A Hard Science Spoken Word List

A Hard Science Spoken Word List (HSWL) was developed and validated to help second language learners of hard sciences better comprehend academic speech at English-medium universities. It consists of the 1,595 most frequent and wide ranging word families in a 6.5-million running word hard science spoken corpus which represents 12 subjects across two equally-sized sub-corpora. Its coverage in different discourse types indicates that the HSWL truly reflects the language in hard science academic speech. The comparison between the HSWL with Dang, Coxhead, and Webb's (2017) Academic Spoken Word List shows that the HSWL focuses more on specialized vocabulary in hard science speech. Depending on their vocabulary levels, learners may achieve 93%-96% coverage of hard science academic speech with knowledge of the HSWL words.

Key words: hard sciences, corpus, academic spoken discourse, vocabulary, word lists

1. Introduction

To achieve academic success at English-medium universities, second language (L2) learners planning to study hard sciences (e.g., Mathematics, Biology, Engineering, Medicine) need to comprehend not only their reading materials but also lectures, seminars, labs, and tutorials (Becker, 2016; Biber, 2006). A good vocabulary knowledge enhances listening comprehension (Matthews & Cheng, 2015; van Zeeland & Schmitt, 2013). Therefore, mastering the most important words in academic speech from their disciplines is crucial for these learners. To meet this need, Dang, Coxhead, and Webb (2017) developed an Academic Spoken Word List (ASWL) for students from different disciplines who study

in the same English for Academic Purposes (EAP) programs. This study expands on Dang et al.'s (2017) study by developing a Hard Science Spoken Word List (HSWL) for EAP programs where all learners plan to study hard science subjects. Together with the ASWL, the HSWL should provide more choices for hard science students in different EAP programs.

1.1.Why do we need a specialized spoken wordlist for hard science students?

Research on the variation in the subject matter characteristics, structure, output, teaching styles (Biglan, 1973a, 1973b), and learning styles (Kolb, 1981) in a wide range of university academic areas has suggested that academic disciplines can be divided into groups based on three dimensions: hard/soft, pure/applied, and life/non-life. The hard/soft dimension is concerned with the existence of a paradigm, the pure/applied dimension is related to application, and the life/non-life dimension is concerned with life system. Of the three dimensions, the hard/soft division is the strongest. Hard sciences have greater consensus about content and methods than soft sciences (Biglan, 1973a, 1973b). According to Neumann (2001) and Neumann, Parry, and Becher (2002), the content of hard sciences is fixed, and the teaching in these disciplines has a greater emphasis on helping students to acquire and apply accepted scientific facts, principles, and concepts. In contrast, soft sciences place greater importance on building critical thinking skills and individual interpretations of the world of human experience. Hence, the content of soft subjects is more free-ranging with the teaching and learning activities being constructive and interpretative.

Studies investigating how many words learners need to know to comprehend a certain discourse type also reveal the distinction between hard and soft sciences. Drawing on the close relationship between comprehension and lexical coverage (Laufer, 1989; Laufer & Ravenhorst-Kalovski, 2010; Schmitt, Jiang, & Grabe, 2011; van Zeeland & Schmitt, 2013), these studies examined the vocabulary sizes needed to reach 95% and 98% coverage of different kinds of hard and soft science texts. Lexical coverage is the percentage of known words in a text (Nation & Waring, 1997); 95% and 98% are widely used as the coverage figures to indicate high and stable degrees of listening and reading comprehension (Hu & Nation, 2000; Laufer, 1989; van Zeeland & Schmitt, 2013).

In terms of written discourse, a vocabulary size of 5,000 word families is needed to reach 95% coverage of textbooks in Engineering, and a vocabulary size of 10,000 word families is necessary to achieve 98% coverage (Hsu, 2014). These vocabulary sizes are larger than those needed to reach 95% coverage (3,500 word families) and 98% coverage (5,000 word families) of textbooks in Business (Hsu, 2011). A similar trend is seen in academic spoken English. Dang and Webb (2014) analyzed academic speech from two hard disciplinary groups (Physical Sciences and Life and Medical Sciences) and two soft disciplinary groups (Arts and Humanities and Social Sciences) of the British Academic Spoken English Corpus (BASE). They found that, to reach 95% and 98% coverage of the academic speech from the hard science disciplines, learners need a vocabulary size of 4,000-5,000 word families and 10,000-13,000 word families, respectively. These vocabulary sizes are larger than those needed in the case of soft sciences: 3,000-4,000 word families (95%), and 5,000-7,000 word families (98%). Together, these findings indicate

that, written and spoken texts of hard sciences are more challenging than those of soft sciences in terms of lexical coverage. This then highlights the importance of developing wordlists to support the reading and listening comprehension of hard science students.

In recognition of this need, several specialized wordlists for hard science students have been developed. The majority of them were derived from written text (e.g. Coxhead, 2000; Coxhead & Hirsh, 2007; Gardner & Davies, 2014; Wang, Liang, & Ge, 2008; Ward, 1999, 2009; Watson-Todd, 2017). In contrast, only three studies have attempted to develop a wordlist that is representative of spoken English. All of them are universal specialized wordlists for EAP programs which are made up of both hard and soft science students.

Simpson-Vlach and Ellis (2010) focused on multi-word units by developing a spoken Academic Formulas List. An academic spoken corpus and a non-academic spoken corpus were compiled in that study. The academic spoken corpus had a total size of 2.1-million running words, and was divided into five sub-corpora: Humanities and Arts, Social Sciences, Physical Sciences, and Non-departmental/other. The non-academic spoken corpus consisted of 2.9 million running words. To be included, an academic formula had to be outside the formulas that occurred frequently in both the academic and non-academic spoken corpora. Moreover, it had to occur at least 10 times per million in four out of five sub-corpora. There were 979 formulas satisfying these criteria. A list of multi-word units is beneficial because knowledge of multi-words is significant for fluent processing (Nation & Webb, 2011; Simpson-Vlach & Ellis, 2010). However, knowledge of single words also provides valuable support for the acquisition of multi-words. Hence, there is value in developing lists of single words.

Nesi (2002) investigated single words. Her Spoken Academic Word List (SAWL) was created from the BASE corpus which consists of 1.6-million running words. The SAWL contains items that do not appear in Nation's most frequent 2,000 word families, but have high frequency and wide range in the BASE corpus. Unfortunately, to date, no precise information about the list has been reported, and the list is not available to access. Considering these facts, Dang et al. (2017) further developed a list of single words. Their ASWL was created from a 13-million running word corpus or academic spoken English. The corpus had four sub-corpora which represented academic speech from four disciplinary groups: hard-pure (e.g., Mathematics, Physics), hard-applied (e.g., Engineering, Medicine), soft-pure (e.g., Arts, History), and soft-applied (e.g., Law, Business). Each disciplinary sub-corpus was made up of materials from six subject areas. The disciplinary divisions followed Becher's (1989) classification of academic disciplines in higher education which was based on the findings of Biglan (1973a, 1973b) and Kolb (1981). Unlike Nesi (2002), Dang et al. (2017) did not remove general high-frequency words (i.e. the most frequent 2,000 words of general vocabulary such as know, therefore, determine, and approach) from their lists if these words fulfilled the three following selection criteria. First, the ASWL word families had to occur in all four sub-corpora and in at least 50% of the subject areas. Second, they had to have a frequency of at least 26.9 times per millions in the corpus. Third, they had to have a Juilland and Chang-Rodrigues's (1964) dispersion D of at least 0.6. As a result, 1,741 word families met these criteria and were included in the ASWL. The list provided 90% coverage of the corpus from which it was developed and around the same amount of coverage in each disciplinary sub-corpus. When tested against an independent academic spoken corpus of a similar size and structure, the list provided about 90% coverage. The

consistent coverage of the ASWL indicates that it is a useful list for both hard and soft science students. With knowledge of proper nouns and marginal words, these learners can achieve from 92% to 96% coverage of academic spoken English depending on their vocabulary levels.

It is important to note that while these universal specialized wordlists are valuable resources for hard science students studying in the same EAP programs with soft science students, there are programs with all learners planning to study hard sciences (Coxhead & Hirsh, 2007; Valipouri & Nassaji, 2013; Ward, 1999, 2009; Watson-Todd, 2017)). In such programs, discipline-specific wordlists that are specifically developed for hard science students may be more useful. These lists will focus these students more on specialized words in their field, especially items that have high frequency and wide range in the speech of hard sciences but are absent from universal academic wordlists due to their low frequency and narrow range in the speech of soft sciences.

1.2.How are existing wordlists for hard science students *adaptable to learners'* proficiency?

Different approaches have been taken to identify specialized vocabulary for hard science students. The most common approach is to assume that learners already know a certain number of words and look for items outside these words that have high frequency and wide range in the specialized corpora. Therefore, some specialized wordlists did not include general high-frequency vocabulary (i.e., the most frequent 2,000 or even 3,000 words of general English) (Browne, Culligan, & Phillips, n.d.; Coxhead, 2000; Hsu, 2013, 2014; Nesi, 2002; Valipouri & Nassaji, 2013; Wang et al., 2008). The other (Coxhead & Hirsh,

2007) even excluded general academic vocabulary (i.e. the shared vocabulary across multiple subject areas and disciplines such as minimize, ambiguous, paradigm) apart from general high-frequency vocabulary. Another approach is not to assume learners already know any words and develop specialized wordlists from scratch (Gardner & Davies, 2014; Lei & Liu, 2016; Ward, 1999, 2009).

According to Dang et al. (2017), each approach has its own strength. The first approach allows teachers and learners to avoid repeatedly teaching and learning known items. In contrast, the second approach enables the specialized wordlists to avoid the limitations related to the general high-frequency wordlists and general academic wordlists that they were based on. However, Dang et al. (2017) also point out that these methods share the same limitation; that is, they assume that all learners have the same vocabulary level when using their lists. To address this limitation, Dang et al. (2017) developed their ASWL from scratch but graded the list into four levels according to Nation's (2012) BNC/COCA frequency levels. Levels 1, 2, and 3 represent the ASWL items which appear at the 1st, 2nd, and 3rd 1,000 BNC/COCA frequency levels, respectively. Level 4 is made up of ASWL words outside the most frequent 3,000 BNC/COCA word families. Depending on their current vocabulary levels, learners can skip certain levels of the ASWL. This approach is innovative. It makes the best use of the strengths of the two approaches towards developing specialized wordlists, and results in a list which is more adaptable to learners' proficiency. It also enables teachers to incorporate the ASWL with Nation's (2012) BNC/COCA lists in organizing a systematic vocabulary program for L2 learners as the learning sequence presented in Figure 1. According to this sequence, depending on their learners' vocabulary

levels and learning purposes, teachers can identify the relevant levels of the ASWL and BNC/COCA lists to focus on. For these reasons, the development of the HSWL in this study followed Dang et al.'s (2017) approach.

[FIGURE 1 NEAR HERE]

1.3. Research questions

1. Which lexical items occur frequently and are evenly distributed in a wide range of academic speech in hard science subjects?
2. What is the coverage of these items in independent collections of academic speech of hard science, soft science, academic writing, and non-academic speech?
3. How do these items compare with those from Dang et al.'s (2017) ASWL?
4. With knowledge of these words, how much coverage of academic speech in hard sciences may be reached by learners with different vocabulary levels?

2. Methodology

2.1. Developing the corpora

Five corpora were developed in the present study (Table 1). The first hard science spoken corpus was used to develop the HSWL while the other four corpora were used to validate the list from different perspectives. This is a common approach to validate specialized wordlists (Coxhead, 2000; Coxhead & Hirsh, 2007; Gardner & Davies, 2014). Each corpus has around 6.5-million running words, which satisfies Nation and Webb's (2011) guideline that a validating corpus should have a similar size as the corpus from which the list is

developed so that it can provide an accurate assessment about the occurrences of items in the list in the target discourse.

[TABLE 1 NEAR HERE]

Tables 2 and 3 present the composition of the two hard science spoken corpora while Table 4 shows the structure of the soft science spoken corpus. These corpora have a similar size and structure. Each corpus is divided into two sub-corpora: pure and applied, each of which consists of around 3.2-million running words. These corpora were made up of naturally occurring academic speech from a wide range of academic subjects recorded in various universities in different parts of the world (the U.S, the U.K, Hong Kong, New Zealand). They represent four kinds of speech events (lectures, seminars, labs, and tutorials) and at least seven varieties of English (American-English, Australian-English, British-English, Canadian-English, Hong Kong-English, Irish-English, and New Zealand-English). The sources of these corpora are presented in Appendix A.

[TABLES 2, 3, AND 4 NEAR HERE]

Table 5 demonstrates the components of the hard science written corpus. This corpus includes different kinds of hard science written texts (book chapters, journal articles, student writings, research reports, and textbooks) from the Massachusetts Institute of Technology open courseware, the British Academic Written English corpus (BAWE), and the Corpus of Contemporary American English. This academic written corpus has a similar structure as the two academic spoken corpora. It contains around 6.5-million running words and is divided into two sub-corpora, each of which has more than 3-million running words.

Table 6 presents the components of the non-academic spoken corpus. This corpus represents different kinds of general spoken English (e.g., TV programs, movies, telephone conversation) and 10 varieties of English. It also has a total size of around 6.5-million running words.

[TABLES 5 AND 6 NEAR HERE]

2.2. Determining the unit of counting for the HSWL

A great effort has been made to argue whether lemmas or word families should be the suitable unit of counting for specialized wordlists (Gardner, 2007; Gardner & Davies, 2014; Nation, 2013). Nation (2016), however, points out that the lemma, in fact, is a level in Bauer and Nation's (1993) scale of word families. In this scale, word families can be classified into seven levels according to the frequency, productivity, predictability, and regularity of the affixes. Word families at Level 1 consist of the most elementary and transparent members while those at Level 7 consist of the least transparent members. A lemma is relevant to a Level 2 word family; that is, it is made up of the stem itself (e.g. evaluate) together with its inflections (e.g., evaluated, evaluating, evaluates). A Level 6 word family includes the stem (e.g. evaluate), its inflections (e.g., evaluated, evaluating, evaluates), and closely related derivations with affixes up to Level 6 (e.g., evaluation, evaluations, evaluative, evaluator, evaluators). Therefore, according to Nation (2016), the question is not whether lemmas or word families are the best unit of counting, but which word family level is the most suitable for a particular group of learners.

Word families up to Level 6 were chosen as the unit of counting for the HSWL for two reasons. First, this level is the most common unit of counting in specialized wordlists

(Coxhead & Hirsh, 2007; Hsu, 2013, 2014; Valipouri & Nassaji, 2013; Wang et al., 2008). Second, following earlier studies (e.g., Coxhead, 2000; Nation, 2013), this study does not consider knowledge of word families as something that can be acquired all at the same time, but is gradually picked up during the learning process. Knowledge of a known word form (e.g. happy) may provide support for the acquisition of other word forms from the same word family (e.g., unhappy, happily). This assumption is supported by studies which reported an incremental increase in L2 learners' derivational knowledge over time (Mochizuki & Aizawa, 2000; Schmitt & Zimmerman, 2002). In other words, a Level 6 word family should be considered as a guide rather than a handbook for teachers and learners to strictly follow. However, given that learners' morphological knowledge increases incrementally, following Dang et al. (2017), another version of the HSWL was also developed. This version listed the HSWL lemmas within each Level 6 HSWL word family. Presenting the HSWL in different formats allows the list to better suit learners with different proficiency levels.

2.3. Developing and validating the HSWL

To be selected, an HSWL word family had to satisfy the range, frequency, and dispersion criteria. These are common criteria in the construction of corpus-based wordlists so that the lists can capture the words that occur frequently and distribute evenly in a wide range of target texts (Nation, 2016; Nation & Webb, 2011). The range and frequency criteria were based on Dang et al.'s (2017) criteria when developing their ASWL. With respect to dispersion, unlike Dang et al. (2017), this study used Gries's (2008) DP rather than Juilland and Chang-Rodrigues's (1964) D because DP seems to be better at distinguishing well-

dispersed and not well-dispersed items in a corpus with a large number of sub-sections (Biber, Reppen, Schnur, & Ghanem, 2016). The detailed selection criteria are as follows:

(1) Range: a selected word family had to occur in both sub-corpora of the first hard science spoken corpus, and in at least 50% of the subjects in this corpus (six out of 12 subjects). This criterion ensures that the HSWL benefits learners from different hard science subjects.

(2) Frequency: a selected word family had to occur at least 26.9 times per million running words in the first hard science spoken corpus. As the first hard science spoken corpus has 6.5-million running words, this means that a selected word family had to have a frequency of at least 175 times in the whole corpus. This criterion makes sure that the HSWL includes items that hard science students are likely to encounter often in academic speech.

(3) Dispersion: a selected word family had to have Gries's (2008) DP below 0.6. The DP value indicates how evenly a word family distributes across the corpus. It can range from 0 (perfectly even distribution) to 1 (extremely uneven distribution). The DP cut-off point of 0.6 is the result of extensive experimentation which compared the items included or excluded from the HSWL when different DP cut-off points (from 0.1 to 0.9) were chosen. Unlike lower DP cut-off points (0.1-0.5), 0.6 resulted in a list which provided higher coverage in the two hard academic spoken corpora than Dang et al.'s (2017) ASWL. Unlike higher DP cut-off points, 0.6 resulted in a list with a smaller number of items than the ASWL. This study aims to draw hard science students' attention to the most important words in their specific areas and provide a shortcut to

reduce the amount of learning for these students. Choosing the DP cut-off point of 0.6 means that the HSWL has a smaller size but still provides higher coverage in hard science spoken English than the ASWL.

Items that satisfied these three criteria were included in the HSWL. Following Dang et al. (2017), the HSWL words were then divided into levels according to Nation's (2012) BNC/COCA lists so that the list is suitable for learners with different vocabulary levels. Levels 1, 2, and 3 represent HSWL words from the 1st, 2nd and 3rd 1,000 BNC/COCA frequency levels, respectively. Level 4 are HSWL words that are outside the most frequent 3,000 BNC/COCA word families. Given the difference in the nature of function words (e.g., through, unless) and lexical words (e.g., equation, fibre) (Dang & Webb, 2016), like the ASWL, each level of the HSWL is broken down into one list of function words and sub-lists of lexical words. Each sub-list of lexical words consists of around 50 items.

The coverage of the HSWL and its levels was examined against the first hard science spoken corpus and the four validating corpora. This was done by running these corpora in turn through Heatley, Nation, and Coxhead's (2002) RANGE with the HSWL and its levels serving as the base wordlists. The RANGE program was downloaded from Paul Nation's website: http://www.victoria.ac.nz/lals/about/staff/paul-nation. Also, the HSWL was compared with the ASWL in two aspects: (1) the coverage provided by the each list in the two hard science spoken corpora, and (2) the overlap between items from the two lists.

2.4. Determining the potential coverage for different groups of learners

The potential coverage that learners may reach with the aid of the HSWL was the combination of the coverage provided by (1) the word families that they already know and

(2) the HSWL word families that they may not know (Figure 2). Items in the first group are represented by the BNC/COCA word families that are at learners' existing vocabulary levels. Items in the second group are HSWL words that are outside the BNC/COCA words in the first group. For example, pre-intermediate learners may have the vocabulary level of the most frequent 1,000 words. Therefore, the potential coverage that they may reach with the support of the HSWL is the sum of the coverage of the 1st 1,000 BNC/COCA word families and the coverage of the HSWL word families from Level 2 to Level 4.

[FIGURE 2 NEAR HERE]

3. Results

In answer to the first research question about the lexical items occurring frequently in a wide range of academic speech in hard science subjects, there are 1,595 word families that met the selection criteria. Although six was fixed as the range cut-off point, 83.01% of the HSWL words appear in all 12 subjects. In fact, 99.50% of the HSWL words appear in at least nine subjects. Similarly, although 0.6 was set as the maximum cut-off point for the dispersion criterion, 88.46% of the HSWL had DP lower than 0.5. The lexical profile of these words in the HSWL and its levels are presented in Table 7. See Appendices B-F for the HSWL headwords in each sub-list within each level.

[TABLE 7 NEAR HERE]

It can be seen that 449 HSWL words are outside general high-frequency words. These words account for 28.15% of the words in the HSWL. This proportion is larger than the proportion of words outside general high-frequency words in Dang et al.'s (2017) ASWL

(26.13%). The HSWL covers 90.94% of the whole corpus. This coverage is higher than the coverage provided by the most frequent 2,000 BNC/COCA word families (88.06%) although the HSWL has 405 fewer word families. Also, the HSWL consistently provides around the same amount of coverage in the hard-pure (90.12%) and hard-applied (91.80%) sub-corpora.

In answer to the second research question about the coverage of the HSWL in independent validating corpora, the HSWL covers 90.82% of the words in the second hard science spoken corpus, which is similar to the coverage of the list in the first hard science spoken corpus. In contrast, its coverage in the three other validating corpora is lower: 88.48% (soft science spoken corpus), 83.81% (non-academic spoken corpus), and 80.12% (hard science written corpus).

The third research question is about the comparison between the HSWL and Dang et al.'s (2017) ASWL. The ASWL covers 90.24% of the first hard science spoken corpus and 89.84% of the second hard science spoken corpus. These coverage figures are lower than the coverage provided by the HSWL in these two corpora (90.94% and 90.82%). It should be noted that the HSWL has 146 fewer word families than the ASWL. In terms of overlap, 1,438 HSWL word families (90.16%) occur in the ASWL while 157 word families (9.84%) are unique to the HSWL. Noticeably, 76.91% of the shared items are among the most frequent 2,000 BNC/COCA word families (e.g., investigate, technology, research) whereas 74.52% of the items unique to the HSWL are outside the most frequent 2,000 BNC/COCA word families (e.g., cell, vector, molecule). Moreover, the examination of the 157 word families unique to HSWL showed that a number of them have high frequency, wide range,

and even distribution in the hard science subjects but have either low frequency or narrow range in soft science speech. For example, seven HSWL word families (amplitude, quadratic, epsilon, cosine, micron, cubed, theta) appeared in nine subject areas of the first hard science spoken corpus with a frequency from 31.08 to 152.62 times per million running words. These word families, however, appeared in no more than three subject areas and had a frequency of no more than 2.46 times per million running words in this corpus. All of these seven word families are at low BNC/COCA frequency levels: 8th 1,000 (amplitude), 10th 1,000 (micron, epsilon), 11th 1,000 (cubed), 12th 1,000 (theta), 13th 1000 (quadratic), and 16th 1,000 (cosine).

In answer to the fourth research question about the amount of coverage of academic speech in hard sciences which may be reached by learners with different vocabulary levels, Table 8 shows the potential coverage that learners of different vocabulary levels may reach with the support of the HSWL. The number of HSWL words that are beyond learners' existing vocabulary level is presented in the third column of the table. The coverage that learners may gain if they study the HSWL is presented in the next two columns. Coverage provided by proper nouns (e.g., James, Helen) and marginal words (e.g., oh, hm) is shown in the last two rows of the table. Earlier research on the vocabulary load of spoken English (Dang & Webb, 2014; Nation, 2006; Webb & Paribakht, 2015; Webb & Rodgers, 2009a, 2009b) added the coverage by proper nouns and marginal words to the potential coverage because they assumed that these words have minimal learning burden for learners.

The potential coverage achieved by beginner learners (i.e. those having not mastered the most frequent 1,000 words) is demonstrated in the first row of the table. Their insufficient

vocabulary knowledge means that these learners are unlikely to know the HSWL words. Learning all 1,595 word families from the HSWL may allow them to reach around 91% of the words in the two hard science academic spoken corpora. If they know proper nouns and marginal words, these learners may gain potential coverage of around 93%. This coverage is much larger than the coverage provided by the most frequent 2,000 BNC/COCA2000 word families plus proper nouns and marginal words (90.09%, 89.44%).

The second row of the table presents the potential coverage for pre-intermediate learners (i.e. those with the vocabulary level of the most frequent 1,000 BNC/COCA word families). Their existing knowledge means that these learners only need to study 833 HSWL words which are beyond their level. These word families, however, may enable them to achieve coverage of more than 91% (without proper nouns and marginal words) and about 94% (with proper nouns and marginal words). This potential coverage figure is higher than the potential coverage provided by the most frequent 2,000 BNC/COCA word families. It should be noted that learning the HSWL also allows the beginner and pre-intermediate learners to achieve reasonable coverage of the non-academic spoken corpus: 89.15% (beginner learners) and 90.98% (pre-intermediate learners).

[TABLE 8 NEAR HERE]

The next two rows of the table demonstrate the number of HSWL words that the intermediate learners (i.e. those having mastered the most frequent 2,000 BNC/COCA word families) and advanced learners (i.e. those with the vocabulary level of the most frequent 3,000 BNC/COCA word families) need to learn and the potential coverage that they may reach with the aid of the HSWL. These learners only need to study a small number of

words: 449 (intermediate learners) and 153 (advanced learners). Yet, they may achieve potential coverage of 92%-94% (without proper nouns and marginal words) and 95%-96% (with proper nouns and marginal words). These amounts of coverage are larger than (in the case of intermediate learners) or as large as (in the case of advanced learners) the potential coverage that these learners may gain from learning 1,000 word families from the subsequent BNC/COCA frequency levels (see Table 9).

[TABLE 9 NEAR HERE]

4. Discussion

4.1. The HSWL is a useful resource for hard science students

This study suggests that the HSWL effectively supports the vocabulary development of hard science students for four reasons. First, the list accurately reflects the vocabulary in hard science speech. Its consistent coverage in the two academic spoken corpora (one to develop and one to validate the list) indicates that the HSWL truly represents the most frequent, wide ranging, and evenly distributed words in hard science academic speech. Moreover, the HSWL provides higher coverage in the two hard science spoken corpora than in the soft science spoken corpus, the non-academic spoken corpus, and the hard science written corpus. This indicates that the list better represents vocabulary in hard sciences rather than soft science, academic rather than non-academic, and spoken rather than written English. This finding is in line with the findings of previous studies validating specialized wordlists (e.g., Coxhead, 2000; Coxhead & Hirsh, 2007; Gardner & Davies, 2014). These studies found the coverage of their lists in the corpora from which they were created was similar to their coverage in corpora of a similar genre but higher than in

corpora of different genres. Together, these findings suggest that the HSWL truly captures the items that hard science students are likely to encounter often in a wide range of academic speech.

Second, the HSWL benefits learners from a wide range of hard science subjects. The list was derived from academic speech in 12 hard science subjects, but it still provides similar coverage in the two sub-corpora (hard-pure and hard applied). This indicates that the HSWL can offer fairly equal benefits for students planning to study hard science subjects. Moreover, the division of the subject areas in the sub-corpora of the two hard science spoken corpora was based on Becher's (1989) classification of academic subjects in higher education. This classification has been validated in numerous contexts (Biglan, 1973a, 1973b; Kolb, 1981), and has been widely used to classify academic disciplines in higher education (Jones, 2011) as well as organizing academic corpora such as the BASE and BAWE. Considering the high validity and wide transference of Becher's (1989) classification, the HSWL is expected to be useful for hard science students irrespective of their specific academic areas and the administrative structure of their institutions.

Third, the HSWL is more specialized than Dang et al.'s (2017) ASWL. It has 146 fewer word families but provides higher coverage in the two hard science spoken corpora (about 91%) than the ASWL (around 90%). Moreover, the HSWL has a higher proportion of words outside general high-frequency words (28.15%) than the ASWL (26.13%). The HSWL words outside general high-frequency words are from a wider range of frequency levels (3^{rd}- 16^{th} 1,000 word levels) than those from the ASWL (3^{rd}-10^{th} 1,000 word levels).

Additionally, most items unique to the HSWL are outside general high-frequency words, and many of them do not have high frequency or wide range in soft science speech.

The fourth reason why the HSWL is a useful list for hard science students is that it benefits these learners regardless of their vocabulary levels. Intermediate and advanced learners only need to study 449 word families and 153 word families from the HSWL, respectively. Yet, they can achieve around 95%-96% coverage of hard science speech, which is larger than the coverage they may gain from learning 1,000 word families at the subsequent BNC/COCA frequency level. This finding is even more meaningful when compared with Dang and Webb's (2014) result. These researchers found that a vocabulary size of 4,000-5,000 word families is needed to reach 95% coverage of hard science speech. This means that the number of BNC/COCA word families beyond their levels that these learners would have to study is 2,000-3,000 word families (mid-level learners) and 1,000-2,000 word families (high-level learners). The HSWL better serves intermediate and advanced learners because these learners have to learn a much smaller number of words and still allows them to reach more than 95% coverage, which should provide a high and stable degree of listening comprehension (van Zeeland & Schmitt, 2013).

Beginner and pre-intermediate learners may study the most frequent 2,000 and even 3,000 BNC/COCA words before moving to the HSWL so that they can reach 95% or more coverage of hard science speech. However, this may be too daunting a goal for a proportion of L2 learners. Research on the vocabulary growth rate of English as a Foreign Language (EFL) learners has shown that these learners can acquire an average of 400 word families per year (Webb & Chang, 2012). This means that it may take beginner learners more than

six years to learn the most frequent 2,000 BNC/COCA word families plus 499 extra HSWL word families, and nearly eight years to learn the most frequent 3,000 BNC/COCA word families and the extra 153 word families. Research on the vocabulary knowledge of learners in various EFL contexts (Henriksen & Danelund, 2015; Matthews & Cheng, 2015; Nguyen & Webb, 2016; Webb & Chang, 2012) suggested that some learners may have even slower vocabulary growth rates. A reasonable proportion of learners in these studies had not mastered the most frequent 2,000 words, and even the most frequent 1,000 words after a long period of formal English instruction.

Therefore, for these beginner and pre-intermediate learners, learning the ASWL words that are beyond their current level may be more reasonable. It may allow them to reach around 91% coverage (without proper nouns and marginal words) and 93%-94% coverage (with proper nouns and marginal words) of hard science speech. These figures mean that beginner and pre-intermediate learners need to study a much smaller number of items but may achieve higher coverage of hard science speech than learning the subsequent 1,000-word levels of general vocabulary. While listening comprehension may not be as easy as at the 95% coverage figure, 91%-94% coverage may enable learners to achieve at least basic comprehension of academic speech. Van Zeeland and Schmitt (2013) reported no significant difference in L2 listening comprehension between the 90% and 95% coverage figures. Moreover, in real life academic speech, students receive support from various sources such as reading materials, visual aids, interaction with course instructors and other students, which may enable them to compensate for their inefficient vocabulary knowledge and enhance their listening comprehension of hard science speech (MacDonald, Badger, &

White, 2000; Mulligan & Kirkpatrick, 2000). Noticeably, if beginner and pre-intermediate learners study the HSWL words that are beyond their current levels, they may reach around 90% coverage of general spoken English, allowing them to achieve basic comprehension of this important discourse type. In sum, the HSWL is an effective shortcut for beginner and pre-intermediate learners to achieve basic comprehension of both hard science speech and general spoken English.

4.2. Wordlists should suit the context

This study provides further insight into the debate over the value of universal academic wordlists versus discipline-specific wordlists. One view suggests that there is a core vocabulary across multiple academic disciplines, and supports the development of universal academic wordlists for L2 learners irrespective of their academic disciplines (e.g., Coxhead, 2000; Gardner & Davies, 2014). A second view questions the existence of a core academic vocabulary from different academic disciplines and argues that frequency, range, meanings, functions, and collocations of a certain word change across disciplines due to the variations in the practice and discourse of disciplines (Hyland & Tse, 2007). Hence, it promotes the idea of developing discipline-specific wordlists.

As mentioned, the HSWL is more specialized than Dang et al.'s (2017) ASWL, which supports the value of discipline-specific wordlists. However, around 90% of the HSWL words appear in the ASWL, and the ASWL provides around 90% coverage of the two hard science spoken corpus. This indicates that, although not as great a tool for hard science students as the HSWL, the ASWL is still an effective tool for these learners. Together the findings of the present study suggest that there is no one-size-fits-all specialized wordlist.

The more specialized a wordlist, the narrower its application, but the greater its benefit in a specific context (Coxhead & Hirsh, 2007). Depending on the particular teaching and learning context, either a discipline-specific list or a general academic wordlist can be a valuable resource for L2 learners. This idea supports Hyland (2016), who points out that the general and specific EAP approaches should be seen as ends of a continuum rather than a dichotomy. In other words, specificity in wordlist construction should be implemented with flexibility and consideration of the circumstances of particular students in a class. It also echoes Nation's (2016) suggestion that wordlists should suit the characteristics of a particular group of learners.

4.3. Model of learning sequence for EAP learners

Expanding on Dang et al.'s (2017) sequence of learning the ASWL, this study proposes a model which assigns together general high-frequency wordlists, universal academic wordlists, and discipline-specific wordlists (see Figure 3). Teachers can set the learning goal and sequence to match the target academic subjects, vocabulary levels, and learning purposes of the learners in a particular language program by following the two steps in this model. The ASWL, HSWL, and potentially, a soft science spoken wordlist (SSWL)[2], and a medical spoken wordlist (MSWL)[3] are used as the illustrations for the model. Although these lists are spoken wordlists, this model can also be applied to written wordlists.

[FIGURE 3 NEAR HERE]

In the first step, teachers can determine the relevant wordlist for a particular group of learners based on their target academic subjects. In an English for General Academic Purposes (EGAP) program which is made up of both hard and soft science students, it is

usually challenging for EAP teachers to address the specific needs of every learner due to the great variation in learners' target subject areas. Hence, universal academic wordlists such as the ASWL are more practical than a discipline specific list.

It is important to note that drawing EGAP learners' attention to the core vocabulary does not mean a lack of focus on the discipline-specific meanings of a word. As suggested by Nation (2013), the core meaning and discipline-specific meanings should not be seen as different from each other. Knowledge of the core meaning provides an excellent scaffolding for the acquisition of discipline-specific meanings (Crossley, Salsbury, & McNamara, 2010). Highly frequent meanings are more likely to be stored as separate entries in the brain while less frequent meanings are more likely to be inferred from the context. Therefore, knowledge of the core meaning of an academic word will help learners to gradually become aware of its discipline-specific meanings if they meet the word very often in texts from their specific disciplines. These multiple encounters of the items from universal academic wordlists in different contexts help to enrich learners' knowledge of the discipline-specific meanings and help storage.

However, in an English for Specific Academic Purposes (ESAP) or English for Specific Purposes (ESP) program where learners have highly specific needs and plan to study the same discipline (e.g., hard discipline) or even the same subject area (e.g., Medicine), discipline-specific lists may better serve learners' needs than general academic wordlists. Specialized vocabulary tends to occur more often in specialized texts (Chung & Nation, 2004; Nation, 2016). Hence, compared with universal academic wordlists, discipline-specific wordlists, which are solely developed from texts in learners' target disciplines, are

better at drawing their attention to the most important words in their specific areas and providing a shortcut to reduce the amount of learning (Nation, 2013). Moreover, learners are motivated to learn items from these lists because they can clearly see the relationship between what they study in their English courses and their subject courses (Basturkmen, 2003; Hyland, 2016). Additionally, the similarities in the learners' academic discipline may make it easier for teachers to focus on more specialized vocabulary in a specific discipline. Therefore, a wordlist specially developed for hard science students like the HSWL is the most suitable for ESAP programs that consist of only hard science students, and a potential SSWL may be the most relevant for ESAP programs with only soft science students. Similarly, in ESP programs where all learners plan to study the same specific subject areas, for example, Medicine, a specialized wordlist such as a potential MSWL may be the most appropriate for this group of learners.

Once teachers have identified the relevant wordlist for their learners, the next step is identifying the learning goals and sequences for the learners based on their current vocabulary levels and learning purposes. Let us take the beginner learners (i.e., those having not mastered the most frequent 1,000 BNC/COCA word families) in an ESAP program whose learners all plan to study hard sciences as an example. If these learners would like to go straight to the most frequent and wide ranging lexical items in hard science speech, they can start learning items from Level 1 of the HSWL. This sequence is efficient in terms of time because learners' attention would be drawn to the lexical items that are most relevant to their academic subject areas. The trade-off is that they would miss the items that are useful for engaging in general conversation. For example, 21 survival words

(Nation & Crabbe, 1991) are not in the HSWL: bus, delicious, excuse, goodbye, hospital, ladies, police, sick, thirteen, town (1st 1,000 BNC/COCA word level), gents, hotel, newspaper, restaurant, stamps, ticket, toilet, tourist, welcome (2nd 1,000 BNC/COCA word level), entrance (3rd 1,000 BNC/COCA word level), and exit (4th 1,000 BNC/COCA word level). If the learners would like to master items that are useful for general use first, they can learn items from Nation's (2012) BNC/COCA lists, and then move to the relevant levels of the HSWL once they are happy with their knowledge of general vocabulary. While these sequences take more time than going straight to the HSWL, they would enable learners to effectively engage in both general conversation and academic spoken discourse. Dividing specialized wordlists into levels and integrating it into lists of general vocabulary is innovative. This approach makes it possible for learners to start learning specialized vocabulary at any level of general proficiency, and therefore, gives more flexibility to teachers and learners. Students can consider the pros and cons of each sequence and choose the one that best suits learners' learning purposes and proficiency level.

Taken together, considering learners' target academic subject areas, proficiency levels, and learning purposes in the determination of the learning goal and sequence offers several benefits. First, it ensures that the list draws as much as possible learners' attention to specialized vocabulary in their academic disciplines but still matches the context of their EAP programs. Second, it avoids repeatedly teaching and learning known items and allows learners and teachers to spend their time effectively. It should be noted that once the relevant list and the learning sequence for the learners have been identified, teachers and material designers can use Nation's (2007) four strands to guide the design of learning

activities and materials so that their learners can repeatedly encounter and use the target words in different contexts related to their subject areas. This allows learners to acquire, consolidate, and expand their knowledge of these words in a meaningful way.

5. Limitations and future research

The present study has a number of limitations. First, like previous corpus-driven studies into the vocabulary load of spoken discourse (Dang & Webb, 2014; Nation, 2006; Webb & Paribakht, 2015; Webb & Rodgers, 2009a, 2009b), the coverage figures in this study were calculated with the assumption that learners are able to recognize the spoken forms of the words and proper nouns. Although L2 learners' aural and orthographic knowledge are closely related, the gap between the two kinds of knowledge may vary according to the learning contexts and learners' characteristics (Milton, 2009). Additionally, it may be overoptimistic to expect that L2 learners need little effort to recognize proper nouns in listening (Kobeleva, 2012). Second, this study focuses on single word units while knowledge of multi-words is also essential for fluent processing (Simpson-Vlach & Ellis, 2010). Third, like most previous corpus-based wordlists (e.g., Coxhead, 2000; Gardner & Davies, 2014), this study used lexical coverage as the only indicator of the list value while there are many factors influencing the value of a wordlist for L2 learners.

There are a few directions for future research. First, it is beneficial to develop a SSWL for ESAP programs which consist of only soft science students. Such research provides not only a useful tool for soft science students in these programs, but also further insight into the similarities and differences between hard and soft spoken vocabulary. Second, it is useful to create spoken wordlists of specialized vocabulary in a specific subject area (e.g.,

Mathematics, Medicine, or Engineering). Third, the importance of multi-word units in fluency development also means that there is value in developing discipline-specific lists of multi-words. Such wordlists together with the ASWL and the HSWL will provide teachers and learners with a wide range of options so that they can choose the list which best suits their learning and teaching contexts. Fourth, it is beneficial for future studies to investigate learners' knowledge and teachers' perceptions of the value of items in these corpus-based wordlists so that these lists can better serve the need of learners in a particular context.

6. Conclusion

This study is among several attempts to explore the nature of academic spoken vocabulary to better support L2 learners' vocabulary development. On one hand, it confirms the value of Dang et al.'s (2017) ASWL for EGAP programs. On the other hand, it indicates the value of discipline-specific wordlists for ESAP programs. The HSWL developed in this study consists of 1,595 word families which are the most frequent, wide ranging, and evenly distributed items in hard science speech. It benefits hard science students in ESAP programs irrespective of their target subject areas, language proficiency levels, and university administrative structures. Compared with the ASWL, the HSWL better focuses hard science students' attention to specialized vocabulary in their field. Therefore, this study highlights the fact that there is no one-size-fits-all specialized wordlist, and provides a model in which teachers can choose the specialized wordlist and learning sequence to match the target academic subjects, current proficiency levels, and learning purposes of the learners in their programs.

Notes

[1] so-called is outside the BNC/COCA lists because these lists do not in include hyphenated items

[2], [3] These wordlists have not been developed yet. They are potential wordlists that were used to illustrate for the model.

Acknowledgements

I would like to thank Professor Stuart Webb and Dr. Averil Coxhead for encouraging me to conduct this research project, and Dr. Deborah Laurs for acting as a reader of the early version of this paper.

My thanks to the following publishers and researchers for their generosity in letting me use their materials to create my corpora: Cambridge University Press, Pearson, Dr. Lynn Grant, Assistant Professor Michael Rodgers, the lecturers at Victoria University of Wellington, the researchers in the British Academic Spoken English corpus project, the British Academic Written English corpus project, the International Corpus of English project, the Massachusetts Institute of Technology Open courseware project, the Open American National corpus project, the Santa Barbara Corpus of Spoken American-English project, the Stanford Engineering Open courseware project, the University of California, Berkeley Open courseware project, and the Yale University Open courseware project.

References

Basturkmen, H. (2003). Specificity and ESP course design. RELC Journal, 34(1), 48–63.

Bauer, L., & Nation, P. (1993). Word families. International Journal of Lexicography, 6(4), 253–279.

Becher, T. (1989). Academic tribes and territories. Bristol: The Society for Research into Higher Education and Open University Press.

Becker, A. (2016). L2 students' performance on listening comprehension items targeting local and global information. Journal of English for Academic Purposes, 24, 1–13.

Biber, D. (2006). University language: a corpus-based study of spoken and written registers. Amsterdam: John Benjamins Publishing.

Biber, D., Reppen, R., Schnur, E., & Ghanem, R. (2016). On the (non)utility of Juilland's D to measure lexical dispersion in large corpora. International Journal of Corpus Linguistics, 21(4), 439–464.

Biglan, A. (1973a). Relationships between subject matter characteristics and the structure and output of university departments. Journal of Applied Psychology, 57(3), 204–213.

Biglan, A. (1973b). The characteristics of subject matter in academic areas. Journal of Applied Psychology, (57), 195–203.

Browne, C., Culligan, B., & Phillips, J. (n.d.). A new academic word list. Retrieved from http://www.newacademicwordlist.org/

Chung, T. M., & Nation, P. (2004). Identifying technical vocabulary. System, 32(2), 251–263.

Coxhead, A. (2000). A new academic word list. TESOL Quarterly, 34(2), 213–238.

Coxhead, A., & Hirsh, D. (2007). A pilot science word list for EAP. Revue Française de Linguistique Appliqueé, XII, 2, 65–78.

Crossley, S., Salsbury, T., & McNamara, D. (2010). The development of polysemy and frequency use in English Second Language speakers. Language Learning, 60(3), 573–605.

Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The academic spoken word list. Language Learning, 67(4).

Dang, T. N. Y., & Webb, S. (2014). The lexical profile of academic spoken English. English for Specific Purposes, 33, 66–76.

Dang, T. N. Y., & Webb, S. (2016). Making an essential word list. In I. S. P. Nation (Ed.), Making and using word lists for language learning and testing (pp. 153–167). Amsterdam: John Benjamins.

Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. Applied Linguistics, 28(2), 241–265.

Gardner, D., & Davies, M. (2014). A new academic vocabulary list. Applied Linguistics, 35(3), 305–327.

Gries, S. T. (2008). Dispersions and adjusted frequencies in corpora. International Journal of Corpus Linguistics, 13(4), 403–437.

Heatley, A., Nation, I. S. P., & Coxhead, A. (2002). Range: A program for the analysis of vocabulary in texts. Retrieved from http://www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx

Henriksen, B., & Danelund, L. (2015). Studies of Danish L2 learners' vocabulary knowledge and the lexical richness of their written production in English. In P. Pietilä, K. Doró, & R. Pipalová (Eds.), Lexical issues in L2 writing (pp. 1–27). Newcastle upon Tyne: Cambridge Scholars Publishing.

Hsu, W. (2011). The vocabulary thresholds of business textbooks and business research articles for EFL learners. English for Specific Purposes, 30, 247–257.

Hsu, W. (2013). Bridging the vocabulary gap for EFL medical undergraduates: the establishment of a medical word list. Language Teaching Research, 17(4), 454–484.

Hsu, W. (2014). Measuring the vocabulary load of engineering textbooks for EFL undergraduates. English for Specific Purposes, 33, 54–65.

Hu, M., & Nation, I. S. P. (2000). Vocabulary density and reading comprehension. Reading in a Foreign Language, 13(1), 403–430.

Hyland, K. (2016). General and specific EAP. In K. Hyland & P. Shaw (Eds.), The Routledge handbook of English for Academic Purposes (pp. 17–29). London: Routledge.

Hyland, K., & Tse, P. (2007). Is there an "academic vocabulary"? TESOL Quarterly, 41(2), 235–253.

Jones, W. A. (2011). Variation among academic disciplines: An update on analytical frameworks and research. The Journal of the Professoriate, 6(1), 9–27.

Juilland, A. G., & Chang-Rodrigues, E. (1964). Frequency dictionary of Spanish words. London: Mouton & Co.

Kobeleva, P. P. (2012). Second language listening and unfamiliar proper names: comprehension barrier? RELC, 43(1), 83–98.

Kolb, D. A. (1981). Learning styles and disciplinary differences. In A. W. Chickering (Ed.), The modern American college (pp. 232–255). San Francisco: Jossey Bass.

Laufer, B. (1989). What percentage of text lexis is essential for comprehension? In C. Laurén & M. Nordman (Eds.), Special language: From humans thinking to thinking machines (pp. 316–323). Clevedon: Multilingual Matters.

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(1), 15–30.

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.

MacDonald, M., Badger, R., & White, G. (2000). The real thing?: Authenticity and academic listening. English for Specific Purposes, 19(3), 253–267.

Matthews, J., & Cheng, J. (2015). Recognition of high frequency words from speech as a predictor of L2 listening comprehension. System, 52, 1–13.

Milton, J. (2009). Measuring second language vocabulary acquisition. Bristol, UK: Multilingual Matters.

Mochizuki, M., & Aizawa, K. (2000). An affix acquisition order for EFL learners: An exploratory study. System, 28(2), 291–304.

Mulligan, D., & Kirkpatrick, A. (2000). How much do they understand? Lectures, students and comprehension. Higher Education Research & Development, 19(3), 311–335.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? Canadian Modern Language Review, 63(1), 59–82.

Nation, I. S. P. (2007). The four strands. Innovation in Language Learning and Teaching, 1(1), 1–12.

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

Nation, I. S. P., & Waring, R. (1997). Vocabulary size, text coverage, and word lists. In N. Schmitt & McCarthy, M. (Eds.), Vocabulary: Description, acquisition and pedagogy (pp. 6–19). Cambridge: Cambridge University Press.

Nation, I. S. P., & Webb, S. (2011). Researching and analyzing vocabulary. Boston: Heinle, Cengage Learning.

Nation, P., & Crabbe, D. (1991). A survival language learning syllabus for foreign travel. System, 19(3), 191–201.

Nesi, H. (2002). An English Spoken Academic Word List. In A. Braasch & C. Povlsen (Eds.), Proceedings of the Tenth EURALEX International Congress (Vol. 1, pp. 351–358). Copenhagen, Denmark.

Neumann, R. (2001). Disciplinary differences and university teaching. Studies in Higher Education, 26(2), 135–146.

Neumann, R., Parry, S., & Becher, T. (2002). Teaching and learning in their disciplinary contexts: A conceptual analysis. Studies in Higher Education, 27(4), 405–417.

Nguyen, T. M. H., & Webb, S. (2016). Examining second language receptive knowledge of collocation and factors that affect learning. Language Teaching Research, 1 –23.

Rodgers, M. P. H., & Webb, S. (2011). Narrow viewing: The vocabulary in related television programs. TESOL Quarterly, 45(4), 689–717.

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

Schmitt, N., & Zimmerman, C. B. (2002). Derivative word forms: What do learners know? TESOL Quarterly, 36(2), 145–171.

Simpson-Vlach, R., & Ellis, N. C. (2010). An Academic Formulas List: New methods in phraseology research. Applied Linguistics, 31(4), 487–512.

Valipouri, L., & Nassaji, H. (2013). A corpus-based study of academic vocabulary in chemistry research articles. Journal of English for Academic Purposes, 12(4), 248–263.

van Zeeland, H., & Schmitt, N. (2013). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? Applied Linguistics, 34(4), 457–479.

Wang, J., Liang, S., & Ge, G. (2008). Establishment of a Medical Academic Word List. English for Specific Purposes, 27(4), 442–458.

Ward, J. (1999). How large a vocabulary do EAP engineering students need? Reading in a Foreign Language, 12(2), 309–323.

Ward, J. (2009). A basic engineering English word list for less proficient foundation engineering undergraduates. English for Specific Purposes, 28(3), 170–182.

Watson-Todd, R. (2017). An opaque engineering word list: Which words should a teacher focus on? English for Specific Purposes, 45, 31–39.

Webb, S., & Chang, A. C.-S. (2012). Second language vocabulary growth. RELC Journal, 43(1), 113–126.

Webb, S., & Paribakht, T. S. (2015). What is the relationship between the lexical profile of test items and performance on a standardized English proficiency test? English for Specific Purposes, 38, 34–43.

Webb, S., & Rodgers, M. P. H. (2009a). The lexical coverage of movies. Applied Linguistics, 30(3), 407–427.

Webb, S., & Rodgers, M. P. H. (2009b). Vocabulary demands of television programs. Language Learning, 59(2), 335–366.

Table 1. Five corpora in the present study

Corpus	Purposes	Size
1st hard science spoken corpus	Develop the HSWL	6,515,717
2nd hard science spoken corpus	Validate the hard science, academic, and spoken nature of the list	6,397,458
Soft science spoken corpus	Validate the hard science nature of the list	6,513,944
Hard science written corpus	Validate the spoken nature of the list	6,631,403
Non-academic spoken corpus	Validate the academic nature of the list	6,505,382

Table 2. First hard science spoken corpus

Hard-pure		Hard-applied	
Subject	Running words	Subject	Running words
Astronomy	593,062	Chemical Engineering	563,938
Biology	552,452	Computer Sciences	555,175
Chemistry	556,138	Cybernetics	555,401
Ecology & Geology	555,312	Electrical Engineering	550,181
Mathematics	450,481	Health & Medical Sciences	470,795
Physics	554,178	Mechanical Engineering	558,604
Total	3,261,623	Total	3,254,094

Table 3. Second hard science spoken corpus

Hard-pure		Hard-applied	
Subjects	Running words	Subjects	Running words
Biology	699,286	Applied Statistics*	14,179
Chemistry	761,025	Civil Engineering*	33,718
Ecology & Geology	15,459	Computer Sciences	347,348
Mathematics	924,437	Construction*	20,358
Physics	768,409	Cybernetics	904,854
		Electrical Engineering	1,581,306
		Engineering Graphics*	22,409
		General Engineering*	118,751
		Industrial & Operation Engineering*	10,722
		Manufacturing*	63,912
		Marine Engineering*	46,567
		Mechanical Engineering	21,279
		Meteorology*	43,439
Total	3,168,616	Total	3,228,842

* Subjects that are not represented in the first hard science spoken corpus

Table 4. Soft science spoken corpus

Soft-pure		Soft-applied	
Subjects	Running words	Subjects	Running words
Art	553,160	Business	513,133
Cultural Studies	498,393	Economics	610,998
History	554,214	Education	571,023
Philosophy	549,577	Law	616,398
Political Studies	545,059	Management	461,093
Psychology	555,880	Public Policy	485,016
Total	3,256,283	Total	3,257,661

Table 5. Hard science written corpus

Hard pure		Hard applied	
Subjects	Running words	Subjects	Running words
Astronomy	293,720	Agriculture	425,647
Biology	341,250	Civil engineering	430,706
Chemistry	122,283	Computer science	191,735
Ecology & Geology	275,173	Cybernetics	86,208
General Sciences	1,195,124	Electrical engineering	576,810
Mathematics	688,465	General Engineering	720,587
Physics	183,776	Health & Medicine	398,153
		Material Engineering	155,905
		Mechanical engineering	382,337
		Media Art & Science	120,796
		Meteorology	42,728
Total	3,099,791	Total	3,531,612

Table 6. Non-academic spoken corpus

Corpus	Main variety of English	Running words
International Corpus of English (spoken, non-academic)	Indian, Pilipino, Singapore, Canadian, Hong Kong, Irish, Jamaican & New Zealand	5,262,502
TV program corpus (Rodgers & Webb, 2011)	British & American	943,058
Santa Barbara Corpus of Spoken American-English (non-academic)	American	299,822
Total		6,505,382

Table 7. Lexical profile of the HSWL

HSWL level	BNC/COCA word level	Number of word-families	Additional coverage (%)	Examples
Level 1	1st 1,000	762	81.47	machine, gas
Level 2	2nd 1,000	384	5.06	metal, laboratory
Level 3	3rd 1,000	296	3.14	molecule, element
Level 4	4th 1,000	73	0.65	matrix, magnitude
	5th 1,000	36	0.28	analogy, equilibrium
	6th 1,000	17	0.15	vector, invert
	7th 1,000	10	0.07	gradient, gamma
	8th 1,000	5	0.03	iterate, algebra
	9th 1,000	1	0.01	exponential
	10th 1,000	5	0.04	sine, epsilon
	11th 1,000	2	0.01	lambda, cubed
	12th 1,000	1	0.02	theta
	13th 1,000	1	0	quadratic
	16th 1,000	1	0.01	cosine
	Outside BNC/COCA25000	1	0	so-called[1]
Total		1,595	90.94	

Table 8. Potential coverage gained by learners with the aid of the HSWL (%)

Group	Existing vocabulary level	Number of HSWL beyond	Without PN & MW		With PN & MW	
			1st corpus	2nd corpus	1st corpus	2nd corpus
Beginner	Less than 1,000	1,595	90.94	90.82	92.97	93.35
Pre-Intermediate	1,000	833	91.48	91.15	93.51	93.68
Intermediate	2,000	449	92.47	91.81	94.50	94.34
Advanced	3,000	153	93.55	92.54	95.58	95.07
Proper nouns (PN)			0.66	0.74		
Marginal words (MW)			1.37	1.79		

Table 9. Potential coverage gained by the intermediate and advanced learners from the subsequent BNC/COCA frequency levels (%)

Group of learners	Existing vocabulary level (BNC/COCA word families)	Extra words from the BNC/COCA list	Without PN & MW		With PN & MW	
			1st corpus	2nd corpus	1st corpus	2nd corpus
Intermediate	2,000	3rd 1,000	92.28	90.87	94.31	93.40
Advanced	3,000	4th 1,000	93.79	92.32	95.82	94.85

Figure 1. Dang et al.'s (2017) learning sequence of the ASWL

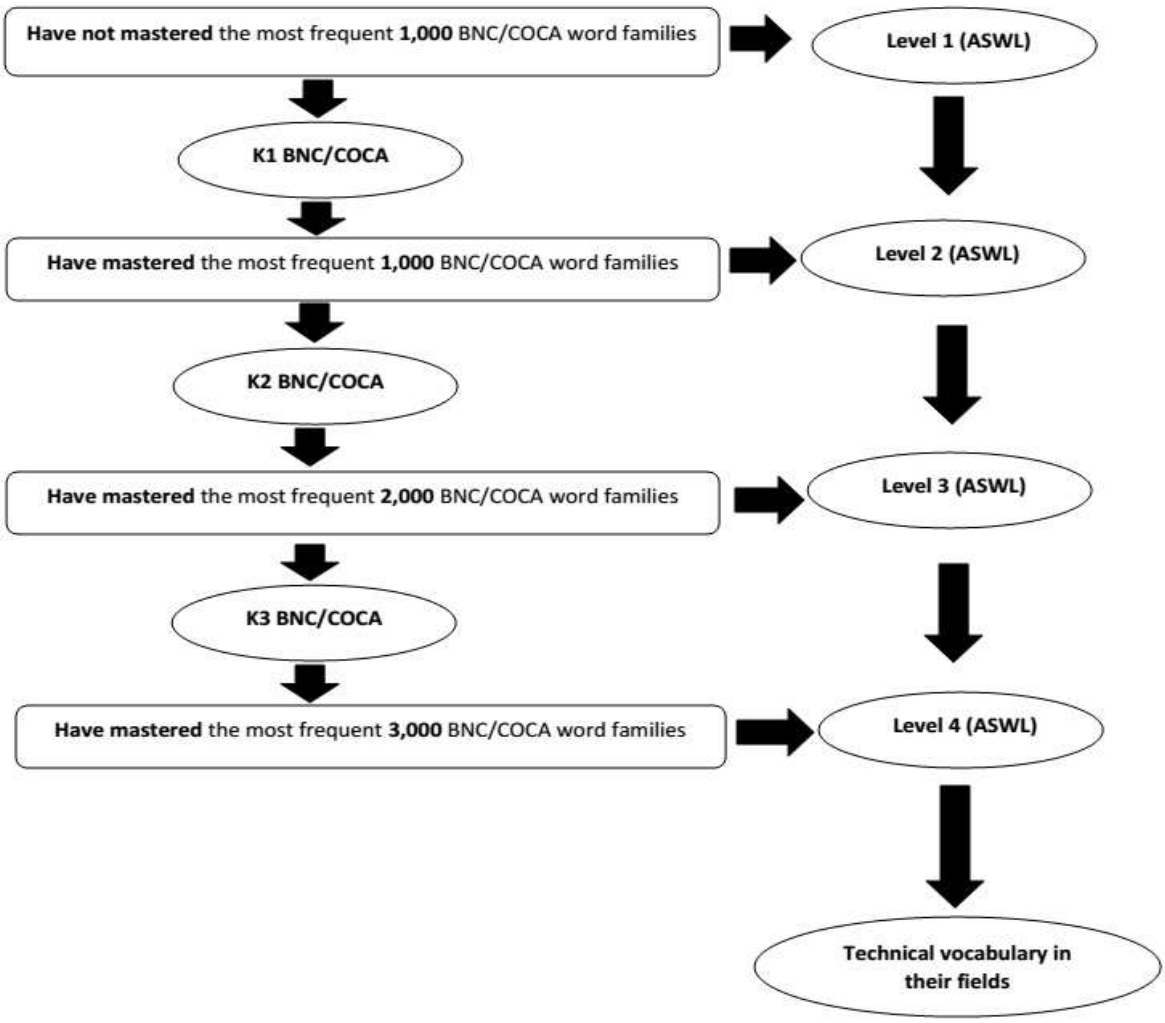

Figure 2. Component of the potential coverage reached by learners of different vocabulary levels with the aid of the HSWL

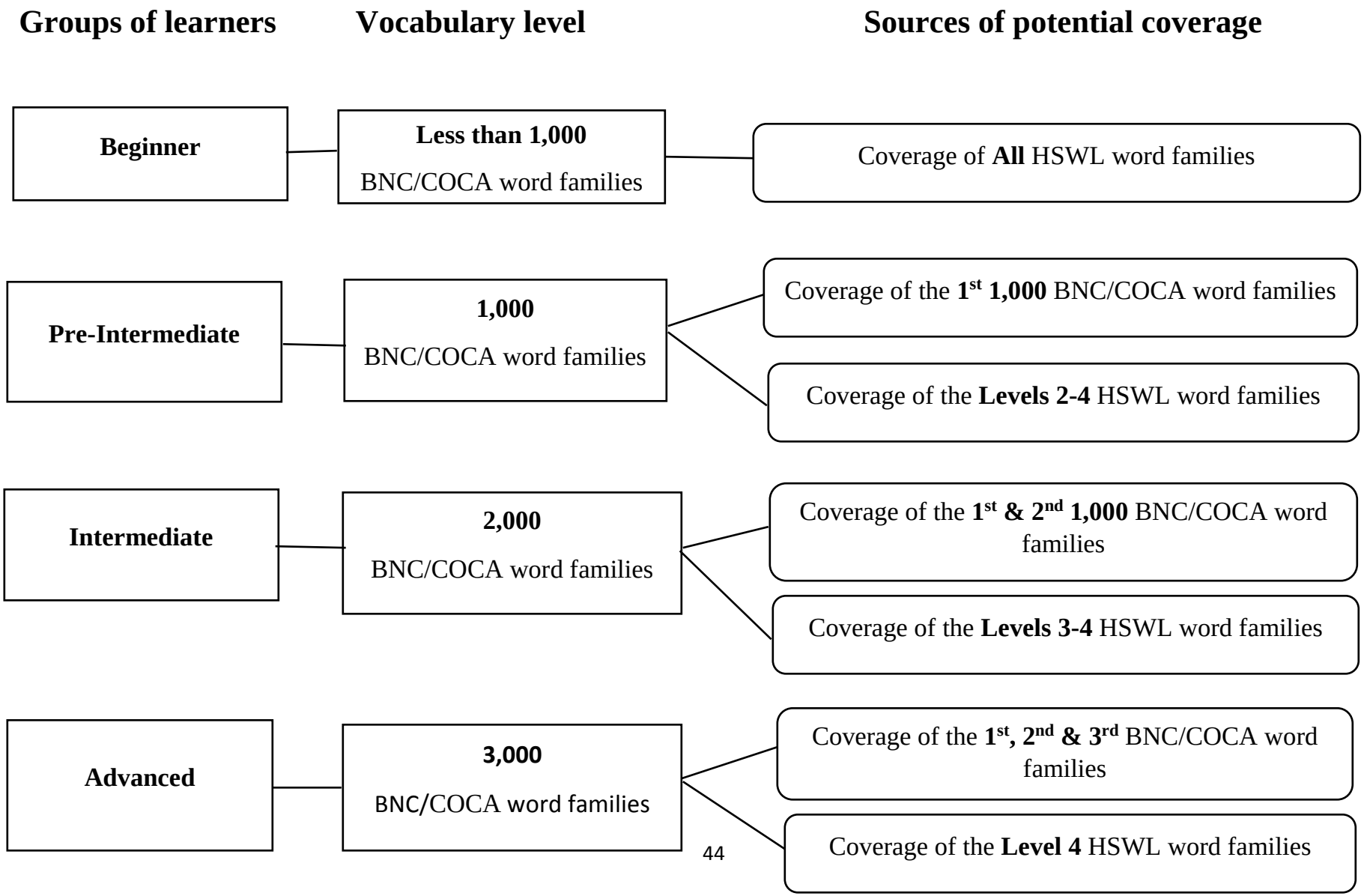

Figure 3. Options in wordlists and learning sequence for different EAP/ESP programs