Vocabulary Distribution in a High-School English Textbook in Vietnam

Linh Ai Tran
Phu Quoc High School, Vietnam

Abstract
This paper aims to describe the distribution of the vocabulary presented in reading texts and listening transcripts from the new English textbook for high school students in 11th grade in Vietnam. A mini-corpus of all the reading and listening texts from the textbook was analyzed and compared with Google Books and data from the Corpus of Contemporary American English (COCA). The analysis focused on the distribution of the most frequent function and content words as well as the collocation of the three most typical content words (*have, parents* and *people*) appearing in the textbook. An analysis of the concordance from the COCA data is carried out to see whether the use of the vocabulary in the new English textbook for 11th grade students is in line with the English used in the United States.

Vocabulary in Language Acquisition

Vocabulary acquisition has a great effect on higher-level language processing (Adams and Collins, 1977; Chall, 1987). McCarthy and Cart (2013) considered the learning of vocabulary as the heart of language acquisition as it is the "social and linguistic structure of language." In other words, learners of language need a foundation of basic vocabulary in order to acquire a language (Spada, 2006).

It is clear that vocabulary is essential in language learning, but does a language learner have to acquire every word? The answer comes from the studies which focus on the significance of word frequency in language acquisition. Nation (2009) discussed the role of teaching vocabulary in intensive reading. He suggested that one of the principles for vocabulary teaching is the priority of highly frequent words. Therefore, language learners and teachers should pay considerable attention to vocabulary with high frequency (Nation, 2001).

Moreover, successful vocabulary acquisition requires a high level of exposure to the items. Spada (2006) believed "[among] the factors that make new vocabulary more easily learnable by second language learners is the frequency with which the word is seen, heard and understood" (p. 98). She suggested an estimate of at least sixteen times of exposure to a word for learners to acquire it. Nation (2001)'s study also suggested that frequent engagement in new words is necessary for learners to acquire language. As a result, high frequency words and the frequent exposure to them should be emphasized for language learners.

The New English Textbooks for Grade 11 in Vietnam

The new English textbook is written for students with the A2 English level and to be taught by teachers with C1 level of English. The textbook was published by Vietnamese Publish House of Education with the professional and technical cooperation of Pearson and British Council. It aims to develop students' communication skills in English through improving such skills as reading, speaking, listening, writing and grammar. The learner's own life experiences are the main focus in this book. "Our lives" and "our society" are the two themes employed in the book, which is divided into 10 units: The Generation Gap, Relationships, Becoming Independent, Caring For Those In Need, Being Part Of ASEAN, Global Warming, Further Education, Our World Heritage Sites, Cities Of The Future, and Healthy Lifestyle and Longevity (Hoang et al., 2014). The book is in the early days of adoption, so it is currently used in only

Tran, L. A. (2015). Vocabulary distribution in a high-school English textbook in Vietnam. *Hawaii Pacific University TESOL Working Paper Series 13*, 77-85.
Website: http://www.hpu.edu.
* Email: linhtranhpu@gmail.com. Address: TESOL Program, MP 441, 1188 Fort Street Mall, Honolulu, HI 96813, USA.

some high schools throughout Vietnam. In the near future, this new English textbook is going to be applied widely around Vietnam after some revision. This paper aims to describe and analyze the vocabulary in this textbook.

Research Questions

1. What are the most frequent content words in the new English 11th-grade textbook in Vietnam?
2. What are the collocations of the most frequent content words in this textbook?

Data and Method

A corpus is an electronically stored collection of language which occurs naturally in mainly spoken and/or written form (Reppen, 2010). For the analysis of language distribution in the new textbook of English for Vietnamese students in grade 11, three corpora were employed. The first one is a mini corpus consisting of reading and listening passages taken from Grade 11 English (author, year), totaling 2,717 words and analyzed by using the software Antconc.3.2.4m (Anthony, 2011). Second, the Google Books collection (retrieved from https://books.google.com/ngrams) is consulted to get a sense of the usage of the three most frequent content words. Google Books is an electric database of full text of over 25 million books and magazines (in October 2015). And finally, data from the Corpus of Contemporary American English (COCA) (retrieved from http://corpus.byu.edu/coca/) is employed for the comparison of the frequency and collocation of the target words. COCA is the largest electric online storage of American English containing over 450 million words from authentic contexts (newspapers, magazines, academic texts, transcripts of TV and radio shows, and novels) between 1990 and present.

Findings

Vocabulary Frequency

Function Words

The mini corpus shows that in the new Grade 11 English textbook, function words such as *the*, *and*, *to*, *of*, *in*, and *a* are the top six words in the vocabulary frequency list. While *the*, *and*, *to*, *of*, *a* and *in* are the most frequent in reading texts (by frequency rank), those in the listening ones are *and*, *to*, *the of*, *in*, and *a* (See Figure 1).

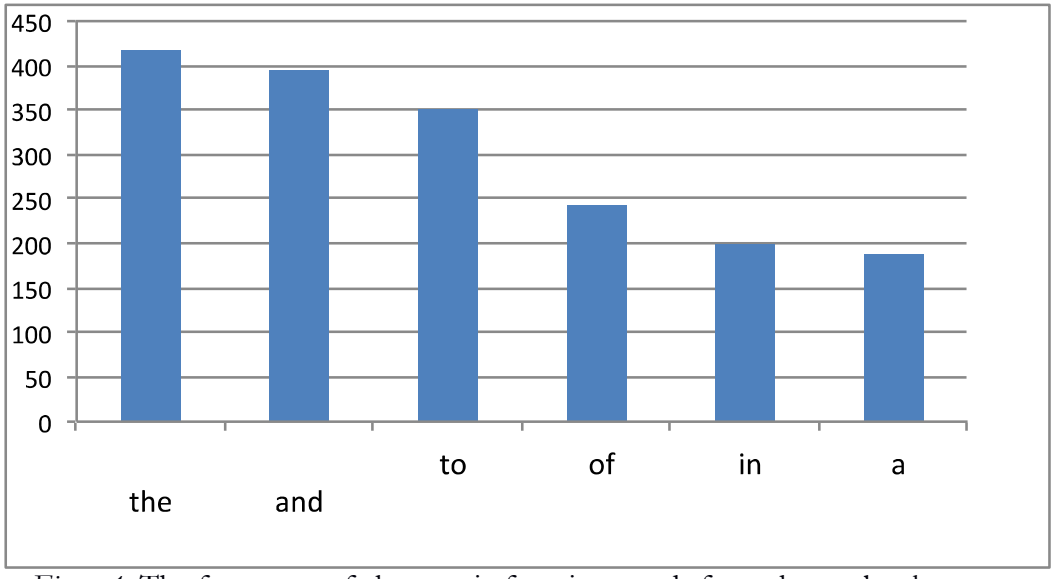

Figure 1. The frequency of the top six function words from the textbook corpus

Meanwhile, the top seven words in the COCA word frequency list (retrieved from http://www.wordfrequency.info/free.asp?s=y) include *the*, *be*, *and*, *of*, *a*, *in*, and *to* (Figure 2).

Frequency			
rank	word/lemma	PoS	frequency
1	the	a	22038615
2	be	v	12545825
3	and	c	10741073
4	of	i	10343885
5	a	a	10144200
6	in	i	6996437
7	to	t	6332195
8	have	v	4303955
9	to	i	3856916
10	it	p	3872477

Figure 2. The top ten words in COCA

Thus, the most frequent function words in the textbook appear in the top seven of the COCA word frequency list. In this regard, the textbook vocabulary seems to pattern with the American English corpus.

Content Words

Analysis of the textbook reading and listening passages shows that the top content words are *have, people, and parents*. Figure 3 shows the frequency of these words in Google's book collection. Generally, *have* is used far more than *people* and *parents*, especially in the 1820s. Its frequency dropped dramatically from 1840 to 1980, and it has appeared consistently since 1980. Meanwhile *people* and *parents* are steadily used in books through time, with *people* being more frequently used than *parents*.

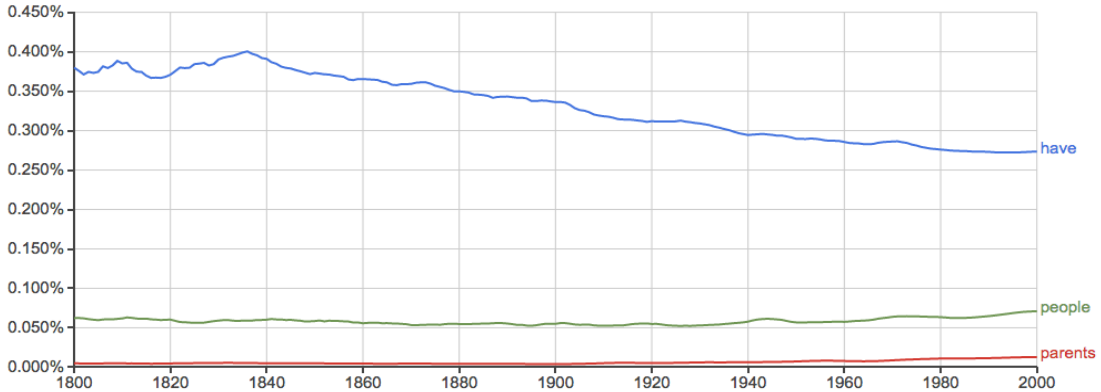

Figure 3. The frequency of *have, people,* and *parents* in Google Books (Ngram Viewer, accessed 2015)

In the textbook, *have* is the first ranked content word (with the frequency of 78 over 11,818). *Have* is also the first content word to appear in the COCA word frequency list and ranked 8th following the top seven frequent function words with the frequency of 4,303,955.

In the textbook corpus, *have* is used in three ways. First, it is used as a content verb (with the meaning of owning something) the most frequently (43 over 78 tokens). Second, it is used as an auxiliary verb in the present perfect tense (27 out of 78 tokens). And finally, *have* occasionally refers to the necessity of doing something (followed with a *to* infinitive) (8 out of 78 tokens) (see Figure 4).

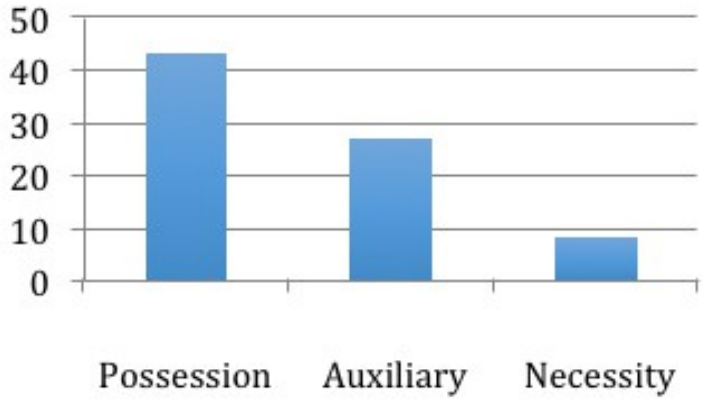

Figure 4. The frequency of *have* by meaning in the textbook

However, data from COCA show a different frequency. *Have* as an auxiliary verb in the present perfect tense is the most frequently used, with 582,559 tokens. Next frequent is *have* as a modal of necessity (followed by a *to* infinitive), with 239,877 times. The least frequent usage of *have* is its possessive meaning, with only 48,945 instances (Figure 5). Thus, it seems that the textbook's presentation of *have*'s usage differs in frequency compared to COCA.

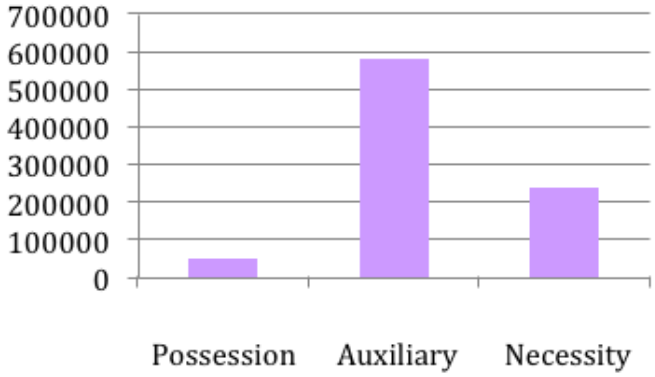

Figure 5. The frequency of *have* by meaning in COCA data

The next two content words are *parents* and *people,* with about the same frequency, 64 and 63 tokens respectively (among the 11,811 tokens from the textbook) or 5,418 and 5,334 per million respectively. However, there are considerable differences in the frequency of these two words in COCA. Specifically, *parents* (in its forms as *parent, parents, to parent*, and *parenting*) occurs at 293.08 per one million words while

people is used at 1,692.4 per one million words. This means that (1) in COCA, *people* is far more frequent than *parent* and (2) the textbook uses both *people* and *parents* more frequently than general American texts, with *parent* being used with much higher frequency. This suggests that the textbook and the texts in COCA may have difference content focuses: the textbook has a heavier focus on topics related to parents compared to texts from American media and academic sources.

Further, the analysis from the mini corpus shows some interesting collocations of *parents* and *people* used in the textbook. First of all, there is a trend in which some possessives and modal verbs tend to go before and after *parents*. Possessives such as *my* (17 times), *their* (12 times), and *your* (10 times) precede *parents* the most frequently in the textbook (Figure 6).

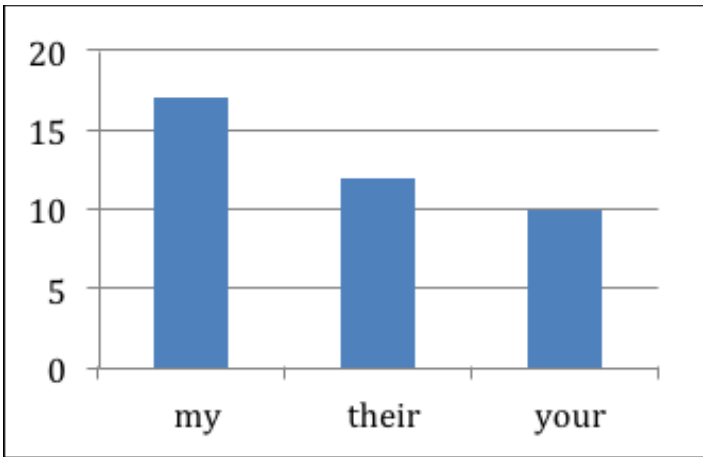

*Figure 6.*The distribution of *my*, *their*, and *your* before *parents* in the textbook

This order remains the same in the COCA list where *my* ranks first with 9,825 tokens, *their* ranks second with 9,309 tokens, and *your* ranks fifth with 3310 tokens after *his* and *her (*Figure 7). In this regard, the textbook language can be said to pattern with general American English.

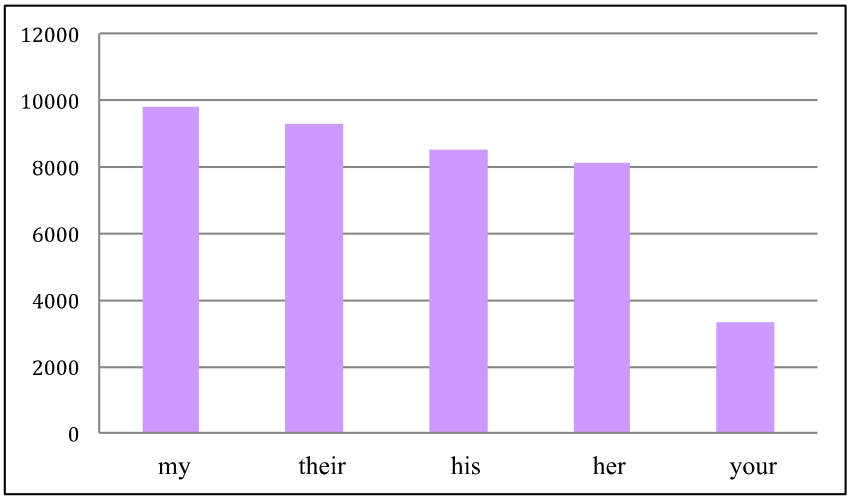

Figure 7. The distribution of the top five possessives before *parents* from COCA

Certain modal verbs such as *should* and *may* go after *parents* frequently in the textbook (Figure 8).

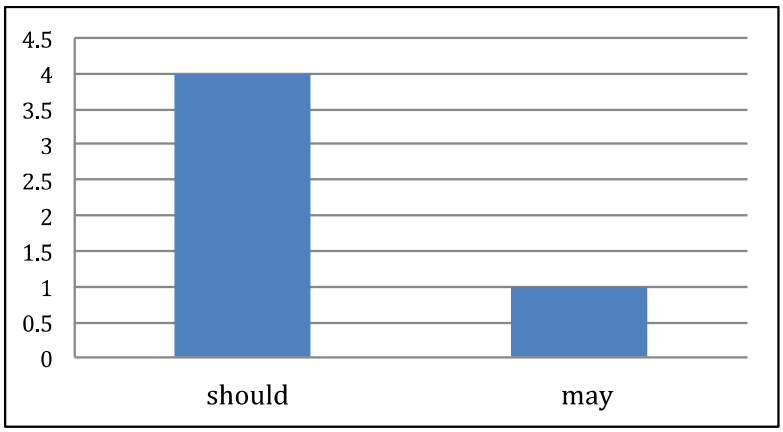

Figure 8. The distribution of *should* and *may* after *parents* from the textbook.

In the COCA data, *should* is more frequently used than *may* after *parents* (Figure 9). However, in COCA, two other modals are more frequent than even *should*: *can* and *may*.

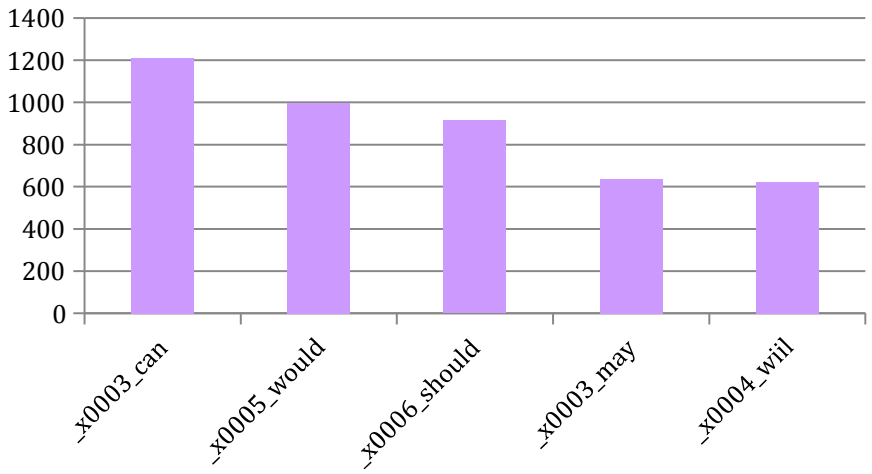

Figure 9. The distribution of the top five modal verbs following *parents* from COCA.

Additionally, *people* is frequently preceded in the textbook by adjectives such as *young* and *disabled,* with a frequency of 10 and 5 respectively (Figure 10).

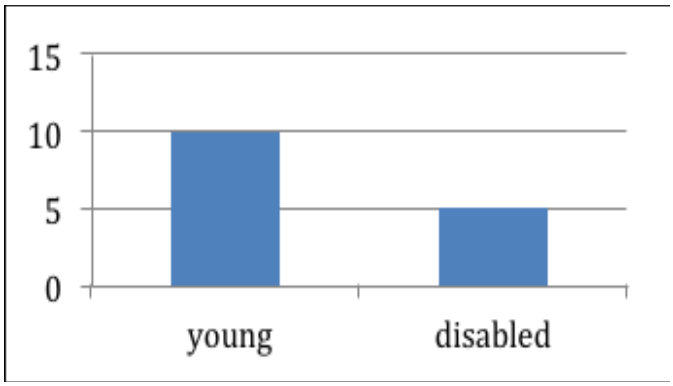

*Figure 10.*The distribution of *young* and *disabled* before *people* from the textbook.

Data from COCA show the same frequency order for *young* (with 13,681 tokens, after *other* and *American),* and *disabled* (with 496 tokens) as words that go before *people*. However, in COCA, *disabled* is much less frequent than *young,* and there are other words that precede *people* with higher frequency than *young: other* and *American* (see Figure 11). This difference in collocation frequencies may again suggest that the textbook corpus and COCA have different content focuses. While the textbook focuses more on young and disabled people, COCA, being an American corpus, contains texts about American people as well as "other people."

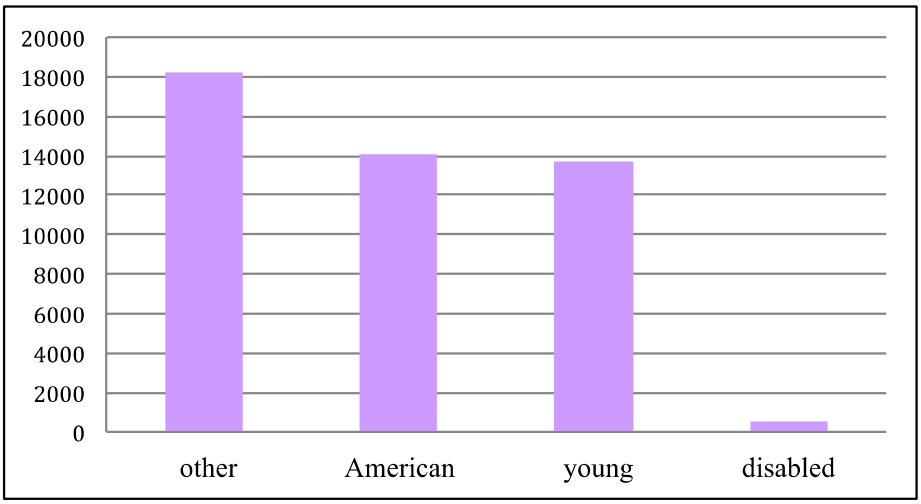

Figure 11. The distribution of *other*, *American*, *young*, and *disabled* preceding *people* from COCA.

Finally, prepositions are also found to follow *people* most of the time. *With*, *in* and *of* are the most frequent ones to go after *people* in the textbook whereas this order changes in COCA data (see Figure 12). In COCA, the most frequent preposition following *people* is *in* while in the textbook, it is *with* (Figures 12 and 13).

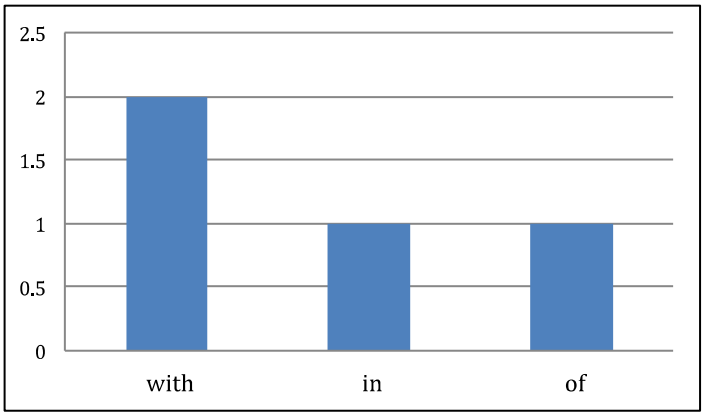

Figure 12. The distribution of *with*, *in*, and *of* following *people* from the textbook.

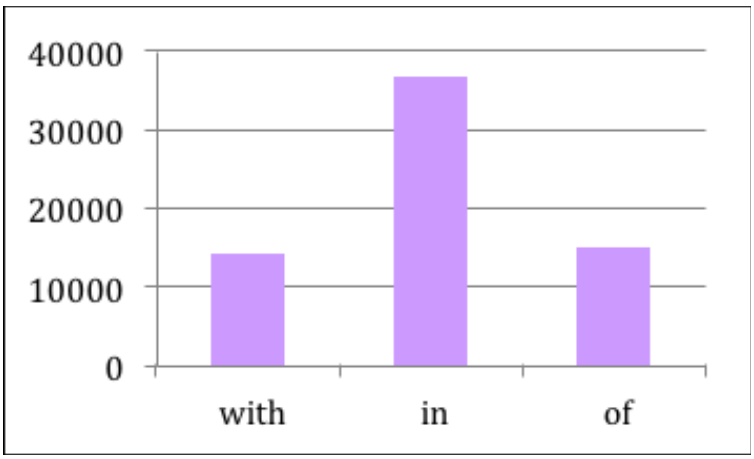

Figure 13. The distribution of *with*, *in*, and *of* following *people* from COCA.

Conclusion

The analysis above shows that while there are some similarities in the distribution of the most frequent lexical items in the Grade 11 English textbook and general American English, there are also some important differences. While some frequency differences have to do with the two corpora's different content focus (such as the frequency of *parent, people* and the collocations with the word *people*), some other frequency differences have to do with grammatical patterns (such as the different functions of *have* and the different prepositions following *people*). This second type of difference may bear consequences on language learning. Given Vietnamese learners' limited exposure to English, seeing more frequent use of the possessive meaning of *have,* for example, may lead learners to overuse that meaning, in contrast to the more common auxiliary meaning of the word in general American English. Similarly, learners may overuse the phrase *people with* due to the textbook's higher frequency of this phrase compared to the more common phrase *people in* in general American English. Certainly, the goal of English education in Vietnam is not to reproduce American English. Thus, the lexical patterns that reveal differences in content focuses may be suitable in this regard. However, if the goal is to enable our students to use English naturally, the differences in the frequency of grammatical patterns between textbook materials and general English can be a concern. At the very least, teachers and students need to be aware of the differences between the language provided in the textbook and the language used in the media and academic texts in American English. This small study hopes to have informed teachers and students about these differences.

References

Adams, M. J., & Collins, A. (1977). A schema-theoretic view of reading. In *Center for the Study of Reading: Technical Report* (Vol. 32). Cambridge, MA: Bolt Beranek and Newman Inc.

Anthony, L. (2011). AntCont (version 3.2.4m) [software]. Retrieved from http://www.antlab.sci.waseda.ac.jp/antconc_index.html

Chall, J. (1987). Two vocabularies for reading: Recognition and meaning. In M.G. McKeown, & M.E. Curtis (Eds).*The nature of vocabulary acquisition* (pp. 7-17). Hillsdale, NJ: Erlbaum.

Davies, M. (2011). Google books (American English) corpus (155 billion words, 1810-2009). Available online at http://googlebooks.byu.edu/.

Google (2004). Google books. Retrieved from https://en.wikipedia.org/wiki/Google_Books

Hoang, V. V., Phan, H., Hoang, T. H. H, Hoang, T. X. H., Kieu, T. T. H., Vu, T. L., Dao, N. L., Chung, T. Q. (2014). *Tieng Anh 11*. Ha Noi: Publish House of Education.

Lightbown, P. M., & Spade, N. (2006). *How languages are learned* (3rd ed.). Oxford, UK: Oxford University Press.

McCarthy M. & Cart R. (2013). *Vocabulary and language teaching*. Routledge: New York, USA
Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge, UK: Cambridge University Press
Nation, I. S. P. (2009). Teaching ESL/EFL reading and writing. New York, NY: Routledge, Taylor and Francis.
Reppen, R. (2010). *Using corpora in the language classroom*. New York, NY: Cambridge University Press.