Article:
Dang, TNY orcid.org/0000-0002-3189-7776 (2020) High-frequency words in academic spoken English: Corpora and learners. ELT Journal, 74 (2). pp. 146-155. ISSN 0951-0893

https://doi.org/10.1093/elt/ccz057

Dang, T. N. Y. (in press). High-frequency words in academic spoken English: Corpora and learners. *ELT Journal.* Accepted on 26 October 2019. doi: 10.1093/elt/ccz057

High-frequency words in academic spoken English: Corpora and learners

Thi Ngoc Yen Dang

EAP teachers and course designers usually assume that learners have already mastered the most frequent words of general language before entering their courses. Therefore, they focus on words that are outside high-frequency vocabulary but common in academic written English. This fixed vocabulary benchmark is questionable given EAP learners' varied language proficiency. A good understanding of academic spoken English is certainly important for students' academic success. Concentrating only on academic written vocabulary for comprehending academic spoken English may therefore be problematic. No attempts have been made to address these issues in a single study. This study compared the occurrences of high-frequency words in academic spoken and written English. It also tested EAP learners' receptive knowledge of these words. Results showed that high-frequency words play a very important part in academic spoken English, but most participants failed to master them receptively. Suggestions for enhancing EAP learners' knowledge of high-frequency words are provided.

Introduction

High-frequency words, which are represented in Nation's (2012) list of the most frequent 2,000 BNC/COCA words (BNC/COCA2000), are words that L2 learners may encounter and use very often in different contexts of everyday language such as newspapers, telephone conversations, emails, and television programmes (Nation 2013). Examples of high frequency words are *think, know,* and *good*. Because high-frequency words are the basis for general conversation, EAP teachers and course designers usually assume that learners already have a good knowledge of these words before entering EAP courses. Therefore, many EAP courses are likely to focus on words that are outside high-frequency vocabulary but appear very often in academic written English. For example, many EAP programmes use Coxhead's (2000) Academic Word List as a guide to set the learning goals and design learning materials for their students. The Academic Word List consists of 570 words that are not high-frequency words but occur very often in academic written English.

However, this practice can be questioned in several ways when we consider the current development of English-medium programmes. First, students in these programmes have to comprehend not only their reading materials but also lectures, seminars, lab discussions, and tutorials. Yet L2 learners expressed difficulty in comprehending academic spoken English, and insufficient vocabulary knowledge was frequently cited as a big reason for this difficulty (Soruç and Griffiths 2018). Given this situation, it is important for EAP programmes to help learners master the words that they are likely to encounter often in academic spoken English. Dang and Webb (2014) investigated the occurrence of Coxhead's (op.cit.) Academic Word List in academic lectures and seminars and suggested that knowledge of words that are common in academic written English but are outside high-frequency words may not be sufficient for students to comprehend academic spoken

English. Dang, Coxhead, and Webb (2017) found that high-frequency words accounted for more than 70 per cent of the most frequent words in academic spoken English. This situation results in questions around the EAP practice of focusing only on words that are common in academic written English but are outside high-frequency vocabulary. In other words, it is essential to investigate the relative value of high-frequency words in academic spoken and written English.

Second, English is widely used as a medium of instruction at universities throughout the world. This means that students arrive on EAP courses with varied vocabulary knowledge (Clarke 2018), and many of them may not have sufficient knowledge of high-frequency words. For example, Akbarian (2010) measured the receptive vocabulary knowledge of 112 EAP learners in Iran and found that over 76 per cent of them had not mastered the most frequent 2,000 words in West's (1953) General Service List. Similarly, Matthews and Cheng's (2015) study with 167 EAP students in China revealed that these learners only receptively knew about 77 per cent of the most frequent 2,000 BNC/COCA words. These findings indicated that it is worth reconsidering the assumption that every EAP learner has mastered high-frequency words receptively before entering EAP programmes. As EAP programs vary according to context (Hyland and Shaw 2016), research with learners in other contexts would provide a further insight into EAP learners' receptive knowledge of high-frequency words.

No attempts have been made to address these two issues in a single study. The present study was conducted to fill this gap. It compared the occurrence of high-frequency words in academic spoken English with that in academic written English. Also, it measured the receptive knowledge of high-frequency words of 66 learners in an EAP programme at a university in Vietnam. In other words, the study aims to address three questions:

1. What percentage of words in academic *spoken* English are high-frequency words?
2. What percentage of words in academic *written* English are high-frequency words?
3. To what extent do EAP students know high-frequency words receptively before attending their EAP course?

The findings would provide EAP teachers and course designers with a better insight into the value of high-frequency words for their learners, especially those in EFL contexts.

Methodology

Corpus analysis

The RANGE programme was used to analyze the occurrences of high-frequency words in academic spoken English and academic written English. This is a popular program to count frequency of words in texts, which is available at Paul Nation's website: http://www.vuw.ac.nz/lals/staff/paul-nation/ nation.aspx. Nation's (2012) list of the most frequent 2,000 BNC/COCA words (BNC/COCA2000) was used to represent high-frequency words. This list represents the most common words in a range of spoken and written discourses such as movies, telephone conversation, TV programmes, newspapers, and texts for young children.

The academic spoken corpus and the academic written corpus developed by Dang et al. (op.cit.) were used to represent academic spoken English and academic written English, respectively. As can be seen from Tables 1 and 2, each corpus consists of over 13 million words and is divided into four disciplinary sub-groups. Each sub-group has around 3.25 million words. Materials included in the academic spoken corpus were selected from 11 sources which represent lectures, seminars, lab discussion, and tutorials as recorded in various institutions around the world and spoken in multiple varieties of English. Materials included in the academic written corpus were textbooks, journal articles, book chapters, students writing, and research reports from courses at various institutions. Further information about the corpora can be found at https://osf.io/tp687/.

[TABLES 1 & 2 NEAR HERE]

Vocabulary test

A total of 66 first-year students (aged 18 years old) in a one-year EAP course at a university in Vietnam took part in the study. After completing the one-year course, these students were going on to study 40 per cent of their academic subjects (Engineering and Technology) in English. To measure these students' receptive knowledge of high-frequency words, a vocabulary test was delivered at the beginning of their EAP course as part of the entry test. The vocabulary test was comprised of the 1,000 and 2,000 word level sections in Webb, Sasao, and Ballance's (2017) Updated Vocabulary Levels Test. The Updated Vocabulary Levels Test has five levels (1000, 2000, 3000, 4000, and 5,000). The 1,000 and 2,000 word levels measure receptive knowledge of the most frequent 2,000 BNC/COCA words, which represent high-frequency vocabulary. The remaining levels measure receptive knowledge of words at the lower frequency. Each level has 10 clusters. Each cluster has six words and three definitions. Test takers have to select three words to match the three definitions (see Figure 1 for the example of test items). To master a level, test takers need to score at least 29 out of 30 correct answers. The Updated Vocabulary Levels Test has recently been validated and widely used to measure L2 learners' receptive vocabulary knowledge. Therefore, it is expected that results of the vocabulary test in the present study can provide an accurate assessment of the students' receptive knowledge of high-frequency words.

[FIGURE 1 NEAR HERE]

Results & Discussion

Nature of vocabulary in academic spoken English

In answer to the first and second research questions, the BNC/COCA2000 accounted for 88.61 per cent of the words in the academic spoken corpus. This figure was much higher than the percentage of these words in the academic written corpus (76.39 per cent), which indicates that high-frequency words made up a much larger proportion of words in academic spoken English than in academic written English. To illustrate this finding, let us consider Excerpt 1, which was taken from the academic spoken corpus, and Excerpt 2, which was from the academic written corpus.

> [1] *So the one sentence summary of this class is that this is about efficient procedures for solving problems on large inputs. And when I say large inputs, I mean things like*

*the US highway system, a map of all of the highways in the United States; the human **genome**, which has a billion letters in its **alphabet**; a social network responding to Facebook, that I guess has 500 million **nodes** or so. So these are large inputs.*

*[2] Logical **deductions** or **inference** rules are used to **prove** new **propositions** using previously proved ones. A fundamental **inference** rule is **modus ponens**. This rule says that a **proof** of P together with a **proof** that P **implies** Q is a proof of Q. **Inference** rules are sometimes written in a funny **notation**. For example, **modus ponens** is written as follows.*

High-frequency words, which were unmarked in these excerpts, account for a larger proportion of words in Excerpt 1 than in Excerpt 2. The larger proportion of known words in a text, the better comprehension is (Nation, op.cit.).The finding of the present study suggests that from the vocabulary perspective, high-frequency words play a much more important role in enhancing L2 learners' comprehension of academic spoken English than they do in academic written English.

There are several possible reasons for the significant role of high-frequency words in academic spoken English. To begin with, technical concepts in academic speech tend to consist of one or more high-frequency words. Additionally, in academic spoken English, speakers have to transfer dense and abstract information under real time production circumstances while listeners have to unpack such information quickly. The density of the information and the pressure of transferring such information in a quick but easy for listeners to understand may make speakers rely more on high-frequency words. As can be seen from Excerpt 1, the technical concept *large inputs* is made up of two high-frequency words *large* and *inputs*. The lecturer used examples from everyday life to explain this concept and relied mainly on high-frequency words to express his ideas.

The great importance of high-frequency words in academic spoken English can also be explained by the function of academic speech. One important function is classroom management, which refers to situations where course instructors provide students with important information about the courses and learning tasks. This feature can be demonstrated in Excerpt 3:

[3] *What I want to do today is spend literally a minute or so on administrative details, maybe even less. What I'd like to do is to tell you to go to the **website** that's listed up there and read it. And you'll get all information you need about what this class is about from a **standpoint** of **syllabus**; what's expected of you; the problem set schedule; the **quiz** schedule; and so on and so forth.*

In this excerpt, the lecturer covered important administrative details of the course, and most of his words were high-frequency words (unmarked words).

EAP learners' knowledge of high-frequency words

In answer to the last research question about EAP learners' receptive knowledge of high-frequency words, results of the vocabulary test showed that only less than one-fifth of the participants had mastered the most frequent 2,000 words receptively. Of particular

concern, more than 20 per cent of the participants had not mastered even the most frequent 1,000 words receptively (see Table 3).

[TABLE 3 NEAR HERE]

It is important to note that the vocabulary test only measured students' ability to match the forms of high-frequency words with their meanings. Knowing a word involves many other aspects, not just the receptive knowledge of form-and-meaning relationship (Nation op.cit.). For example, learners need to know which words the selected word tends to combine with, how the word changes its forms, and how it is used in different contexts. Because the receptive knowledge of form-and-meaning relationship is the most basic aspect of vocabulary knowledge (Nation op.cit.), these learners' knowledge of other aspects of high-frequency words is likely to be even lower. The finding of the present study is consistent with findings of studies with EAP learners in other contexts such as Iran (Akbarianop.cit.) and China (Matthews and Cheng op.cit.). Together they suggest that it may be more realistic to expect a reasonable proportion of EAP learners, especially those in EFL contexts, to have insufficient knowledge of high-frequency words when entering EAP programmes.

Taken as a whole, the corpus analysis and the vocabulary test indicate that knowledge of high-frequency words is essential for comprehension of academic spoken English; therefore, EAP learners should have a strong knowledge of these lexical items. Yet a reasonable number of EAP learners lack knowledge of these words. This situation calls for support from EAP researchers, teachers, course designers, and material designers to help these learners deal with vocabulary in academic spoken English. One useful tool to solve this problem is Dang et al.'s (op.cit.) Academic Spoken Word List. To the best of my knowledge, this list and Simpson-Vlach and Ellis's (2010) Academic Formulas List are among the very few resources to help EAP learners deal with vocabulary in academic spoken English. In the next section, I will describe some key features of the ASWL and provide some suggestions for implementing the ASWL in EAP programmes.

Pedagogical implications

The Academic Spoken Word List

The Academic Spoken Word List, which is available at https://osf.io/gwk45/, was developed from a 13 million word corpus of academic spoken English and carefully validated on other corpora of similar sizes. The list represents 1,741 words that occur frequently in academic speech from a wide range of academic subject areas. The development of the ASWL took into account the role of high-frequency words in academic spoken English by including these words in the list, as long as they occur frequently in academic speech. In particular, the ASWL words are divided into four levels based on their frequency in general language (see Table 4).

[TABLE 4 & 5 NEAR HERE]

ASWL words at Level 1 are also in the most frequent 1,000 general words from the BNC/COCA. ASWL words at Level 2 are also in the 1001st-2000th most frequent general words from the BNC/COCA. Those at Level 3 are also in the 2001st-3000th most frequent

general words from the BNC/COCA. ASWL words at Level 4 are outside the most frequent 3,000 words of general vocabulary. Including high-frequency words in the ASWL highlights the significant role of these words in academic spoken English. Depending on their current knowledge of general vocabulary, learners may be able to recognize from 92 to 96 per cent of words in academic spoken English (see Table 5). Let us consider how the ASWL words can be used to help EAP learners with insufficient knowledge of high-frequency words to learn the vocabulary needed for comprehension of academic spoken English.

ASWL in setting learning goals and designing learning materials

At the beginning of EAP programmes, teachers and course designers should identify students who have insufficient knowledge of high frequency words so that they can provide these students with necessary support. They can use Webb et al.'s (2017) Updated Vocabulary Levels Test to test students' receptive knowledge of general vocabulary. Based on the test results and the learning contexts, they can set relevant learning goals for their students by using Dang et al.'s (op.cit.) model as a guide (Figure 2).

[FIGURE 2 NEAR HERE]

Although Dang et al.'s (op. cit) model can be applied for learners with higher levels of general vocabulary, in this article I will focus on those having not mastered the most frequent 2,000 words receptively. Their insufficient knowledge of high-frequency words means these learners need special support from teachers and course designers to successfully deal with academic speech in their future study. Based on their scores in the Updated Vocabulary Levels Test, these learners can be divided into two groups: (a) those having not yet mastered the most frequent 1,000 words and (b) those having mastered the most frequent 1,000 words but yet to master the 1001st-2000th most frequent general words.

Let us look at Figure 2 to see possible learning sequences for each group. As shown in the left hand side of the figure, given their poor vocabulary knowledge, ideally these students should focus on the first, second, and even third 1,000 BNC/COCA words before moving on to the relevant ASWL word levels. This learning sequence would enable them to build up a firm vocabulary foundation for any further vocabulary development, not just academic spoken vocabulary.

However, considering the time constraints of some English language programmes such as the programme of the participants in the present study, another option is that students can go straight to the ASWL level which is relevant to their current vocabulary level (see the right hand side of Figure 2). That is, students who have not mastered the most frequent 1,000 words receptively can go straight to academic spoken vocabulary by progressing gradually through Level 1 to Level 4 of the ASWL. Similarly, students who have mastered the most frequent 1,000 words but yet to master the 1001st-2000th most frequent general words receptively may start learning words from Level 2 of the ASWL and then go to Level 3 and Level 4. Because ASWL words at Levels 1 and 2 are also high-frequency words, this approach allows these students to acquire the most important words in academic speech while still helping them to build up their knowledge of high-frequency vocabulary.

After the learning goals have been identified, course designers and teachers need to ensure that their classroom programmes include a clear focus on the development of this vocabulary both receptively and productively. Nation's (2007) Four Strands of meaning-focused input (to develop receptive knowledge), meaning-focused output (to develop productive knowledge), language-focused learning, and fluency development (to develop both receptive and productive knowledge) can be used to guide the learning materials and activities (see Coxhead, 2014 for activities following the four strands).

EAP textbooks, material designers, and teachers can also use the ASWL words to guide the selection and design of learning materials. They can run the transcripts of their listening materials with the ASWL as the base word lists though RANGE, which is available at Paul Nation's website https://www.victoria.ac.nz/lals/about/staff/paul-nation or AntWord Profiler, which is available at Laurence Anthony's website http://www.laurenceanthony.net/software/antwordprofiler/. In this way, they can see the occurrences of the ASWL in their materials and evaluate how well the vocabulary in these materials reflects the vocabulary in academic spoken English. This can help to guide them in designing materials and textbooks for EAP learners whose target academic programmes consist of lectures, seminars, labs, and tutorials.

Raising learners' awareness of the importance of ASWL words

Apart from using the ASWL as a guide in setting learning goals, selecting and designing learning materials and activities, it is equally important for teachers to raise learners' awareness of the value of the ASWL words for comprehending academic speech. First, they can provide students with opportunities to share their assumptions about the most frequent words in academic spoken English. For example, to what extent is it similar to/different from vocabulary in academic written English and everyday language?

Then, they can refer to the findings of the present study as well as Tables 4 and 5 to raise learners' awareness of the importance of high-frequency words in academic spoken English and how the ASWL can help them enhance their knowledge of high-frequency words in academic spoken English. Also, they can take an excerpt from academic speech, highlight the ASWL there and show this to their students. This way helps students realise how often the ASWL words appear in the text. Provide research-based evidence as well as examples from the text will help their suggestions more convincing.

To raise learners' awareness of the specialized uses of the ASWL, teachers should encourage learners to use the ASWL words as search terms and check the concordance of these words in the British Academic Spoken English corpus (https://www.sketchengine.eu/british-academic-spoken-english-corpus/) or the Michigan Corpus of Academic Spoken English (https://quod.lib.umich.edu/cgi/c/corpus/corpus?c=micase;page=simple). This way helps learners to be aware of the technical meanings and uses of these words. In the context where learners' language proficiency is low or technology is a problem, teachers can select some target ASWL words, search for the concordances of the words and select a number of concordance lines which are relevant to learners and include them in the worksheets.

Conclusion

This study challenges an assumption about high-frequency words that is held by many EAP teachers and course designers. It reveals that knowledge of high-frequency words is essential for comprehending academic spoken English, but a reasonable number of learners have insufficient receptive knowledge of these words when entering EAP programmes. Therefore, this study calls for a greater emphasis on high-frequency words from EAP teachers and course designers. It also (a) introduces the ASWL, a useful vocabulary resource to help EAP learners with insufficient receptive knowledge of high-frequency words to deal with academic spoken English and (b) provides some suggestions for implementing the ASWL words in EAP programmes. This study, however, cannot address the demand for hand-on materials, which is a potential area for future research and material development.

References

Akbarian, I. 2010. 'The relationship between vocabulary size and depth for ESP/EAP learners'. *System* 38: 391–401.

Clarke, D. 2018. 'Student responses to vocabulary learning strategies on an ESAP course'. *ELT Journal Volume* 72/3: 319–328.

Coxhead, A. 2000. 'A new academic word list'. *TESOL Quarterly* 34/2: 213–238.

Coxhead, A. 2014. *New ways in teaching vocabulary*. Alexandria: TESOL Inc.

Dang, T. N. Y., **A., Coxhead**, and **S. Webb.** 2017. 'The academic spoken word list'. *Language Learning* 67/4: 959-997.

Dang, T. N. Y. and **S. Webb.** 2014. 'The lexical profile of academic spoken English'. *English for Specific Purposes* 33: 66-76.

Hyland, K. and **P. Shaw.** 2016. 'Introduction' in K. Hyland & P. Shaw (eds.). *The Routledge handbook of English for Academic Purposes* (pp. 1–13). London: Routledge.

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

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* (Second edition). Cambridge: Cambridge University Press.

Simpson-Vlach, R., and **N. C. Ellis.** 2010. 'An academic formulas list: New methods in phraseology research'. *Applied Linguistics* 31: 487–512

Soruç, A., and C. Griffiths. 2018. 'English as a medium of instruction: Students' strategies'. *ELT Journal*, 72/1: 38–48.

Webb, S., Y. Sasao. and **O. Ballance.** 2017. 'The updated Vocabulary Levels Test'. *ITL – International Journal of Applied Linguistics* 168/1: 34–70.

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

The author

Thi Ngoc Yen Dang is a Lecturer in Language Education at the University of Leeds. She obtained her PhD from Victoria University of Wellington, New Zealand. Her research interests include vocabulary studies and corpus linguistics. Before joining Leeds, she had many years teaching General English and English for Academic/Specific Purposes courses in Vietnam. Her articles have been published in *Language Learning, English for Specific Purposes,* and *Journal of English for Academic Purposes*.

Email: T.N.Y.Dang@leeds.ac.uk

Tables

Table 1. Composition of the academic spoken corpus

Hard-pure		Hard-applied		Soft-pure		Soft-applied	
Subject	Words	Subject	Words	Subjects	Words	Subjects	Words
Astronomy	593,062	Chemical Engineering	563,938	Art	553,160	Business	513,133
Biology	552,452	Computer Sciences	555,175	Cultural Studies	498,393	Economics	610,998
Chemistry	556,138	Cybernetics	555,401	History	554,214	Education	571,023
Ecology & Geology	555,312	Electrical Engineering	550,181	Philosophy	549,577	Law	616,398
Mathematics	450,481	Health & Medical Sciences	470,795	Political Studies	545,059	Management	461,093
Physics	554,178	Mechanical Engineering	558,604	Psychology	555,880	Public Policy	485,016
Total	3,261,623	Total	3,254,094	Total	3,256,283	Total	3,257,661

Table 2. Composition of the academic written corpus in terms of disciplines

Hard pure		Hard applied		Soft pure		Soft applied	
Subjects	Words	Subjects	Words	Subjects	Words	Subjects	Words
Astronomy	293,720	Agriculture	425,647	Anthropology	110,084	Architecture	20,449
Biology	341,250	Civil engineering	430,706	Archeology	184,828	Business	319,167
Chemistry	122,283	Computer science	191,735	Classic Studies	201,195	Economics	214,940
Ecology & Geology	275,173	Cybernetics	86,208	Cultural Studies	211,260	Education	1,249,258
General Sciences	1,195,124	Electrical engineering	576,810	English	262,155	Law	405,044
Mathematics	688,465	General Engineering	720,587	History	286,184	Management	738,946
Physics	183,776	Health & Medicine	398,153	Linguistics	253,306	Public Policies	56,479
		Material Engineering	155,905	Philosophy	247,281		
		Mechanical engineering	382,337	Political Studies	1,585,357		
		Media Art & Science	120,796	Psychology	195,616		
		Meteorology	42,728	Sociology	276,632		
Total	3,099,791	Total	3,531,612	Total	3,813,898	Total	3,004,283

Table 3. Vocabulary levels of the 66 participants

Vocabulary level	Number of students
Mastered the most frequent 2,000 words	13
Mastered only the most frequent 1,000 words	39
Yet to master the most frequent 1,000 words	14

Table 4. The lexical profile of the Academic Spoken Word List

ASWL level	BNC/COCA word level	Number of items	Examples
Level 1	1st 1,000	830	*alright, agree, though, maybe, stuff*
Level 2	2nd 1,000	456	*example, identify, determine, therefore*
Level 3	3rd 1,000	380	*define, potential, focus, versus, achieve*
Level 4	Outside the most 3,000 BNC/COCA words	75	*arbitrary, scenario, synthesis, maximise*
Total		1,741	

Table 5. Potential coverage gained by learners with the aid of the ASWL (%)

Current vocabulary knowledge (BNC/COCA word-families)	Potential coverage
Less than 1,000	92%-93%
1,000	93%
2,000	94%
3,000	95%-96%

Figure 1. Example of items in the vocabulary test

	game	island	mouth	movie	song	yard
land with water all around it		✓				
part of your body used for eating and talking			✓			
piece of music					✓	

Figure 2. Dang, Coxhead, & Webb's (2017, p.29) vocabulary learning sequences for different groups of learners