The Language Learning Journal

ISSN: 0957-1736 (Print) 1753-2167 (Online) Journal homepage: www.tandfonline.com/journals/rllj20

Student talk as a source for incidental vocabulary learning: what can corpora tell us?

Thi Ngoc Yen Dang & Phung Dao

To cite this article: Thi Ngoc Yen Dang & Phung Dao (2025) Student talk as a source for incidental vocabulary learning: what can corpora tell us?, The Language Learning Journal, 53:4, 521-534, DOI: 10.1080/09571736.2024.2397658

To link to this article: https://doi.org/10.1080/09571736.2024.2397658

THE LANGUAGE LEARNING JOURNAL
2025, VOL. 53, NO. 4, 521–534
https://doi.org/10.1080/09571736.2024.2397658

Routledge
Taylor & Francis Group

Student talk as a source for incidental vocabulary learning: what can corpora tell us?

Thi Ngoc Yen Dang [a] and Phung Dao [b]

[a]School of Education, University of Leeds, Leeds, United Kingdom; [b]Faculty of Education, University of Cambridge, Cambridge, United Kingdom

ABSTRACT

While experimental studies have confirmed the value of student talk for incidental L2 vocabulary learning, no studies have evaluated it from the corpus linguistics perspective. Taking the Vietnamese EFL university context as a case, this study used corpus linguistics to investigate the value of student talk for incidental vocabulary learning. Three corpora were created: a cross-sectional corpus of 381 student-student interactions in tertiary English language courses, and two longitudinal corpora derived from all lessons in each course. The analyses of words, represented by Nation's (2012) BNC/COCA lists, and core formulaic sequences, represented by Martínez and Schmitt's (2012) Phrasal Expressions List, were respectively undertaken with RANGE and Sketch Engine. Results show that knowledge of the most frequent 1000-word families is needed for reasonable comprehension of student talk. 5.23%−23% of the high-frequency word families and 0.59%−13.47% of the core formulaic sequences were encountered multiple times during each course. The results suggest student talk is an excellent source for learning and reinforcing high-frequency vocabulary and a good source for learning and reinforcing core formulaic sequences. Our study highlights the value of student talk as an input source for low-proficiency learners to learn vocabulary incidentally and showcases how corpus linguistics can inform L2 pedagogy.

ARTICLE HISTORY
Received 8 January 2024
Accepted 23 August 2024

KEYWORDS
Student talk; incidental vocabulary learning; corpus linguistics; student-student interaction

Introduction

High-frequency words (e.g. *effort*), mid-frequency words (e.g. *destination*), and core formulaic sequences (e.g. *in order to*) are those that occur frequently in various contexts of target language use (Nation 2022). Knowledge of these lexical items is essential for effective communication in the target language. However, research has consistently found that students in different EFL contexts have limited knowledge of high and mid-frequency words and formulaic sequences and need further support in their vocabulary development (e.g. Nguyen and Webb 2017). In addition to deliberate vocabulary learning activities, teachers should encourage students to engage with various sources of L2 input to learn these lexical items incidentally (Nation 2007). For incidental vocabulary learning to occur, students should have exposure to a large amount of input (Webb 2020). Therefore, an important strand of vocabulary studies is to count the occurrences of lexical items in corpora representing a certain kind of input to evaluate its value for incidental vocabulary

CONTACT Thi Ngoc Yen Dang T.N.Y.Dang@leeds.ac.uk

learning. Corpora are principled collections of a large number of texts in electronic format (O'Keeffe, McCarthy, and Carter 2007). Corpus-driven vocabulary research has suggested diverse sources for incidental vocabulary learning (e.g. movies, novels, songs). However, most of them require a knowledge of 3000 or more word families[1] for incidental learning to happen. Such knowledge is beyond the vocabulary level of a considerable number of EFL learners (e.g. Webb and Chang 2012).

Student talk generated in learner-learner interaction (or peer interaction) is considered essential and conducive to L2 learning when viewed through multiple theoretical frameworks, such as cognitive-interactionist, socio-cultural, and socio-constructive peer learning perspectives. From the cognitive-interactionist perspective, student talk in peer interaction serves as a source of input that facilitates the process of learning an L2 across various dimensions, including grammatical, phonological, and lexical aspects (Gass and Mackey 2020; Long 1996; see also Krashen 1985 for the essential role of student talk as comprehensible input in L2 learning). Additionally, student talk in peer interaction represents L2 learning opportunities in which learners may (a) receive modified input from peers, (b) notice language errors in their own and their peers' talk, (c) negotiate for meaning and form, and (d) practise producing output (see Swain 2000 for the discussion of three functions of student talk or output) – all of these elements are considered central and conductive to the learning of an L2 (including vocabulary) (Mackey 2012). From the socio-cultural perspective, student talk in peer interaction is indicative of learner-centered, dialogic, and purposeful talk (Mercer 1996, 2019) and/or and collaborative dialogue (Swain 2006, 2010; Swain and Watanabe 2012) throughout which learners may scaffold each other's talk, provide assistance, and pool each other's linguistic and non-linguistic resources to co-construct L2 knowledge. In socio-cultural terms, student talk is a mediation tool that learners use for languaging – 'a process of making meaning and shaping L2 knowledge and experience through language' (Swain 2006, 98) and it is 'a social mode of thinking – a tool for the joint construction of knowledge among learners' (Mercer 2019, 63). From the socio-constructive peer learning perspective, student talk is representative of an active process where learners construct their L2 knowledge with others through social interaction (i.e. peer interaction) (Mercer 2019) and personal experience with the environment elements (e.g. task, content, peer, and teachers) in order to actively construct knowledge rather than passively receiving it.

Apart from the crucial role of student talk in L2 learning as explained from the three perspectives above, student talk may be considered among the major sources of L2 input for EFL learners (Lu and Dang 2023), especially for those with low proficiency levels, bearing in mind the current recognition or advocacy of learner-centered L2 teaching approaches (e.g. communicative language teaching and Task-Based Language Teaching, TBLT) in different educational contexts. Characteristically, student talk, produced by learners themselves, may be less lexically demanding and thus may create less cognitive burden on learners' working memory than other kinds of input. This may have the advantage of allowing learners to devote more cognitive resources to noticing and learning vocabulary from the input.

It is often assumed that student talk contains language errors and infelicities, and thus is not perceived as a reliable source for L2 learning. However, no empirical research to date has documented the negative impact of language errors in student talk on the learning of an L2. Thus, theoretically speaking, student talk can arguably be considered as a facilitative source/input conductive to L2 learning. Indeed, a considerable amount of L2 interaction research to date has documented the positive impact of student talk in peer interaction, not only on learning vocabulary (e.g. Newton 2013) but also on other language-related dimensions (see Sato and Ballinger 2016 and Philp, Adams, and Iwashita 2014 for reviews of empirical evidence for the positive impact of student talk in peer interaction for L2 learning).

Despite the potential of student talk for language learning, little research has employed corpus linguistics as a method to evaluate the value of student talk for incidental vocabulary learning, especially the learning of core formulaic sequences. This lack of research could be because of the difficulty of creating spoken corpora. It could also be due to the assumption that student talk might contain errors and may not be a good source for L2 learning. However, empirical L2 research has shown

that student talk produced during peer interaction is conductive to incidental L2 learning, including lexical items (Kim 2008, 2011; Sato and Ballinger 2016; see also Philp, Adams, and Iwashita 2014).

This study therefore aims to explore the potential of talk produced by Vietnamese EFL university students for incidental learning of lexical items using corpus linguistics. Its findings should allow us to contribute further to understanding of the value of student talk for vocabulary learning and generate useful implications for designing activities to facilitate EFL learners' vocabulary development.

High-frequency word families, mid-frequency word families, and core formulaic sequences

Vocabulary in a language can be classified into 1000-word levels according to their frequency (Nation 2022). High-frequency word families are those from the 1st to the 3rd most frequent 1000-word family levels (e.g. *famous*). Mid-frequency word families are those from the 4th to the 9th most frequency 1000-word family levels (e.g. *obesity*). Low-frequency word families refer to the words outside the most frequent 9000-word families (e.g. *litany*). High-frequency word families are the smallest in number but make up the largest proportion of texts (75%−92%). Meanwhile, mid-frequency word families are larger in number, but account for a smaller proportion of texts (2% −14%). Low-frequency word families are the largest in number, but account for the smallest proportion of texts (less than 2%). Because the percentage of known words in a text positively contributes to text comprehension (e.g. van Zeeland and Schmitt 2013), vocabulary researchers have suggested that a good knowledge of high and mid-frequency word families is essential for successful communication and that learners should learn high-frequency word families before moving on to mid-frequency word families. As a result, several high-frequency word lists have been developed. Research (Dang and Webb 2017; Dang, Webb, and Coxhead 2022b) has found that Nation's (2012) BNC/COCA2000 list appears to be the most relevant to EFL learners, or at least to Vietnamese EFL learners. Fewer lists of mid-frequency word families have been created. Nation's (2012) BNC/COCA lists of mid-frequency word families are the most popular. They have been adopted to represent mid-frequency word families in various vocabulary studies (e.g. Nation 2014).

Knowledge of core formulaic sequences is also important for EFL learners. Core formulaic sequences are the sequences of words that occur frequently in different kinds of texts in the target language (e.g. *as a result*). Knowledge of these sequences helps learners to process the target language faster, which enables them to become more proficient language users (Siyanova-Chanturia and Pellicer-Sánchez 2019). Lists of core formulaic sequences are limited in number. To date, Martínez and Schmitt's (2012) Phrasal Expressions List is probably the best-known list of core formulaic sequences. It consists of 505 formulaic sequences that occur frequently in English and has been suggested as a useful learning goal for L2 learners (Schmitt and Schmitt 2020).

Potential sources for incidental vocabulary learning

Incidental vocabulary learning means learning vocabulary through participating in meaning-focused activities (e.g. reading books, watching films) (Webb and Nation 2017). As incidental learning is a slow and gradual process, learners need repeated exposure to a large amount of comprehensible input. Therefore, an important line of vocabulary research is to use information from corpora to estimate the potential of different kinds of input for L2 learners. This line of research investigates the potential for incidental learning from two perspectives: the comprehensibility of the input and the degree of repetition of target vocabulary.

Estimating the number of words necessary to reach particular levels of comprehension is useful because input needs to be comprehensible for incidental vocabulary learning (Nation 2007). If learners understand the content of the input, they are more likely to notice unknown vocabulary or unknown aspects of vocabulary knowledge in meaning-focused input and subsequently learn

them. If the input is too difficult, learners may struggle with understanding its content and may not have sufficient cognitive resources to notice and process the new lexical items in the input.

A typical approach to evaluating input comprehensibility is to estimate the amount of vocabulary required to reach a specific lexical coverage cut-off point. Lexical coverage is the proportion of the words in a text that learners are likely to know (Nation and Webb 2011). A reasonable number of studies have examined the lexical coverage needed for comprehension of written discourses. Laufer (1989) found that 95% coverage was needed for reasonable comprehension of an academic text while Hu and Nation (2000) found that 98% was necessary for adequate unassisted comprehension of easy L2 fiction text. However, Schmitt, Jiang, and Grabe (2011) found that although lexical coverage and reading comprehension significantly correlated with each other, there was no coverage threshold for comprehension. Similarly, Laufer (2020) found no significant difference in the reading comprehension and inferencing scores of L2 learners reading the 95% text coverage version and the 98% version. Therefore, subsequent studies often follow Laufer and Ravenhorst-Kalovski's (2010) suggestions by adopting two lexical coverage cut-off points to indicate different degree of reading comprehension: 95% for minimal and 98% for optimal comprehension.

Fewer studies have examined lexical coverage for comprehension of spoken texts. van Zeeland and Schmitt (2013) replicated Hu and Nation's study with listening and focused on informal monologue narratives as the input. Their findings indicated that unlike reading, L2 learners may not need to aim for 98% coverage for good listening comprehension. Both 90% and 95% coverage led to relatively high listening comprehension. While the comprehension scores were less varied among participants in the case of 95% coverage, there was no significant difference in listening comprehension between 90% and 95% coverage. Later, Giordano (2021) replicated van Zealand and Schmitt's study with dialogues and found similar results. Therefore, he suggested that teachers could aim for wider range of lexical coverage (90%–98%) when selecting listening materials for L2 learners.

Taken together, previous studies have indicated that the lexical coverage needed for comprehension of texts may vary according to modality, kind of texts and degree of comprehension. Lower lexical coverage tends to be needed in the case of listening (90%–98%) compared with reading (95%–98%). The lower lexical coverage for listening could be because in spoken communication, other factors may support listeners' comprehension such as contextual clues (e.g. facial expressions) (Harris 2003) and co-construction of knowledge during the interaction (Vandergrift 2003). Despite the difference in the lexical coverage for listening and reading comprehension, previous corpus-driven studies only employed 95% and 98% as the lower and upper lexical coverage indicators of different comprehension levels. No studies have examined the vocabulary size needed for 90% coverage.

Nation's (2006) corpus-driven study found that knowledge of the most frequent 6000–7000 word families and the most frequent 8000–9000 word families is necessary for reaching 98% coverage of both spoken and written texts. Subsequent studies found that for 95% coverage, students would have to know the most frequent 2000 word families in the case of graded readers, 3000 word families in the cases of television programmes, movies, and songs, 4000 word families in the case of news, and 5000 word families in the case of L1 children's literature (see Nurmukhamedov and Webb 2019). A larger number of words are needed to achieve 98% coverage: the most frequent 3000 word families for graded readers, 6000 word families for movies and songs, 7000 word families for television programmes, and 10,000 word families for L1 children literature. Taken together, except for graded readers, L2 learners should know the most frequent 3000 word families or more for reasonable comprehension of these sources. It means that these sources may not always be appropriate for learners who have limited knowledge of high-frequency word families. Although some studies (e.g. Green 2022a) have found that with careful selection, teachers may be able to identify some authentic books as comprehensible input for learners with smaller vocabulary sizes, the number of these books is limited and may not always match learners' interests. Therefore, it is important to explore other resources for low-level learners to learn vocabulary incidentally.

Apart from comprehensible input, repetition (i.e. the number of encounters with a certain lexical item) is also important for incidental learning to occur. While there are no definite frequency cut-off points for incidental vocabulary learning to occur, the chance for a lexical item to be learned increases with the number of times it is encountered (e.g. Dang, Lu, and Webb 2022a, 2022b, 2023). Thus, a trend in corpus-driven incidental learning studies is to examine the re-occurrences of certain kinds of lexical items in input. Several studies have examined the potential for learning high frequency word families from graded readers (Claridge 2005; Wan-a-rom 2008), mid-frequency word families from novels (Green 2022a; Nation 2014) and news (Hsu 2019) and low-frequency word families from texts for L1 children, L2 learners, and L1 older readers (Webb and Macalister 2013), movies (Webb 2010), and television programmes (Rodgers and Webb 2011). Other studies have investigated the potential for learning specialised word families from novels (e.g. Rolls and Rodgers 2018), and online news (Dang and Long 2024), specialised flemmas[2] from novels, television programmes, and movies (Green 2022b), and specialised word types[3] from L1 children's texts (Gardner 2008) and discipline-related television programmes (e.g. Dang 2020). Very few studies have examined the learning of core formulaic sequences. Dang and Long (2024) used Martínez and Schmitt's (2012) Phrasal Expressions List to represent core formulaic sequences and found that online news is a valuable resource for students to learn these sequences.

Compared to out-of-class input, a smaller number of studies have used corpus linguistics to examine the potential for incidental vocabulary learning from in-class input. Most of them have examined the potential of English language textbooks for incidental vocabulary learning (e.g. Coxhead, Rahmat, and Yang 2020; Nguyen 2021; Yang and Coxhead 2020). Their findings indicate that depending on the examined textbooks, a vocabulary size of between 2500–13,000 word families is required to achieve 95% coverage. Moreover, textbooks do not appear to offer enough opportunities for incidental vocabulary learning. Meanwhile, two studies have looked at the potential of teacher talk. Horst (2010) found that knowledge of the most frequent 2000-word families and the most frequent 4000-word families is needed to reach 95% and 98% coverage of teacher talk in ESL class in Canada. She also found very few unfamiliar word families being recycled in talk. Coxhead (2017) found that knowledge of the most frequent 2000–3000 word families is needed to reach 95% coverage of a corpus of teacher talk at an international school in Germany, and 4000–7000 word families are needed to achieve 98% coverage. The vocabulary load of teacher talk went up from 2000-word families at the beginning of the course to 3000-word families in the middle of the course and remained stable at the end of the course.

By identifying a range of potential sources for incidental L2 vocabulary learning, previous research has made significant contributions to our understanding of incidental vocabulary learning. However, several areas need further attention. First, to comprehend most of these input types, learners would have to know at least the most frequent 3000-word families. Yet a number of EFL learners do not have sufficient knowledge of the most frequent 3000-word families (e.g. Olmos 2009). Therefore, these resources may not be appropriate for these low-level learners.

Student talk may be a potential resource for them. It is among the most frequent sources of L2 input to which learners are exposed in EFL contexts (Lu and Dang 2023). Research (e.g. Nguyen 2021; Kim 2008, 2011) has also found that vocabulary can be learned through peer interaction. Moreover, when organising learning activities, teachers are likely to prompt students to use target lexical items during task performance (Dao and Iwashita 2018); thus, when performing learning tasks, learners are more likely to use lexical items and structures that have been 'seeded' or are required by the design of the learning task (e.g. focused tasks). These conditions enable students to encounter linguistic items (e.g. words) repeatedly in peers' talk, which can then result in subsequent L2 learning (Dao and McDonough 2018). Despite its potential, to the best of our knowledge, few studies have employed corpus linguistics as a method to explore the value of student talk as a resource for incidental L2 vocabulary learning.

Furthermore, earlier corpus-driven studies typically only examined repetition from the perspective of the total occurrences of lexical items in the input. However, various levels of repetition

may facilitate learning differently (Webb and Nation 2017). First, spaced repetition tends to lead to greater long-term retention than massed repetition (Nakata 2015). Therefore, repetition should be spaced within a learning session and across sessions. Second, apart from verbatim repetition (i.e. meeting the same word in the same form and context), varied repetition (i.e. meeting the same word in the different forms or contexts, or both) is also important to strengthen and expand learners' vocabulary knowledge.

Previous corpus-driven studies have also focused mainly on the learning of single words; only a few studies have examined the potential for learning of formulaic sequences (Coxhead, Rahmat, and Yang 2020; Dang and Long 2024) and these studies related respectively to textbooks and online news. The modest number of studies on formulaic sequences is surprising given that knowledge of these lexical items is important for L2 learners, but a considerable number of EFL learners have limited knowledge of these lexical items (e.g. Lu and Dang 2023).

To address these gaps, this corpus-driven study aims to examine the potential of student talk in the Vietnamese EFL university context for the incidental learning of high-frequency word families, mid-frequency word families, and core formulaic sequences from two perspectives: comprehensibility and repetition (spaced repetition and varied repetition). The Vietnamese EFL context was chosen as a case for two reasons. First, similar to learners in other EFL contexts such as China and Spain (e.g. Lu and Dang 2023; Olmos 2009), a large number of Vietnamese EFL learners do not have sufficient knowledge of high-frequency word families and formulaic sequences (e.g. Nguyen and Webb 2017), and thus need further support. Second, the Vietnamese EFL context shares features with many other EFL contexts; in particular, English is the most frequently studied foreign language subject in Vietnamese schools, where it is a compulsory subject at secondary level from the age of 11. Learners' contact with English is almost entirely within the language classroom. At the age of 18, Vietnamese students complete a compulsory English language test as part of the National High School Graduation Exam. To be admitted to study at a higher education institution, English-major students are required to pass this English test or equivalent, although this is not always obligatory for non-English major students. During their first/second year of higher education study, students need to take compulsory English courses to reach the language exit requirements of their institutions. In addition to these compulsory courses, many students take extra English courses at private language centres to improve their English language proficiency. Given these features, the findings of the present study are likely to be relevant to other similar EFL contexts.

Research questions

(1) How many word families are required to understand student talk in the Vietnamese EFL university context?
(2) To what extent do high-frequency word families, mid-frequency word families, and core formulaic sequences occur in student talk in this context?

Methodology

Corpora

A cross-sectional student talk corpus and two longitudinal student talk corpora were developed in this study. Analysis of the cross-sectional corpus should provide information about the occurrences of high-frequency word families, mid-frequency word families, and core formulaic sequences in student talk from a range of courses. Meanwhile, analysis of the longitudinal corpora should provide in-depth insight into the changes in the occurrences of these words over the duration of a specific course. Such information should thus enable us to evaluate the value of student talk for incidental vocabulary learning from different perspectives.

The cross-sectional corpus was derived from the transcripts of 381 student-student interactions in compulsory and optional English language courses for EFL university students in Vietnam. This corpus (190,786 words) was larger than the teacher talk corpora of Horst's (2010) (120,553 words) and Coxhead's (2017) (107,587 words) studies. Table 1 presents the structure of the corpus. It consists of two sub-corpora: compulsory courses at university/college and optional courses at private language centres.

The input for the compulsory courses was collected from both English major and non-English major courses. For each group, we sampled data from one university-level institution and one college-level institution, thus representing the two typical kinds of higher education institutions in the Vietnamese EFL context. The language proficiency levels of the English-major and non-English major students were, respectively, around the intermediate/upper-intermediate (B1 and B2), and pre-intermediate (A2) levels based on the CEFR. We collected samples from one class per institution and one lesson per class. Meanwhile, the input for the optional courses at private English language centres was collected from two English language centres, from the typical kind of General English courses delivered by private language centres in the Vietnamese EFL context. Courses in the first language centre aimed to enhance students' four language skills (reading, writing, speaking, and reading) and grammar knowledge. Courses in the second language centre aimed to develop learners' English conversational skills, with a few lessons focusing on grammar knowledge and pronunciation. In each center, we collected data from two classes, representing different language proficiency levels based on the results of the language centre's internal placement tests. The language proficiency of these students was mixed, ranging roughly from pre-intermediate (A2) to upper-intermediate (B2) level based on the CEFR. Samples of student talk were collected from one lesson per class.

The two longitudinal corpora were created to provide further insights into the lexical profile and the occurrences of high-frequency word families, mid-frequency word families, and core formulaic sequences in student talk over the duration of a specific course. Data for the first longitudinal corpus were collected over five lessons from a university course whereas those of the second corpus were collected over seven lessons from an optional course at a private language centre.

Data analysis

To analyse the occurrences of high and mid-frequency word families in student talk, we used the RANGE program (Heatley, Nation, and Coxhead 2002) to count the occurrence of items from Nation's (2012) BNC/COCA 25 1000-word family lists (Nation 2012) in each student talk corpus. Items from the 1st to the 3rd 1000 BNC/COCA-word family lists represent high-frequency word families; those from the 4th to the 9th 1000 BNC/COCA-word family lists are mid-frequency word families; those from the 10th to the 25th 1000 BNC/COCA-word family list are low-frequency word families. Nation's lists were chosen because they have been widely used in corpus-driven vocabulary studies to represent high, mid, and low-frequency word families. To analyse the occurrences of core formulaic sequences, we searched items[4] from Martínez and Schmitt's (2012) Phrasal Expression List (PHRASE) in each corpus with the concordance function of Sketch Engine. The

Table 1. Components of the cross-sectional student talk corpus (190,786 words).

Compulsory courses at university/college (132,627 words)				Optional courses at private language centres (58,159 words)			
English major		Non-English major		4 skills plus grammar		Conversation	
Sub-components	Words	Sub-components	Words	Sub-components	Words	Sub-components	Words
University level	61,747	University level	27,255	High proficiency	8,241	High proficiency	23,651
College level	9,130	College level	34,495	Low proficiency	6,616	Low proficiency	19,651
Total	70,877	Total	61,750	Total	14,857	Total	43,302

PHRASE was used to represent the core formulaic sequences because it is the best known list of core formulaic sequences (Schmitt and Schmitt 2020).

To identify the comprehensibility of student talk, we estimated the number of word families needed for 90%, 95% and 98% coverage of the cross-sectional corpus of student talk and the longitudinal corpora. Expanding on previous corpus-driven studies, this study employed a 90% cut-off point in addition to the 95% and 98% cut-off points. These lexical coverage cut-off points are in line with Giordano's (2021) suggestions that teachers could aim for wider range of lexical coverage (90%–98%) when selecting spoken materials. We added up the coverage of each 1000 BNC/COCA-word family list until we achieved the 90%, 95%, and 98% coverage. Proper nouns, marginal words, compounds, abbreviations, and foreign words (Vietnamese words) were also counted toward the cumulative coverage because they tend to have less learning burden than the other kinds of word (Nation and Webb 2011).

To examine the occurrences of high-frequency word families, mid-frequency word families, and core formulaic sequences in each corpus, we referred to the RANGE and Sketch Engine output to count the repetition of each group of vocabulary in each lesson and each corpus. Following Dang (2020), we set 20 or more, 15 or more, 10 or more, and 7 or more occurrences as the frequency cut-off points indicating the learning of the form-meaning link of new vocabulary, whereas 5 or more occurrences as the frequency cut-off point indicating the learning of other aspects of known vocabulary.

Results

RQ1. The amount of vocabulary needed for comprehension of student talk in the Vietnamese EFL university context

The analysis of the cross-sectional corpus showed that if knowledge of proper nouns, marginal words, compounds, abbreviations, and foreign words is included, the most frequent 1000-word families covered nearly 95% (94.83%) of the cross-sectional corpus while knowledge of the most frequent 2000-word families provided more than 98% of this corpus (98.37%). The analysis of the two longitudinal corpora showed that the lexical demand of those from the private language centre was similar to that found in the analysis of the cross-sectional corpus. The most frequent 1000-word families and the most frequent 2000-word families covered 94.82% and 98.01% of the total number of running words in the language centre corpus. However, student talk in the compulsory university course was slightly more demanding than the optional courses in the language centre. Knowledge of the most frequent 1000-word families and the most frequent 2000-word families only covered 93.62% and 97.53% of this corpus. To reach 98% coverage, the most frequent 3000-word families were necessary.

RQ2a. Occurrences of high-frequency word families in student talk in the Vietnamese EFL context

The analysis of the two longitudinal corpora shows that irrespective of the course, a substantial number of high-frequency word families occurred in student talk, and the number of these word families increased over the duration of the courses. Only 418 and 406 high-frequency word families occurred in the first lesson of the university course and the language centre course. However, by the end of these courses, these figures had doubled in the case of the university course (851 word families) and tripled in the case of language centre course (1332 word families).

In terms of repetition, a reasonable percentage of high-frequency word families occurred multiple times in the student talk. 13.03% and 22.33% of the high-frequency word families occurred at least 5 times in the university and language centre courses, respectively. If higher cut-off points of repetition were applied, the percentages of high-frequency word families occurring multiple times in the two

courses were still 5.23%–10.10% (university courses) and 9.60%–18.50% (private language centre course). Moreover, the percentage of high-frequency word families occurring multiple times also increased from the beginning of the university courses (1.93%–5.57%) and the language centre course (2.07%–5.27%) to the end of these courses (5.23%–13.03% and 9.60%–22.33%)

In terms of spaced repetition, over half of the number of the high-frequency word families that occurred in the student talk appeared in more than one lesson in both the university course (52.88%) and the language centre course (58.18%). Importantly, more than 12% and 9% of the high-frequency word families occurred in all lessons of the university course and the language centre course, respectively.

In terms of varied encounters, around half of the high-frequency word families that occurred multiple times in the courses appeared in varied forms: 53.56% (university course) and 46.06% (language centre).

RQ2b. Occurrences of mid-frequency word families in student talk in the Vietnamese EFL context

In terms of occurrence, a reasonable number of mid-frequency word families occurred in student talk regardless of the course type. This number increased to more than three times over the duration of the university course (from 86 to 276 words) and nearly 7.5 times in the language centre course (from 77 to 569 words).

In terms of repetition, irrespective of the repetition cut-off points, only a very small percentage of mid-frequency word families occurred multiple times in the university course (0.05%–0.25%) and the language centre course (0.05%–0.47%). Moreover, although the percentage of mid-frequency word families with multiple occurrences increased over the two courses, the growth was very slow from the first lesson (0.02%–0.10% in the university course; 0%–0.08% in the language centre course) to the last lesson (0.05%–0.25% in the university course; 0.05%–0.50% in the language centre course). Additionally, although 18.89% and 12.15% of the mid-frequency word families occurred in at least one lesson in, respectively, the university data and the language centre data, none of them occurred in all lessons.

In terms of varied encounters 25.55% of the mid-frequency word families found in the university corpus and 22.10% of the mid-frequency word families in the language learning center corpus had different forms.

RQ2c. Occurrences of core formulaic sequences in student talk in the Vietnamese EFL context

This study found that only a relatively small number of core formulaic sequences appeared in the first lesson of the two courses: 28 (university course) and 52 (language centre course). However, by the end of the courses, these figures had increased by nearly three times in the case of the university course (73 formulas) and nearly 3.5 times in the case of the language centre course (179 formulas).

In terms of repetition, the percentage of core formulaic sequences with multiple occurrences was lower than that of high-frequency word families but higher than that of mid-frequency word families. Nearly 5% and nearly 14% of the core formulaic sequences occurred 5 times or more in, respectively, the university and language centre courses. Even with higher cut-off points of repetition, the percentage of core formulaic sequences with multiple occurrences in both the university course and the language centre course was only 0.59%–2.77% and 3.56%–9.70%, respectively.

The percentage of core formulaic sequences also increased throughout the university course and the language centre course, but at a lower rate: from 0%–1.39% to 0.59%–4.95% in the university course and from 0%–1.98% to 3.56%–11.88% in the language centre course. Moreover, a considerable percentage of core formulaic sequences in the student talk occurred in more than one lesson in

the university course (39.73%) and the language centre course (56.42%). Regardless of course type, more than 5% of these sequences appeared in all lessons.

In terms of varied encounters, although the core formulas have fixed structures, these sequences were used in varied contexts in the two longitudinal student talk corpora (see Appendix, online supplementary material for the use of *kind(s) of* as an example).

Discussion

Based on analysis of the vocabulary in both the cross-sectional corpus and the longitudinal corpora of student talk, this study provides useful insight into the value of student talk for incidental vocabulary learning. This section will highlight key findings for the potential of student talk for incidental vocabulary learning.

The analysis of the cross-sectional corpus and the longitudinal corpora consistently indicated that the most frequent 1000-word families covered nearly 95% of the running words in the student talk, while the most frequent 2000-word families provided around 98% of the running words. This suggests that L2 learners would likely need to know the most frequent 1000-word families and the most frequent 2000-word families for, respectively, a minimum and maximum level of acceptable comprehension of student talk. These findings are meaningful. For incidental learning to occur, the input should be comprehensible to learners (Webb 2015). Earlier corpus-driven studies on in-class and out-of-class input have indicated that learners would need to know at least the most frequent 2000–3000 word families for a minimum level of acceptable comprehension of these input sources and a much larger number of word families for a maximum level of acceptable comprehension (e.g. Nurmukhamedov and Webb 2019). Yet a number of EFL learners, including Vietnamese EFL learners, have limited knowledge of the vocabulary beyond the most frequent 1000-word family level (e.g. Nguyen and Webb 2017). This means that the number of word families needed for comprehension of teacher talk, textbooks, and out-of-class input is typically beyond the vocabulary levels of these learners. Consequently, these resources are less likely to be appropriate for them. In contrast, student talk appears to be less lexically demanding. As found in this study, students would need to know only the most frequent 1000 and 2000-word families for a minimum and maximum level of acceptable comprehension of student talk, respectively. And to achieve basic comprehension of student talk as represented by the 90% coverage, they are likely to need to know less than the most frequent 1000 word families. These vocabulary levels are more aligned with the vocabulary levels of most EFL learners, indicating that student talk may be a more relevant incidental vocabulary learning source for lower-level EFL students than other sources of input.

The current study also shed light on the relative value of student talk for learning different kinds of vocabulary. First, it suggests that student talk is a potential source for incidental learning of high-frequency word families. In addition to having a vocabulary load relevant to students' vocabulary levels, student talk also consisted of a substantial number of high-frequency word families. The analysis of the longitudinal corpora shows that by the end of the university course and the language centre course, students were likely to have encountered, respectively, 28.37% and 44.4% of the high-frequency word families. Moreover, a reasonable percentage of high-frequency word families occurred multiple times in both the university course (5.23%−13%) and the language centre course (9.60%−23%). Importantly, both the percentage of high-frequency word families occurring in these courses and the percentage of high-frequency word families with multiple occurrences increased steadily over the duration of these courses. Additionally, more than 50% of the high-frequency word families that occurred in the student talk appeared in more than one lesson and around 50% of the high-frequency word families that occurred multiple times in the corpus appeared in varied forms. Spaced repetition and varied repetition are important conditions for vocabulary learning to happen (Webb and Nation 2017). Thus, our study suggests that student talk could be a useful resource for incidental learning of high-frequency word families. Previous studies have reported the insufficient knowledge of high-frequency word families among various groups of EFL learners and

have called for investigation of appropriate resources for these learners to learn high-frequency word families incidentally in order to supplement deliberate learning (e.g. Nguyen and Webb 2017). By showing that student talk is a potential source for EFL learners to learn high-frequency word families, this study effectively responds to this call.

The number of mid-frequency word families and core formulaic sequences occurring in the student talk was not as large as that of high-frequency word families, which is entirely predictable. Given the participants' low proficiency level, it would not be expected that they produce many mid-frequency word families and core formulaic sequences in their talk. Despite being less frequent than high-frequency word families, a reasonable number of core formulaic sequences nevertheless occurred multiple times in the student talk corpora (0.59%–4.95% in the university course and 3.56%–13.47% in the language centre course). Moreover, a considerable percentage of these sequences occurred in multiple lessons and in varied contexts. This suggests good opportunities for learning these formulas.

This study found a reasonable number of mid-frequency word families occurring in the student talk, and this number increased over the duration of the courses. However, the number of mid-frequency word families occurring multiple times (0.05%–0.25% in the university course and 0.05%–0.47% in the language centre course) was very small. This suggests that the chances for these word families to be learned through listening to student talk are relatively small. However, readers should not be discouraged by this finding. Nation (2022) advises that students with insufficient knowledge of high-frequency word families should learn these words before moving on to mid-frequency word families; the analysis undertaken in this study suggests that student talk provides valuable input for incidental learning and reinforcement of high-frequency vocabulary. Once learners have well-established knowledge of high-frequency word families, they can more effectively make use of other kinds of input (e.g. movies, television programs) to learn mid-frequency word families. This study thus contributes further to existing L2 interaction research, to confirm that value of student talk in peer interaction, especially in terms of L2 lexical learning (Kim 2008, 2011; Newton 2013).

Several limitations should be acknowledged. First, this study only examined student talk in the Vietnamese EFL context. Although this context shares features with other EFL contexts, replicating studies with learners in other contexts would be valuable. Second, this study only investigated the potential of student talk for vocabulary learning from the receptive perspective (i.e. the potential for EFL learners to learn vocabulary from listening to their peers' talk). It may be useful to investigate also the impact of producing speech on speakers' vocabulary learning in the context of peer interaction.

Pedagogical implications

This study has several pedagogical implications. First, student talk was found to have potential for learning high-frequency word families and core formulaic sequences. To optimize this, teachers could design tasks that prompt students to use high-frequency word families and core formulaic sequences repeatedly in various contexts and encourage them to use different members of high-frequency word families to broaden and deepen their knowledge. Second, despite its value, student talk should obviously not be deemed as the only source for learning vocabulary. Incidental learning needs to be seen as complementary to explicit teaching and learning. Third, the modest opportunities for learning mid-frequency word families highlighted by our analysis suggests that in mixed-proficiency level classes, it may be important to direct students with a good command of high-frequency word families to more effective resources for learning mid and low-frequency word families (e.g. television programmes, novels) and encourage them to look actively for opportunities for vocabulary learning outside the classroom.

Conclusion

This is the first corpus-driven study that has examined the potential of student talk for L2 vocabulary learning, and among the few studies exploring resources for incidentally learning formulaic

sequences. Based on the analysis of both cross-sectional and longitudinal corpora of student talk, it indicates that student talk is an excellent source for incidental learning of high-frequency word families and a good source for learning core formulaic sequences. This is meaningful when considering the limited sources for low-level EFL learners to learn these lexical items incidentally.

Notes

1. A word family (*talk*) covers the base form (*talk*), its inflected forms (e.g. *talks, taking, talked)*, and its derived forms (e.g. *talkative*, *talker*).
2. A flemma (*talk*) covers the base form (*talk*) and its inflected forms (*talks, talking, talked*). It does not distinguish between word classes. That is, *talk* (v) and *talk*(n) are considered as one flemma.
3. Word types refer to unrepeated separate word form. For example, *talk, talks, taking, talked*, *talkative*, *talker* are six word types.
4. The number of occurrences of inflectional variations (*consist of*, *consisting of*, *consisted of*) of the same sequence (*consist of*) was checked separately and then added up together for the total number of occurrences of that sequence for further analysis.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Thi Ngoc Yen Dang http://orcid.org/0000-0002-3189-7776
Phung Dao http://orcid.org/0000-0002-8612-5589

References

Claridge, G. 2005. Simplification in graded readers: measuring the authenticity of graded texts. *Reading in a Foreign Language* 17, no. 2: 144–158.
Coxhead, A. 2017. Academic vocabulary in teacher talk: challenges and opportunities for pedagogy. *Oslo Studies in Language* 9, no. 3: 29–44. doi:10.5617/osla.5845.
Coxhead, A., Y. Rahmat, and L. Yang. 2020. Academic single and multiword vocabulary in EFL textbooks: case studies from Indonesia and China. *TESOLANZ Journal* 28: 75–88.
Dang, T. N. Y. 2020. The potential for learning specialized vocabulary of university lectures and seminars through watching discipline-related TV programs: insights from medical corpora. *TESOL Quarterly* 54, no. 2: 436–459. doi:10.1002/tesq.552.
Dang, T.N.Y., and X. Long. 2024. Online news as a resource for incidental learning of core academic words, academic formulas, and general formulas. *TESOL Quarterly* 58, no.1: 32–62. doi:10.1002/tesq.3208.
Dang, T.N.Y., C. Lu, and S. Webb. 2022a. Incidental learning of single words and collocations through viewing an academic lecture. *Studies in Second Language Acquisition* 44, no. 3: 708–736. doi:10.1017/S0272263121000474.
Dang, T.N.Y., C. Lu, and S. Webb. 2023. Open access academic lectures as sources for incidental vocabulary learning: examining the role of input mode, frequency, type of vocabulary, and elaboration. *Applied Linguistics* 44, no. 4: 747–770. doi:10.1093/applin/amac044.
Dang, T.N.Y., and S. Webb. 2017. Evaluating lists of high-frequency words. *ITL – International Journal of Applied Linguistics* 167, no. 2: 132–158. doi:10.1075/itl.167.2.02dan.
Dang, T.N.Y., S. Webb, and A. Coxhead. 2022b. Evaluating lists of high-frequency words: teachers' and learners' perspectives. *Language Teaching Research* 26, no. 4: 617–641. doi:10.1177/1362168820911189.
Dao, P., and N. Iwashita. 2018. Teacher mediation in L2 classroom task-based interaction. *System* 74: 183–193.
Dao, P., and K. McDonough. 2018. Effect of proficiency on Vietnamese EFL learners' engagement in peer interaction. *International Journal of Educational Research* 88: 60–72.
Gardner, D. 2008. Vocabulary recycling in children's authentic reading materials: a corpus-based investigation of narrow reading. *Reading in a Foreign Language* 20, no. 1: 92–122.
Gass, S.M., and A. Mackey. 2020. Input, interaction, and output in L2 acquisition. In *Theories in Second Language Acquisition*, eds. B. VanPatten, G.D. Keating, and S. Wulff, 192–222. New York: Routledge.
Giordano, M.J. 2021. Lexical coverage in dialogue listening. *Language Teaching Research* 27, no. 6: 1504–1531. http://dx.doi.org/10.1177/1362168821989869.
Green, C. 2022a. Extensive reading for a 9,000-word vocabulary: evidence from corpus modeling. *Reading in a Foreign Language* 34, no. 2: 232–248.

Green, C. 2022b. Computing curriculum time and input for incidentally learning academic vocabulary. *Language Learning and Technology* 26, no. 1: 1–21.
Harris, T. 2003. Listening with your eyes: the importance of speech-related gestures in the language classroom. *Foreign Language Annals* 36, no. 2: 180–187. doi:10.1111/j.1944-9720.2003.tb01468.x.
Heatley, A., I.S.P. Nation, and A. Coxhead. 2002. *Range: A Program for the Analysis of Vocabulary in Texts*. http://www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx.
Horst, M. 2010. How well does teacher talk support incidental vocabulary acquisition? *Reading in a Foreign Language* 22, no. 1: 161–180.
Hsu, W. 2019. Voice of America news as voluminous reading material for mid-frequency vocabulary learning. *RELC Journal* 50, no. 3: 408–421. doi:10.1177/0033688218764460.
Hu, H.C., and P. Nation. 2000. Unknown vocabulary density and reading comprehension. *Reading in a Foreign Language*, *13*(1), 403–430.
Kim, Y. 2008. The contribution of collaborative and individual tasks to the acquisition of L2 vocabulary. *The Modern Language Journal* 92, no. 1: 114–130. doi:10.1111/j.1540-4781.2008.00690.x.
Kim, Y. 2011. The role of task-induced involvement and learner proficiency in L2 vocabulary acquisition. *Language Learning* 61, no. s1: 100–140. doi:10.1111/j.1467-9922.2011.00644.x.
Krashen, S.D. 1985. *The Input Hypothesis: Issues and Implications*. London: Longman.
Laufer, B. 1989. What percentage of text lexis is essential for comprehension? In *Special Language: From Humans Thinking to Thinking Machine*, eds. C. Lauren, and M. Nordman, 316–323. Clevedon: Multilingual Matters.
Laufer, B. 2020. Lexical coverages, inferencing unknown words and reading comprehension: How are they related?. *TESOL Quarterly* 54, no. 4: 1076–1085. http://dx.doi.org/10.1002/tesq.v54.4.
Laufer, B., and G.C. Ravenhorst-Kalovski. 2010. Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension. *Reading in a Foreign Language* 22, no. 1: 15–30.
Long, M.H. 1996. The role of linguistic environment in second language acquisition. In *Handbook of Second Language Acquisition*, eds. W.C. Ritchie, and T.K. Bhatia, 413–468. Sandiego: Academic Press.
Lu, C., and T.N.Y. Dang. 2023. Effect of L2 exposure, length of study, and L2 proficiency on EFL learners' receptive knowledge of form-meaning connection and collocations of high-frequency words. *Language Teaching Research*. doi:10.1177/13621688231155820.
Mackey, A. 2012. *Input, Interaction and Corrective Feedback in L2 Learning*. Oxford: Oxford University Press.
Martínez, R., and N. Schmitt. 2012. A phrasal expressions list. *Applied Linguistics* 33, no. 3: 299–320. doi:10.1093/applin/ams010.
Mercer, N. 1996. The quality of talk in children's collaborative activity in the classroom. *Learning and Instruction* 6, no. 4: 359–378. doi:10.1016/S0959-4752(96)00021-7.
Mercer, N. 2019. *Language and the Joint Creation of Knowledge*. New York: Routledge.
Nakata, T. 2015. Effects of expanding and equal spacing on second language vocabulary learning: Does gradually increasing spacing increase vocabulary learning? *Studies in Second Language Acquisition* 37, no. 4, 677–711. https://doi.org/10.1017/S0272263114000825
Nation, I.S.P. 2006. How large a vocabulary is needed for reading and listening? *The Canadian Modern Language Review* 63, no. 1: 59–82. doi:10.3138/cmlr.63.1.59.
Nation, I.S.P. 2007. The four strands. *Innovation in Language Learning and Teaching* 1, no. 1: 1–12. doi:10.2167/illtedi_1.0.
Nation, I.S.P. 2012. The BNC/COCA word family lists. http://www.victoria.ac.nz/lals/about/staff/paul-nation.
Nation, P. 2014. How much input do you need to learn the most frequent 9,000 words? *Reading in a Foreign Language* 26, no. 2: 1–16.
Nation, I.S.P. 2022. *Learning Vocabulary in Another Language*. Cambridge: Cambridge University Press.
Nation, I.S.P., and S. Webb. 2011. *Researching and Analyzing Vocabulary*. Boston: Heinle Cengage Learning.
Newton, J. 2013. Incidental vocabulary learning in classroom communication tasks. *Language Teaching Research* 17, no. 2: 164–187. doi:10.1177/1362168812460814.
Nguyen, C.D. 2021. Fostering L2 academic listening by audio-visual input, input repetition and text-based output tasks. *TESOLANZ Journal* 29: 1–14.
Nguyen, T.M.H., and S. Webb. 2017. Examining second language receptive knowledge of collocation and factors that affect learning. *Language Teaching Research* 31, no. 3: 98–320.
Nurmukhamedov, U., and S. Webb. 2019. Lexical coverage and profiling. *Language Teaching* 52: 188–200.
O'Keeffe, A., M. McCarthy, and R. Carter. 2007. *From Corpus to Classroom*. Cambridge: Cambridge University Press.
Olmos, C. 2009. An assessment of the vocabulary knowledge of students in the final year of secondary education. Is their vocabulary extensive enough? *Journal of English Studies*, Special issue, 3–90.
Philp, J., R. Adams, and N. Iwashita. 2014. *Peer Interaction and Second Language Learning*. New York: Taylor and Francis.
Rodgers, M.P.H, and S. Webb. 2011. Narrow viewing: the vocabulary in related television programs. *TESOL Quarterly* 45, no. 4: 689–717. doi:10.5054/tq.2011.268062.
Rolls, H., and M.P.H. Rodgers. 2018. Science-specific technical vocabulary in science fiction-fantasy texts: a case for 'language through literature'. *English for Specific Purposes* 48: 44–56. doi:10.1016/j.esp.2017.07.002.

Sato, M., and S. Ballinger. 2016. *Peer Interaction and Second Language Learning: Pedagogical Potential and Research Agenda*. Amsterdam: John Benjamins.
Schmitt, N., X. Jiang, and W. Grabe. 2011. The percentage of words known in a text and reading comprehension. *The Modern Language Journal* 95: 26–43.
Schmitt, N., and D. Schmitt. 2020. *Vocabulary in Language Teaching*. Cambridge: Cambridge University Press.
Siyanova-Chanturia, A., and A. Pellicer-Sánchez. 2019. *Understanding Formulaic Language: A Second Language Acquisition Perspective*. London: Routledge.
Swain, M. 2000. The output hypothesis and beyond: mediating acquisition through collaborative dialogue. In *Sociocultural Theory and Second Language Learning*, ed. J.P. Lantolf, 97–114. Oxford: Oxford University Press.
Swain, M. 2006. Languaging, agency and collaboration in advanced second language learning. In *Advanced Language Learning: The Contributions of Halliday and Vygotsky*, ed. H. Byrnes, 95–108. London: Continuum.
Swain, M. 2010. 'Talking-it-through': languaging as a source of learning. In *Sociocognitive Perspectives on Second Language Learning and Use*, ed. R. Batstone, 112–129. Oxford: Oxford University Press.
Swain, M., and Y. Watanabe. 2012. Languaging: collaborative dialogue as a source of second language learning. In *The Encyclopedia of Applied Linguistics*, ed. C.A. Chapelle, 1-8. Chichester: Wiley.
Vandergrift, L. 2003. Listening: theory and practice in modern foreign language competence. Accessed August 14, 2024 from https://web-archive.southampton.ac.uk/www.llas.ac.uk/resources/gpg/67.html.
van Zeeland, H., and N. Schmitt. 2013. Lexical coverage in L1 and L2 listening comprehension: the same or different from reading comprehension? *Applied Linguistics* 34, no. 4: 457–479. doi:10.1093/applin/ams074.
Wan-a-rom, U. 2008. Comparing the vocabulary of different graded-reading schemes. *Reading in a Foreign Language* 20, no. 1: 43–69.
Webb, S. 2010. A corpus driven study of the potential for vocabulary learning through watching movies. *International Journal of Corpus Linguistics* 15, no. 4: 497–519.
Webb, S. 2015. Extensive viewing: Language learning through watching television. In *Language Learning Beyond the Classroom*, eds., D. Nunan, and J.C. Richards, 159–168. New York: Routledge.
Webb, S. 2020. Incidental vocabulary learning. In *The Routledge Handbook of Vocabulary Studies*, ed. S. Webb, 225–239. London: Routledge.
Webb, S.A., and AC-S. Chang. 2012. Second language vocabulary growth. *RELC Journal* 43, no. 1: 113–126. doi:10.1177/0033688212439367.
Webb, S., and J. Macalister. 2013. Is text written for children useful for L2 extensive reading?. *Quarterly* 47, no. 2: 300–322.
Webb, S., and I.S.P. Nation. 2017. *How Vocabulary is Learned*. Oxford: Oxford University Press.
Yang, L., and A. Coxhead. 2020. A corpus-based study of vocabulary in the New Concept English textbook series. *RELC Journal* 53, no. 3: 597–611.