Reading in a Foreign Language
ISSN 1539-0578

January 2024, Volume 36, No. 1
pp. 1–16
https://hdl.handle.net/10125/67453

Corpus Lists for English Learners: Supporting Reading Comprehension of Digital Science Resources

Rebeca Arndt
University of Central Florida
United States

Abstract

This study explored the lexical coverage of corpus-based vocabulary lists (general, academic, and content-specific) across several million tokens gathered from digital science resources (DSR) for middle school (6–8 grade) students in the United States. The goal was to estimate the extent to which a combination of well-known word lists, mostly designed for the needs of L2 learners, might help students reach text coverage that could result in reasonable comprehension of science texts. The findings of this study show that: (a) the top 570 word families in the newer Academic Vocabulary List (AVL) provide 75% more lexical coverage in the corpus than the 570 word families in the older Academic Word List (AWL), (b) a series of lists designed for second language learners, such as General Service List (GSL), Academic Word List (AWL), and the English for Academic Purposes (EAP) Science List (about 2,900 word families) offer 88.33% coverage, and (c) the GSL and the Middle School Vocabulary List (MSVL) for Science (fewer than 2,500 word families) provide 87.77% coverage.

***Keywords*:** vocabulary, science digital texts, lexical coverage, corpus-based vocabulary lists, middle school English L2 learners, middle school science texts, L2 science reading comprehension, English for Academic Purposes

It is well known that vocabulary knowledge is important to receptive and productive language comprehension (van Zeeland & Schmitt, 2013; Verhoeven et al., 2011). Vocabulary knowledge can be estimated in terms of vocabulary breadth (or size), vocabulary depth (how well a learner knows a lexical item) (Webb, 2020), and fluency (the degree of automaticity to which a learner can recognize, process, and use a lexical item) (Daller et al., 2007). From a reading research standpoint, vocabulary knowledge or lexical access is presented as one of the lower-level processes that support reading comprehension together with other dimensions such as syntactic parsing and semantic proposition formation (Grabe & Stoller, 2020). Although there are substantial linguistic and processing differences between first language (L1) and second language (L2) readers in terms of reading abilities, they "share almost all of the same component skills, and the reading construct is very similar in terms of underlying cognitive and linguistic components" (Grabe & Stoller, 2020, p.66). In fact, the functioning mechanisms of vocabulary

knowledge in achieving comprehension are similar for L1 and L2 readers, however, vocabulary knowledge marks a different starting point for a beginner L2 reader compared with a beginner L1 reader, particularly as a L1 kindergartner may have already acquired a vocabulary size of approximately 6,000 words by the time reading instruction begins (Grabe & Stoller, 2020).

Lexical Coverage Needed for Comprehension

Vocabulary knowledge has been examined in vocabulary research from several perspectives such as the lexical coverage of texts (how many words students need to know to comprehend texts) and the lexical mastery level of learners (vocabulary size of the learner) to reveal that texts vary in difficulty and require different vocabulary sizes to enable comprehension (Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006). For instance, fiction texts targeting adolescent readers require knowledge of 9,000 word families plus proper nouns to reach 98% coverage whereas newspapers require knowledge of 8,000 word families and proper nouns to reach the same coverage (Nation, 2006). Likewise, authentic texts require knowledge of 5,000 word families and proper nouns to reach 95% coverage (Laufer, 2021).

The 98% threshold has been referred to as providing optimal coverage for reading comprehension, whereas the 95% threshold has been referred to as offering minimal coverage (Laufer, 2021). Recently, Laufer (2020) found that minimal coverage could be achieved by knowing 90% of words in the text and inferring the meaning of 5% of the words in a text. In science contexts, however, middle school learners need to know at least the first 8,000 of Nation's (2012) British National Corpus (BNC)/Corpus of Contemporary American English (COCA) (Davies, 2008) word families together with supplementary lists to reach the 95% reading threshold and at least 11,000 word families plus supplementary lists to reach the 98% reading threshold (Coxhead & Bourtorwick, 2018).

Knowing that L2 learners or English learners (ELs) require large vocabulary sizes to comprehend authentic texts (Coxhead & Bourtorwick, 2018; Laufer & Ravenhorst-Kalovski, 2010; Nation, 2006), one line of inquiry in vocabulary research explored the potentiality of acquiring a vocabulary size of 9,000 word families via incidental learning possibly because "a vocabulary size of 9,000 words or more is a sensible long-term goal for unassisted reading of unsimplified texts" (Nation, 2014, p.2) (Green, 2022; McQuillan, 2016).

Since acquiring a vocabulary of 9,000 words would require a learner to read one hour each day for 6.3 years or two hours each day for 3 years for incidental vocabulary learning to occur (Green, 2022; McQuillan, 2016), it is important to explore other vocabulary learning alternatives such as leveraging corpus-based lists for intentional vocabulary teaching and learning purposes. Barcroft (2009) showed that L2 intentional vocabulary learning on word form yielded better results than incidental vocabulary learning. Therefore, it is sensible to assess the potential value of corpus-based lists in academic vocabulary teaching and learning contexts.

Coverage of Corpus Lists Across Digital Science Resources (DSR) for Middle School Learners

Insofar, only a few studies explored the coverage of various general, academic, and content-specific (specialized) corpus-based lists across science textbooks (Coxhead et al., 2010; Coxhead & Boutorwick, 2018; Greene & Coxhead, 2015). Surprisingly, no study has explored the coverage of general, academic, and content-specific corpus lists across digital science resources

(DSR) for middle school that science teachers use more often in their science instruction than interactive and non-interactive textbooks (Gray & Lewis, 2021). This line of inquiry is important because profiling general and specialized corpus-based lists, including lists designed specifically for L2 learners across state-of-the-art DSR, can reveal the percentage of general and specialized lexical items across these resources, and provide insight into the merits of teaching and learning lexical items in these lists to support middle school English learners with text comprehension. Since ELs struggle with academic and science-specific vocabulary (Ardasheva et al., 2017) it is particularly important to examine if corpus-based lists have pedagogical value and might mitigate the "significant vocabulary, including science vocabulary, and general language skills gaps among current ELs" (Ardasheva et al., 2021, pp. 932–933).

The findings of this study can be equally valuable for learners and other educational stakeholders (e.g., teachers, ESOL, ESL, L2 professionals, coaches, administrators, and parents) involved in science language teaching and learning. This is because the study presents the percentage of general, academic, and content-specific lexical items that students would recognize in DSR if they would learn general high-frequency, academic, and content-specific words enclosed in the premade vocabulary lists examined. The information about the lexical coverage of these lists across DSR is useful when comparing the coverage of lists across a collection of texts and the minimal and optimal reading comprehension thresholds (95% and 98% lexical coverage). For instance, examining the lexical coverage of the General Service List (GSL) (West, 1953), a list assembled for L2 learners, across the entire corpus can reveal the value of teaching lexical items in this list to beginner and intermediate EL learners. Simultaneously, profiling the lexical coverage of the corpus using academic and content-specific vocabulary lists created for L2 learners, such as the Academic Word List (AWL) (Coxhead, 2000) and the English for Academic Purposes (EAP) Science List (Coxhead & Hirsh, 2007) can highlight the instructional value of these lists for intermediate and advanced L2 learners.

Research Purpose

This study focuses on finding the lexical coverage of several corpus-based lists (most of them designed for L2 learners) across DSR for middle school learners to reveal the usefulness of teaching and learning lexical items from these lists. The three sub-questions under scrutiny in this study are as follows:

What lexical coverage in digital science resources (DSR) for middle school is provided by:

(a) the GSL (West, 1953) + AWL (Coxhead, 2000) + EAP Science List (Coxhead & Hirsh, 2007)?
(b) the top 570 word families in the Academic Vocabulary List (AVL) (Gardner & Davies, 2014)?
(c) the GSL (West, 1953) + Middle School Vocabulary List (MSVL) for Science (Greene & Coxhead, 2015)?

Literature Review

Starting from the premise that vocabulary is a contributing factor to science and science language learning (Ardasheva, 2017; Ardasheva et al., 2017; Ardasheva et al., 2021) it is important to examine the composition of science vocabulary including general, academic, and content-specific or technical vocabulary. In the last decades, several general, academic, and content-specific vocabulary lists were designed considering different inclusion and exclusion criteria and targeting specific audiences. A clear explanation of how these lists were designed is warranted to better understand the benefits of using general, academic, and technical corpus-based lists. Given the scope of this research, it is also important to examine studies that explored the lexical coverage of such lists across secondary science school texts. The sections that follow include: (a) a description of vocabulary categories, (b) the purpose and process of extracting the vocabulary lists employed in this study, and (c) an overview of the vocabulary needed to develop L2 language.

General Vocabulary and the General Science List (GSL)

General vocabulary is essential for all students, whether native speakers of English or EFL or ESL learners. However, not all lexical items are equally valuable for learners (Webb & Nation, 2017). The value of knowing various lexical items depends on the communicative or learning needs of a learner. For pedagogical and academic purposes, it is useful to focus on learning the lexical items based on frequency. From the frequency perspective, vocabulary is generally categorized into high-frequency vocabulary and low-frequency vocabulary bands (Webb & Nation, 2017). General vocabulary contains lexical items that are most used in a language in all discourse types. This kind of vocabulary is generally referred to as Tier 1 vocabulary (Beck et al., 2013) in contexts. For L2 learners, general vocabulary is essential for the development of basic interpersonal communication skills (BICS) (Cummins (1979; 2000).

In vocabulary research, the General Service List (GSL) designed by West (1953) for ESL and EFL learners from a corpus compiled almost 90 years ago, has been used as the yardstick to delineate general, high-frequency vocabulary. The GSL contains approximately 2,000 word families (Nation, 2004) and has been increasingly used as a “starting point for vocabulary" (Webb & Nation, 2017, p. 25). The general nature of this list is evident through the GSL's composition, which is high-frequency lexical items that are universally encountered across various foreign languages used in communicative discourses on multiple topics (Gilner, 2011). In terms of lexical coverage, previous studies revealed that GSL’s coverage is 71.52% to 78.85% across science texts (Coxhead et al., 2010; Coxhead & Hirsch, 2007).

Academic and Technical (Specialized) Vocabulary and Lists

It is important to note that both academic and technical lexical items have been referred to as specialized vocabulary (Coxhead, 2018; Nation, 2013). While the GSL focuses on vocabulary for general purposes, vocabulary can further be categorized into academic vocabulary and technical vocabulary. These two kinds of vocabulary are comprised of lexical items that frequently occur in academic and specialized discourse but quite infrequently in everyday language (Webb & Nation, 2017).

Academic vocabulary, also called sub-technical vocabulary or Tier 2 vocabulary in K–12 circles (Beck et al., 2013), consists of lexical items identified across multiple academic fields (Webb & Nation, 2017). For instance, the word *occur* can be encountered in fields such as English Language Arts, Biology, Economics, Linguistics, Mathematics, Science, etc. The role of academic vocabulary is to provide contextual information about technical words, which are generally essential to the topic.

Academic vocabulary (semi-technical) works in tandem with technical vocabulary. By nature, technical vocabulary is generally formed by low-frequency words in everyday language but is predominantly frequent in specialized areas (Webb & Nation, 2017). Moreover, technical vocabulary, which is often referred to as Tier 3 vocabulary among K–12 educators (Beck et al., 2013), "represent[s] specialized knowledge that is essential to learning a particular topic" (Webb & Nation, 2017, p. 30) and can be up to 30% of the running words in a discipline-specific text (Nation, 2013). Interestingly, technical vocabulary can also be comprised of general high-frequency (e.g., *interest*) words and academic words (e.g., *tension*), and it includes not solely single words but multiword units and acronyms such as *flux theorem* (Liu & Lei, 2020). In this study, specialized vocabulary will be used to define "an umbrella term for vocabulary which relates in some way to its particular discipline, whether the word is high, mid or low frequency" (Coxhead, 2018, p. 21).

An essential characteristic of specialized vocabulary is its dominant role in understanding and learning content in a specific discipline, particularly in written science texts, which contain a large proportion of specialized vocabulary. For instance, middle school science textbooks may contain as much as 10.17% technical words (Greene & Coxhead, 2015). For L2 learners to comprehend specialized texts, they must acquire Cognitive Academic Language Proficiency (CALP), which Cummins (1979) suggests can take between 5–7 years. CALP takes longer to acquire than BICS (basic interpersonal communication skills) as academic language is usually acquired in formal settings. In such settings, opportunities for incidental learning is decreased due to a lower amount of input due to less repeated exposure to lexical items (Miralpeix, 2020).

Coxhead's Academic Word List (AWL)

A seminal academic vocabulary list is the Academic Word List (AWL) (Coxhead, 2000), assembled for the needs of L2 learners, particularly for English for Academic Purposes (EAP). This list was compiled from the Academic Corpus containing 3,513,330 running words collected from 28 subject areas and four main disciplines (Arts, Commerce, Law, Science). The words included in each of the equally weighted sub-corpora in the four main disciplines were retrieved mainly from college-level journal articles, texts, book chapters, textbooks, and manuals for English speakers in New Zealand, the United Kingdom, the United States, Australia, and Canada. The software used to process the Academic Corpus was the Range Program (Heatley & Nation, 1996). This software profiled the vocabulary across the corpus and extracted the word families that eventually made it to the AWL based on several criteria.

In determining which word families to add to the AWL, Coxhead (2000) made an operational decision to exclude GSL word families and include only words with specialized occurrences. The reasoning behind this exclusion was based on the idea the AWL was designed for college-level learners who would have (or at least should have) already acquired general vocabulary. In

her study, the AWL word families had to meet three criteria: range, frequency, and specialized occurrence (Folse, 2020). Range-wise, AWL word families had to appear in all four sub-corpora with a frequency of at least ten occurrences and in at least half of the subject area (i.e., in at least 15 of the 28 subject areas represented). Frequency-wise, AWL accepted word families had to appear at least 100 times overall in the Academic Corpus.

Once Coxhead identified the 570 word families that make up her AWL, she calculated the coverage offered by the AWL. It turns out the AWL provided 10% coverage across academic writing. The validation on the list across a fiction corpus yielded 1.4% coverage and roughly 8.5% coverage across a similarly built academic corpus (678,000 running words). Since its publication, the AWL has been validated numerous times (Li & Qian, 2010; Coxhead et al., 2010; Paribakht & Webb, 2016). Multiple studies have confirmed that AWL covers approximately 10% of academic texts.

Gardner and Davies' Academic Vocabulary List (AVL)

Created from a different perspective than the AWL (Coxhead, 2000), the Academic Vocabulary List (AVL) (Gardner & Davies, 2014) also considered the needs of L2 learners as this list is intended to be used in EAP contexts. The researchers chose lemma as the unit for counting considering that the learners for which this list was intended may display basic knowledge of general English vocabulary. Unique to the methodology employed is that the list was not based on any premade vocabulary lists. Instead, it was based on statistical power that dictated the selection of the lexical items within the academic corpus from which the list was assembled. At the same time, the 3,015 lemma list was derived from one of the largest corpora compiled available at the time of this study: 120 million running words from the academic portion of the Corpus of Contemporary American English (COCA) (Davies, 2008). This corpus encompassed contemporary language gathered from nine different sub-corpora of varying sizes, each of which contains roughly 8 to 22 million running words from Education, Humanities, History, Social Science, Philosophy, Religion, Psychology, Law and Political Science, Science and Technology, Medicine and Health, and Business and Finance.

The inclusion criteria of lemmas within the AVL were clearly laid out. Firstly, a ratio/frequency criterion was adopted, meaning that the frequency of the lemmas in the entire corpus must be at least 50% higher than in the non-academic corpus. Secondly, a range criterion was selected. This criterion ensured the lemmas included in the AVL occur at least 20% of the expected frequency in seven or more of the nine subjects. Thirdly, the lemmas had to be dispersed as close as possible to 1.00, which depicts perfect dispersion, thus not lower than 0.8. Finally, a discipline measure criterion was put in place: AVL lemmas must appear a maximum of three times the expected frequency in any of the nine subject areas. This last criterion prevented the inclusion of technical words.

For the sake of comparison, the AVL lemmas were converted into word families. As part of the study, the researchers compared 570 randomly chosen word families from the AVL with the AWL. The coverage of these two word-families-based lists was analyzed across two academic corpora. One of the corpora was the COCA, containing 120,847,709 running words. The other corpus was the British National Corpus (BNC), which counted 32,828,961 running words. The

analysis showed that the AVL covered 13.8% of the COCA while the AWL covered only 7.2%. It also revealed that the AVL covered 13.7% of the BNC while the AWL covered 6.9%.

The Middle School Vocabulary List for Science (MSVL)

The Middle School Vocabulary List (MSVL) for Science (Greene & Coxhead, 2015) is a 435 word families (or 859 word types) list designed from 3,725,422 running words retrieved from 6th, 7th, and 8th grade science textbooks for learners in the United States. This list, which is part of a set of five lists developed from the MS-CAT Corpus (109 textbooks corpus), was developed using "much of the same methodology" (Greene & Coxhead, 2015, p 123) as the AWL (Coxhead, 2000). Due to the methodology employed, it can be inferred that the list is valuable for L2 learners.

The MSVL for Science was built on top of the GSL, meaning all the GSL words were removed from the lists. Following a multi-level top-down procedure, the researchers used several inclusion criteria, such as range and frequency. First, the researcher added to the list AWL word family members that met two criteria (i.e., range and frequency). This meant that the base of the MSVL for Science was comprised of AWL words that occurred 11.4 times per million words in the science sub-corpora and at least 28.5 times per million words in the larger corpus, the MS-CAT Corpus. Next, in the second step, the AWL word family members were included if they had frequencies of at least 28.5 times per million words in the MS-CAT Corpus and 11.4 times per million words in the science sub-corpora. Then, in the third step, non-AWL word families with the same range and frequency as determined in the first step were added to the list. Lastly, the researchers selected non-AWL words with a frequency of at least 100 times per million words in the science-sub-corpora. This last step ensured that technical words were accurately accounted for. In terms of coverage, the MSVL for Science provided 10.17% coverage across the science sub-corpus and 9.48% across a parallel corpus created from other sets of middle school texts.

The English for Academic Purposes Science List (EAP)

The English for Academic Purposes (EAP) Science List (Coxhead & Hirsch, 2007) was assembled following Coxhead's (2000) methodology to extract technical words outside of the GSL and the AWL. Two sub-corpora formed the corpus used in this study. One portion of the corpus was a previously compiled science corpus used by Coxhead (2000) to assemble the AWL. Another portion of the corpus was a newly developed sub-corpus. The existing science-based sub-corpora used by Coxhead (2000) contained 875,846 tokens harvested from several subject areas (e.g., Biology, Chemistry, Geography, Geology, Mathematics, Computer Science, and Physics). The newly assembled sub-corpus was formed from seven science-based subject areas (e.g., Agricultural Science, Ecology, Horticultural Science, etc.). Roughly 125,000 tokens were collected from each of the seven science-based subject areas. In total, the final corpus that was used to design the EAP Science List contained 1,761,380 tokens.

After the compilation of the corpus, the instrument Range Program was used to obtain the words included in the EAP Science List based on three inclusion criteria: (1) range (occuring in at least half of the subject areas included in the corpus), (2) frequency (appearing at least 50 times in the entire corpus), and (3) dispersion (at least 35 dispersion factor). Proper nouns, symbols, and abbreviations were not included in the specialized list. The final EAP Science List consisted of

318 word families divided into six sub-lists. This specialized list offered coverage of 3.79% across the 1,761,380 token corpora. On a broader scale, the GSL, AWL, and the EAP Science List offered coverage of 84.27% across the corpus. To verify its accuracy, the EAP Science List was run across four non-science-based corpora containing between 874,723 tokens (Commerce) and 3,763,733 tokens (Fiction). It turned out that the coverage provided by the specialized science list was between 0.27% (Fiction) and 0.61% (Arts), values considerably lower than the coverage over the science corpus. List validation proved that the EAP Science List meets the purpose for which it was designed.

Method

The goal of this study was to explore the lexical coverage of several corpus-based lists across Digital Science Resources (DSR) for middle school learners to show the pedagogical value of teaching and learning lexical items in these lists in science contexts. Supplemental digital resources have been reportedly used by middle school teachers in the United States at higher rates than interactive and non-interactive science textbooks (Gray & Lewis, 2021), which warrants a closer examination of these resources with pedagogical value. To accomplish this goal, a digital corpus was first assembled: the Digital Science Corpus (DSC). Then five premade lists were profiled against the DSC.

Corpus

The Digital Science Corpus (DSC) was composed of 3,456 DSR gathered from ten science-based platforms considering the representativeness of the corpus, coherence, the balance of sub-corpora in the corpus, and the size of the corpus (Dang, 2020; Nation & Sorell, 2016). The corpus was representative (Dang, 2020) because it was assembled from digital content on three science topics taught in grades 6–8 across the US: Physical Science (PS), Life Science (LS), and Earth and Space Science (ESS) and categorized based on the New Generation Science Standards (NGSS)(NGSS Lead States, 2013).To balance the sub-corpora and the corpus (Nation & Sorell, 2016), the DSR were divided into three categories based on the size of the resources: small (S), medium (M), and large (L). The small resources were 750 tokens or less, the medium resources were between 751 and 1500 tokens and the large resources were larger than 1,501 tokens. Coherence-wise, the resources were gathered from ten platforms (e.g., National Geographic, Newsela, Read Actively, etc.) used by middle school teachers across the United States, as found via a questionnaire administered to 91 teachers. The size of the corpus was also considered (Nation & Sorell, 2016) as 2,734,595 usable tokens were included in the DSC.

Instrument and Lists

The lexical profiling analysis of the DSC was conducted to answer the main research question with its three sub-questions. Lexical profiling analysis explored five premade general and academic word lists in the compiled digital corpus. The freely available software AntWordProfiler (Anthony, 2021) version 1.5.1 was used along with the GSL, the AWL, the top 570 word families in the AVL, the EAP Science List, and the MSVL for Science. Each list (the size of each list is available in Table 1) was used as a base list to evaluate their lexical coverages across the DSC.

Table 1
Word Families at Level 6 In Various Lists

Vocabulary List	Number of Word Families in the List
GSL (1st GSL and 2nd GSL)	1,986
AWL	570
EAP Science List	318
AVL*	1,991*
MSVL for Science	435

*Only the top 570 word families in the AVL

Results

The First Research Sub-Question

Table 2 shows that the broadest lexical coverage across the Digital Science Corpus (DSC) is provided by general lexical items, namely lexical items included in the 1st and 2nd General Service List (GSL) (79.21%). Similarly, the Academic Word List (AWL) provides a lexical coverage of 5.76% across the DSC, and the English for Academic Purpose (EAP) Science List offers a lexical coverage of 3.36% across the same corpus. Therefore, 88.33% of the DSC is covered by 2,841 word families or 9,806 types included in the GSL, AWL, and the EAP Science List.

Table 2
Lexical Coverage of the GSL, AWL, and EAP Science List Across DSC

List	Token	Token%	Total Token%	Type	Type%	Total Type%	Group	Group%	Total Group%
1st GSL	1,982,559	72.5	72.5	3,652	9.17	9.17	998	3.04	3.04
2nd GSL	183,624	6.71	79.21	2,955	7.42	16.59	966	2.94	5.98
AWL	157,507	5.76	84.97	2,269	5.7	22.29	564	1.72	7.7
EAP SC	91,752	3.36	88.33	930	2.34	24.63	313	0.95	8.65
Not in List	319,153	11.67	100	30,016	75.38	100	30,016	91.35	100
Total	2,734,595	100		39,822			32,857		

These findings are aligned with Greene and Coxhead (2015) who found that the 1st and 2nd GSL grants 79.09% coverage across the science sub-corpora of the MS-CAT. However, since Greene and Coxhead (2015) only reported the coverage of AWL across the entire MS-CAT (5.37%), it is unreasonable to conduct a comparison between the coverage of AWL across a specialized corpus (DSC) and a general academic corpus (MS-CAT). For reference, Greene and Coxhead (2015) found that the AWL covered 5.37% of the MS-CAT and that the lexical coverage of GSL and AWL across the entire MS-CAT was 84.93%.

Another comparison of these findings with earlier research points out that the EAP Science List provides 3.79% lexical coverage across an academic corpus comprised of college-level texts (Coxhead & Hirsh, 2007) which is comparable to the coverage of the same list across the DSC (3.36%). Although the size and composition of DSC are different from the science corpus used by Coxhead and Hirsh (2007) (i.e., the latter had a corpus of 1,761,380 words while the DSC had 2,734,595 tokens), this analysis suggests the EAP Science List could potentially be used with secondary learners.

The Second Research Sub-Question

As evident in Table 3, the top 570 word families (1,816 types) in the Academic Vocabulary List (AVL) provide a lexical coverage of 10.07% in the DSC. Surprisingly, these results suggest that the top 570 word families in the AVL offer 4.31% more lexical coverage in the DSC than the 570 word families in AWL across the same corpus, meaning that it provides an additional 74.82% lexical coverage in the DSC than AWL.

Table 3
Lexical Coverage of the First 570 Word Families Across DSC

List	Token	Token%	Total Token%	Type	Type%	Total Type%	Group	Group%	Total Group%
AVL 570	275,311	10.07	10.07	1,816	4.56	4.56	566	1.47	1.47
Not in List	2,459,284	89.93	100	38,006	95.44	100	38,006	98.53	100
Total:	2,734,595		39,822			38,572			

These findings are in alignment with Newman (2016) and Hernandez (2017), whose research showed that the top 570 word families in the AVL had a higher coverage in academic corpora than the 570 word families included in the AWL (16.14% and 15.39%, respectively). Notably, Hernandez's (2017) study suggests that for the 8th grade academic sub-corpora, the 570 word families in the AVL provided 23.05% more coverage than the 570 word families included in the AWL. However, several hundred-word families are shared by the GSL and the AVL. Therefore, it may be asserted that some of the academic lexical items in the AVL could be general in nature. The overlap between the GSL and the AVL could explain the greater coverage of the AVL compared with the AWL, a list that does not share any word families with the GSL.

The Third Research Sub-Question

The analysis revealed that the 1st and 2nd GSL together with the Middle School Vocabulary List (MSVL) for Science gives 87.77% coverage across the DSC. Specifically, the MSVL for Science covers 8.56% of the DSC (see Table 4). In other words, the MSVL for Science offers slightly less coverage (0.56%) across the DSC than the AWL and the EAP Science List together. However, when considering the size of the lists, 888 word families (AWL and the EAP Science List) versus 435 families (MSVL for Science), it can be noticed that acquiring fewer MSVL for Science word families can produce similar coverage of middle school digital text compared with acquiring more than double the word families in the AWL and the EAP Science List.

Table 4
Lexical Coverage of GSL and the MSVL for Science Across DSC

List	Token	Token%	Total Token%	Type	Type%	Total Type%	Group	Group%	Total Group%
1st GSL	1,982,559	72.5	72.5	3,652	9.17	9.17	998	2.87	2.87
2nd GSL	183,624	6.71	79.21	2,955	7.42	16.59	966	2.78	5.65
MSVL SC	234,103	8.56	87.77	852	2.14	18.73	434	1.25	6.9
Not in List	334,309	12.23	100	32,363	81.27	100	32,363	93.1	100
Total:	2,734,595			39,822		34,761			

These results are comparable with Greene and Coxhead's (2015) results. Their study showed that the 1st and 2nd GSL provided 79.09% coverage across the MS-CAT and the MSVL for Science granted 10.17% coverage across the Science Sub-corpora of MS-CAT for a total of 89.26%. This evidence can constitute the basis for arguing that the GSL and MSVL for Science can be comparable in usefulness when using paper-based texts as well as digital texts.

Discussion and Implications

This study examined the coverage of five premade corpus-based lists across a specialized digital corpus. Four of these targeted lists were designed for L2 learners (e.g., GSL, AWL, AVL, EAP Science List). However, the fifth one, namely, MSVL for Science, although not explicitly designed for L2 learners used a similar methodology as for the AWL. This study showed that if secondary-level L2 learners acquire lexical items in the scrutinized lists, they would increase their vocabulary size and would be able to comprehend a substantial portion of the science texts they read (Ardasheva, 2017; Coxhead & Boutorwick, 2018; Coxhead et al., 2010; Greene & Coxhead, 2015).

Previous research studies explored the lexical coverage of premade word lists in science contexts (Arndt, 2022; Coxhead et al., 2010; Coxhead & Boutorwick, 2018; Greene & Coxhead, 2015; Hirsch & Coxhead, 2007). This study contributes to the existing body of knowledge as the findings indicate that the GSL, AWL, and EAP Science List together offer considerable lexical coverage (i.e., 88.33%), suggesting that purposefully teaching 2,841 word families from these three lists would help L2 learners approach the 90% threshold. Laufer (2020) suggests that reaching minimal reading comprehension (95%) can be achieved with existing knowledge of 90% of the words in the text and inferring 5% of the words. However, the results of this study do not indicate which word families within these three lists should be taught. This study solely

provides instructional guidance by suggesting that knowledge of the words in the GSL, AWL, and the EAP Science List is beneficial to achieving a lexical coverage of almost 90% across Digital Science Resources (DSR) for middle school (MS) learners.

This study also highlights the pedagogical value of purposefully teaching the GSL and MSVL for Science which together provide about 88% coverage. Notably, learning the lexical items in these lists involves a lower learning burden since the size of the MSVL for Science is much lower than the size of the AWL and EAP for Science Lists together. Although teaching and learning all the words in these two lists is not sufficient to reach the 95% threshold, there is evidence that teaching and learning the words in these lists is a robust starting point to support minimal text comprehension of L2 learners.

Moreover, this study is helpful for learners and practitioners seeking information about the academic list with the highest coverage across digital science texts. This study points out that the first 570 AVL word families offer much higher coverage (about 75% higher) than the AWL across the inspected corpus. However, it is worth noting that there is an overlap between AVL word families and the GSL which can potentially explain the high coverage provided by the first 570 word families of the AVL.

While teaching all word families in these lists may seem strenuous and seemingly out of the scope of a science teacher, it is sensible to consider conducting small-scale Lexical Frequency Profiling (LFP) on specific science texts used in the classroom. Such analysis enables the extraction of general, academic, and content-specific words. These words could be taught, reviewed, and learned in science contexts. The platform Compleat Web Classic v4 (Cobb, 2021) provides the opportunity for text analysis against the GSL, AWL, and MSVL. Boulton and Cobb's (2017) meta-analysis highlighted the importance of actively engaging with corpus-based lists, found under the data-driven learning (DDL) umbrella (Boulton & Cobb, 2017). Research suggests that the studies involving the use of students' active use of technology showed considerable effect sizes (Boulton & Cobb, 2017), therefore, engaging students in using corpus-based tools for vocabulary learning could be considered.

Conclusion

Since digital science texts are used in middle school classrooms more often than science textbooks (Gray & Lewis, 2021), it is essential to explore the composition of these materials in terms of general, academic, and content-specific vocabulary in relation to minimal and optimal reading comprehension. Previous studies have contributed important insights by examining corpora and assembling robust corpus-based lists designed for vocabulary teaching and learning purposes.

The findings of this research show that five general, academic, and science-specific corpus-based lists, among which four were assembled with the needs of L2 learners in mind, provide substantial lexical coverage of digital science texts for secondary students in the United States. Practically speaking, teaching, and learning the word families in only three general, academic, and science-specific corpus lists (GSL, AWL, and EAP for Science List) with fewer than 2,900-word families provide almost 89% lexical coverage of digital science texts. This research also points out that employing DDL via corpus-based lists is a useful approach to embed in science and language instruction as it can help learners approach 90% coverage. If learners are to achieve

90% coverage, Laufer (2020) shows that an additional 5% coverage can be achieved inferring, meaning that learners could, potentially, reach 95% minimal reading comprehension.

Data Availability Statement

The data supporting this study's findings are available from the corresponding author upon request

References

Anthony, L. (2021). AntWordProfiler (Version 1.5.1) [Computer Software]. Tokyo, Japan: Waseda University. Retrieved from https://www.laurenceanthony.net/software

Ardasheva, Y. (2017). Scaffolding science vocabulary for middle school newcomer ELLs. In de Oliveira L., Campbell Wilcox K. (Eds.), *Teaching science to English language learners* (pp. 219–241). Palgrave Macmillan: Cham. https://doi.org/10.1007/978-3-319-53594-4_10

Ardasheva, Y., Newcomer, S. N., Firestone, J. B., & Lamb, R. L. (2017). Mediation in the relationship among EL status, vocabulary, and science reading comprehension. *The Journal of Educational Research*, *110*(6), 665–674. https://doi.org/10.1080/00220671.2016.1175407

Ardasheva, Y., Crosson, A., Carbonneau, K. & French. B. (2021). Unpacking contributions of morphosyntactic awareness and vocabulary to science reading comprehension among linguistically diverse students. *TESOL Quarterly*, *55*(3), 931–958. https://doi.org/10.1002/tesq.3039

Arndt, R. (2022). Vocabulary in digital science resources for middle school learners. *Applied Corpus Linguistics*, *2*(3). https://doi.org/10.1016/j.acorp.2022.100023

Barcroft, J. (2009). Effects of synonym generation on incidental and intentional L2 vocabulary learning during reading. *TESOL Quarterly, 43*, 79–103. https://doi.org/10.1002/j.1545-7249.2009.tb00228.x

Beck, I., McKeown, M. & Kucan, L. (2013). *Bringing words to life: Robust vocabulary instruction (second edition)*. New York: Guildford Press.

Boulton, A. & Cobb, T. (2017). Corpus use in language learning: A meta-analysis. *Language Learning*, *65*(2),1–46. https://doi.org/10.1111/lang.12224

Cobb, T. (2021). Compleat Web VP Classic (Version 4) [Computer software]. Retrieved from https://www.lextutor.ca/vp/eng/

Coxhead, A. (2018). *Vocabulary and English for specific purposes research: Quantitative and qualitative perspectives*. London and New York: Routledge.

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

Coxhead, A., & 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), 588–610. https://doi.org/10.1002/tesq.450

Coxhead, A., & Hirsh, D. (2007). A pilot science-specific word list. *Revue Française de Linguistique Appliqueé, 12*(2), 65–78. https://www.cairn-int.info/journal--2007-2-page-65.htm.

Coxhead, A., Stevens, L., & Tinkle, J. (2010). Why might secondary science textbooks be difficult to read? *New Zealand Studies in Applied Linguistics, 16*(2), 35–52.

Cummins, J. (1979). Cognitive/academic language proficiency, linguistic interdependence, the optimum age question and some other matters. *Working Papers on Bilingualism, 19*, 121–129.

Cummins, J. (2000). Academic language learning, transformative pedagogy, and information technology: Towards a critical balance. *TESOL Quarterly, 34*(3), 537–47. https://doi.org/10.2307/3587742

Daller, H., Milton, J., & Treffers-Daller, J. (2007). *Modelling and assessing vocabulary knowledge*. Cambridge, UK: Cambridge University Press.

Dang, T. N. Y. (2020). Corpus-based word lists in second language vocabulary research, learning, and teaching. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 288–303). New York, NY: Routledge.

Davies, M. (2008). Corpus of Contemporary American English (1990–2019) available at https://www.english-corpora.org/coca/

Folse, K. (2020). Academic word lists (E-single): *What every teacher needs to know Kindle Edition*. The University Michigan Press [Kindle version].

Heatley, A., & Nation, P. (1996). Range [Computer software]. Wellington, New Zealand: Victoria University of Wellington. Retrieved from http://www.vuw.ac.nz/lals

Hernandez, M. M. (2017). *Comparing the AWL and AVL in textbooks from an intensive English program*. (Master's thesis). Retrieved from All Theses and Dissertations, Brigham Young University, BYU Scholars Archive. No 6458.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. *Applied Linguistics, 35*, 305–327. https://doi.org/10.1093/applin/amt015

Gilner, L. (2011). A primer on the General Service List. *Reading in a Foreign Language, 23*(1), 65–83. http://hdl.handle.net/10125/66658

Grabe, W. & Stoller, F. (2020). *Teaching and researching reading (3rd ed.)*. New York and London: Routledge.

Gray, L., & Lewis, L. (2021). *Use of Educational Technology for Instruction in Public Schools: 2019–20* (NCES 2021-017). U.S. Department of Education. Washington, DC: National Center for Education Statistics. Retrieved from https://nces.ed.gov/pubsearch/pubsinfo.asp?pubid=2021017

Green, C. (2022). Extensive Reading for a 9,000-Word Vocabulary: Evidence from Corpus Modeling. *Reading in a Foreign Language*, *34*(2), 232–248. http://hdl.handle.net/10125/67424

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

Laufer, B., & Ravenhorst–Kalovski, G. C. (2010). Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language*, *22*, 15−30.

Laufer, B. (2020). Lexical coverages, inferencing unknown words and reading comprehension: How are they related? *TESOL Quarterly*, *54*(4), 1076–1085. https://doi.org/10.1002/tesq.3004

Laufer, B. (2021). Lexical thresholds and alleged threats to validity: A storm in a teacup? *Reading in a Foreign Language*, *33*(2), 238–246. http://hdl.handle.net/10125/67402

Li, Y., & Qian, D. D. (2010). Profiling the Academic Word List (AWL) in a financial corpus. *System*, *38*(3), 402–411. https://doi.org/10.1016/j.system.2010.06.015

Liu, D. & Lei, L. (2020). Technical vocabulary. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 111–123). New York, NY: Routledge.

Miralpeix, I. (2020). L1 and L2 Vocabulary size and growth. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 189–202). New York, NY: Routledge.

McQuillan, J. (2016). Time, texts, and teaching in vocabulary acquisition: A rebuttal to Cobb (2016). *Reading in a Foreign Language*, *28*(2), 307–318. https://doi.org/10125/66905

Nation, I. S. P. (2004). A study of the most frequent word families in the British National Corpus. In P. Bogaards & B. Laufer (Eds.), *Vocabulary in a second language: Selection, acquisition, and testing* (pp. 3–13). Amsterdam, Netherlands: John Benjamins.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening*? Canadian Modern Language Review*, *63*, 59–82.

Nation, I. S. P. (2012). The BNC/COCA word family lists. Retrieved from http:// www.victoria.ac.nz/lals/about/staff/paul-nation

Nation, I. S. P. (2013). *Learning vocabulary in another language (2nd ed.)*. Cambridge: Cambridge University Press.

Nation, P. (2014). How much input do you need to learn the most frequent 9,000 words? *Reading in a Foreign Language*, *26*(2), 1–16. http://hdl.handle.net/10125/66881

Nation, P. & Sorell, J. (2016). Corpus selection and design. In I.P.S. Nation (Ed), *Making and using word lists for language learning and testing*. Amsterdam: John Benjamins.

Newman, J. A. (2016). *A corpus-based comparison of the academic word list and the academic vocabulary list* (Unpublished master's thesis). Brigham Young University, 6080.

N.G.S.S. Lead States (2013). N*ext generation science standards: For states, by states*. The National Academies Press, Washington, DC (vol. 2, Appendices).

Paribakht, T. S., & Webb, S. (2016). The relationship between academic vocabulary coverage and scores on a standardized English proficiency test. *Journal of English for Academic Purposes*, *21*, 121–132. https://doi.org/10.1016/j.jeap.2015.05.009

Van Zeeland, H., & Schmitt, N. (2013). Incidental vocabulary acquisition through L2 listening: A dimensions approach. *System*, *41*, 609–624. https://doi.org/10.1016/j.system.2013.07.012

Verhoeven, L., van Leeuwe, J., & Vermeer, A. (2011). Vocabulary growth and reading development across the elementary school years. *Scientific Studies of Reading*, *15*(1), 8–25. https://doi.org/10.1080/10888438.2011.536125

Webb, S. (2020). Incidental vocabulary learning. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 225–240). New York, NY.

Webb, S., & Nation, I. S. P. (2017). *How vocabulary is learned*. Oxford: Oxford University Press.

West, M. (1953). *A general service list of English words*. Longman: Longman, Green. Yilin Education.

About the Author

Rebeca Arndt received her Ph.D. in Education, TESOL track, from the University of Central Florida. She is a former schoolteacher currently working as a district administrator. Her research interests include corpus analysis, vocabulary, reading comprehension, and literacy. E-mail: re730121@ucf.edu