Article:
Dang, TNY (2018) The nature of vocabulary in academic speech of hard and soft-sciences. English for Specific Purposes, 51. pp. 69-83. ISSN: 0889-4906

https://doi.org/10.1016/j.esp.2018.03.004

The nature of vocabulary in academic speech of hard and soft-sciences

Source: Dang, T. N. Y. (2018). The nature of vocabulary in academic speech of hard and soft sciences, *51*, 69–83.

ABSTRACT

Little is known about the similarities and differences between the vocabulary in hard-sciences (e.g., Maths, Engineering, Medicine) and soft-sciences (e.g., Business, Law, History), especially in spoken discourse. To address this gap, a Soft Science Spoken Word List (SSWL) was developed for second language learners of soft-sciences at English-medium universities. The list consists of the 1,964 most frequent and wide-ranging word-families in a 6.5 million word corpus of soft-science speech, which represents 12 subjects across two equally-sized sub-corpora. The list may allow learners to recognize 94%-97% of the words in academic speech of soft-sciences. A comparison of the SSWL with Dang's (2018) Hard Science Spoken Word List revealed that although the most frequent 3,000 words are important for comprehending academic speech of both soft- and hard-sciences, the value of these words in soft-sciences is greater than in hard-sciences. Pedagogical implications related to this nature of vocabulary in hard- and soft-science speech are provided.

Key words: Hard science, soft science, academic spoken discourse, vocabulary, word list, corpora.

1. Introduction

English has been widely used as the medium of instruction in academic courses at tertiary levels—in face-to-face, distance learning, and online contexts—in both English speaking and non-English speaking countries. Second language (L2) learners in these courses have to comprehend not only reading materials such as textbooks and research articles but also lectures, seminars, labs, and tutorials (Biber, 2006). Because vocabulary knowledge and comprehension are closely related (Laufer & Ravenhorst-Kalovski, 2010; Schmitt, Jiang, & Grabe, 2011; van Zeeland & Schmitt, 2013), it is essential for learners to master the words that they are likely to encounter often in a wide range of academic written and spoken texts. A large number of wordlists have been developed to assist L2 learners' comprehension of academic writing (e.g.,

Coxhead, 2000; Coxhead & Hirsh, 2007; Gardner & Davies, 2014; Liu & Han, 2015; Martínez, Beck, & Panza, 2009; Wang, Liang, & Ge, 2008; Watson-Todd, 2017). Yet little has been done to help these learners comprehend academic speech. In fact, understanding academic spoken English is a great challenge for L2 learners in different contexts (Flowerdew & Miller, 1992; Mulligan & Kirkpatrick, 2000). Given this fact, it is crucial to create wordlists that capture the most frequent and wide-ranging words in academic speech. Together with written wordlists, these spoken wordlists are valuable resources for English for Academic Purposes (EAP) programs to support L2 learners' comprehension of academic English.

Depending on the target subject areas of their learners, EAP programs can be divided into two types: English for Specific Academic Purposes (ESAP) and English for General Academic Purposes (EGAP)[1]. Learners in ESAP programs are fairly homogeneous in terms of target subject areas. That is, they all plan to study hard-science subjects (e.g., Mathematics, Physics, Engineering or Medicine) or soft-science subjects (e.g., Linguistics, Law, Business, or Education). The hard/soft division refers to the existence of a paradigm, or 'a body of theory which is subscribed to by all members of the field' (Biglan, 1973b, p.201). Earlier research on students' learning strategies and scholars' behavior and opinions about various aspects of academic disciplines (e.g., teaching, learning, research styles) (e.g., Becher, 1989; Biglan, 1973a, 1973b; Neumann, Parry, & Becher, 2002) has indicated that the hard/soft division is the strongest dimension[2] that distinguishes academic subjects in higher education. Hard-sciences (e.g., Mathematics, Engineering) are likely to have a single paradigm that allows scholars working in these areas to reach a wide consensus on research methods and key concepts. In contrast, soft-sciences (e.g., Law, Philosophy) are likely to lack a single paradigm, and scholars working in these areas seem to argue over methods and key concepts. While the hard/ soft division cannot fully reflect the complexity and variation in inquiry processes and knowledge structures in various disciplines (Becher & Trowler, 2001; Nesi, 2002), this division is a useful shorthand from attempting to explain the complexity and diversity of academic discourse. Learners in EGAP programs, however, are more heterogeneous in terms of their target disciplines. In other words, there is a mixture of hard-science and soft-science students in these programs. EGAP programs can also be the programs where (a) learners have not yet identified their target subject areas, (b) learners plan to study interdisciplinary subject areas, (c) or teachers lack background knowledge of learners' specific subject areas.

According to Hyland (2016), general and specific EAP approaches should be considered as a continuum rather than a dichotomy. Depending on the teaching and learning context of a particular EAP program, either a general academic wordlist or a discipline-specific wordlist is more suitable than the other (Dang, Coxhead, & Webb, 2017). A general academic wordlist is more relevant to EGAP programs. The diversity in learners' academic subject areas may make it challenging for teachers in these programs to satisfy the specific needs of every learner. Meanwhile, a discipline-specific wordlist is more suitable for ESAP programs. As specialized vocabulary tends to occur more often in specialized texts (Chung & Nation, 2004), discipline-specific wordlists focus learners' attention on items that occur very often in their specific areas and provide a shortcut to reduce the amount of learning (Nation, 2013). Learners are more motivated to learn items from discipline-specific wordlists because they can see clearly the link between what they learn in their ESAP course and their subject courses (Coxhead & Hirsh, 2007; Hyland, 2016). Additionally, the similarities between learners' academic disciplines may make it easier for teachers to focus on specialized vocabulary in a particular discipline.

Several general academic spoken wordlists have been developed for EGAP programs such as Nesi's (2002) Spoken Academic Word List, Simpson-Vlach and Ellis's (2010) Academic Formulas List, and Dang et al.'s (2017) Academic Spoken Word List. As a result, we have a fairly good understanding about the shared spoken vocabulary across hard- and soft-sciences. In contrast, only one spoken discipline-specific wordlist has been developed for ESAP programs, and it focuses on spoken vocabulary in hard-sciences: Dang's (2018) Hard Science Spoken Word List (HSWL). No attempts have been made to identify the most frequent and wide-ranging words in soft-science speech. The lack of such a list makes it challenging to compare the vocabulary in academic speech of the hard- and soft-sciences.

The present study was conducted with two aims. The first aim was to develop a Soft Science Spoken Word List (SSWL) for ESAP programs which consist of solely soft-science students. The second aim was to compare this list with Dang's (2018) HSWL to see the similarities and differences between the most frequent and wide-ranging words in hard disciplines and soft disciplines. The research thus provides soft-science students in ESAP programs with a useful instrument to achieve better comprehension of academic speech, and sheds light on the nature of vocabulary in hard- and soft-science speech.

1.1. Background

Studies that investigated the nature of vocabulary in spoken texts of hard- and soft-sciences either examined the vocabulary load of the two disciplines or focused on the shared vocabulary between these disciplines. Vocabulary load studies (Coxhead, Dang, & Mukai, 2017; Dang & Webb, 2014) determined the number of words required to understand academic speech. They looked at the lexical coverage of different 1,000-word frequency levels of general vocabulary in academic speech and estimated the number of words needed to reach 95% and 98% coverage of these texts. Lexical coverage is the percentage of words covered by items from a particular wordlist in a text (Nation & Waring, 1997). The 95% and 98% figures have been widely used as the coverage cut-off points to indicate high and stable degrees of comprehension (van Zeeland & Schmitt, 2013).

Dang and Webb (2014) examined the vocabulary load of lectures and seminars of the hard- and soft-sciences represented in the British Academic Spoken English Corpus (BASE). They found that a vocabulary size of 3,000-4,000 word families is needed to reach 95% coverage of soft-science speech while a vocabulary size of 5,000-7,000 word families is necessary to achieve 98% coverage. These coverage figures are lower than those needed to reach 95% coverage (4,000-5,000 word families) and 98% coverage (10,000-13,000 word families) of hard-science speech. A similar pattern was reported by Coxhead et al. (2017) when investigating the vocabulary load of university tutorials (a distinctive speech event of soft-sciences) and labs (a distinctive speech events of hard-sciences). A vocabulary size of 2,000 word families and 4,000 word families was necessary to reach nearly 95% and 98% coverage of tutorials, respectively. These vocabulary sizes are much smaller than those needed to reach 95% coverage (3,000 word families) and 98% coverage (7,000 word families) of labs. Together, Dang and Webb's (2014) and Coxhead et al.'s (2017) findings indicate that comprehending soft-science speech is less demanding than comprehending hard-science speech in terms of lexical coverage. However, the corpora used in these studies were fairly small (137,000 to 400,000 words) and only presented a limited number of speech events: lectures and seminars (Dang & Webb, 2014) or labs and tutorials (Coxhead et al., 2017). Importantly, they examined the occurrences of words at different frequency levels of general vocabulary in specialized texts. EAP/ESP research examining various kinds of academic and general discourse from different perspectives such as phraseology, terminology, and grammatical structures (e.g, Biber, 2006; Biber, Conrad, & Cortes, 2004; Cabré, 1999; Csomay,

2006; Hyland, 2000; Resche, 2012) has indicated that the linguistic features vary according to discourse types. Therefore, a comparison of the most frequent and wide-ranging words in spoken texts of hard- and soft-sciences may provide a better insight into the nature of vocabulary in hard and soft disciplines.

In recognition of this need, several studies have been conducted to identify the most frequent and wide-ranging lexical items in spoken texts of hard- and soft-sciences. Three studies have investigated multi-word units in academic speech. Biber et al. (2004) analyzed a 1.2-million word corpus of university classroom teaching from both hard- and soft-sciences. They then developed a list of 84 most frequent lexical bundles in the corpus (e.g., *I mean you know, you need to know*). Simpson-Vlach and Ellis (2010) developed a spoken Academic Formulas List (AFL). The top 200 spoken AFL formulas (e.g., *you know what I mean, how do you know*) were selected from a 2.1-million word spoken corpus which represented academic speech from both hard- and soft-sciences. Coxhead et al. (2017) focused on labs and tutorials specifically and came up with lists of the most frequent sequences in each kind of speech event as well as those in tutorials and labs combined (e.g., *do you know what, you don't know why)*. Lists of multi-word units have great value because knowledge of multi-word units is essential for fluency development (Nation & Webb, 2011; Simpson-Vlach & Ellis, 2010). Yet knowledge of single words is also important for learners to comprehend academic speech because it provides valuable support for the acquisition of multi-word items in the discourse. Specific phrasing of multi-word items may vary across different lists, but these items share a reasonable number of core single words (Coxhead et al., 2017; Shin & Nation, 2008). This pattern can be seen clearly from the above examples of Biber et al.'s (2004), Simpson-Vlach and Ellis's (2010) and Coxhead's (2017) lists. Hence, developing lists of single words is equally as important as developing lists of multi-word units. Given this fact, Nesi (2002) and Dang et al. (2017) developed lists of single words.

Nesi's (2002) Spoken Academic Word List (SAWL) was created based on the common approach taken to develop academic written wordlists (e.g., Coxhead, 2000; Xue & Nation, 1984). According to this approach, general vocabulary is seen as a series of layers, each of which represents a 1,000-item frequency level. Words at the first 1,000-word level are the most frequent and wide ranging whereas items at the second 1,000-word level are less frequent and

narrower ranging. Words at the first and second 1,000-word levels, or even the third 1,000-word level are general high-frequency words (Nation, 2013; Schmitt & Schmitt, 2014). Academic words are then defined as items that fall outside general high-frequency word levels, but have high frequency and wide range in academic texts. In other words, this approach assumes that learners have already mastered general high-frequency words, and, therefore, academic wordlists do not include general high-frequency words but only lower frequency words that have wide range and high frequency in academic texts. In the case of Nesi's (2002) list, the selected SAWL words were outside Nation's most frequent 2,000 word-families, occurred more than three times in each sub-corpus of the BASE, and had high frequency in the corpus. Unfortunately, no precise information is written about the development and validation of the SAWL, and this list is not available to access.

Dang (2017) Academic Spoken Word List consists of 1,741 word-families which were selected from an academic spoken corpus with four equally-sized sub-corpora: hard-pure (e.g., Mathematics, Physics), hard-applied (e.g., Medicine, Engineering), soft-pure (e.g., History, Philosophy), and soft-applied (e.g., Law, Business). Each sub-corpus represents six subject areas of around 500,000 running words. In other words, the whole corpus was derived from academic speech of 24 subject areas which made up a total of about 13 million running words. Unlike Nesi (2002) and Simpson-Vlach and Ellis (2010), Dang et al. (2017) made the best use of the two approaches towards developing general academic wordlists.

Following Gardner and Davies (2014), Dang et al. (2017) considered academic vocabulary as a separate kind of vocabulary that cuts across different 1,000-word levels of general vocabulary and created the ASWL from scratch. That is, they selected all word-families that met the following criteria: (1) appearing in all sub-corpora of the corpus and in at least 50% of the subject areas, (2) occurring at least 350 times in the whole corpus of 13 million words (i.e., 26.9 times per million words) and (3) having a Juilland and Chang-Rodrigues's (1964) dispersion (D) of at least 0.6. Creating the ASWL as an entirely new list avoids the limitations related to existing lists of general high-frequency vocabulary such as West's (1953) General Service List and directs learners' attention to the most frequent and wide-ranging words in academic spoken English. However, while Gardner and Davies only included items that have wider range and higher frequency in academic texts than non-academic texts in their Academic Vocabulary List,

Dang et al.'s (2017) ASWL consists of items that have wide range and high frequency in academic texts regardless of the ratio between their frequency in academic texts and in non-academic texts. This approach then allows the inclusion of items that are common in both general and academic speech (e.g., *investigate, think, issue*).

Following other earlier studies (e.g., Coxhead, 2000; Xue & Nation, 1984), Dang et al. (2017) considered academic vocabulary in relation to general vocabulary. Nevertheless, instead of setting a fixed benchmark for the number of general words that every learner should know before learning items from the ASWL, the list was divided into four levels based on Nation's (2012) BNC/COCA lists. Levels 1 to 3 represent ASWL words appearing at the first, second, and third 1,000 BNC/COCA-word levels, respectively. ASWL words at Level 4 are those outside the most frequent 3,000 BNC/COCA words. Depending on their current levels of general vocabulary, learners can skip certain levels of the ASWL. Dividing the list into levels makes it adaptable to learners' proficiency levels.

The ASWL covered 90.13% of the whole corpus and around the same amount of coverage in each sub-corpus. When tested against a second academic spoken corpus of similar size and structure, the ASWL provided around 90% coverage. Moreover, its coverage in the academic spoken corpora was higher than the coverage in the academic written corpus and non-academic spoken corpus of a similar size. If proper nouns and marginal words are known, knowledge of the ASWL may enable learners to reach 92%-96% coverage of academic spoken English depending on their proficiency levels.

It is important to note that Nesi's (2002), Simpson-Vlach and Ellis's (2010), and Dang et al.'s (2017) lists are general academic wordlists. By developing these lists, these researchers confirm the existence of a core vocabulary across hard- and soft-science speech. Yet, these general academic wordlists only represent the lexical items that have high frequency and wide range in both hard- and soft-sciences. Items that have high frequency and wide range in one discipline but do not have high frequency and wide range in the other are not included in these lists. Moreover, from the pedagogical perspective, while general academic wordlists are valuable resources for EGAP programs, wordlists developed specifically for hard-science students and for soft-science students are more useful for ESAP programs.

To address this gap, Dang (2018) developed a HSWL from a 6.5 million word corpus which was solely made up of spoken texts from hard-sciences. This corpus had two equally-sized sub-corpora: hard-pure and hard-applied. Each sub-corpus consisted of six subject areas. Each subject area contained 500,000 words. Following Dang et al.'s (2017) approach, Dang (2018) created the HSWL from scratch by including in the list items that satisfied the range, frequency, and dispersion criteria. The range and frequency criteria were adopted from those used to select Dang et al.'s (2017) ASWL words. However, unlike Dang et al. (2017), Dang (2018) used Gries's (2008) dispersion (DP) rather than Juilland and Chang-Rodríguez's (1964) dispersion (D) to measure dispersion, because DP is likely 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). In particular, to be included, an HSWL word-family had to have a DP of 0.6 or lower. There are 1,595 word-families satisfying these criteria. They accounted for 90.94% coverage of the hard-science spoken corpus and around the same amount of coverage in an independent hard-science spoken corpus of similar size and structure. These coverage figures are higher than those in the soft-science spoken corpus, hard-science written corpus, and non-academic spoken corpus of similar sizes. The HSWL has fewer items than Nation's (2012) most frequent 2,000 BNC/COCA word-families and Dang et al.'s (2017) ASWL, but provides higher coverage than these lists in the hard-science spoken corpora. Additionally, the HSWL consists of a larger proportion of words outside general high-frequency words (i.e., the most frequent 2,000 words, Nation, 2013) (28.15%) than the ASWL (26.13%). The range of frequency levels of these items is wider in the case of the HSWL (3^{rd}- 16^{th} 1,000 word levels) than in the case of the ASWL (3^{rd}-10^{th} 1,000 word levels). These findings suggest that the HSWL better captures the most frequent and wide-ranging words in hard-science speech than a general academic wordlist like the ASWL. Given the lack of a wordlist that represents the most frequent and wide-ranging words in soft-science speech, it is important for further research to develop a soft-science spoken wordlist taking the same approach as the HSWL. Such research would provide soft-science students in ESAP program with a useful resource to enhance their comprehension of academic speech as well as shed light on the similarities and differences between vocabulary in the speech of hard- and soft-sciences.

1.2. Research questions

1. Which lexical items occur frequently and are evenly distributed in a wide range of academic speech in soft-sciences?
2. With knowledge of these words, how much coverage of academic speech in soft-sciences may be reached by learners with different vocabulary levels?
3. How do these items compare with those from Dang's (2018) Hard Science Spoken Word List (HSWL)?

2. Methodology

2.1. Developing the corpora

Five corpora were developed in the present study (Table 1). The first soft-science spoken corpus was used to develop the SSWL while the other four corpora were used to validate different aspects of the list. This approach has been widely used to validate specialized wordlists (e.g., Coxhead, 2000; Gardner & Davies, 2014). Each corpus consists of around 6.5 million running words. This meets Nation and Webb's (2011) guideline; that is, to achieve a valid assessment of the occurrences of items in a wordlist, the list should be validated in independent corpora of similar size as the corpus from which it was developed.

[TABLE 1 NEAR HERE]

The composition of the two soft-science spoken corpora and the hard-science spoken corpus is presented in Tables 2-4. These corpora have the same size and structure. Each of them had around 6.5 million running words, and is divided into two sub-corpora: pure and applied. Each sub-corpus contains around 3.25 million running words. Four kinds of speech events are presented in these sub-corpora (lectures, seminars, labs, and tutorials) which represent naturally-occurring academic speech recorded in various institutions in different parts of the world (the U.S, the U.K, Hong Kong, New Zealand) and at least seven main varieties of English (American English, Australian English, British English, Canadian English, Hong Kong English, Irish English, and New Zealand English). Further information about the composition and sources of the two soft-science spoken corpora and the hard-science spoken corpus is provided in Appendices A and B.

[TABLES 2-4 NEAR HERE]

Table 5 shows the structure of the soft-science written corpus. This corpus represents different kinds of academic writing from different subject areas: book chapters, journal articles, research reports, and textbooks. These materials were from the British Academic Written English corpus (BAWE) and courses at a university in New Zealand. The soft-science written corpus has a structure similar to those of the three academic spoken corpora. It contains about 6.5 million running words and is divided into two equally-sized sub-corpora. Table 6 presents the components of the non-academic spoken corpus. This corpus represents a range of general spoken English (e.g., telephone conversations, TV programs, movies) and 10 varieties of English (American English, British English, Canadian English, Hong Kong English, Indian English, Irish English, Jamaican English, New Zealand English, Filipino English, and Singapore English). The target users of the SSWL come from different learning contexts in both English speaking and non-English speaking countries. The fact that the corpora used to develop and validate the SSWL include materials representing different varieties of English therefore ensured that the list presents as closely as possible the words that soft-science students from different contexts are likely to encounter in their academic study.

[TABLES 5 & 6 NEAR HERE]

2.2. Determining the unit of counting of the SSWL

Bauer and Nation's (1993) Level-6 word families were chosen as the unit of counting for the SSWL. A Level-6 word family (*generate*) is made up of a stem (*generate*), its inflections (*generates, generated, generating*), and closely related derivations (*generative, generatively*). This unit of counting was chosen for two reasons. First, following Coxhead (2000) and Nation (2013), the present study considers knowledge of word families is gradually picked up during the learning process rather than knowledge of a family's words all being acquired at the same time, and learners are given training on word part knowledge and word building skills. It means that knowledge of a known word-family member (*sad*) may facilitate the acquisition of other members (*sadly*). This assumption is supported by previous studies which found that L2 learners' derivational knowledge increased incrementally over time (Mochizuki & Aizawa, 2000; Schmitt & Meara, 1997; Schmitt & Zimmerman, 2002), and learners' vocabulary knowledge expands along with instruction on different inflections and derivations (Schmitt & Meara, 1997; Wei, 2014). Second, this study aims to (a) compare the SSWL with Dang's (2018) HSWL and Dang et al.'s (2017) ASWL, and (b) integrate the SSWL with Nation's (2012) BNC/COCA lists to

organize a vocabulary learning program for soft-science students. The Level-6 word-family is the unit of counting of the HSWL, ASWL, and BNC/COCA. Choosing the Level-6 word-family as the unit of counting of the SSWL ensures a fair comparison between this list and the HSWL and the ASWL, and a systematic integration between the SSWL and the BNC/COCA.

However, considering that L2 learners' vocabulary knowledge develops over time, another version of the SSWL is also developed. This version lists all flemmas within each Level-6 SSWL word family. Similar to lemmas, flemmas (*generate*) only consist of the stem (*generate*) and its inflections (*generates, generated, generating*) (Pinchbeck, 2014). It is a smaller unit of counting than Level-6 word families. For example, the Level-6 word family *generate* comprises three flemmas (*generate, generative, generatively*). However, unlike lemmas, flemmas do not distinguish between parts of speech. For example, *smile (v) and smile (n)* are considered as one flemma but two separate lemmas. This study developed a flemma list rather than a lemma list because distinguishing parts of speech may overestimate the learning burden of very closely related items like *smile* (v) and *smile* (n), but cannot distinguish homonyms with the same part of speech like *bank* (for money) and *bank* (for river) (Nation, 2016). Making the SSWL available in different formats helps the list to better support different groups of L2 learners.

2.3.Developing the SSWL and comparing it with the HSWL

To be included in the SSWL, a word family had to meet the range, frequency, and dispersion criteria. Frequencies are the number of occurrences of a word in the first soft-science spoken corpus. The higher the frequency, the more likely that learners are to meet the word in soft-science speech. Yet some words that are not widely used across many subject areas within the corpus still have high frequencies in the entire corpus just because they appear very often in a certain subject area. For example, *basilica* met the SSWL frequency criterion, but occurred in only 2 out of 12 subject areas of this corpus. Using range (i.e., the number of different subject areas in which a word occurs) as one selection criterion allows researchers to eliminate such items as *basilica*. However, range can only reveal whether the words occur in a subject area or not. It cannot discriminate between words having different distribution within multiple subject areas. For example, *marble* satisfied the range and frequency criteria of the SSWL, but it was not evenly distributed across the soft-science spoken corpus. *Marble* occurred 602 times per million words in Arts but no more than 8 times per million words in the remaining 11 subject areas.

Dispersion allows us to address this issue because it indicates how evenly a word is distributed across a corpus. Given their power, frequency, range, and dispersion are common criteria for selecting items for corpus-based specialized wordlists (Nation, 2016; Nation & Webb, 2011).

The present study aims to compare the SSWL with Dang's (2018) HSWL. To provide a valid comparison, the selection criteria of the HSWL were adopted to determine items for the SSWL:

(1) *Range:* a selected word family had to occur in both sub-corpora of the first soft-science spoken corpus, and in at least 50% of the subjects in this corpus (six out of 12 subjects).

(2) *Frequency:* a selected word family had to occur at least 26.9 times per million words. This criterion was originally used by Dang et al. (2017) to select items for their ASWL. This frequency figure was the result of extensive experimentation which compared the items included in or excluded from the ASWL at different frequency cut-off points. A word list with the frequency cut-off point of 26.9 times per million words had fewer items than lists with lower frequency cut-off points. Meanwhile, the list with the frequency cut-off point of 29.6 times per million still provided higher coverage in the ASWL corpus and its sub-corpora than the most frequent 2,000 BNC/COCA word-families. Additionally, a list with the frequency cut-off point of 26.9 times per million words included more high frequency, wide ranging, and evenly distributed items outside the most frequent 2,000 BNC/COCA word-families (e.g., *perception, infer*) than lists with higher frequency cut-off points. Given these findings, Dang et al. (2017) chose 26.9 times per million as the minimum frequency of the ASWL words. This criteria was later adopted by Dang (2018) to select the HSWL words. To be consistent with Dang et al. (2017) and Dang (2018) and to allow a fair comparison among lists, in this study, a selected SSWL word family had to appear with a frequency of at least 26.9 times per million words. As the first soft-science spoken corpus has 6.5 million running words, the selected word family then had to have a frequency of at least 175 times in the whole corpus (=26.9 x 6.5).

(3) *Dispersion:* a selected word family had to have Gries's (2008) dispersion (DP) lower than 0.6. This dispersion cut-off point was adopted from the criterion used to select Dang's (2018) HSWL words. Dang compared the items included in or excluded from the HSWL when different DP cut-off points (from 0.1 to 0.9) were chosen. A DP of 0 means perfectly even distribution while a DP of 1 means extremely uneven distribution. Of the various pilot lists,

the list with DP lower than 0.6 had a smaller size but still provided higher coverage in hard-science spoken English than Dang et al.'s (2017) ASWL. Because Dang (2018) aimed to provide hard-science students with a wordlist which is superior to the ASWL, 0.6 was chosen. The current study aims to compare the SSWL with the HSWL, and the dispersion criterion of the HSWL was adopted to select items for the SSWL.

Following Coxhead (2000) and Dang et al. (2017), proper nouns (e.g., *Lucy, John*) and marginal words (e.g., *oh, ah*) were not included in the SSWL but were listed separately because the learning burden of these words may be lighter than other words (e.g., Dang & Webb, 2014; Nation, 2006). The proper nouns and marginal words included (a) items listed in Nation's (2012) supplementary lists of proper nouns and marginal words which are available in the RANGE program package and (b) 1,207 proper nouns (e.g., *Antoni, Issac*) and three marginal words (*amh, blah, haha*) which were not found in these lists but appeared in the corpus.

Items that satisfied these criteria were included in the SSWL. Similar to Dang et al.'s (2017) ASWL and Dang (2018) HSWL, the SSWL was divided into four levels according to Nation's (2012) BNC/COCA lists. Levels 1, 2, and 3 represent the SSWL words from the first, second, and third 1,000 BNC/COCA-word levels. Level 4 includes the SSWL words that are outside the most frequent 3,000 BNC/COCA words. Each level has separate lists of function words (e.g., *but, about*) and lexical words (e.g., *concept, precise*). The lexical words were further divided into 50-item sub-lists based on their frequency in the first soft-science spoken corpus. Previous word lists—Coxhead's (2000) Academic Word List, Dang and Webb's (2016) Essential Word List, Dang et al.'s (2017) ASWL, and Dang's (2018) HSWL—are also divided into sub-lists of 60 or 50 items. The manageable sizes of the sub-lists make it possible for learners and teachers to set short term learning goals as well as incorporating these lists into language learning programs (Nation, 2016). Teaching and learning based on the rank order of sub-lists ensures that the most frequent and wide ranging words are learned first and allow programs to prepare a curriculum that covers all sub-lists and avoids introducing the same items between courses (Dang & Webb, 2016; Coxhead, 2000).

To determine the coverage of the SSWL in the corpus from which it was developed and the four validating corpora, each corpus was run through Heatley, Nation, and Coxhead's (2002) RANGE

program with the SSWL serving as the baseword list. The RANGE program was downloaded from Paul Nation's website: http://www.victoria.ac.nz/lals/about/staff/paul-nation.

The potential coverage that learners may gain with the support of the SSWL was calculated in two ways: (a) when the coverage of proper nouns and marginal words was not counted and (b) when the coverage of these words was counted. The potential coverage (without proper nouns and marginal words) is made up of (a) the coverage provided by the word families that learners already know and (b) the coverage provided by the SSWL word families that are beyond learners' existing levels (Figure 1). Items in the first group are represented by the BNC/COCA word families that are relevant to learners' current vocabulary levels. Items in the second group are the SSWL words that are outside the BNC/COCA words in the first group. Let us take learners with knowledge of the most frequent 1,000 words as an example. The potential coverage is the combination of the coverage provided by word families from the first 1,000 BNC/COCA frequency level and from the SSWL Levels 2-4. The potential coverage (with proper nouns and marginal words) also includes the coverage provided by proper nouns and marginal words. It is important to report this kind of potential coverage because earlier research on the vocabulary load of spoken English (e.g., Dang & Webb, 2014; Nation, 2006) usually added the coverage provided by proper nouns and marginal words to the potential coverage with the assumption that these words have minimal learning burden for learners.

The comparison of the SSWL and HSWL involves examining (1) the number of SSWL and HSWL words at each BNC/COCA level and (2) the overlapping and non-overlapping items between the two lists.

[FIGURE 1 NEAR HERE]

3. Results

3.1. RQ1. Which lexical items occur frequently and are evenly distributed in a wide range of academic speech in soft-sciences?

In the first soft-science spoken corpus, 1,964 word families satisfied the range, frequency, and dispersion criteria (see Appendices C-G for the SSWL levels and sub-lists). Although six was set as the range cut-off point, all SSWL word families appeared in at least eight out of 12 subjects. In fact, 92.67% of the words occurred in all 12 subjects, and 99.54% of the words occurred in at

least 10 subjects. Similarly, although 0.6 was set as the maximum of Gries's (2008) DP, 93.64% of the SSWL word-families had a DP lower than 0.5. Table 7 presents the lexical profile of the SSWL. It shows that 28.31% (556 out of 1,964 word-families) of the SSWL is outside the most frequent 2,000 BNC/COCA words, which represent general high-frequency vocabulary. This proportion is greater than the proportion of words outside the most frequent 2,000 BNC/COCA words in Dang et al.'s (2017) ASWL (26.13%). The SSWL covered 91.73% of the first soft-science spoken corpus, which is much higher than the coverage provided by the most frequent 2,000 BNC/COCA word-families (89.17%) although the former list has 36 items fewer than the latter list. The SSWL also provided around the same amount of coverage in the two sub-corpora: 90.91% (soft-pure) and 92.62% (soft-applied).

[TABLE 7 NEAR HERE]

When tested against an independent soft-science spoken corpus of similar size and structure, the SSWL provided around the same coverage (90.86%) as its coverage in the first soft-science spoken corpus. The most frequent 1,741 SSWL words provided 91.06% and 90.18% of the first and second soft-science spoken corpus, respectively. These coverage figures are higher than the coverage of the 1,741 ASWL words (89.94%, 89.31%). Compared with its coverage in soft-science spoken corpora, the coverage of the SSWL in other validating corpora is lower: 90.49% (hard-science spoken corpus), 86.81% (non-academic spoken corpus), and 84.86% (soft-science written corpus).

3.2. RQ2. With knowledge of these words, how much coverage of academic speech in soft-sciences 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 SSWL. The number of SSWL words that are beyond learners' existing vocabulary level is presented in the second column of the table. The potential coverage that learners may gain if they study the SSWL is presented in the next four columns. Coverage provided by proper nouns (e.g., *Lucy, John*) and marginal words (e.g., *oh, ah*) is shown in the last two rows of the table.

[TABLE 8 NEAR HERE]

Considering their insufficient vocabulary knowledge, learners with knowledge of fewer than the most frequent 1,000 words are not likely to know the SSWL words. If they learn all 1,964 SSWL words, they may be able to reach 91%-92% coverage of soft-science speech. If they know proper nouns and marginal words, the potential coverage for these learners may go up to 94%-95%. This coverage is much larger than the coverage provided by the most frequent 2,000 BNC/COCA word families. Together with proper nouns and marginal words, the most frequent 2,000 BNC/COCA word families only covered 92.07% of the first soft-science spoken corpus and 91.66% of the second soft-science spoken corpus.

With their existing knowledge, learners having mastered the most frequent 1,000 word families may need to study only 1,086 SSWL words which are beyond their level. These word families, however, may enable them to achieve coverage of 91%-92% (without proper nouns and marginal words) and 94%-95% (with proper nouns and marginal words). These potential coverage figures are higher than those achieved by the 1,000 word families from the second 1,000 BNC/COCA word level. It should be noted that learning the SSWL words that are beyond their existing vocabulary levels may also enable these low level learners to achieve reasonable coverage of the non-academic spoken corpus: 92.15% (those having not mastered the most frequent 1,000 words) and 92.74% (those having mastered the most frequent 1,000 words).

Learners having mastered the most frequent 2,000 word families only need to study 556 word families from the SSWL. Yet they may reach potential coverage of 92%-93%. If proper nouns and marginal words are counted, the potential coverage will be 95%-96%. These coverage figures are around the same as the coverage that these learners may gain if they study 1,000 word families from the third 1,000 BNC/COCA word level (96.43%, 95.69%).

Learners having mastered the most frequent 3,000 word families only have to learn 56 SSWL word families, but may achieve potential coverage of 93%-94% (without proper nouns and marginal words) and 96%-97% (with proper nouns and marginal words). If these learners study 1,000 word families from the fourth 1,000 BNC/COCA-word level, they may gain similar coverage: 97.56% and 96.71%. Taken together, if proper nouns and marginal words are counted, the potential coverage ranges from 94% to 97%.

3.3.RQ3. How do these items compare with those from Dang's (2018) HSWL?

Applying the same selection criteria, the SSWL contains a larger proportion of the most frequent 3,000 words (97.15%) than the HSWL (90.41%) (Appendix H). At the first, second, and third 1,000 BNC/COCA word levels, the SSWL had more items than the HSWL while it is the other way around at lower frequency levels. Additionally, the range of frequency levels of items in the SSWL (1^{st}-10^{th} 1,000 word levels) is narrower than in the HSWL (1^{st}- 16^{th} 1,000 word levels).

There were 1,344 word families appearing in both the HSWL and SSWL. Most of them (98.81%) were among the most frequent 3,000 BNC/COCA word families. Of the non-overlapping items, 620 word families are unique to the SSWL, and 251 word families are unique to the HSWL. It is noteworthy that 93.55% of the words unique to the SSWL are among the most frequent 3,000 BNC/COCA word families. In contrast, 54.58% of the words unique to the HSWL are outside the most frequent 3,000 BNC/COCA word families. This trend can be seen clearly from the top 100 lexical words that are unique to each list (Appendix I). All of the top 100 unique SSWL lexical words (e.g., *business, society, debate*) are among the most frequent 3,000 BNC/COCA words while nearly half of the top 100 unique HSWL lexical words (e.g., *gradient, epsilon, sine*) are outside the most frequent 3,000 BNC/COCA words.

4. Discussion

4.1. Is the SSWL a useful list for soft-science students in ESAP programs?

The SSWL is a valuable resource for soft-science students in ESAP programs for four reasons. First, this list provides an accurate reflection of the most frequent and wide-ranging words in soft-science speech. The list consistently provided around 91% coverage in the corpus from which it was developed as well as in an independent corpus of similar size and structure. This indicates that the list truly represents the most frequent and wide-ranging word families in soft-science speech. Additionally, the coverage of the SSWL in the two soft-science spoken corpora was higher than those in the hard-science spoken corpus, the soft-science written corpus, and the non-academic spoken corpus. This finding is consistent with those from previous studies (e.g., Coxhead, 2000; Coxhead & Hirsh, 2007; Gardner & Davies, 2014) which found specialized wordlists provide similar coverage in corpora of similar genres but lower coverage in corpora of different genres.

Second, the SSWL offers benefits to learners from a wide range of soft-science subjects. The list was developed from spoken data from 12 soft-science subjects. Yet, it still covered around the same amount of coverage in the two sub-corpora (soft-pure and soft-applied). Furthermore, the division of subject areas in these sub-corpora was based on Becher's (1989) classification of academic subjects in higher education. Becher's classification has been validated in a wide range of contexts (see Jones, 2011 for a review) and has been adopted to structure academic corpora such as the BASE, BAWE, and Hyland's (2000) academic written corpus. Given the high validity and utility of Becher's (1989) classification, it is expected that the SSWL will be a valuable resource for soft-science students regardless of their specific subject areas and institutional structure.

Third, the SSWL can benefit learners with different vocabulary levels. The list is a shortcut for learners with the vocabulary knowledge of 1,000 word families or lower to achieve adequate listening comprehension. Learning the most frequent 2,000 BNC/COCA words may allow learners with knowledge of fewer than the most frequent 1,000 words to achieve only around 92% coverage of soft-science speech. In contrast, learning only 1,964 word-families from the SSWL may enable these learners to reach 95% coverage. Similarly, learners with knowledge of the most frequent 1,000 words only need to study 1,086 SSWL word families which are beyond their existing vocabulary level to reach 95% coverage. This is much higher than the coverage that these learners may reach if they learn 1,000 word families from the 2nd 1,000 BNC/COCA word families. Research has reported that considerable proportions of L2 learners in different contexts—China (Matthews & Cheng, 2015), Denmark (Henriksen & Danelund, 2015), Indonesia (Nurweni & Read, 1999), Israel (Laufer, 1998), Taiwan (Webb & Chang, 2012), and Vietnam (Nguyen & Webb, 2016)—have not mastered the most frequent 2,000 words, and even the most frequent 1,000 words, after a long period of formal English instruction. Given this fact, the SSWL is a valuable resource for learners with the vocabulary knowledge of 1,000 word families or lower. It is even more meaningful when considering the fact that knowledge of the SSWL also allows these learners to recognize 92%-93% of non-academic speech, an important discourse type.

Learners with knowledge of the most frequent 2,000 words need to study only 556 word families and learners with knowledge of the most frequent 3,000 words only need to study 56 word

families from the SSWL. Yet they can achieve 95%-97% coverage of academic speech, which is similar to the coverage they may gain from learning 1,000 word families at the third 1,000 BNC/COCA word level (for those with knowledge of the most frequent 2,000 words) and 1,000 word families at the fourth 1,000 BNC/COCA word level (for those with knowledge of the most frequent 3,000 words). This is even more meaningful when the results of this study are compared with Dang and Webb's (2014) study. These researchers found that a vocabulary size of 3,000-4,000 word-families is needed to reach 95% coverage of soft-sciences. This means that learners with knowledge of the most frequent 2,000 words would have to study 1,000-2,000 BNC/COCA word families and learners with knowledge of the most frequent 3,000 words would have to study 1,000 word families. Compared with the subsequent BNC/COCA lists, the SSWL better serves these two groups of learners. These learners have to learn a much smaller number of words but are still able to reach more than 95% coverage and even nearly 98% coverage, which are important coverage points for high and stable comprehension (van Zeeland & Schmitt, 2013).

The fourth reason why the SSWL is a valuable resource for soft-science students is that it is more specialized than Dang et al.'s (2017) ASWL. The SSWL has a higher proportion of words outside general high-frequency words (28.31%) than the ASWL (26.13%). Moreover, the most frequent 1,741 SSWL words provided higher coverage in the two soft-science spoken corpora than the ASWL words. Importantly, the SSWL may allow soft-science students to reach around 95% coverage of soft-science speech regardless of their proficiency levels. In contrast, the ASWL only allows learners with knowledge of the most frequent 2,000 words to achieve 95% coverage. It is understandable given the differences in the purposes of the two lists and the nature of the corpora from which they were developed. The ASWL aims to benefit both hard- and soft-science students in EGAP programs. Therefore, the ASWL corpora consisted of data from both hard and soft-sciences. In contrast, the SSWL only targets soft-science students in ESAP programs. Hence, only data from soft-sciences were included in the SSWL corpora. The more homogeneous the texts in a corpus, the larger proportion of specialized vocabulary makes up of the corpus (Chung & Nation, 2004; Nation, 2016). Therefore, it is less challenging to meet 95% coverage of the corpora with the SSWL than with the ASWL.

4.2. What is the nature of vocabulary in academic speech of hard and soft-sciences?

This study sheds light on the nature of vocabulary in hard- and soft-science speech. The common assumption is that vocabulary in academic speech is full of specialized words that are different to general high-frequency vocabulary. Surprisingly, the present study found that more than 90% of the SSWL and HSWL words are among the most frequent 3,000 BNC/COCA word families. Additionally, nearly 99% of the overlapping items between the two lists are at the first, second, and third 1,000 BNC/COCA-word levels. These findings indicate that knowledge of the most frequent 3,000 word families of general English is important for comprehension of spoken texts from both hard- and soft-sciences. This linguistic feature can be explained by considering the nature of academic speech.

First, one distinctive feature of academic speech is 'on-line informational elaboration' (Csomay, 2006). This feature characterizes the situation in which the speakers and listeners share the same contexts. Speakers are under pressure to transfer dense and abstract information in real time production circumstances while the listeners are under pressure to quickly unpack that information. The density of the information and the demands of real time production require academic speakers to quickly choose the vocabulary and linguistic structures[3] that help them to convey the information in a way that is easy for their listeners to process (Biber, 2006; Deroey & Taverniers, 2011). As a result, academic speakers tend to rely on a relatively small set of words (e.g., *look, see, make, guess, need, use*) but use them with extremely high frequencies (Biber, 2006). In this way, they do not need to spend time finding a different word to replace an already used word and can facilitate listeners' comprehension. It explains why the most frequent 3,000 words of general English accounted for a large proportion of the most frequent and wide-ranging words in spoken texts of both hard- and soft-sciences.

Second, one important function of academic speech is classroom management, a function in which the speakers manage organizational matters to ensure that listeners have necessary information about the courses and learning tasks (Biber, 2006; Deroey & Taverniers, 2011). As the language of classroom management is very close to general conversation (Biber, 2006; Csomay, 2006), this may be the reason why the majority of the most frequent and wide-ranging words in hard- and soft-science speech are the most frequent 3,000 words of general English.

This feature can be seen clearly from the following examples of classroom management which were taken from the academic spoken corpora in the present study.

[1] *I'm planning on posting the exam at seven p.m. tonight*

[2] *I hope you can play an active role in learning rather than sit down and wait for the material I give you you have to go to the library and search for the reference books*

Example [1] is the lecturer's exam announcement while Example [2] is the lecturer's expectation about the students' participation in the course. An analysis with the RANGE program showed that all words in these examples are among the most frequent 3,000 BNC/COCA words.

The third reason why the most frequent 3,000 words made up a large proportion of academic speech is related to multi-word units. These items are more common in academic speech than academic writing (Biber et al., 2004; Simpson-Vlach & Ellis, 2010). Coxhead et al. (2017) analyzed multi-word sequences in university labs and tutorials and found that the majority of these sequences were made up of high-frequency words, especially the most frequent 1,000 words. The same patterns were found in the current study. An analysis of the 10 most frequent two-word terms in the first soft-science spoken corpus and the hard-science spoken corpus revealed that most of them have at least one high-frequency word as a component (e.g., *solar system, kinetic energy, teacher development*, *global warming*) (see Appendix J).

The important role of the most frequent 3,000 words of general English in hard- and soft-science speech provides further insight into the ultimate cut-off point of high-frequency vocabulary. The traditional approach suggested that 2,000 items should be the cut-off point (Nation, 2001). However, Schmitt and Schmitt (2014) reviewed research on the vocabulary load of different general written and spoken texts (novels, newspapers, audio narrative stories, general conversation, movies, TV programs) and proposed that the 2,000 cut-off point should be extended to 3,000. By examining the vocabulary in academic spoken discourse, the present study provides further evidence for this extension.

While this study suggests that knowledge of the most frequent 3,000 words is essential for comprehension of spoken texts from both hard- and soft-sciences, it also shows that these words appear to play a more important part in soft-sciences than in hard-sciences. First, the proportion of the most frequent 3,000 words in the SSWL (97.15%) is larger than in the HSWL (90.41%).

Second, the SSWL always had more items at the first, second, and third 1,000 BNC/COCA levels than the HSWL. In contrast, beyond the third 1,000-word level, the SSWL always consists of fewer items than the HSWL. Third, the range of the SSWL words beyond the most frequent 3,000 words (1^{st}-10^{th} 1,000-word levels) is narrower than those of the HSWL (1^{st}-16^{th} 1,000-word levels). Fourth, nearly all items unique to the SSWL are among the most frequent 3,000 BNC/COCA word-families whereas more than half of the number of items unique to the HSWL are outside the most frequent 3,000 words.

The difference in the role of the most frequent 3,000 words and words at lower frequency levels may be the result of the difference in the nature of soft and hard sciences. As previously mentioned, although the hard/soft division cannot fully reflect the complexity and variation in inquiry processes and knowledge structures in various disciplines (Becher & Trowler, 2001; Nesi, 2002), this division is a useful shorthand from attempting to explain the complexity and diversity of academic discourse. While both hard and soft sciences have complex concepts whose meanings may be uncommon in everyday language, the components of these terms may be different between hard and soft sciences. Terms in soft-sciences (e.g., *invisible hand*) may be more likely to comprise of words which are common in everyday language (*invisible, hand*). In contrast, terms in hard sciences (e.g., *radial velocity*) may be more likely to comprise of words which are rare in everyday language but shared among multiple hard science disciplines (*radial, velocity*)[4]. This claim is further supported by the analysis of the 10 most frequent two-word combinations in each subject area of the first soft-science spoken corpus and the hard-science spoken corpus (see Appendix J). Soft-science terms are more likely to be the combinations of two high-frequency words (e.g., *demand curve, fat man, green revolution*) while hard-science terms are normally the combinations of a high-frequency word and words at lower frequency level (e.g., *hash function, convex problem, hydraulic system*) or combinations of low-frequency words (e.g., *harmonic oscillator, amino acid*). The difference in the nature of vocabulary in soft- and hard-science speech supports the findings of previous research that there was substantial variation in the lexical items across disciplines (Durrant, 2014, 2016; Hyland & Tse, 2007). It then highlights the value of discipline-specific wordlists such as the HSWL and the SSWL for learners in highly specific learning contexts.

While the variation in the vocabulary of academic speech between hard- and soft-sciences leads us to favor discipline-specific wordlists, it does not mean that we should dismiss the idea of general academic wordlists. As found in this study, a considerable number of word-families appeared in both the SSWL and the HSWL. Moreover, a further comparison of items in the SSWL and the ASWL showed that 82.28% of the SSWL words are also ASWL words. Dang (2018) also reported that around 90% of the HSWL appeared in the ASWL. These findings mean that the similarities between the spoken vocabulary of hard- and soft-sciences are greater than the differences. Therefore, in the contexts where learners are more heterogeneous or unclear about their disciplinary areas, a general academic wordlist such as Dang et al.'s (2017) ASWL is more practical. In such a context, it is usually challenging for EAP teachers to satisfy the specific needs of every learner in their classes. This study then supports Dang and colleagues' idea that depending on the particular teaching and learning context, either a discipline-specific list or a general academic wordlist can be a more valuable resource for L2 learners (Dang, 2018; Dang et al., 2017) and Hyland's (2016) suggestion that specificity in EAP studies should be implemented with the consideration of the circumstances of particular students in a class. Given that little has been done to compare the vocabulary in academic speech of hard- and soft-sciences and given that there are no spoken wordlists for soft-science students in ESAP programs, the present study has great theoretical and pedagogical value.

5. Pedagogical implications

The SSWL can be integrated with Nation's (2012) BNC/COCA lists to organize a systematic learning program for soft-science students in ESAP programs by following Dang's (2018) model (see Figure 2). According to this model, at the beginning of the programs, teachers can use Webb, Sasao, and Ballance's (2017) New Vocabulary Levels Test (NVLT) to diagnose their learners' vocabulary knowledge. Based on learners' current vocabulary knowledge and their learning purposes, teachers can determine the learning goal and sequence for their learners. For example, if learners would like to go straight to items that occur often in academic speech of soft-sciences, they can start learning items in subsequent levels of the SSWL that are beyond their current vocabulary level. However, if learners would like to expand their knowledge of general vocabulary before moving to specialized vocabulary of soft-sciences, they can learn items from the subsequent levels of the BNC/COCA lists. Once they are satisfied with their knowledge of general vocabulary, they can move to the relevant levels of the SSWL. This model

gives teachers and learners flexibility in selecting the learning goal and sequence that match their particular teaching and learning contexts. Moreover, it avoids repeatedly learning and teaching known items and draws learners' attention to useful words that are beyond their current vocabulary levels.

[FIGURE 2 NEAR HERE]

It is important to note that the NVLT only diagnoses learners' receptive knowledge of form and meaning and indicates when learners are ready to move on to learn new items from the next level of the SSWL and HSWL. It does not mean that once learners have mastered Levels 1-3 of these lists, they should stop enriching their knowledge of these words. Knowing a word involves many aspects (Nation, 2013; Nation & Webb, 2011). As found in this study, the most frequent 3,000 words tend to combine with each other (in the case of soft-sciences) or with words at lower frequency level (in the case of hard-sciences) to reveal complex disciplinary concepts. Additionally, the 'on-line informational elaboration' nature of academic spoken discourse also requires learners to become fluent in their use of these words so that they can convey and process the information quickly. Therefore, apart from helping learners to acquire new words from the SSWL and HSWL Level 4, it is equally important for teachers to create many opportunities for learners to encounter and use the Level 1-3 SSWL and HSWL words in different contexts related to their target disciplines by following Nation's (2007) Four Strands principles. Especially, they should draw learners' attention to the words that these items tend to collocate with such as those identified in Appendix J and their specific meanings. This will allow learners to consolidate and expand their knowledge of these words and make a link between their general use and technical use as well as their collocations.

6. Limitations and future research

This study used hypothetical calculation to estimate the amount of coverage of academic speech that L2 learners may reach with the aid of the SSWL. It assumes that if learners have reached the first, second, and third 1,000-word levels according to the NVLT results, they may have mastered high-frequency words. However, there may be cases in which even advanced learners have not mastered the full range of uses of high-frequency words given the highly polysemous nature of these words. Additionally, this study only investigated the influence of vocabulary on

comprehension in the form of lexical coverage while other factors (e.g., background knowledge, interaction) may also affect comprehension.

Intervention studies with real learners would provide further insight into the actual coverage that learners may achieve with the aid of the SSWL. It would also be useful to further examine multi-words in hard- and soft-science speech and the relationships between vocabulary, comprehension, and other factors influencing listening comprehension (e.g., background knowledge, interaction). Another direction for future research would be to develop spoken wordlists for a particular subject (e.g., engineering, business) for ESP programs.

7. Conclusion

A 1,964 word-family SSWL was developed for L2 learners of soft-sciences in ESAP programs. Knowledge of the SSWL words may allow these learners to achieve 94%-97% coverage of soft-science speech. The list is a useful resource for soft-science students irrespective of their current vocabulary levels and subject areas. This study also shed light on the nature of vocabulary in academic speech of hard- and soft-sciences. It revealed that knowledge of the most frequent 3,000 words of general English is important for comprehension of spoken texts from both hard disciplines and soft disciplines; however, the value of these words is greater in soft-sciences than in hard-sciences.

Acknowledgements

I would like to thank Professor Stuart Webb and Dr. Averil Coxhead for encouraging me to conduct this research project, and Mark Toomer for acting as a reader of the early version of this paper. My thanks to the Editor and the Reviewers for their constructive feedback and 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, Professor Stuart Webb, 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 Michigan Corpus of Academic Spoken English, 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.

Notes

[1] These are the definitions of EGAP and ESAP in the present study. There is some variation in the understanding of these terms in the field of EAP, however.

[2] The other dimensions are pure/applied and life/non-life, which refer to the concern of the areas with (a) application to practical problems and (b) life systems, respectively.

[3] Linguistic structures are beyond the scope of this study.

[4] *radial* is at the 8th 1,000 BNC/COCA word levels but appeared in 11 out of 12 subjects of the hard science corpus; *velocity* is at the 5th 1,000 BNC/COCA word levels but occurred in 10 out of 12 subject areas of the hard-science corpus.

References

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.

Becher, T., & Trowler, P. R. (2001). *Academic tribes and terrritories* (2nd ed.). Philadephia: Open University Press.

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

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at…: Lexical bundles in University teaching and textbooks. *Applied Linguistics*, *25*(3), 371–405.

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 different academic areas. *Journal of Applied Psychology*, *57*(3), 195–203.

Cabré, M. T. (1999). *Terminology: Theory, methods and applications*. Amsterdam: John Benjamins.

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., Dang, T. N. Y., & Mukai, S. (2017). Single and multi-word unit vocabulary in university tutorials and laboratories: Evidence from corpora and textbooks. *Journal of English for Academic Purposes*, *30*, 66–78.

Coxhead, A., & Hirsh, D. (2007). A pilot science-specific word list. *Revue Française de Linguistique Appliqueé*, *12*(2), 65–78.

Csomay, E. (2006). Academic talk in American university classrooms: Crossing the boundaries of oral-literate discourse? *Journal of English for Academic Purposes*, *5*(2), 117–135.

Dang, T. N. Y. (2018). The hard science spoken word list. *ITL – International Journal of Applied Linguistics*, *169*(1), 44–71.

Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The academic spoken word list. *Language Learning*, *67*(4), 959–997.

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.

Deroey, K. L. B., & Taverniers, M. (2011). A corpus-based study of lecture functions. *Moderna Språk*, *2*, 2–22.

Durrant, P. (2014). Discipline and level specificity in university students' written vocabulary. *Applied Linguistics*, *35*(3), 328–356.

Durrant, P. (2016). To what extent is the Academic Vocabulary list relevant to university student writing? *English for Specific Purposes*, *43*, 49–61.

Flowerdew, J., & Miller, L. (1992). Student perceptions, problems and strategies in second language lecture comprehension. *RELC*, *23*(2), 60–80.

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.

Hyland, K. (2000). *Disciplinary discourses: Social interactions in academic writing*. London: Longman.

Hyland, K. (2016). General and specific EAP. In Ken 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-Rodríguez, E. (1964). *Frequency dictionary of Spanish words*. London: Mouton.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: same or different? *Applied Linguistics*, *19*(2), 255–271.

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.

Liu, J., & Han, L. (2015). A corpus-based environmental academic word list building and its validity test. *English for Specific Purposes*, *39*, 1–11.

Martínez, I. A., Beck, S. C., & Panza, C. B. (2009). Academic vocabulary in agriculture research articles: A corpus-based study. *English for Specific Purposes*, *28*(3), 183–198.

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

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. (2001). *Learning vocabulary in another language* (1st ed.). Cambridge: Cambridge University Press.

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 & M. McCarthy (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.

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.

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

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.

Nurweni, A., & Read, J. (1999). The English vocabulary knowledge of Indonesian university students. *English for Specific Purposes*, *18*(2), 161–175.

Pinchbeck, G. G. (2014). Lexical frequency profiling of a large sample of Canadian high school diploma exam expository writing: L1 and L2 academic English. Presented at the

Roundtable presentation at American Association of Applied Linguistics, Portland, OR, USA.

Resche, C. (2012). Towards a better understanding of metaphorical networks in the language of economics: The importance of theory-constitutive metaphors. In H. Herrera-Soler & M. While (Eds.), *Metaphor and mills: Figurative language in business and economics* (pp. 77–102). Berlin: De Gruyter Mouton.

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., & Meara, P. (1997). Researching vocabulary through a word knowledge framework. *Studies in Second Language Acquisition*, *19*(01), 17–36.

Schmitt, N., & Schmitt, D. (2014). A reassessment of frequency and vocabulary size in L2 vocabulary teaching. *Language Teaching*, *47*(4), 484–503.

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

Shin, D., & Nation, P. (2008). Beyond single words: the most frequent collocations in spoken English. *ELT Journal*, *62*(4), 339–348.

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

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.

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., Sasao, Y., & Ballance, O. (2017). The updated Vocabulary Levels Test. *ITL – International Journal of Applied Linguistics*, *168*(1), 34–70.

Wei, Z. (2014). Does teaching mnemonics for vocabulary learning make difference? Putting the keyword method and the word part technique to the test. *Language Teaching Research*, *19*(1), 43–69.

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

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication*, *3*(2), 215–229.

Table 1

Five corpora in the present study

Corpus	Purposes	Size
1st soft-science spoken corpus	Develop the SSWL	6,513,944
2nd soft-science spoken corpus	Validate the soft, academic, spoken nature of the SSWL	6,343,161
Hard-science spoken corpus	Validate the soft science nature of the SSWL	6,515,717
Soft-science written corpus	Validate the spoken nature of the SSWL	6,818,181
Non-academic spoken corpus	Validate the academic nature of the SSWL	6,505,382

Table 2

First soft-science spoken corpus

Soft-pure		Soft-applied	
Subjects	*Words*	*Subjects*	*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 3

Second soft-science spoken corpus

Soft-pure		Soft-applied	
Subject	*Words*	*Subject*	*Words*
Anthropology*	53,903	Architecture & Design*	103,417
Archeology*	10,382	Economics	1,907,942
Art	204,759	Education	63,910
Classic Studies*	290,367	Film, Theater, Music*	71,022
Communication*	124,335	Law	895,181
Cultural Studies*	345,588	Nursing*	119,903
English & Literature	220,456	Textiles & Clothing*	13,829
Gender Studies*	34,451		
Geography*	204,806		
History	782,628		
Journalism*	34,463		
Linguistics*	86,692		

Philosophy	65,370		
Political Science	67,401		
Psychology	28,181		
Religious Studies*	398,652		
Sociology*	215,523		
Total	*3,167,957*	*Total*	*3,175,204*

* *Subjects that are not represented in the first academic spoken corpus*

Table 4

Hard-science spoken corpus

Hard-pure		Hard-applied	
Subject	Size	Subject	Size
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 5

Soft-science written corpus

Soft pure		Soft applied	
Subjects	Words	Subjects	Words
Anthropology	110,084	Architecture	20,449
Archeology	184,828	Business	319,167
Classic Studies	201,195	Economics	214,940
Cultural Studies	211,260	Education	1,249,258
English	262,155	Law	405,044
History	286,184	Management	738,946
Linguistics	253,306	Public Policies	56,479
Philosophy	247,281		
Political Studies	1,585,357		
Psychology	195,616		
Sociology	276,632		
Total	*3,813,898*	*Total*	*3,004,283*

Table 6

Non-academic spoken corpus

Corpus	Main variety of English	Words
International Corpus of English (spoken, non-academic)	Indian, Pilipino, Singapore, Canadian, Hong Kong, Irish, Jamaican & New Zealand	5,262,502
TV program corpus	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 Soft Science Spoken Word List

SSWL level	BNC/COCA word level	Number of word-families	Additional coverage (%)	Examples
Level 1	1st 1,000	878	82.23	*Court, business, alright,*
Level 2	2nd 1,000	530	5.85	*demand, approach*
Level 3	3rd 1,000	500	3.43	*fundamental, principle*
Level 4	4th 1,000	45	0.17	*norm, hierarchy, dilemma*
	5th 1,000	5	0.02	*Optimal, plausible*
	6th 1,000	3	0.01	*Intuition*
	7th 1,000	1	0	*Syllabus*
	10th 1,000	1	0.01	*Semester*
	Outside BNC/COCA25000	1	0.01	*So-called*
	Total	1,964	91.73	

Table 8

Potential coverage gained by learners with the aid of the SSWL (%)

Learners with knowledge of	Number of SSWL beyond learners' level	Without proper nouns & marginal words		With proper nouns & marginal words	
		1st corpus	2nd corpus	1st corpus	2nd corpus
Fewer than 1,000 words	1,964	91.73	90.86	94.63	93.85
1,000 words	1,086	92.06	91.23	94.96	94.22
2,000 words	556	92.82	91.94	95.72	94.93
3,000 words	56	93.75	92.85	96.65	95.84
Proper nouns		1.80	2.07		
Marginal words		1.10	0.92		

Figure 1

Sources of the potential coverage for different groups of learners

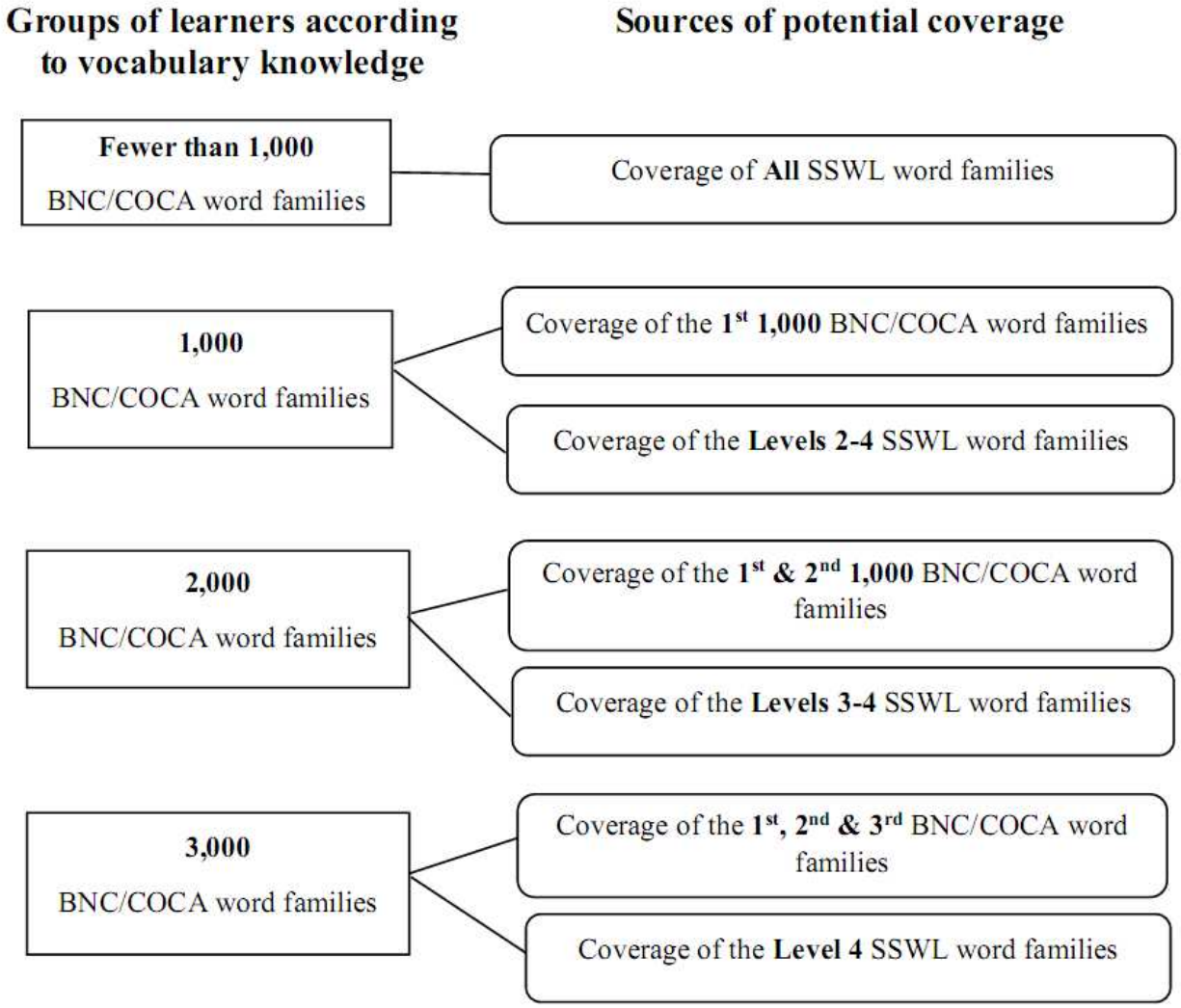

Figure 2

Learning sequence for soft-science students in ESAP programs