tesol QUARTERLY
tesol international association

Online News as a Resource for Incidental Learning of Core Academic Words, Academic Formulas, and General Formulas

THI NGOC YEN DANG AND XU LONG
University of Leeds
Leeds, UK

Abstract

Knowledge of academic words, academic formulas and general formulas is essential for second language learners, but resources for incidental learning of these lexical items are very limited, especially in English as a Foreign Language contexts. This study employed a corpus-driven approach to examine the potential of online news for incidental learning of core academic words, academic formulas, and general formulas. Twenty-six corpora were created from voice of America (VOA) news to represent the amounts of reading that learners with the reading speed of 200 words per minute (wpm) and those with the reading speed of 150 wpm could complete within certain periods of time if they read VOA news regularly for 40 minutes per day and 5 days per week. Then, occurrences of core academic words, academic formulas, and general formulas represented by well-known lists in each corpus were counted with corpus-based tools. Analysis showed that the number of core academic words, academic formulas, and general formulas increased steadily as more input was added. Depending on the kind of target vocabulary and the reading speed, reading VOA news regularly for 3 months to 3.5 years would likely offer learners opportunities to encounter nearly all target words/formulas and encounter at least 80% of them 12 or more times in their reading. This finding suggests that online news is a useful resource for incidental vocabulary learning. It also indicates the minimal amounts of reading needed for incidental learning of core academic words, academic formulas, and general formulas.

doi: 10.1002/tesq.3208

INTRODUCTION

There are a large number of words and formulas that second language (L2) learners need to know. Explicitly teaching all of them is challenging given the limited class time (Webb & Nation, 2017). Therefore, apart from deliberate learning, incidental learning has been widely recommended as a way for L2 learners to expand their vocabulary knowledge. Incidental vocabulary learning refers to learning vocabulary through meaning-focused activities (e.g., reading news, listening to songs, or watching television programs) (Ellis, 1999; Webb & Nation, 2017). This means learners mainly focus on understanding the message in the input rather than deliberately learn a set of lexical items from the activities. For incidental learning to happen, learners need to repeatedly meet target lexical items in a large amount of comprehensible input of their interests so that these items could be attended to and learned (Nation, 2007). However, research has shown that the amount of L2 input in English as a Foreign Language (EFL) contexts is very limited and mainly comes from textbooks (Alsaif & Milton, 2012; Jordan & Gray, 2019). Unfortunately, EFL textbooks provide very few chances for learners to incidentally learn general words (Sun & Dang, 2020), academic words, and academic formulas (Coxhead, Rahmat, & Yang, 2020). This leads to the need for exploring sources for incidental learning of these lexical items. Yet earlier studies have mainly focused on resources for incidental learning of general words. Very few studies have investigated resources for incidental learning of formulas and academic vocabulary. This is surprising because formulaic knowledge enables L2 learners to reach higher levels of language proficiency (Siyannova-Chanturia & Pellicer-Sánchez, 2019) while knowledge of academic vocabulary is essential for those planning to study in English-medium or English for Academic Purposes (EAP) programs (Coxhead, 2018).

Online news might be a potential source for incidental learning of formulas and academic vocabulary for several reasons. First, as authentic materials, online news may include a considerable number of formulas. Moreover, because online news covers a wide range of topics (e.g., science, heath, technology, art, and culture), it may contain a substantial number of academic words and formulas. Reading a large amount of online news, therefore, may offer learners opportunities to repeatedly encounter academic words (e.g., *terminate*, *maximize*, *utilize*), academic formulas (e.g., *according to the*, *in terms of*, *in response to*), and general formulas (e.g., *no doubt*, *a great deal*, *kind of*) in various contexts. As previous research (Hsu, 2019; Teng, 2015) found that mid-frequency words could be learned through reading online news, the same patterns may be found in the case of academic words, general

formulas, and academic formulas. Another reason why online news may be a potential source for learning academic words, general formulas, and academic formulas is its lighter vocabulary load compared to academic texts. To achieve reasonable comprehension,[1] learners would need to know 4000 words in the case of online news (Hsu, 2019), but 11,000 words in the case of academic texts (Lu & Coxhead, 2019). The lighter vocabulary load may make it easier for learners to attend to unfamiliar lexical items and incidentally learn them when online news is used as the input than when academic texts are used. Last but not least, online news covers a wide range of topics. It is much shorter than academic texts and is freely available for everyone through a cell phone or a computer connecting to the Internet. Therefore, learners could access this kind of input easily, choose the news to read according to their interests, and can finish reading a piece of news in a short period of time. This will motivate them to read regularly and make it possible for teachers to set reading news as both in-class and outside classroom activities.

Despite the potential of online news for incidental learning of academic words, academic formulas, and general formulas, no attempts have been made to test this hypothesis. To address this gap, this corpus-driven study aimed to estimate the extent to which learners would encounter academic words, academic formulas, and general formulas if they read voice of America (VOA) news, a popular kind of online news, over certain periods of time. The finding of the study would provide further insight into the value of online news for incidental vocabulary learning and useful implications for English language teachers and learners, especially those working in English-medium programs and EAP programs.

BACKGROUND

Academic Words, Academic Formulas, and General Formulas

Vocabulary can be counted as single words and formulas (Webb & Nation, 2017). A formula is a string of words that 'is perceived by the

[1] Following Nation's (2006) seminal work, corpus-driven research has widely adopted the 98% coverage as the optimal lexical coverage for comprehension of written texts. However, a lower coverage cut-off point (95%) was selected in the present study to indicate reasonable comprehension because there are chances that learners could read for pleasure at this lexical coverage level. First, Laufer' (2020) recent experiment with EFL learners found that the comprehension and inferencing scores of participants who read the 95% lexical coverage version of a text were not significantly different from the scores of those reading the 98% lexical coverage version. Therefore, it could be expected that there would be no significant difference in students' reading speed whether the 95% or the 98% coverage was adopted. Second, lexical coverage is not a single factor affecting comprehension of texts, other factors (e.g., background knowledge on the topic, reading strategies) also play a role (e.g., Nergis, 2011).

agent (i.e., learners, researchers, etc) to have an identity or usefulness as a single lexical unit' (Wray, 2019, p. 267). Single words and formulas can be further classified into academic and general items. Academic words (e.g., *accumulate*, *significant*) and academic formulas (e.g., *in terms of*, *in response to*) occur frequently in academic discourse whereas general words (e.g., *holiday*, *cake*) and general formulas (e.g., *feel like*, *no doubt*) occur frequently in various kinds of discourse.

Knowledge of academic words, academic formulas, and general formulas[2] is particularly important for learners studying in English medium university programs and EAP programs. Academic vocabulary is an essential component of university discourse and knowledge of this vocabulary significantly contributes to students' comprehension of academic text and academic performance (Coxhead, 2020). Meanwhile, formulaic knowledge is essential for L2 learners to achieve higher language proficiency levels (Siyannova-Chanturia & Pellicer-Sánchez, 2019). As formulas make up a large proportion of language, knowledge of these items would allow learners to improve their comprehension and production of the language, which would then develop their fluency (Simpson-Vlach & Ellis, 2010).

Despite the value of academic vocabulary and formulas, L2 learners have insufficient knowledge of these lexical items (e.g., Nguyen & Webb, 2017; Read & Dang, 2022). Therefore, several lists of core academic vocabulary and formulas have been developed to support vocabulary learning of L2 learners, especially EAP learners. Reviewing all of them is beyond the scope of a single study (please see Dang, 2020a for a systematic review). Therefore, this section will focus on four vocabulary lists: Academic Word List (Coxhead, 2000), the Academic Spoken Word List (Dang, Coxhead, & Webb, 2017), the Phrasal Expressions List (PHRASE list) (Martinez & Schmitt, 2012), and the Academic Formulas List (Simpson-Vlach & Ellis, 2010). There are several reasons for this focus. First, these lists have clear pedagogical purposes. Second, they have clear selection criteria, which makes it possible for replication. Third, they have gone through a rigorous process of development and validation and has been widely recommended by vocabulary researchers (e.g., Coxhead, 2018, 2020; Nation, 2016). Last but not least, these lists are all publicly available, which makes it possible for teachers, learners, and researchers to make use of them.

Appendix 1 presents the key features of each list. The Academic Word List (AWL) (Coxhead, 2000) and the Academic Spoken Word List (ASWL) (Dang et al., 2017) are lists of core academic words. The

[2] General words can be further divided into high-frequency words, mid-frequency words, and low-frequency words. Knowledge of high-frequency words and mid-frequency words is essential for L2 learners. As Hsu (2019) has already examined the potential of VOA news for learning these words, the present study would not focus on these words.

AWL was developed from an academic written corpus of 3.5-million words and the ASWL was derived from an academic spoken corpus of 13-million words to support EAP learners' comprehension of academic written and spoken English, respectively. Statistical criteria (range, frequency, and dispersion) were used to ensure these lists capture core items that occur frequently in a wide range of academic texts. Moreover, Coxhead (2000) and Dang et al. (2017) also consider learners' prior knowledge of general words and excluded from their lists words that are likely to be known by learners. The AWL and ASWL consistently covered 10% of the words in academic written English and 90% of the words in academic spoken English, respectively.

The Phrasal Expressions List (PHRASE List) (Martinez & Schmitt, 2012) and the Academic Formulas List (Simpson-Vlach & Ellis, 2010) are lists of core formulas. The PHRASE List was developed with the aim to guide L2 vocabulary learning and teaching and to measure progress in vocabulary acquisition while the AFL was developed to support EAP learners in their academic study. The PHRASE List was derived from the British National Corpus of 100-million words, which represents both spoken and written English. Meanwhile, the AFL was created from an academic spoken corpus and an academic written corpus, each of which has around 2.1-million words. Statistical criteria (range, frequency, and N-grams) were used in the development of these lists to ensure that they capture core formulas that occur frequently in general English (PHRASE List) and academic English (AFL). Other criteria were also used to make these lists better meet their end-users' need. A series of qualitative criteria (e.g., morpheme equivalence, semantically transparency, potentially deceptive transparency) were applied to filter items in the PHRASE List so that only meaningful formulas were kept. Meanwhile, the AFL formulas were ranked according to their value for learning and teaching as rated by experienced EAP instructors.

As all items in the AWL, the ASWL, the AFL, and the PHRASE list have met the rigorous selection criteria, they well represent core academic written words, academic spoken words, academic formulas, and general formulas. Their developers and other vocabulary researchers have recommended that if learners would like to have a solid knowledge of these core lexical items, they should know all items from these lists (e.g., Coxhead, 2000, 2018; Dang et al., 2017; Martinez & Schmitt, 2012; Nation, 2016; Simpson-Vlach & Ellis, 2010). Given the pedagogical value of the four lists for L2 learners, it is important to identify various ways to support L2 learners' acquisition of items from these lists (Coxhead, 2018; Nation, 2016). In addition to explicit teaching, exploring resources for learners to incidentally learn these core academic vocabulary and formulas is also important. The next section discusses research on incidental vocabulary learning further.

Factors Affecting Incidental Second Language Vocabulary Learning

Incidental vocabulary learning from meaning-focused input is essential for L2 vocabulary development (Nation, 2007). To communicate effectively, learners need to know a large number of words and formulas, but the limited class time makes it challenging for teachers to explicitly teach all of them to their learners. Moreover, vocabulary learning is an incremental process; therefore, apart from explicit teaching, teachers should create opportunities for learners to repeatedly encounter the target lexical items in varied contexts so that they can deepen knowledge of these items (Webb & Nation, 2017).

Research has indicated that incidental learning of single words and formulas through written input could be affected by various vocabulary-related factors (e.g., frequency of occurrences, cognateness, congruency, concreteness), learner-related factors (e.g., prior vocabulary knowledge, topic familiarity, language proficiency, engagement), and input-related factors (e.g., input modes, text enhancement) (see Boers, 2020 and Peters, 2020 for detailed reviews). Among these factors, frequency of occurrence has received the greatest attention from previous studies. Research has found that frequency of occurrence has a positive impact on the learning of both single words (e.g., Pellicer-Sánchez & Schmitt, 2010; van Zeeland & Schmitt, 2013) and formulas (Webb, Newton, & Chang, 2013). However, no agreement has been reached about the frequency threshold at which incidental vocabulary learning occurs. Vidal (2011) found that learning increased the greatest between two and three repetitions in the case of reading and five to six repetitions in the case of listening. However, other studies suggested that a larger number of repetitions are needed: 10 or more times (Webb, 2007) and 20 or more times (Brown, Waring, & Donkaewbua, 2008) in the case of written input, and 15 or more times in the case of spoken input (van Zeeland & Schmitt, 2013). The inconsistent findings are probably because of the variation in the measures of vocabulary knowledge and the kind of input being examined across studies. Moreover, as previously mentioned, apart from frequency of occurrence, other factors also affect incidental vocabulary learning (Boers, 2020; Peters, 2020). In fact, in a meta-analysis of the effects of frequency on incidental vocabulary learning, Uchihara, Webb, and Yanagisawa (2019) found that frequency and learning gains moderately correlated with each other and other factors (e.g., learners' age and engagement) also contributed to vocabulary learning. These findings suggest that while frequency of occurrence is important, it is not the only factor that affects vocabulary learning. Thus, Webb and

Nation (2017) pointed out that although the more frequently a lexical item occurs in the input, the more likely it is to be learned, there is no frequency threshold at which learning occurs.

Given the lack of a frequency threshold for incidental vocabulary learning to happen, corpus-driven studies have adopted certain arbitrary frequency cut-off points to estimate the potential for incidental learning from input. Knowledge of form-meaning connection is a key aspect of knowing a word, but knowledge of other aspects (e.g., word parts, associations) is also important (Nation, 2013). Therefore, most studies (e.g., Csomay & Petrovíc, 2012; Rodgers & Webb, 2011; Webb, 2010) set 10 or more times as the point at which incidental learning of the form-meaning connection of new words happened and 5 or more times as the point at which knowledge of other aspects of known words was acquired. Meanwhile, other studies (Hsu, 2019; Nation, 2014) chose 12 or more times as the point at which incidentally learning of the form-meaning connection of new vocabulary occurred. Recently Dang (2020b) and Sun and Dang's (2020) adopted a range of frequency cut-off points: 5 or more times, 10 or more times, 15 or more times, and 20 or more times. This approach provides a better insight into the potential of learning from L2 input. The present study followed Dang's (2020b) and Sun and Dang's (2020) approach by using multiple frequency cut-off points: 5 or more times, 12 or more times, and 20 or more times. The cut-off point of 5 or more times was chosen to indicate the point at which partial knowledge of different aspects of known vocabulary could be acquired and expanded. The cut-off point of 12 or more times was chosen to indicate the point at which knowledge of the form-meaning connection of new vocabulary could be learned incidentally. As this cut-off point has been used by Hsu (2019) to examine the potential of VOA news for learning mid-frequency words, it would allow a direct comparison of the findings of the present study with those of Hsu's study. The cut-off point of 20 or more times was chosen because it indicates the point at which the effect of frequency of occurrence on incidental learning of new vocabulary is still stable (Uchihara et al., 2019).

Potential Sources for Incidental Vocabulary Learning

Although incidental vocabulary learning is essential for L2 vocabulary development, learners in EFL contexts have limited amounts of input (Webb & Nation, 2017). Therefore, a major trend in vocabulary studies is to explore potential sources of input for incidental vocabulary learning to happen. Based on systematic analysis of the occurrences of vocabulary in large amounts of texts representing certain

discourse types, corpus linguistics offers an innovative approach to identify these sources.

Earlier corpus-driven studies have indicated that textbooks in EFL contexts do not create ideal conditions for incidental learning of both single words and formulas. Sun and Dang's (2020) analysis of a set of senior high school EFL textbooks in China revealed that 86.7% of the second 1000 words and 62.8% of the third 1000 words of general English occurred in these textbooks; importantly, only 50.2% of the second 1000 words and 27% of the third 1000 words occurred 10 or more times in the textbooks. This indicates that these textbooks provided poor conditions for incidental learning of the second and third 1000 words of general English. As for academic vocabulary, Coxhead et al. (2020) used the AWL to represent core academic written words and the AFL to represent core academic formulas. They found that only 3.51% of the academic words occurred at least 10 times in a set of three EFL textbooks in Indonesia, and only 1.23% occurred at least 10 times in a set of two EFL textbooks in China. Meanwhile, 57% of the academic formulas occurred in one Chinese EFL textbook and 45.89% occurred in the other. However, Coxhead et al. (2020) did not report the percentage of core academic formulas that occurred 10 times or more times in their textbook corpora. Despite this fact, together Coxhead et al.'s findings indicated that the examined EFL textbooks did not provide sufficient exposure for incidental learning of core academic words and formulas.

As EFL textbooks provide few opportunities for incidental vocabulary learning, research has identified several potential resources for EFL learners to incidentally learn vocabulary. Most research has focused on general words. Webb and Rodgers conducted a series of studies exploring the potential of television programs (e.g., Rodgers & Webb, 2011; Webb & Rodgers, 2009a) and movies (Webb, 2010; Webb & Rodgers, 2009b) for learning low-frequency words. These researchers set 10 or more encounters as the point at which incidental learning of the form-meaning connection of new words happened and 5 or more encounters as the point at which knowledge of other aspects of known words was learned. They found that a reasonable number of low-frequency words occurred at least five or more times in television programs and movies, suggesting that television programs and movies are potential sources for incidental learning of low-frequency words.

Unlike Webb and Rodger, Nation (2014) and Hsu (2019) identified sources for incidental learning of mid-frequency words. Setting 12 or more occurrences as the frequency cut-off point for incidental learning to happen, Nation (2014) found that 80% of the words from the second to the nineth 1000 words occurred at least 12 times in his 3-million-word corpus of novels, and therefore suggested that to

incidentally learn mid-frequency words, learners would need to read at least 3-million words from novels. He also pointed out that if EFL learners read at the speed of 200 words per minute and spend 40 minutes per day and 5 days per week, they would be able to complete this reading amount within a year and are likely to have enough exposure to most of the mid-frequency words. Following Nation's approach, Hsu (2019) also set 12 or more times as the cut-off point for incidental learning of mid-frequency words. She found that 80% of the words from the first to the nineth 1000-word levels occurred at least 12 times in a 6-million word corpus of VOA news. This finding indicates that to incidentally learn most of the mid-frequency words, learners would need to read 6-million words from VOA news. Hsu's study also revealed that 4000 words are needed for reasonable comprehension of VOA news (i.e., to know 95% of the words in the news). Together, Nation's (2014) and Hsu's (2019) studies indicate that novels and VOA news are potential sources for incidental learning of mid-frequency words.

Several corpus-driven studies have explored potential sources for incidental learning of technical words. Rolls and Rodgers (2017) analyzed the occurrences of scientific specific technical words represented by Coxhead and Hirsh's (2007) EAP Science List in a corpus of science fiction-fantasy texts. The results showed that the EAP science list covered 0.5% of the words in the corpus, which was 46% higher and 70% higher than its coverage in fiction texts (0.27%) and academic science journals (1.68%), respectively. Rolls and Rodgers also found that 21% of the items from the EAP Science list occurred 10 or more times at the 500,000-word reading level and suggested that science fiction-fantasy could be a potential source for learners to learn specialized vocabulary in science.

Two studies have explored the potential for learning technical words through watching discipline-related television programs (Csomay & Petrovíc, 2012; Dang, 2020b). Csomay and Petrovíc defined technical words as those appearing in discipline-related movies and television programs and having specialized meaning in a specialized dictionary. They then developed a list of technical words in law-related television programs and movies and analyzed the occurrences of these words in these movies and programs. The findings showed that words with occurrences of 10 or more encounters made up 73.8% of the words in the corpus of movies and programs. Dang (2020b), however, defined technical vocabulary as those occurring frequently in specialized texts. She created a Medical Spoken Word List (MSWL) which represented technical words in medical lectures and seminars, and analyzed the occurrences of these words in a corpus of 37 medical television programs. The results revealed that watching all 37 programs

would enable learners to be exposed to all MSWL words, and repeatedly meet 99.44% of the MSWL words 20 or more times. Despite the different approaches, both Csomay and Petrovíc's (2012) and Dang's (2020b) findings suggested that discipline-related television programs have potential for learning technical words.

It can be seen that previous research on non-textbook resources for incidental vocabulary learning only examined the learning of single words, especially general words (mid and low frequency words) or technical words in a specific discipline (science, law, and medicine). Moreover, only three studies have explicitly estimated the minimal amounts of input needed for incidental vocabulary learning to happen. Yet they only focused on mid-frequency words (Hsu, 2019; Nation, 2014) and technical words in a specific subject (Rolls & Rodgers, 2017). General formulas (e.g., *as well as*, *of course*), academic formulas (e.g., *in terms of*, *in response to*), and academic words (e.g., *accumulate*, *significant*) are also important for L2 learners. Knowledge of core formulas such as general formulas and academic formulas is essential for higher language proficiency levels (Siyannova-Chanturia & Pellicer-Sánchez, 2019) while knowledge of core academic vocabulary (e.g., the lexical items that occur frequently in a range of academic disciplines) is also essential for L2 learners, especially those studying in English-medium programs and EAP programs (Coxhead, 2018). Given the importance of core general formulas, academic formulas, and academic words and the lack of research exploring the potential sources for incidental learning of these items, further research is warranted.

The Present Study and Research Questions

The literature review has shown that knowledge of core academic words, academic formulas, and general formulas is essential for L2 learners, especially those studying in English-medium or EAP programs. Yet no studies have identified the potential sources for incidental learning of these lexical items. Nor have any studies estimated the minimal amounts of inputs which could potentially lead to incidental learning of core academic words, academic formulas, and general formulas. To address these gaps, the present study employed a corpus-driven approach and used VOA news as an example to investigate the potential of online news for incidental learning of the core academic words, academic formulas, and general formulas represented by well-known vocabulary lists. Apart from this primary aim, this study also has a secondary aim of comparing the relative value of online news for learning each group of core vocabulary. VOA news was chosen because

earlier research has shown that reading VOA news could help learners to incidentally learn mid-frequency words (Hsu, 2019). Moreover, VOA news is freely available and has been widely used as the learning materials in many EFL classes. These features make it possible to incorporate this kind of input in real classrooms. In particular, this study aimed to answer the following research questions:

1. To what extent are core academic words encountered in different amounts of VOA news reading?
2. To what extent are core academic formulas encountered in different amounts of VOA news reading?
3. To what extent are core general formulas encountered in different amounts of VOA news reading?

METHODOLOGY

Research Design

A corpus-driven approach was employed to find the answer to the three research questions. First, we created 26 corpora representing the amounts of reading that learners with the reading speed of 200 words per minute (wpm) and 150 wpm could complete within certain periods of time if they read VOA news regularly for 40 minutes per days and 5 days per week. These reading speeds were recommended by Nation (2014). Nation's review of speed reading research indicates that university EFL students with speed reading training can easily read texts at an average rate of 200 wpm. As speed reading is a common activity in many language courses, the speed of 200 wpm was chosen in the present study. However, considering that not all EFL learners receive speed reading training, we also examined the reading speed of 150 wpm, which is considered by Nation as a conservative figure to account for low proficiency learners. Adopting both reading speeds to estimate the reading amounts that learners are likely to complete in certain periods of time would provide an in-depth evaluation of the potential of online news for incidental vocabulary learning. The amount of time spent on regular reading (40 minutes per days and 5 days per week) is also based on Nation's suggestion.

Once the corpora had been created, we used corpus tools to count the occurrences of core academic words, academic formulas, and general formulas presented by well-known vocabulary lists in each corpus in turn. After that, we compared the occurrences of these lexical items in different corpora to see the degree to which core academic words,

academic formulas, and general formulas were encountered in different reading amounts. Detailed information of the corpora, vocabulary lists, and data analysis is presented below.

Corpora

Twenty-six corpora of VOA news were created in the present study. Materials for these corpora were developed from VOA news published online from April 2009 to December 2021. The news was manually downloaded from the VOA news website https://www.voanews.com/. Headlines and captions of images were included in the corpus because they were related to the main content of the news and students could refer to them when reading the news. Authors' names and dates of publication were removed because they were not closely related to the content of the news and students may not pay much attention to them when reading. Each corpus has five sub-corpora representing five main topical sections in the VOA news websites: (1) Arts & Culture, (2) Economy & Business, (3) Politics, (4) Science & Health, and (5) Silicon Valley & Technology.

The steps of creating these corpora are as follows. First, the size of each corpus was decided based on two factors:

1. The size of the corpus should be roughly around the amount of reading that EFL learners with different reading speeds could read within certain periods of time as estimated by Nation (2014); that is, they could read 200 wpm or 150 wpm and read 40 minutes per day and 5 days per week.
2. The number of words in each sub-corpus should be roughly the same to minimize the bias caused by a particular topic on the lexical analysis.

Second, a series of trials were conducted in which news from five main topical sections in the VOA news website was gradually added up until 80% of the academic spoken words, academic written words, academic formulas, and general formulas in the selected lists occurred at least 12 times in the whole corpus. As a result, reading amounts relevant to 2.5 years[3] and 3.5 years[4] were chosen for the reading speed of

[3] The 2.5-year corpus (200 wpm) consists of 7478 pieces of news (1459 from Art & Culture, 1732 from Economy & Business, 1059 from Politics, 1524 from Science & Health, and 1704 from Silicon Valley & Technology).

[4] The 3.5-year corpus (150 wpm) includes 7834 pieces of news (1596 from Art & Culture, 1752 from Economy & Business, 1140 from Politics, 1609 from Science & Health, and 1737 from Silicon Valley & Technology).

200 and 150 wpm, respectively. After that, the corpora representing these reading amounts and their topical sub-corpora were split into sections representing different reading amounts (e.g., 1 day, 3 months, 6 months) with AntFile Splitter (Anthony, 2017) (see Appendices 2 and 3). This approach was consistent with earlier studies (Hsu, 2019; Nation, 2014; Rolls & Rodgers, 2017) and ensured the analysis would not be biased toward a certain topic or reading amount.

Vocabulary Lists

Two lists of single words and two lists of formulas were used for the analysis (see Appendix 1). The lists of single words are the AWL (570 words) (Coxhead, 2000) and the ASWL (1741 words) (Dang et al., 2017). These lists represent core academic written words and academic spoken words, respectively. The lists of formulas are the AFL (207 formulas) (Simpson-Vlach & Ellis, 2010) and the PHRASE List (505 formulas) (Martinez & Schmitt, 2012). These lists in turn represent core academic formulas and general formulas. The four vocabulary lists were chosen for the reasons mentioned in the "Background" section.

Analysis

The occurrences of core academic words, academic formulas, and general formulas in each corpus representing the amount of VOA news being read in a certain period of time were investigated from two perspectives: (1) the number of these items occurring in the news and (2) the number of these items occurring multiple times in the news. To investigate the occurrences of academic words, each corpus was run through the RANGE program (Heatley, Nation, & Coxhead, 2002). The AWL and ASWL were in turn used as the base word lists. To examine the occurrences of core academic formulas and general formulas, each corpus was uploaded to Sketch Engine. Then, the frequency of each item in the AFL and the PHRASE lists in the corpus was computed with the N-gram function and each list of formulas serving as the base list. Based on these analyses, items in each list were classified into four groups: (1) items with one or more occurrences, (2) items with 5 or more occurrences, (3) items with 12 or more occurrences, and (4) items with 20 or more occurrences. Then, the

percentage of each group in the list[5] was calculated by dividing the number of items in the group by the total number of items in the list and multiplying by 100%. For example, the AWL had 570 words, but only 3 words occurred 20 or more times in the 1-day reading amount (200 wpm). It means that 0.53% (=3 ÷ 570 × 100%) of the AWL words occurred 20 or more times in this reading amount.

RESULTS

Occurrences of Core Academic Words in VOA News

In answer to Research question 1, Appendices 4–7 present the occurrences of core academic spoken words and academic written words represented by the AWL and the ASWL in different reading amounts when the reading speeds were 200 wpm (Appendices 4 and 5) and 150 wpm (Appendices 6 and 7). At the speed of 200 wpm, more than 46% of the core academic spoken words and nearly 32% of

[5] As the selected lists have different length, readers may think that it is better to compare the most frequent items in each list so that each list contains an identical number of items. While this approach is useful, it does not work in the present study for two reasons. First, all items in the four examined lists, not just the most frequent items, deserve attention from teachers and learners. If we only evaluate the potential of online news for learning the most frequent items in the selected lists, we cannot provide a thorough evaluation of the potential of online news for learning all core academic words, academic formulas, and general formulas. For example, analysis of the top 100 items of each selected list in the 3-month reading amount (at the reading speed of 200 wpm) showed that all top 100 ASWL words and top 100 AWL words appeared in the examined online news. Moreover, all of the top 100 ASWL words and 92 of the top 100 AWL words occurred at least 12 times. However, only 61 of the top 100 AFL formulas and 91 of the top 100 PHRASE formulas occurred in the examined online news, and only 29 of the top 100 AFL formulas and 60 of the top PHRASE formulas occurred 12 or more times. This analysis enables us to compare the relative value of online news for learning the top 100 items of each list. However, it missed to reflect the fact that other core items in the selected lists, though not among the top 100 items, also occurred in the reading (1602 ASWL words, 545 AWL words, 61 AFL formulas, and 216 PHRASE formulas) and that many of them occurred 12 or more times (1344 ASWL words, 310 AWL words, 13 AFL formulas, 60 PHRASE formulas). In other words, it failed to reflect the fact that reading online news could potentially help learners to learn other core lexical items, not just the top 100 items. Second, the original number of items in each list was driven by the number of items that met the selection criteria set by the list developers. Removing any items from a list would consequently change the list nature and negatively affect its pedagogical value. Comparing the percentage of items in each list, though not perfect, is a more reasonable solution. It helps to deal with the uneven number of items in different lists and at the same time takes the relative value of all items in the list into account. It also allows us to provide precise implications to teachers and learners; that is, if learners spend a certain amount of time reading online news regularly, what percentage of core academic words, academic formulas, and general formulas could be learned through this activity. This finding would help teachers to evaluate the relative value of reading online news for learning core lexical items.

the core academic written words occurred in the 1-day reading amount. If the speed was 150 wpm, a smaller percentage of core academic spoken words (nearly 45%) and academic written words (more than 26%) occurred in the reading. A relatively small percentage of core academic spoken words occurred multiple times in both the cases of 150 wpm and 200 wpm: 11% and 15% (5 or more occurrences), and 3% and 5% (12 or more occurrences), and 2% and 2% (20 or more occurrences). Likewise, regardless of the reading speed, the percentage of core academic written words with multiple occurrences was small: 3%–5% (5 or more occurrences), 0.18%–2% (12 or more occurrences), and 0%–0.5% (20 or more occurrences).

As more input was added, the percentage of core academic words being encountered increased accordingly. If the reading speed was 200 wpm, there was a sharp rise in the percentage of academic spoken words and academic written words in the VOA news in the first 3 months. Appendices 4 and 5 show that nearly 98% of core academic spoken words and more than 97% of core academic written words appeared in the 3-month reading amount. Within the first 3 months of reading, the percentage of core academic spoken words occurred 5 or more times, 12 or more times, and 20 or more times also went up rapidly to more than 90%, more than 80%, and nearly 75%, respectively. Similarly, more than 85%, more than 70%, and more than 55% of core academic written words occurred 5 or more times, 12 or more times, and 20 or more times.

The percentage of core academic words in VOA news kept rising in the next 3 months. Nearly all core academic spoken words occurred in the 6-month reading amount. More than 95% of core academic spoken words had occurrences of 5 or more, more than 90% had occurrences of 12 or more, and more than 85% had occurrences of 20 or more. Meanwhile, more than 98% of core academic written words occurred in the 6-month reading amount and most of them occurred multiple times: more than 90% (5 or more times), more than 80% (12 or more times), and nearly 75% (20 or more times).

After the first 6 months, the percentage of academic words in VOA news still went up, but the growth was very small. Reading VOA news for an extra 2 years only resulted in a growth of 1.27% of core academic spoken words and 1.58% of core academic written words. Moreover, the percentage of core academic spoken words and academic written words with multiple occurrences only rose by 3.85% and 7.72% (5 or more times), 7.76% and 14.73% (12 or more times), and 11.84% and 20.17% (20 or more times) over this two-year period. Despite the slow growth, by the end of 2.5 years, nearly all core academic spoken words and all academic written words appeared in the

reading. Importantly, nearly all core academic spoken words and academic written words had multiple occurrences: 99% and 99% (5 or more times), 99% and 97% (12 or more times), 98% and 94% (20 or more times).

If the reading speed of 150 wpm was adopted, a longer period of time would be needed for learners to encounter most core academic words multiple times. Appendices 6 and 7 show that the percentage of core academic spoken words and academic written words represented by the AWL and ASWL in the VOA news rose sharply in the first 6 months. More than 98% of core academic spoken words appeared in the six-month reading amount (Appendix 6). The percentage of core academic spoken words occurred 5 or more times, 12 or more times, and 20 or more times also went up rapidly to nearly 95%, more than 85%, and more than 80% within the first 6 months of reading, respectively. The percentage of core academic written words in the VOA news also increased during the first 6 months, but the increase was slightly slower than that of academic spoken words (see Appendix 7). Nearly 98% of core academic written words appeared in the 6-month reading amount, with more than 85%, more than 75%, and more than 65% having occurrences of 5 or more times, 12 or more times, and 20 or more times, respectively.

The percentage of core academic words in VOA news kept increasing in the next 3 months. Nearly all core academic spoken words and academic written words occurred in the 9-month reading amount. Most of them occurred multiple times in both the case of core academic spoken words – more than 95% (5 or more occurrences), more than 90% (12 or more occurrences), and more than 85% (20 or more occurrences) – and core academic written words – more than 90% (5 or more occurrences), more than 80% (12 or more occurrences), and more than 75% (20 or more occurrences).

After the first 9 months, the percentage of core academic words in VOA news still increased, but the growth was very small. Reading VOA news for an extra of 2 years and 9 months only resulted in an increase of 1.1% of core academic spoken words and 1.23% of core academic written words. Moreover, the percentage of core academic spoken words and academic written words with multiple occurrences only rose by 3.33% and 7.01% (5 or more times), 7.12% and 13.69% (12 or more times), and 10.23% and 18.07% (20 or more times) over this 2-year-and-9-month period. Despite the slow growth, by the end of 3.5 years, nearly all core academic spoken words and all core academic written words appeared in the reading. Importantly, nearly all core academic spoken words and written words had multiple occurrences: 99% and 99% (5 or more times), 99% and 97% (12 or more times), 98% and 94% (20 or more times).

Occurrences of Core Academic Formulas in VOA News

In answer to Research question 2, Appendices 8 and 9 show that more than 2% (200 wpm) and more than 1% (150 wpm) of the core academic formulas represented by the AFL appeared in the 1-day reading amount. If the speed was 200 wpm, nearly 0.5% of the core academic formulas occurred 5 or more times, and none of them occurred 12 or more times. If the reading speed is 150 wpm, no core academic formulas occurred multiple times.

However, the percentage of core academic formulas appearing in VOA news kept rising steadily over a long period of time. Appendix 8 shows that if the reading speed was 200 wpm, by the end of 2 years and 3 months, nearly 97% of the core academic formulas occurred in the reading with nearly 95%, more than 80%, and nearly 70% of the core academic formulas occurred 5 or more times, 12 or more times, and 20 or more times, respectively. After this period, the percentage of core academic formulas occurring one or more times did not increase while the percentage of those with more occurrences increased slightly. Reading VOA news for an extra of 3 months did not improve the percentage of core academic formulas being encountered 1 or more times. Yet it helps to increase the percentage of core academic formulas encountered 5 or more times, 12 or more times and 20 or more times by 0.48%, 1.93% and 3.38%, respectively.

Appendix 9 shows that if the reading speed is 150 wpm, by the end of 3 years and 3 months, nearly 97% of the core academic formulas occurred in the reading. Nearly 95% of the core academic formulas, more than 80%, and more than 70% of the core formulas had 5 or more occurrences, 12 or more occurrences, and 20 or more occurrences, respectively. After that, the percentage of core academic formulas encountered one or more times in the reading did not change, but there was a small increase in the percentage of those with more occurrences. Reading VOA news for an extra of 3 months did not increase the percentage of core academic formulas being encountered one or more times. However, it allowed the percentage of core academic formulas encountered 5 or more times, 12 or more times and 20 or more times to go up by 0.03%, 2.41%, and 2.42%, respectively.

Occurrences of Core General Formulas in VOA News

In answer to Research question 3, Appendices 10 and 11 present the occurrences of core general formulas represented by the PHRASE list in different amounts of VOA news reading when the reading speed is 200 and 150 wpm, respectively. Similar to academic formulas, only a

small percentage of core general formulas (more than 3% and nearly 3%) occurred in the 1-day reading amount and a very small percentage of them (0.40% and 0%) occurred five or more times. None of them had occurrences of 12 or more.

However, there was a stable increase in the percentage of core general formulas over the period of time. Compared to academic formulas, the growth in the percentage of core general formulas was slightly slower. It was just by the end of 2.5 years (in the case of 200 wpm) and 3.5 years (in the case of 150 wpm) that the percentage of core general formulas in the reading reached nearly 97%, and more than 90%, 80%, and around 70% of the general formulas occurred 5 or more times, 12 or more times, and 20 or more times, respectively.

DISCUSSION

The present study expands on corpus-driven research on incidental vocabulary learning in several ways. It is the first study to explore resources for incidental learning of core academic vocabulary and formulas. Importantly, it is the first to investigate the opportunities for incidental learning of a wide range of vocabulary over certain periods of time: both single words and formulas, both academic and general vocabulary, both vocabulary in spoken and written English, and at different reading speeds. It is also among the few incidental vocabulary studies examining online news. Detailed answers to the research questions have already been provided in the "Results" section. This section will draw these findings together to provide in-depth discussion of the potential of online news for incidental learning of core academic words, academic formulas, and general formulas.

Online News Is a Potential Source for Incidental Vocabulary Learning

The present study suggests that online news is a potential source for incidental learning of core academic words, academic formulas, and general formulas for several reasons. To begin with, even reading online news for a very short period of time would be likely to offer learners a fairly good opportunity to incidentally learn a number of core academic words and initially encounter several core general and academic formulas. This study found that reading VOA news for 40 minutes at the speed of 200 wpm would allow learners to meet more than 46% of core academic spoken words, nearly 32% of core academic written words, more than 2% of core academic formulas,

and more than 3% of core general formulas represented by the ASWL, AWL, AFL, and the PHRASE list, respectively. Even if learners read at a slower speed (150 wpm), they would meet nearly 45% of the core academic spoken words, more than 26% of the core academic written words, more than 1% of the core academic formulas, and nearly 3% of the core general formulas represented by the selected lists. In particular, depending on the reading speed, this activity would enable them to encounter 11% and 15% of the core academic spoken words, 3% and 5% of the core academic written words, 0.5% of the core academic formulas (in the case of 200 wpm) and nearly 0.5% of the core general formulas (in the case of 200 wpm) five or more times. If these lexical items are unknown vocabulary, encountering them at least five times could create a fairly good condition for learners to attend to these lexical items later if more input is provided, which would then potentially lead to incidental learning in the future. If learners have known the form-meaning connection of these lexical items, encountering them at least five times would help to consolidate their knowledge of form-meaning connection and learn other aspects of these known items.

What is more encouraging is that if learners read online news regularly, the potential for learning core academic words, general formulas, and academic formulas is much greater. First, this activity would bring about an excellent opportunity for them to learn academic words. As shown in this study, if learners read at the speed of 200 wpm, regular reading VOA news for 3 months would likely to allow them to encounter nearly all core academic spoken words, and encounter more than 90%, more than 80%, and nearly 75% of them 5 or more times, 12 or more times, and 20 or more times, respectively. Similarly, if learners read VOA news regularly for a period of 6 months, they would be likely to meet most of the core academic written words in the news and encounter more than 90%, more than 80%, and nearly 75% of them 5 or more times, 12 or more times, and 20 or more times. Even if learners read at a slower speed (150 wpm), reading VOA news regularly for 6 months would allow them to meet all core academic spoken words and encounter most of them multiple times: nearly 95% (5 or more times), more than 85% (12 or more times) and more than 80% (20 or more times). Likewise, if they read VOA news for 9 months, learners would meet nearly all core academic written words and meet 90% of the core academic written words 5 or more times, 80% 12 or more time, and 75% 20 or more times. Additionally, after reading VOA news for 2.5 years (in the case of 200 wpm) or 3.5 years (in the case of 150 wpm), learners would encounter nearly all core academic spoken and written words and encounter nearly all of them at least 20 times. The frequent occurrences and reoccurrences of these lexical

items in VOA news would create an excellent condition for learners to incidentally learn unknown academic words and to consolidate and expand their partial knowledge of known items.

Second, reading online news regularly would provide learners with a very good opportunity to learn core academic formulas and general formulas. This study found that if learners read VOA news regularly for 2 years and 3 months (at the speed of 200 wpm) or 3 years and 3 months (at the speed of 150 wpm), they would have an opportunity to encounter nearly 97% of core academic formulas in VOA news and nearly 95%, more than 80%, and more than 70% of these formulas 5 or more times, 12 or more times, and 20 or more times, respectively. Meanwhile, reading VOA news for 2.5 years (at the speed of 200 wpm) and 3.5 years (at the speed of 150 wpm), learners would likely to meet nearly 97% of the core general formulas, more than 90%, 80%, and around 70% of these formulas 5 or more times, 12 or more times, and 20 or more times. The frequent occurrences of core academic formulas and general formulas in VOA news would well facilitate the learning of the form-meaning connection of unknown formulas as well as consolidating knowledge of the form-meaning connection and acquiring other aspects of known formulas. This finding is meaningful because formulaic knowledge is essential for learners to achieve higher levels of language proficiency (Siyannova-Chanturia & Pellicer-Sánchez, 2019), but EFL learners do not have much exposure to these sequences in their textbooks (Coxhead et al., 2020).

Another reason why online news is an excellent resource for incidental vocabulary learning is that by simply reading online news for pleasure, learners may have opportunities to learn various kinds of vocabulary that are essential for their further language development in one go. This study found that reading VOA news, learners are likely to be exposed to and develop both the breadth and depth of different kinds of vocabulary: academic spoken words, academic written words, academic formulas, and general formulas. Hsu (2019) also suggests that VOA news is a potential source for incidental learning of mid-frequency words. Knowledge of general formulas and mid-frequency words is important for learners to achieve higher language proficiency while knowledge of academic words and formulas is essential for them to deal with academic texts (Nation, 2013). The fact that learners are exposed to only one kind of input, but still have excellent opportunities to repeatedly encounter a wide range of useful vocabulary makes the finding of the present study meaningful.

The findings of this study shed light on the potential sources for incidental learning of formulas and academic vocabulary. Coxhead et al. (2020) found that EFL textbooks did not provide a good condition for incidental learning of core academic words and formulas,

which indicates the need for exploring resources for EFL learners to incidentally learn these lexical items. However, previous research only examined resources for incidental learning of mid-frequency words (e.g., Hsu, 2019; Nation, 2014), low-frequency words (e.g., Webb, 2010), and technical words in a specific discipline (e.g., Rolls & Rodgers, 2017). Therefore, the present study effectively addresses this gap. Based on a systematic analysis of the occurrences of various kinds of vocabulary represented by well-known vocabulary lists in different reading amounts, this study provides solid evidence indicating that online news is a potential source for incidental learning of core academic words, academic formulas, and general formulas. Moreover, by finding that reading online news, which is normally considered as a means for entertainment, could help learners to incidentally learn various kinds of academic vocabulary, both single words and formulas, this study provides further evidence supporting the suggestions from earlier studies that certain non-academic genres could be potential bridge resources for EFL learners to incidentally learn specialized vocabulary (Csomay & Petrovíc, 2012; Dang, 2020b; Rolls & Rodgers, 2017).

The Minimal Amounts of Online News Reading for Incidental Learning of Core Academic Words, Academic Formulas and General Formulas

The present study also provides further insight into the minimal reading amounts needed for core academic words, academic formulas, and general formulas to potentially be learned through reading online news. Although the lexical items that learners are likely to encounter in the texts increase with the amount of reading being completed, this study suggests that in the case of online news, the amounts of reading needed for incidental learning to happen may vary according to the kind of vocabulary. Regardless of the adopted reading speed, among the groups of vocabulary being examined, core academic words are likely to require the smallest amount of input. This study found that if VOA news was used as L2 input, the first 3 months (in the case of 200 wpm) or the first 6 months (in the case of 150 wpm) would be a critical period for incidental learning of core academic spoken words, and the first 6 months (in the case of 200 wpm) or the first 9 months (in the case of 150 wpm) would be essential for incidental learning of core academic written words. Although reading VOA news for a day offers learners a fairly good opportunity to incidentally learn a number of core academic words, the potential for incidental learning was only

apparent for core academic spoken words during the period of 3 months (in the case of 200 wpm) and 6 months (in the case of 150 wpm), and for core academic written words during the period of 6 months (200 wpm) and 9 months (150 wpm). Learners would encounter more than 80% of the core academic spoken words 12 or more times after reading VOA news for 3 months (200 wpm) or 6 months (150 wpm) and encounter more than 80% of the core academic written words 12 or more times after a reading period of 6 months (200 wpm) or 9 months (150 wpm). In contrast, reading VOA news for an extra 2 years (200 wpm) or an extra of 2 years and 9 months (150 wpm) only allows learners to meet extra 1%–2% of core academic spoken words and academic written words. Previous research (Hsu, 2019; Nation, 2014) considered the amount of input being sufficient for learning a specific group of lexical items if at least 80% of the items in that group occurred 12 or more times in the input. Following this cut-off point, the findings of the current study suggest that perhaps the first 3 months (200 wpm) or the first 6 months (150 wpm) would be crucial for learning the form-meaning connection of new academic spoken words while the first 6 months (200 wpm) or the first 9 months (150 wpm) would be essential for learning the form-meaning connection of new academic written words. Meanwhile, the time beyond this timeframe is probably more useful for consolidating and expanding knowledge of known academic words.

The present study also indicates that compared to academic words, larger amounts of input would be needed for incidental learning of core general formulas and academic formulas. Although reading VOA news for 1 day had some potential for incidental learning of core general formulas and academic formulas, it was not until learners had read VOA news for 2 years and 3 months (200 wpm) or 3 years and 3 months (150 wpm) that they would likely encounter at least 80% of the core academic formulas for 12 or more times. Similarly, only after 2.5 years (200 wpm) or 3.5 years (150 wpm) would learners meet 80% or more of the core general formulas for 12 or more times. This suggests that depending on the reading speed, about 2 years and 3 months or 3 years and 3 months would be needed for incidental learning of core academic formulas. Meanwhile, 2.5 or 3.5 years would be essential for incidental learning of core general formulas.

Previous research has found that learners would need to read 3-million words from novels (Nation, 2014) and 6-million words from VOA news (Hsu, 2019) to potentially learn mid-frequency words and 500,000 words from science fiction-fantasy to potentially learn scientific technical words (Rolls & Rodgers, 2017). This study found that learners would need to read 480,000 or 576,000 words from VOA news,

which is relevant to the 3-month reading amount (200 wpm) and the six-month reading amount (150 wpm) – and 960,000 words or 864,000 words from VOA news – which is relevant to the 6-month reading amount (200 wpm) and the 9-month reading amount (150 wpm) – to potentially learn core academic spoken words and academic written words, respectively. Meanwhile, they would need to read 4,320,000 words or 3,744,000 words of VOA news – which is relevant to the 2-year-and-3-month reading amount (200 wpm) or the 3-year-and 3 month reading amount (150 wpm) – and at least 4,800,000 words or 5,041,189 words of VOA news – which is relevant to the 2.5-year reading amount (200 wpm) or the 3.5-year reading amount (150 wpm) – to potentially learn core academic formulas and general formulas, respectively. The differences in the findings of the present study and previous studies may be due to the differences in the kind of vocabulary and input being examined. Moreover, the fact that core academic formulas and general formulas required much larger amounts of input than academic words, mid-frequency words, and technical words may be due to the nature of formulas. Unlike single words, formulas do not occur frequently in a single text, and therefore a large amount of input would be needed for the same formulas to be re-encountered (Nation, 2013). By indicating the minimal amounts of reading needed for incidental learning of different kinds of vocabulary through reading online news, this study made a significant contribution to research on incidental vocabulary learning. It also supports the use of authentic materials as supplementary for vocabulary learning.

This study has several limitations which deserve attention from future research. First, to allow a direct comparison of the findings across studies, this study adopted Nation's (2014) suggestions on the reading speed and the amount of reading time per day. However, these figures may vary depending on individuals. Second, this study aims to investigate the potential of reading online news for incidental learning of items in core vocabulary lists. Therefore, we only examined the core academic words, academic formulas, and general formulas represented by the selected lists. We did not examine modifiers of the core formulas. For example, we only examined the occurrences of *a number of*, but not *a great number of*. It is because *a number of* met Simpson-Valch and Ellis's selection criteria and was included in the AFL, but *a great number of* did not. However, some modifiers of the core formulas may occur in online news and learners may incidentally learn them. The potential for incidental learning of these formulas may not be as great as that of the core formulas though, because learners may process and acquire adjacent formulas more easily than non-adjacent formulas (Vilkaite, 2016). Third, the corpora representing smaller amounts of reading were split from corpora representing

the reading amounts in 2.5 and 3.5 years. While this approach helps to avoid the bias toward a certain topic, the different lengths of news mean that in real life the division is not as neat as those in the present study. Third, this study took VOA news as an example of online news because it has been widely used as supplementary materials in many EFL classes, which makes it easy to apply the implications of the present study to real classrooms. It would be useful to examine other kinds of online news. Fourth, like previous studies (e.g., Hsu, 2019; Nation, 2014), this study drew on evidence from the corpus-driven analysis. Experiments with real learners would provide further insight into the potential of online news for incidental learning of academic words, academic formulas, and general formulas. Moreover, the primary aim of this study is to explore the potential of online news for incidental learning of core academic words, academic formulas, and general formulas. It does not mean to suggest that online news is the only source for incidental learning of these lexical items. As 2.5 or 3.5 years are likely to be needed for most core academic formulas and general formulas to be encountered multiple times, further research should explore if there are better materials to learn these formulas. In other words, future research exploring other bridge resources for learning academic words and formulas and comparing the learning from these sources with that from academic texts would be valuable.

PEDAGOGICAL IMPLICATIONS

This study suggests that online news, or at least VOA news, is a useful resource for learners to learn various kinds of vocabulary. Therefore, teachers should encourage learners to read online news regularly. To ensure the effectiveness of this activity, certain principles should be followed in the implementation. To begin with, as incidental learning is an incremental process (Webb, 2020), teachers should encourage learners to read online news at least 40 minutes per day and 5 days per week. Regular reading would allow them to repeatedly encounter core academic words, academic formulas, and general formulas over a certain period of time as well as developing a good reading habit even after the English language course is over.

Incidental vocabulary learning only happens if the input is comprehensible to learners. To achieve reasonable comprehension of VOA news, learners would need to know the most frequent 4000 words (Hsu, 2019). Therefore, this kind of input is probably more relevant to learners with this vocabulary size. Readers may think that with their current vocabulary knowledge, these learners may already know all core academic words, academic formulas, and general formulas and

may not benefit from reading VOA news. It is important to note that well-known diagnostic tests such as the (updated) Vocabulary Levels Test (e.g., Schmitt, Schmitt, & Clapham, 2001; Webb, Sasao, & Balance, 2017) or the Vocabulary Size Test (Nation & Beglar, 2007) sampled test items from 1000-word frequency bands. Therefore, these test scores only provide us with an estimation of how well learners know a certain 1000-word band, not an individual item in that band. There are always chances that learners know most items in the 4000–5000-word bands, but not all of them. It follows that a number of AWL words are likely to be unknown to learners although they fall into the most frequent 4000–5000-word bands. Moreover, having knowledge of single words in the most frequent 4000–5000-words band does not necessarily mean that learners know the core academic formulas and general formulas made up of words from these bands. Nguyen and Webb (2017) found that L2 learners' knowledge of formulas was lagged behind their knowledge of single words at the same frequency band. Additionally, knowing a word involves various aspects (form, meaning, and use) (Nation, 2013). Each aspect develops at different rates and it is unlikely that L2 learners, even advanced learners, have already fully mastered all of them in one go (Schmitt, 2010). Therefore, while it is important for learners to learn new lexical items, it is equally important for them to deepen and expand their knowledge of known items by engaging with a wide range of meaning-focused activities (Nation, 2007). Reading VOAs is a useful meaning-focused activity because it allows learners to learn the form-meaning connection of new academic words, academic formulas and general formulas, but at the same time enables them to consolidate knowledge of the form-meaning connection and acquire other aspects of known items.

Although VOA news is more relevant to learners with knowledge of the most frequent 4000 words, it does not mean that learners with slightly smaller vocabulary sizes do not benefit from this kind of input. First, the lexical demand varies according to individual texts (Webb, 2021). Therefore, learners with the vocabulary sizes of 2000 or 3000 words may select VOA news which requires knowledge of the most frequent 2000–3000 words for reasonable comprehension and start reading these pieces first (see Appendix 12 for an example). Second, news which requires the vocabulary size of 4,000-words can still be comprehensible for learners with smaller vocabulary sizes. Laufer (2020) found that despite not leading to inferencing scores as high as the 95% and 98% coverage, the 90% coverage resulted in fairly similar comprehension scores as the other two coverage points. This indicates that learners with the vocabulary size of fewer than 4000-words could probably achieve as good comprehension of the news as those with the vocabulary size of 4000 words. Third, pre-learning a very small

number of unknown words can make texts which require a vocabulary size of 4000 words suitable to learners with smaller vocabulary sizes. Unlike other kinds of input (e.g., academic books, articles, novels), online news is very short. Webb (2010) found that pre-learning a small number of unknown words that occur frequently in a short text can help to increase the lexical coverage significantly, which in turns may improve comprehension quickly. The digital format of the news allows learners to use the look-up functions in Microsoft Word or the Kindle function which connected with dictionaries or programs such as Read with resources (www.lextutor.ca) to quickly check the meaning of unknown words in the news without interrupting the reading flow (Nation, 2014). This would help to increase the lexical coverage and improve learners' comprehension of the news. For example, Appendix 13 presents a piece of VOA news. Analysis with the VocabProfiler in Lextutor showed that the most frequent 4000 words accounted for 95.4% of the words in this text. However, this text can still be relevant to learners with the vocabulary knowledge of the most frequent 3000 words. Given their current vocabulary knowledge, they are likely to achieve 92.6% of the words in the text. The text has 352 running words. There are four words at lower frequency level but occurred frequently in the text: *dioxide* (occurring four times), *greenhouse* (occurring six times), and *meteorological* (occurring two times). If learners learn these words, it would add 3.41% coverage [= (4 + 6 + 2)) ÷ 352 × 100]. As a result, learners with the vocabulary size of 3000-word families can achieve 96.01% coverage of the text. To guide learners in the selection of pieces of VOA news that matches their current vocabulary knowledge, teachers could instruct learners to use Webb et al.'s (2017) updated Vocabulary Levels Test to measure their vocabulary knowledge and then use Cobb's (n.d.) Lextutor to analyze the lexical profile of the pieces of VOA news that they are interested in (see Dang, 2022 for further instruction).

Apart from setting regular reading activities and selecting relevant news to learners, to help learners keep track of their progress and motivate them to learn further, teachers can draw on the findings of the current study about the minimal amounts of reading to set specific expectations for learners. If learners' reading speed is 200 wpm, at least 3 months, 6 months, 2 years and 3 months, and 2.5 years would be needed for most of the core academic spoken words, academic written words, academic formulas, and general formulas to be learned incidentally, respectively. However, if their reading speed is only 150 wpm, longer periods of time would be needed to learn these lexical items: 6 months (academic spoken words), 9 months (academic written words), 3 years and 3 months (academic formulas) and 3.5 years (general formulas). The much shorter period of time needed in the

case of 200 wpm compared to 150 wpm suggests that speed reading activities is essential in language courses. The figures in this study also provide learners with a rough idea of the gains that they may achieve if they spend a certain amount of time reading. For example, if they read VOA news regularly for 3 months at the speed of 200 wpm, they could potentially learn more than 80% of the core academic spoken words, more than 70% of the core academic written words, more than 20% of the core academic formulas, and nearly 20% of the core general formulas. However, if they read the news for 6 months, the percentage of vocabulary they could potentially learn would increase to more than 90% (academic spoken words), more than 80% (academic written words), more than 30% (academic formulas), and nearly 40% (general formulas). This would motivate learners to read more and help teachers to evaluate the relative value of online news reading for students in their specific courses.

Last but not least, by suggesting online news as a useful source for L2 learners to learn core academic words, academic formulas, and general formulas, we do not mean that it is the one and only source for L2 learners to learn these lexical items. The fairly large amount of time needed for learners to incidentally learn core academic words, academic formulas, and general formulas suggests that deliberately learning these lexical items is important because it helps learners to learn a larger number of words and phrases in a shorter period of time than incidental learning (Webb & Nation, 2017). However, it does not mean that incidental learning does not have any value. Exposure to a large amount of meaning-focused input such as online news allows learners to repeatedly meet core academic words, general formulas and general formulas in various contexts, which helps them to learn the form-meaning connection of unknown items but at the same time expand their knowledge of other aspects of known items (Webb, 2020). In fact, research has suggested that to better support learners' lexical development, teachers should combine both incidental learning and deliberate learning (Webb & Nation, 2017). Therefore, reading online news should not be the only activity for learners to learn core academic words, academic formulas, and general formulas, but should be combined with other activities in Nation's (2007) Four Strands. In this way, learners would have various opportunities to deliberately and incidentally learn core academic words, academic formulas, and general formulas.

CONCLUSION

As the first corpus-driven study exploring sources for incidental learning of formulas and academic vocabulary, this study indicates that

online news is a potential source for incidental learning of core academic words, academic formulas, and general formulas. Given the limited resources for learners in EFL contexts to incidentally learn core formulas and academic vocabulary, the fact that simply reading online news regularly for entertainment can still allow learners to learn a wide range of vocabulary that is essential for their language development makes the finding of the present study meaningful. This study also suggests that smaller amounts of reading would be needed for incidental learning of core academic spoken words and academic written words than for core academic formulas and general formulas. This finding provides teachers and learners with useful implications of the minimal amounts of reading needed when online news is used as the input and highlights the importance of speed reading in language learning programs.

ACKNOWLEDGMENTS

We would like to thank Charlene Polio and the anonymous reviewers for their constructive feedback, which has improved the quality of the article largely. Our thanks to Thuy Bui, Cailing Lu, and Mohammad Ahmadian for their useful comments on the earlier version of the article.

THE AUTHORS

Thi Ngoc Yen Dang is a Lecturer at the University of Leeds. Her research interests include vocabulary studies and corpus linguistics. Her articles have been published in various journals (e.g., *Applied Linguistics*, *Language Learning*, *TESOL Quarterly*, *Language Teaching Research*, *Studies in Second Language Acquisition*, and *System*)

Xu Long is an English teacher and the founder of an education consultation company in China. She got Masters' degree of Translation and Interpreting in Sichuan University of China and Masters' degree of Education in the University of Leeds. She has more than eight years' rich experience in English teaching for EFL learners.

REFERENCES

Alsaif, A., & Milton, J. (2012). Vocabulary input from school textbooks as a potential contributor to the small vocabulary uptake gained by English as a foreign language learners in Saudi Arabia. *Language Learning Journal*, *40*(1), 21–33.

Anthony, L. (2017). *AntFileSplitter (Version 1.0.0) [Computer Software]*. Tokyo, Japan: Waseda University. Retrieved from https://www.laurenceanthony.net/software

Boers, F. (2020). Factors affecting the learning of multiword items. In S. Webb (Ed.), *The Routledge handbook of vocabularies studies* (pp. 143–157). New York: Routledge.
Brown, R., Waring, R., & Donkaewbua, S. (2008). Incidental vocabulary acquisition from reading, reading-while-listening, and listening to stories. *Reading in a Foreign Language*, *20*(2), 136–163.
Cobb, T. (n.d.). Compleat lexical tutor (8.5) [Computer Software]. Retrieved from https://www.lextutor.ca/vp/
Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*, 213–238. https://doi.org/10.2307/3587951
Coxhead, A. (2018). *Vocabulary and English for specific purposes research: Quantitative and qualitative perspectives*. New York: Routledge.
Coxhead, A. (2020). Academic vocabulary. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 97–110). New York: Routledge.
Coxhead, A., & Hirsh, D. (2007). A pilot science-specific word list. *Revue Française de Linguistique Appliqueé*, *12*(2), 65–78.
Coxhead, A., Rahmat, Y., & Yang, L. (2020). Academic single and multiword vocabulary in EFL textbooks: Case studies from Indonesia and China. *TESOLANZ Journal*, *28*, 75–88.
Csomay, E., & Petrovíc, M. (2012). "Yes, your honor!": A corpus-based study of technical vocabulary in discipline-related movies and TV shows. *System*, *40*, 305–315.
Dang, T. N. Y. (2020a). Corpus-based word lists in second language vocabulary research, learning, and teaching. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 288–304). New York: Routledge.
Dang, T. N. Y. (2020b). The potential for learning specialized vocabulary of university lectures and seminars through watching disciplines-related TV programs: Insights from medical corpora. *TESOL Quarterly*, *54*, 436–459. https://doi.org/10.1002/tesq.552
Dang, T. N. Y. (2022). Using VocabProfiers to select texts for extensive reading activities. In V. Viana (Ed.), *Teaching English with corpora*. New York: Routledge.
Dang, T. N. Y., Coxhead, A., & Webb, S. (2017). The academic spoken word list. *Language Learning*, *67*(4), 959–997. https://doi.org/10.1111/lang.12253
Ellis, R. (1999). *SLA research and language teaching*. Oxford: Oxford University Press.
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
Hsu, W. (2019). Voice of America news as voluminous reading material for mid-frequency vocabulary learning. *RELC Journal*, *50*(3), 408–421. https://doi.org/10.1177/0033688218764460
Jordan, G., & Gray, H. (2019). We need to talk about coursebooks. *ELT Journal*, *73*(4), 438–446.
Laufer, B. (2020). Lexical coverages, inferencing unknown words and reading comprehension: How are they related? *TESOL Quarterly*, *54*(4), 1076–1085. https://doi.org/10.1002/tesq.3004
Lu, C., & Coxhead, A. (2019). Vocabulary in traditional Chinese medicine. *ITL International Journal of Applied Linguistics*, *171*(1), 34–61.
Martinez, R., & Schmitt, N. (2012). A phrasal expressions list. *Applied Linguistics*, *33*(3), 299–320.
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. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge: Cambridge University Press.
Nation, I. S. P. (2014). How much input do you need to learn the most frequent 9,000 words? *Reading in a Foreign Language*, *26*(2), 1–16.
Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. Amsterdam, Netherlands, John Benjamins.
Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher*, *31*(7), 9–13.
Nergis, A. (2011). Exploring the factors that affect reading comprehension of EAP learners. *Journal of English for Academic Purposes.*, *12*, 1–9.
Nguyen, T. M. H., & Webb, S. (2017). Examining second language receptive knowledge of collocation and factors that affect learning. *Language Teaching Research*, *21*(3), 298–320. https://doi.org/10.1177/1362168816639619.
Pellicer-Sánchez, A., & Schmitt, N. (2010). Incidental vocabulary acquisition from an authentic novel: Do things fall apart? *Reading in a Foreign Language*, *22*(1), 31–55.
Peters, E. (2020). Factors affecting the learning of single-word items. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 125–142). New York: Routledge.
Read, J., & Dang, T. N. Y. (2022). Measuring depth of academic vocabulary knowledge. *Language Teaching Research*. https://doi.org/10.1177/13621688221105913
Rodgers, M. P. H., & Webb, S. (2011). Narrow viewing: The vocabulary in related television programs. *TESOL Quarterly*, *45*, 689–717.
Rolls, H., & Rodgers, M. P. H. (2017). Science-specific technical vocabulary in science fiction-fantasy texts: A case for 'language through literature'. *English for Specific Purposes*, *48*, 44–56.
Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. New York: Palgrave Macmillan.
Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behaviour of two new versions of the Vocabulary Levels Test. *Language Testing*, *18*(1), 55–88.
Simpson-Vlach, R., & Ellis, N. C. (2010). An Academic Formulas List: New methods in phraseology research. *Applied Linguistics*, *31*(4), 487–512.
Siyannova-Chanturia, A., & Pellicer-Sánchez, A. (2019). *Understanding formulaic language: A second language acquisistion perspective*. New York: Routledge.
Sun, Y., & Dang, T. N. Y. (2020). Vocabulary in high-school EFL textbooks: Texts and learner knowledge. *System*, *93*, 102279. https://doi.org/10.1016/j.system.2020.102279
Teng, T. (2015). EFL vocabulary learning through reading BBC news: An analysis based on the involvement load hypothesis. *English as a Global Language Education Journal*, *1*(2), 63–90.
Uchihara, T., Webb, S., & Yanagisawa, A. (2019). The effects of repetition on incidental vocabulary learning: A meta-analysis of correlational studies. *Language Learning*, *69*(3), 559–599.
van Zeeland, H., & Schmitt, N. (2013). Incidental vocabulary acquisition through L2 listening: A dimensions approach. *System*, *41*, 609–624.
Vidal, K. (2011). A comparison of the effects of reading and listening on incidental vocabulary acquisition. *Language Learning*, *61*(1), 219–258.
Vilkaite, L. (2016). Are nonadjacent collocations processed faster? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *42*(10), 1632–1642. https://doi.org/10.1037/xlm0000259

Webb, S. (2007). The effects of repetition on vocabulary knowledge. *Applied Linguistics*, *28*(1), 46–65.
Webb, S. (2010). A corpus driven study of the potential for vocabulary learning through watching movies. *International Journal of Corpus Linguistics*, *15*(4), 497–519.
Webb, S. (2020). Incidental vocabulary learning. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 225–239). New York: Routledge.
Webb, S. (2021). Research investigating lexical coverage and lexical profiling: What we know, what we don't know, and what needs to be examined. *Reading in a Foreign Language*, *33*(2), 278–293.
Webb, S., & Nation, I. S. P. (2017). *How vocabulary is learned*. Oxford: Oxford University Press.
Webb, S., Newton, J., & Chang, A. C.-S. (2013). Incidental learning of collocation. *Language Learning*, *63*, 91–120.
Webb, S., & Rodgers, M. P. H. (2009a). The lexical coverage of movies. *Applied Linguistics*, *30*(3), 407–427.
Webb, S., & Rodgers, M. P. H. (2009b). Vocabulary demands of television programs. *Language Learning*, *59*(2), 335–366.
Webb, S., Sasao, Y., & Balance, O. (2017). The Updated Vocabulary Levels Test: Developing and validating two new forms of the VLT. *ITL-International Journal of Applied Linguistics*, *168*(1), 34–70.
Wray, A. (2019). Concluding question: Why don't second language learners more proactively target formulaic sequences? In A. Siyanova-Chanturia & A. Pellicer-Sánchez (Eds.), *Understanding formulaic language: A second language acquisition perspective* (pp. 248–269). New York: Routledge.

Supporting Information

Additional Supporting Information may be found in the online version of this article:

Data S1.