree.v	823	phrase.n	834	pressure.n	1920
27	vector.n	1777	peace.n	801	writing.n	829	heat.n	1865
28	draw.v	1679	begin.v	792	reader.n	817	light.n	1792
29	triangle.n	1658	support.n	781	here.adv	793	constant.adj	1778
30	root.n	1654	member.n	777	learn.v	765	potential.adj	1721
31	sum.n	1557	western.adj	686	english.n	757	frequency.n	1670
32	step.n	1456	plan.n	682	tell.v	754	acceleration.n	1664
33	distance.n	1401	ally.n	666	society.n	745	motion.n	1648
34	method.n	1319	eastern.adj	659	ask.v	745	surface.n	1647
35	positive.adj	1290	communism.n	632	view.n	737	coil.n	1616
36	range.n	1278	election.n	624	understand.v	712	travel.v	1508
37	calculator.n	1259	independence.n	617	old.adj	703	length.n	1455
38	set.n	1251	agreement.n	616	character.n	665	magnet.n	1435
39	represent.v	1223	revolution.n	574	woman.n	664	difference.n	1422

Table 4 (*Continued*)

rank	Biology	freq. p/m	Chemistry	freq. p/m	Economics	freq. p/m	Geography	freq. p/m
	lemma		lemma		lemma		lemma	
40	tangent.n	1215	security.n	567	speaker.n	654	earth.n	1418
41	constant.adj	1178	weapon.n	567	sense.n	642	angle.n	1409
42	height.n	1176	nazi.n	563	meaning.n	635	resistor.n	1396
43	centre.n	1166	establish.v	543	speak.v	634	sound.n	1396
44	obtain.v	1136	conflict.n	538	today.adv	591	act.v	1376
45	normal.adj	1130	invasion.n	531	purpose.n	582	weight.n	1302
46	radius.n	1100	history.n	521	seem.v	575	wavelength.n	1253
47	real.adj	1080	communist.n	519	argument.n	565	kinetic.adj	1244
48	expression.n	1078	believe.v	514	topic.n	549	car.n	1235
49	equal.adj	1030	fight.v	501	young.adj	548	electrical.adj	1190
50	substitute.v	1019	minister.n	499	speech.n	512	work.n	1177

ics word problem means they need to show their working out. Thus, these are important discipline-specific words that would have been missed if the GSL was used as an exclusion list (indeed they are not in the Middle School Vocabulary Lists likely for this reason). It is however a trade-off, particularly with the SVL for English, which contains some GSL words unlike *find* and *set* that are closer to general vocabulary, such as *say, go, read*. This reflects that in English, students learn a range of text types, such as procedures, editorials, poetry, and not just academic essays (Ogle et al., 2016). Despite the reasonable inclusion of such words from an objective standpoint (indeed *read,* while simple in lemma form has more pedagogically valuable word family members such as *misreading*, *readability* etc., see section 4.4), we are not arguing that a teacher need consider *say* as of equal value to other words on the English SVL such as *straightforward, descriptive, pronunciation, chronological, personification* etc.

4.3. Development of word association lists

As Firth (1957) noted, a word is often defined by the 'company it keeps', so collocations can help in understanding its meaning and use. DiCerbo et al. (2014) also note that while wordlists are useful for secondary school, there is a hidden curriculum "that is, the norms and patterns of language as it is used within and across academic disciplines is never made explicit" (p. 446). A list of collocations to accompany the SVL words providing their important lexico-grammatical associations could therefore be a useful supplementary resource. Thus, we took an extra step not present in previously developed academic wordlists and created lists of each word's discipline-specific collocates. These word association lists can also be used in pedagogical materials, such as mind-maps, word-searches, weekly spelling lists etc. A set of criteria justified by previous research and explained in Table 5 guided selection of collocations.

Given the sensitivity of collocations to lexico-grammatical patterns (Hunston & Francis, 2000) collocations are reported separately for all forms of a lemma in the SVL. To compute collocations, each inflectional form was searched in a discipline with the tool *Concord* (Scott, 2016). Criterion 3 was applied and MI relations computed, and the results imported to *Excel* and *Notepad++* where a series of functions and regular expression pattern matches were used to order the lists and delete words that did not meet criteria 1 and 2. Some lemma forms did not have collocates that met the criteria, so the word association lists do not include every lemma inflection. Some had only a few collocates while others had more than 100, therefore we limited to the 10 most frequent collocations. The word association lists contain both MI and frequency information, as Simpson-Vlach and Ellis (2010) showed both inform judgements of teaching worth.

Table 6 illustrates the information in the word association lists. The complete resources are provided in the supplementary materials.

The lists capture related concepts, lexico-grammar, and semantic networks (e.g. hyponyms, synonyms, antonyms). For example, in Biology, *adenine* is a *base* in *DNA* that pairs with *thymine*, similar to the bond between bases such as *cytosine* with *guanine*. While an experienced biology teacher likely teaches these relationships, a literacy teacher may not be aware of such technical relationships, yet can draw on the lists for discipline-specific vocabulary activities in confidence that they capture meaningful associations. In English, the SVL contains the words *context*, *purpose*, *audience*, and *culture* and the word associations lists usefully organizes this set of text analysis terms together; similarly, in the context of Physics, *rarefaction* is an antonym of *compression*. As a reviewer pointed out, such sets might often be glossed by textbooks, but the word association lists capture more than this (and, of course, not every teacher prescribes a class textbook). For example, in History, *nations* have *disputes* that they *settle*, rather than arguments they resolve, and in Geography, one writes of *hydraulic action* not movement. For a student to attain disciplinary literacy, they must incorporate such language into their essays.

4.4. Word families

Given the tradition of wordlists organized by word families, including the seminal AWL (Coxhead, 2000), and the value of the information this provides, a word family version of the SVL was developed. Following Lei and Liu (2016), *Familizer Pro* (Cobb, 2012) took as input the lemma lists and matched each lemma to its word family members within the 25,000 most frequent words of the BNC. Once word families were extracted, each member was searched for within the corpus which ensured that the final lists contained only words that occurred in secondary school textbooks for the discipline. The word family version can be usefully used in conjunction with the other resources (note, however, that the word family headword is not always the

Table 5
Word association lists: Inclusion criteria.

1. ***Statistical Threshold***: > Mutual Information score of 3.00.
   MI is a statistic that computes if two words co-occur significantly more with each other than other words. The threshold of 3 is commonly used as indicating a meaningful relationship (Xiao & McEnery, 2006).
2. ***Minimum frequency***: > 5 co-occurrences within a 5 word span.
   MI excludes words that frequently combine with many words (e.g. *the*), but can overemphasize low frequency collocations (Liu, 2010, 2013). Therefore, collocates had to occur minimally 5 times within 5 words to the left or right of the SVL word.
3. ***Range***: > approximately 20% of texts.
   This metric ensured the collocation was not restricted to only a few texts.

Table 6
Extracts from discipline-specific word association lists ordered by MI score.

Biology	Chemistry	Physics	Geography
adenine(N.)	**compound(N.)**	**RAREFACTION (n.)**	**HYDRAULIC (adj.)**
thymine (n.)	formula (n.)	compression (n.)	action (n.)
(Freq. 152, MI 11.99)	(Freq. 703, MI 6.80)	(Freq. 126, MI 12.26)	(Freq. 99, MI 11.63)
cytosine (n.)	ionic (adj.)	pressure (n.)	abrasion (n.)
(Freq. 128, MI 11.86)	(Freq. 471, MI 6.80)	(Freq. 43, MI 7.21)	(Freq. 34, MI 11.77)
guanine (n.)	molecular (adj.)	wave (n.)	erosion (n.)
(Freq. 88, MI 11.78)	(Freq. 248, MI 5.83)	(Freq. 39, MI 6.78)	(Freq. 24, MI 7.68)
base (n.)	element (n.)	sound (n.)	fracturing (n.)
(Freq. 53, MI 8.96)	(Freq. 233, MI 5.75)	(Freq. 38, MI 7.52)	(Freq. 19, MI 13.57)
DNA (n.)	organic (adj.)	regions (n.)	rock (n.)
(Freq. 46, MI 6.23)	(Freq. 216, MI 6.52)	(Freq. 33, MI 10.77)	(Freq. 12, MI 6.86)
English	**Mathematics**	**Economics**	**History**
CONTEXT (n.)	**estimate.n:**	**policy.n:**	**DISPUTES (n.)**
audience (n.)	unbiased (adj.)	monetary (adj.)	international (adj.)
(Freq. 79, MI 8.18)	(Freq. 188, MI 11.88)	(Freq. 874, MI 9.32)	(Freq. 46, MI 7.96)
purpose (n.)	population (n.)	fiscal (adj.)	league (n.)
(Freq. 62, MI 7.92)	(Freq. 142, MI 9.63)	(Freq. 772, MI 9.81)	(Freq. 35, MI 7.02)
word (n.)	variance (n.)	government (n.)	settle (vb.)
(Freq. 42, MI 5.98)	(Freq. 104, MI 9.80)	(Freq. 348, MI 5.67)	(Freq. 31, MI 11.51)
text (n.)	mean (adj.)	expansionary (adj.)	border (n.)
(Freq. 30, MI 5.05)	(Freq. 78, MI 9.86)	(Freq. 229, MI 9.81)	(Freq. 26, MI 8.93)
culture (n.)	sample (n.)	economic (adj.)	nations (n.)
(Freq. 29, MI 7.77)	(Freq. 30, MI 6.67)	(Freq. 130, MI 4.78)	(Freq. 25, MI 6.57)

SVL headword); for example, the lemma version of the SVL for Biology contains *transmit.*v and *transmission.*n, and the word family version groups these together, along with other word family members in this discipline: *transmissible, transmission, transmitter.* Following Coxhead (2000), the word families version contains all spelling variants, e.g. *ionisation*, *ionisation*. An example of the 5 most frequent word families (from chemistry) is provided in Table 7.

The word family lists are ordered by rank frequency, and frequency information for the entire family and each member is provided. Thus the word *react* occurs 2331 times in the Chemistry subcorpus, *reacts* occurs 2195 times etc., and adding all members together, the REACT family occurs 27,991 times throughout Chemistry. The word family lists provide information for teachers about the most common forms in a family, e.g. *reaction* is the most frequent member of its family, and markedness, e.g. *ions* is twice as frequent as *ion* so the concept to be learned.

5. Discussion

The utility of an academic wordlist can be evaluated in a variety of ways, and perhaps the best is a classroom intervention study of the material. However, more commonly a method has been to gauge utility by how much vocabulary the resource covers. The idea is a rough measure, but the underlying question is: If a student knows the wordlist, how much of the vocabulary will they know in a target text? Coxhead (2000) computed that the AWL covered approximately 10% of the academic corpus from which it was derived, arguing that it was therefore more useful than the earlier *University Word List* of Xue and Nation (1984) which only covered 9.8% but included more word families. Gardner and Davies (2014) computed that their AVL covers more academic text than the AWL and concluded that it had higher utility. They report 13.8% coverage of academic texts in COCA, but only 8% for newspapers, and 4% for fiction, reflecting the academic construct of the AVL. Similarly, the discipline-specific wordlist of Lei and Liu (2016) covered 20.18% of vocabulary in their corpus of medical journal articles compared to the 19.44%, and only 6.64% of general academic writing and 3.68% of the BNC.

Table 7
Extracts from word family lists.

Headword	Family Freq.	Family Members and Freq.
REACT	27,991	react (2331) reactant (882) reactants (1256) reacted (467) reacting (546) reaction (14114) reactions (3789) reactive (1260) reactivity (878) reactor (37) reactors (18) reacts (2195) unreactive (218)
ACID	15,833	acid (11779) acidic (1189) acidity (205) acids (2660)
ATOM	12,317	atom (4665) atomic (1635) atomise (1) atomised (2) atomises (1) atomise (1) atoms (5932) subatomic (80)
ION	11,944	ion (3611) ionisation (690) ionise (53) ionised (96) ionises (46) ionising (3) ionisation (101) ionise (11) ionised (12) ionises (6) ionising (1) ions (7314)
FORM	11,094	form (5582) formed (3791) forming (607) forms (1114)

Therefore, coverage of the SVL was computed within the target discipline, as well as the rest of the corpus with the target discipline excluded. Results are reported in Table 8.

Table 8 indicates substantial discipline-specific coverage, comparable to Lei and Liu (2016), and as with previous validation studies, much less coverage generally. The lower coverage for some humanities subjects, e.g. English and History, possibly reflects that the harder sciences have a more specialised vocabulary while humanities subjects have a richer vocabulary (i.e. a wider range of word types). This is supported by (standardized) type-token ratios, with Biology and Chemistry having an (s)TTR of 35.14 and 32.14 respectively in the corpus, and English and History notably higher at 42.58 and 42.78.

The value of the SVL can be further demonstrated by considering the words at work in secondary texts themselves. Consider the extract in (2), analysed by *Lextutor* (Cobb, 2012):

(2) **Codominance** *occurs* when both **alleles** controlling a **trait** are fully expressed in the **heterozygous** condition. Incomplete *dominance occurs* when neither of the two **alleles** controlling a **trait** is completely **dominant** over the other and both express themselves to result in an *intermediate* **phenotype** in the **hybrids** (Perfect Guide: O Level Biology, Lee & Sim, 2014).

In (2), underlined words are in the most frequent 1000 GSL words, those in italics in the AWL, and those in bold in the SVL biology. Therefore, if a student has mastered the most frequent 1000 words of English, they would know 71.74% of the words in the paragraph, a further 8.7% of words are covered by the AWL, while 19.57% of the text is entirely covered by the SVL, i.e. *alleles*, *dominant*, *codominance*, *heterozygous*, *hybrids*, *phenotype*, *trait*. A long term curriculum to enhance disciplinary literacy that followed Nation's (2016) progression from general high frequency words, to general academic words, to discipline-specific words could effectively employ the GSL, AWL and SVL at progressive year levels from primary through middle and secondary school to achieve remarkable vocabulary coverage for students (in the case (2), 100% coverage). Of course, 100% coverage will not always be the case, though in such cases, the words not covered will be those of very low frequency in the discipline, e.g. *pinnae*.

Offering extensive classroom activities is beyond the scope of this paper, but recent book length treatments on the use of wordlists for pedagogy by Nation (2016) and Greene and Coxhead (2015) are recommended. Nevertheless, let us consider some examples to further demonstrate the value of the SVL. One traditional teaching practice for vocabulary is to select study words for a period of time, followed by a test on spelling, definitions, and productive use in sentences . This activity can be enhanced for disciplinary literacy easily with the SVL, e.g., a teacher can target all words from Biology, or all words from Physics, etc., knowing that these words will be relevant to their students content courses (Römer, 2011). Vocabulary items can be drawn from the lemma lists, word association lists, or word families. Other activities can be developed using the supplementary resources; for example, though we noted *set* is in fact in the GSL, though only in mathematics does it relate to words such as *empty*, *universal*, and *ordered*, i.e. the word association lists capture the subtypes of the mathematical sets. Concept maps are popular pedagogically (Lewis & Wray, 2012, p. 32), and a productive activity would be to provide some word associates of *set* to students as a starting point, asking students to map these to the target word and further elaborate on other subtypes that occur to them. Alternatively, guessing from context tasks (Webb & Nation, 2017, p. 637) would help bind these word associations together, reinforcing meaning and lexical priming (Hoey, 2005), e.g. an *ordered* ___ is a [definition]; A genealogical tree is a type of *empty*/*ordered*/*universal* (circle correct) set.

6. Limitations

A potential limitation of the SVL wordlists, collocations lists, and word family lists relates to the teaching value of some items, a common concern discussed with regard to objectively corpus-derived material (Gardner & Davies, 2016). Not all words on the SVL need equal classroom attention, and the teaching worth of any item depends on the educational context, e.g. grade level, language proficiency etc. As Dang et al. (2017) note, wordlists are a guide and resource and their effective use is very much in the hands of education professionals. Methodological limitations meant that a certain amount of noise crept into our data at several points, with implication for count precision: firstly during scanning and optical character recognition, secondly during tagging and lemmatization, and finally via the software and techniques used to analyse the data. Some of the challenges are, of course, perennial problems in corpus linguistics (Sampson, 2002). Nevertheless, the researchers believe that the extensive

Table 8
SVL coverage per discipline.

% Words Covered	Biology	Chemistry	Physics	Geography
Within discipline	23%	25%	22.7%	15.9%
Corpus overall	2%	2.2%	2.9%	2.4%
	English	**Mathematics**	**Economics**	**History**
Within discipline	13%	20.9%	21.8%	14%
Corpus overall	2%	2.9%	2.2%	1.7%

effort devoted to text preprocessing, and the manual inspection and correction of errors, combined with the many rigorous criteria to determine if a word belonged in the SVL have ensured that important are represented.

7. Conclusion

It has been argued that the technical advances in recent EAP corpus-informed wordlists development at the tertiary-level has not yet been fully leveraged in educational contexts such as secondary school, where there is a need for resources to facilitate disciplinary literacy. Further, in the contemporary landscape of secondary education, with growing numbers of second language speakers and international schools, resources informed by EAP research that has developed in the context of ESL education are increasingly relevant. The current project has responded by producing the *Secondary School Vocabulary Lists* (SVL), a set of academic vocabulary resources freely available in the supplementary materials of this journal, including discipline-specific lemma lists, word association lists, and word families for eight subjects: Biology, Chemistry, Economics, English, Geology, History, Mathematics, and Physics. With the addition of the SVL to existing wordlists, e.g. the (New) GSL, AWL, AVL, there are now resources available that come close to representing most of the vocabulary students need across the secondary school curriculum. Also, the gap has been filled between the recent *Middle School Vocabulary Lists* (Greene & Coxhead, 2015) and the existing discipline-specific wordlists available for tertiary education.

Acknowledgments

This research was funded by Grant number: OER DEV 01/16 CGG, by the Office of Educational Research.

Appendix A. Supplementary data

Supplementary data related to this article can be found at https://doi.org/10.1016/j.jeap.2018.07.004.

Uncited references

Hyland, 2008; Martin, 2013; Prosuite, 2016.

References

Airey, J., Lauridsen, K.M., Räsänen, A., Salö, L., Schwach, V., 2017. The expansion of English-medium instruction in the Nordic countries: Can top-down university language policies encourage bottom-up disciplinary literacy goals?. Higher Education 73 (4), 561–576.
Bauer, L., Nation, P., 1993. Word families. International Journal of Lexicography 6 (4), 253–279.
Brezina, V., Gablasova, D., 2013. Is there a core general vocabulary? Introducing the new general Service list. Applied Linguistics 36 (1), 1–22.
Brezina, V., Gablasova, D., 2017. How to produce vocabulary lists? Issues of definition, selection and pedagogical aims. A response to Gabriele Stein. Applied Linguistics 38 (5), 764–767.
Browne, C., Culligan, B., Phillips, J., 2013. A new general service list, Retrieved from http://www.newgeneralservicelist.org.
Chung, T.M., Nation, P., 2004. Identifying technical vocabulary. System 32 (2), 251–263.
Cobb, T., 2012. The compleat lexical tutor for data driven learning on the web. Montreal University of Quebec, available at http://lexutor.ca/.
Coxhead, A., 2000. A new academic word list. Tesol Quarterly 34, 213–238.
Coxhead, A., 2018. Vocabulary and English for specific purposes Research: Quantitative and qualitative perspectives. Routledge, New York.
Dang, T.N.Y., Coxhead, A., Webb, S., 2017. The academic spoken word list. Language Learning 67 (4), 959–997.
DiCerbo, P.A., Anstrom, K.A., Baker, L.L., Rivera, C., 2014. A review of the literature on teaching academic English to English language learners. Review of Educational Research 84 (3), 446–482.
Durrant, P., 2016. To what extent is the Academic Vocabulary List relevant to university student writing?. English for Specific Purposes 43, 49–61.
Firth, J.R., 1957. Papers in linguistics 1934–1951. OUP, Oxford.
Gardner, D., Davies, M., 2014. A new academic vocabulary list. Applied Linguistics 35 (3), 305–327.
Gardner, D., Davies, M., 2016. A response to "To what extent is the Academic Vocabulary List relevant to university student writing?". English for Specific Purposes 43, 62–68.
Garside, R., Leech, G., McEnery, T., 1997. Corpus annotation: Linguistic information from computer text corpora. Longman, London.
Greene, J., Coxhead, A., 2015. Academic vocabulary for middle school students. Brookes, Baltimore.
Ha, A.Y.H., Hyland, K., 2017. What is technicality? A Technicality Analysis Model for EAP vocabulary. Journal of English for Academic Purposes 28, 35–49.
Harwood, N., 2014. English language teaching textbooks: Content, consumption, production. Palgrave, Basingstoke.
Hoey, M., 2005. Lexical priming: A new theory of words and language. Routledge, London.
Hunston, S., Francis, G., 2000. Pattern grammar: A corpus-driven approach to the lexical grammar of English. Computational Linguistics 27 (2), 318–320.
.
Hyland, K., Tse, P., 2007. Is there an academic vocabulary?. Tesol Quarterly 41 (2), 235–253.
Lee, L., Sim, 2014. Perfect guide: O level biology. Marshall Cavendish, Singapore.
Lei, L., Liu, D., 2016. A new medical academic word list: A corpus-based study with enhanced methodology. Journal of English for Academic Purposes 22, 42–53.
Lewis, M., Wray, D., 2012. Literacy in the secondary school. Routledge, London.
Liu, D., 2010. Is it a chief, main, major, primary, or principal concern?: A corpus-based behavioral profile study of the near-synonyms. International Journal of Corpus Linguistics 15 (1), 56–87.
Liu, D., 2013. Salience and construal in the use of synonymy: A study of two sets of near-synonymous nouns. Cognitive Linguistics 24 (1), 67–113.

Liu, J., Han, L., 2015. A corpus-based environmental academic word list building and its validity test. English for Specific Purposes 39, 1–11.
.
Martinez, R., Schmitt, N., 2012. A phrasal expressions list. Applied Linguistics 33 (3), 299–320.
Miller, D., Biber, D., 2015. Evaluating reliability in quantitative vocabulary studies: The influence of corpus design and composition. International Journal of Corpus Linguistics 20 (1), 30–53.
Měchura, M., 2016. Lemmatization list: English, Retrieved from http://www.lexiconista.com/datasets/.
Nation, P., 2006. How large a vocabulary is needed for reading and listening?. Canadian Modern Language Review 63 (1), 59–82.
Nation, P., 2016. Making and using word lists for language learning and testing. John Benjamins, Amsterdam.
Ogle, D., Blachowicz, C., Fisher, P., Lang, L., 2016. Academic vocabulary in middle and high school: Effective practices across the disciplines. Guilford Publications, New York.
Prosuite, 2016. Wordstat [computer software]. Montreal.
Römer, U., 2011. Corpus research applications in second language teaching. Annual Review of Applied Linguistics 31, 205–225.
Sampson, G., 2002. Empirical linguistics. A&C Black, London.
Schmitt, N., 2010. Researching vocabulary. Palgrave Macmillan, Hampshire.
Schmitt, D., Schmitt, N., Mann, D., 2011. Mastering the academic word list. Pearson Longman, New York.
Scott, M., 2016. WordSmith tools (version 6 and 7) [computer software]. Liverpool.
Shanahan, T., Shanahan, C., 2017. Disciplinary Literacy: Just the FAQs. Educational Leadership 74 (5), 18–22.
Simpson-Vlach, R., Ellis, N., 2010. An academic formulas list: New methods in phraseology research. Applied Linguistics 31 (4), 487–512.
Someya, Y., 1998. E-Lemma [Data file], Retrieved from http://www.lexically.net/downloads/e_lemma.zip.
Thorndike, E.L., Lorge, I., 1944. The teacher's word book of 30,000 words. Columbia University Press, Columbia.
Todd, R.W., 2017. An opaque engineering word list: Which words should a teacher focus on?. English for Specific Purposes 45, 31–39.
Webb, S., Nation, P., 2017. How vocabulary is learned. OUP, Oxford.
West, M., 1953. A general service list of English words. Longmans, London.
Wingate, U., 2011. A comparison of 'additional' and 'embedded' approaches to teaching writing in the disciplines. In: Deane, M., O'Neil, P. (Eds.), Writing in the disciplines. Palgrave Macmillan, Basingstoke, pp. 65–87.
Xiao, R., McEnery, T., 2006. Collocation, semantic prosody, and near synonymy: A cross-linguistic perspective. Applied Linguistics 27 (1), 103–129.
Xue, G., Nation, P., 1984. A university word list. Language Learning and Communication 3 (2), 215–229.
Yang, M., 2015. A nursing academic word list. English for Specific Purposes 37, 27–38.

Dr Clarence Green currently lectures in Psycholinguistics, Corpus Linguistics and Research Methods at the National Institute of Education, Nanyang Technological University. He holds a PhD in linguistics and his research interests include the psychology of language, corpus linguistics, disciplinary literacy, stylistics and cognitive-functional grammar. His research has appeared in journals such as *Cognitive Linguistics*, *Language and Literature*, *Functions of Language*, and *Literary and Linguistic Computing,* and he is author of the book *Patterns and Development in the English Clause System.*

Dr James Lambert is an assistant professor at the National Institute of Education, Nanyang Technological University. His research interests include lexicography, corpus linguistics, and World Englishes. Prior to his doctorate he worked as a professional lexicographer for over 15 years with Macquarie Dictionary Publishers, Sydney. He has edited dictionaries of various varieties of Asian Englishes, including Indian, Malaysian, Singapore, and Philippine English, and was involved in creating the corpora these publications were based on. He has published in the journals *English World-wide*, *World Englishes*, the *International Journal of Lexicography*, and the *British Journal of Educational Technology*.