Vocabulary Learning and Instruction

Volume 10, Number 1,

December 2021

VLI Editorial Team

Editor-in-Chief:

Stuart McLean, *Momoyama Gakuin University, Osaka, Japan*

Senior Associate Editors:

Joseph P. Vitta, *Kyushu University, Fukuoka, Japan*
Christopher Nicklin, *Rikkyo University, Tokyo, Japan*

Editorial Board:

Ali H. Al-Hoorie, *Royal Commission for Jubail and Yanbu, Saudi Arabia*
Aaron Batty, *Keio University, Tokyo, Japan*
David Beglar, *Temple University Japan, Osaka, Japan*
Phil Bennett, *University of Niigata Prefecture, Niigata, Japan*
Luke Fryer, *The University of Hong Kong, China*
Aaron Gibson, *Fukuoka University, Fukuoka, Japan*
Phil Hiver, *Florida State University, Tallahassee, USA*
Tomoko Ishii, *Meiji Gakuin University, Tokyo, Japan*
Brandon Kramer, *Kwansei Gakuin University, Osaka, Japan*
Geoffrey G. Pinchbeck, *Carleton University, Ottawa, Canada*
Tim Stoeckel, *University of Niigata Prefecture, Niigata, Japan*
Jeffrey Stewart, *Tokyo University of Science, Tokyo, Japan*
Tatsuya Nakata, *Rikkyo University, Tokyo, Japan*
Kiyomi Yoshizawa. *Kansai University, Osaka, Japan*
Raymond Stubbe, *Kyushu Sangyo University, Fukuoka, Japan*

The Editorial Team expresses a sincere thank you to Mana Ikawa, who designed the logo and the cover for the print version of *VLI*.

Copyright © 2021 *Vocabulary Learning and Instruction*, ISSN: Online 2187-2759. All articles are copyrighted by their respective authors.

Vocabulary Learning and Instruction

Volume 10, Number 1, December 2021

Table of Contents

Articles	Page
Letter from the Editors *Joseph P. Vitta and Christopher Nicklin*	iv
Challenges in the Assumptions of Using a Flemma-based Word Counting Unit *Tomoko Ishii, Phil Bennett, and Tim Stoeckel*	1
Predicting Japanese Primary Schoolchildren's English Vocabulary Knowledge from a Sound-Meaning Recognition Test *Kohei Kanayama*	16
Longitudinal Measurement of the Vocabulary Size of Japanese Junior High School Students: Developing a Vocabulary Size Test for Beginner Learners *Tsuyoshi Sato*	30
"That Sounds About Right"—Lexical Bundle Naturalness Intuitions in Japanese Learners of English *Julian Northbrook and Kathy Conklin*	49
WordUp: An App that Teaches English Words through Extensive Exposure to Authentic Materials *Musa Nushi, Alireza Aghaei, and Maryam Roshanbin*	64

Letter from the Editors

Dear VLI Readers,

Allow us to begin by wishing all of you a very happy New Year and the journal team wishes you a very successful and prosperous 2022. In this issue, we are pleased to present you with five original articles that are sure to stimulate both researchers and teachers with an interest in vocabulary. In the first article, Ishii and colleagues present a mixed-methods study that adds further evidence to the lemma being the most suitable word counting unit when presenting and testing vocabulary to Japanese English (L2) learners. In the second, Kanayama employs advanced statistical modeling to demonstrate a moderate associative relationship between a sound-meaning recognition test and English vocabulary knowledge. The third article of this issue presents a longitudinal study by Sato which shows that the patterns of uptake of target words varies according to their frequency-determined level. In the fourth paper, Northbrook and Conklin present divergent results in relation to how native- and Japanese speakers of English rate the naturalness of lexical bundles and non-lexical control bundles. The fifth and final report presents a review of *WordUp*, a MALL vocabulary learning application that has potential for students.

This issue also marks an important milestone in VLI's history as the last issue published under the supervision of Raymond Stubbe. It is well-known that Raymond has almost single handedly kept the journal going over the last few years and we all owe a debt of gratitude to him for this service. Raymond has graciously agreed to remain in service to the journal as a member of our newly revamped editorial board, which now features academics from around the world.

In closing, we thank you for your support of our journal and hope that these reports inspire both your research and teaching.

With warm regards,

Joseph P. Vitta and Christopher Nicklin (Associate Editors) on behalf of the entire VLI team

Vocabulary Learning and Instruction
Volume 10, Issue 1, December 2021
http://vli-journal.org

Challenges in the Assumptions of Using a Flemma-based Word Counting Unit

Tomoko Ishii[a], Phil Bennett[b], and Tim Stoeckel[b]
[a]Meiji Gakuin University; [b]University of Niigata Prefecture

Abstract

The choice of word counting units (i.e. word family, flemma, or lemma) is of great importance in vocabulary list and test creation, as there are assumptions underpinning the use of each. Flemma-based counting assumes that if a learner can understand the meaning of a word in one part of speech (POS), they can also understand its meaning when the same word form is used in another POS. A previous quantitative study showed that such an assumption is not always valid, but it did not provide reasons as to why. Therefore, this article presents an interview study probing the challenges learners face when they fail to understand the meaning of a known word form used in a new POS. The data were collected through one-on-one interviews with 16 university students in Japan, where they were asked to demonstrate comprehension of target words embedded in short sentences, as well as to explain how they approached the task. The interviews revealed that factors related to both the words and the learners play a role: some words have more complex meaning relationships across different POS, and some learners did not have advanced enough parsing skills to overcome such complexity. Based on these findings, this article concludes that the flemma might not be a suitable word counting unit for Japanese university students with relatively low levels of proficiency.

Keywords: word counting unit, flemma, lemma

1 Background

Decades have passed since vocabulary became one of the focal points of second language (L2) research, and a number of vocabulary lists and tests have been proposed to help learner's development in this area. A fundamental assumption of word lists, and by extension the tests that are based on them, is that word forms can be grouped into units. The words in each unit are treated as equivalent for the purposes of assessing text difficulty and measuring learner knowledge. That is, if a learner demonstrates knowledge of a single member of a unit, then knowledge of the other members can be reasonably assumed. Bauer and Nation (1993) suggested seven levels of word grouping, or word families, two of which are commonly used in word lists and tests. When the term *word family* is used without annotation, it typically refers to Level 6 of their word family system. This level groups a head word, its inflections and some of its derivations (e.g., *act, acts, action, actor, acting*). On the other hand, Level 2 includes headwords and inflected forms within the same grouping. The question of whether knowledge is assumed to cross part of

Please cite this article as: Ishii, T., Bennett, P., & Stoeckel, T. (2021). Challenges in the assumptions of using a flemma-based word counting unit. *Vocabulary Learning and Instruction, 10*(1), 1–15. https://doi.org/10.7820/vli.v10.1.Ishii

speech (POS) boundaries is not explicitly addressed, which has led to two subdivisions within this level: lemma and flemma. In the *lemma*, different POS are treated as distinct units (e.g., verbal *act* and nominal *act* and their respective inflections are separate), while in the form-based lemma, or *flemma* (Pinchbeck, 2014), members of the lexical unit may be from different POS (e.g., *act* as a noun and as a verb are both included).

Among these units, the Level 6 word family (hereafter "word family") has been the unit of choice for the word lists which some widely used vocabulary tests are based on, including the Vocabulary Levels Test (VLT; Schmitt et al., 2001) and the Vocabulary Size Test (VST; Nation & Beglar, 2007). The use of a word family assumes that the knowledge of a head word allows learners to access its inflections and derived forms (e.g., when learners know the word *act,* they can understand the meaning of *actor* as well, or at least they can learn the new word form *actor* with a small amount of effort and time). However, multiple studies have shown that such an assumption might not hold true in all contexts (e.g., Gardner, 2007; Mochizuki & Aizawa 2000). Iwaizumi and Webb (2021) showed how the knowledge of derivations might differ amongst learner groups of different proficiency levels, which again suggests the inadequacy of using word families across all learners. In his study, McLean (2018) demonstrated that Japanese learners ranging from beginner to advanced levels might not possess enough knowledge of derivation. As such, he argued for the use of flemma to avoid the overestimation of vocabulary knowledge. Laufer et al. (2021) also showed different degrees of mastering derived forms between two proficiency level groups; they argued, however, that these differences did not disqualify the use of word families, since the derived forms that are typically not known by less advanced participants do not frequently appear in actual texts.

Although to a lesser extent than word family-based counting, flemma-based counting still assumes that learners successfully use their knowledge of known word forms when decoding the meaning of unknown ones. In the case of flemmas, the assumption is that knowing the meaning of a word form or its inflectional forms in one POS allows learners to understand the meanings of the same word forms when they occur in another POS (e.g., the knowledge of *act* as a verb would lead them to understand what *an act* might mean). Therefore, building on the work of McLean (2018), Stoeckel, et al. (2020) investigated whether such an assumption would hold true among Japanese learners. The study asked participants to demonstrate their understanding of words with two POS (e.g., *twist, rise, fool*) by translating 24 short sentences containing 12 target words in each POS. In this task, there were no penalties for mistranslations of non-target words and only the comprehension of the target words was assessed. The results suggested that participants sometimes failed to demonstrate knowledge of the meaning of a word even though they knew the same orthographic form in another POS. Namely, even when a participant successfully translated "We *edit* it well," it did not always mean that the same participant could translate "We made an *edit* quickly." Amongst the cases where participants knew the word in at least one POS, 44% indicated a lack of understanding of the word in

the other POS. This led to the conclusion that lemma-based counting may be more suitable for Japanese learners.

This phenomenon may seem rather puzzling for native speakers or advanced English users. In theory, all learners have to do is transfer the meaning of a word they already know to make it fit a new context of meaning and grammar. It is also quite common to have a noun become a verb over time (e.g., *to Google*, *to Photoshop*, *to friend/unfriend*) or vice versa. Upon encountering a known word form in a new POS, learners would not seem to require any special skills to cross the boundary of POS. For example, it is hard to imagine how a learner who knows the verb *to drink* would struggle to understand that *a drink* is something to drink.

No matter how simple it might seem, some learners do have difficulty handling such a shift and fail to understand the meaning of known words in an unknown POS. Although Stoeckel et al. (2020) showed this quantitatively, their report on participants' performance as a group does not tell us why some learners miss what seems apparent. We need qualitative examination in order to see the obstacles each learner might face. Therefore, to complement previous findings, this article reports on an interview study to examine what learners know, what they do not know, and what causes difficulty in applying the knowledge of a word form across different word classes.

2 Methods

2.1 Participants

Sixteen students from three universities in Japan participated in this study. They were recruited via e-mail by their teachers and were offered a small monetary incentive for taking part in the interviews, which took around 30–40 minutes each. No attempt was made to control the English level of the participants, and the group was made up of learners ranging from limited proficiency (Eiken pre-2nd grade) to intermediate proficiency (TOEIC 780), which is considered to be equivalent to a range from A2 to B2 on the Common European Framework of References (CEFR) scale.

2.2 Materials

The target items used in this study were exactly the same as those used by Stoeckel et al. (2020). To determine whether the participants understood the meaning of a single word form in a different POS, they were asked to translate short sentences containing each target word. The target words and sentences are listed in Table 1.

All the target words came from the first 3,000 flemmas of the Cambridge English Corpus, and had more than one POS. The 12 words represented the average and range of difficulty of a broader pool of items appearing on the New General Service List test (Stoeckel et al. 2020). The target words in each POS had common, similar meanings, as determined by the definitions of standard dictionaries. As Zipf (1945) showed, high frequency words typically have multiple senses in dictionaries, and therefore a choice of word meaning had to be made. For the selection of meaning, 100 randomly selected concordance lines from the Corpus

Table 1. Target Words and Prompt Sentences

Target item	POS	Prompt sentence
Edit	Noun	We made an edit quickly.
	Verb	We edit it well.
Rise	Verb	They did not rise this year.
	Noun	We did not see the small rise.
Compound	Verb	This compounds the problem.
	Noun	We need those compounds.
Function	Noun	This has several functions.
	Verb	This functions very well.
Fool	Verb	They fool me every time.
	Noun	That fool did it again.
Quote	Verb	Can I quote you?
	Noun	Can you remember this quote?
Result	Verb	Your action results in three things.
	Noun	We don't know the results.
Export	Noun	They buy our most important export.
	Verb	They export many different things.
Variable	Adjective	They have variable prices.
	Noun	This is an important variable.
Twist	Noun	I gave it several twists.
	Verb	He really twists it a lot.
Pause	Verb	We usually pause here.
	Noun	I took a pause here.
Extra	Noun	I have one extra at home.
	Adjective	We did some extra things.

of Contemporary American English (COCA) for each target word in each POS were examined to identify typical meaning senses. A meaning sense that was frequent in both POS was the intended meaning of the target items in each pair of prompt sentences. Note, however, that in some cases, the most frequent meaning sense in the COCA was a sub-entry under a broader, numbered dictionary definition, and in such cases, the broader definition was the intended meaning in the test items. An example is the noun form of *extra.* The most common meaning sense in the COCA for this word pertains to an extra in a movie. In the dictionary, such use of *extra* is listed under a numbered entry with the broader meaning of an item in addition to what is usual or needed. The latter was the intended meaning of *extra* (n) in the test. We must concede that such variation in the selection of intended meaning was a limitation of this study.

Words with substantially different meanings such as *book* (to reserve / a book to read) or *affect* (to have an effect on / emotion) were not included. In addition, words that have become part of the Japanese vocabulary as loanwords (e.g., *copy*, *dry*) were not selected because of the risk that participants could provide responses based on phonological similarity to their native language (L1), which would make it difficult to tell if the participants really grasped the meanings of the words through the experiment.

In order to eliminate the possibility of participants not being able to respond properly due to a lack of grammatical knowledge, Stoeckel et al. (2020) included items to check the participants' grammatical ability by translating sentences of similar structure and complexity as those used in the main study. In the current study, however, such items were skipped, as the sources of difficulty would be naturally revealed during the interviews.

2.3 Procedure

The interviews were conducted in participants' L1 (Japanese) via video chat and using the screen sharing function. The following steps were taken:

1. Prompt sentences were presented on the screen one by one.
2. The participants translated them orally, while the interviewer took notes of their responses.
3. After completing all the items, the interviewer and the participants reviewed the items together, and the participants were invited to explain how they processed the prompt sentences.

Leow and Morgan-Short (2004) presented verbal reports as a way to elucidate the internal processes employed by learners, pointing to introspective-retrospective distinctions in data gathering procedures. The introspective or real-time approach has some advantages, as unlike the retrospective approach, it does not rely on participants' memory. However, for verbalisation to be captured efficiently and to be reliable and complete, established and dependable procedures need to be followed. Additionally, we did not want concurrent interviews to influence the way learners completed the task. Once the issue of POS was broached, for example, participants may have been more alert to it than they would have been otherwise. Therefore, a retrospective approach was adopted in this study. The participants focused on the task first, and then immediately after the task, when their memory was still fresh, they were invited to explain how they completed the task.

In order to help participants recall the process, as well as describe the state of knowledge that they might not be aware of, the interviewer asked if they knew the target word, if they knew the word in the POS as used in the prompt sentence, and if they applied knowledge of the word in another POS when translating the sentence. When the participants missed an item, the interviewer also provided a quick explanation to help them comprehend the item better, and sometimes asked for further ideas after providing some hints regarding the meaning of the word or the grammatical structure of the prompt sentence.

During the second step of the interview process, the responses were coded following the criteria established in Stoeckel et al. (2020). That is, each response was marked dichotomously as correct or incorrect. Although each target word form had a pair of items with a shared, common meaning sense, any correct meaning of the tested word that fit the prompt sentence was judged as correct. This was done by just one coder, which is a limitation of the present study. However, this person achieved a high level of inter-rater reliability (k = 0.787)

Table 2. Summary of Task Performance

Participant ID	Known meanings (max 24)	Word form known in at least one POS (max 12)	Word form known in both POS (max 12)	Word form known in only one POS (max 12)	Target POS known (max 24)	Another POS knowledge applied (max 12)	Othera (max 24)
HS	21	11	10	1	15	4	2
TS	20	11	9	2	13	5	2
NK	20	10	10	0	11	8	1
RS	18	10	8	2	14	4	0
SW	18	10	8	2	16	2	0
NW	16	8	8	0	14	2	0
MK1	12	8	4	4	9	3	0
AU	11	7	4	3	9	2	0
TK	10	7	3	4	8	0	2
SY	10	6	4	2	7	2	1
AS2	9	6	3	3	6	3	0
MK2	7	6	1	5	6	1	0
SM	5	4	1	3	4	1	0
HI	5	4	1	3	3	2	0
YK	4	3	1	2	3	1	0
AS1	3	3	0	3	0	0	3
Mean	11.81	7.13	4.69	2.44	8.63	2.50	0.69
SD	6.23	2.75	3.57	1.36	4.86	2.03	1.01

[a]"Other" refers to the cases where participants applied their knowledge of other words (e.g., they guessed the meaning of *export* from the word *import*).

with a second coder on the same item set in Stoeckel et al. (2020). Additionally, when word comprehension was unclear, the coder had the opportunity to probe learners' understanding of each target word during the subsequent interview. We believe this approach gives reasonable credibility to the consistency of judgements in the present study.

3 Results

Table 2 provides a summary of the overall performance of the 16 participants. As the focus of this study was on the understanding of the target words, translation errors in other parts of the prompt sentences were disregarded.

As is shown Table 2, participants exhibited a great deal of variety in both the number of items they could understand and the extent to which they understood target word forms in both POS. Participant HS translated 21 items out of 24 correctly, while AS1 translated only 3. Participants NW and MK1 both knew 8 words out of 12, but NW provided correct translations for both POS in all cases, whereas MK1 gave correct translations in both POS for just 4 of the 8 words. Overall, among the cases where at least one POS was known, the other POS was also correctly translated 65.78% of the time (4.69/7.13).

3.1 Interview Data

Seeking explanations for why some participants fail to apply their partial knowledge of words when trying to understand a sentence, one-on-one interviews were held with the participants immediately after they completed the translation task. These interviews showed how learners of different levels approached the task and revealed complex issues for both words and participants.

First, not all words seemed to pose the same degree of difficulty across the two POS. L1 speakers or advanced learners of English might recognise the meaning connection between two POS of the same word form as something straightforward, but for the less proficient learners in this study not all connections were equally transparent.

Even when the concepts are clearly connected, there are sometimes multiple possible meanings when a word is used in another POS. For instance, *Oxford Learner's Dictionaries* (https://www.oxfordlearnersdictionaries.com) defines the noun form of *export* as both: 1) *the selling and transportation of goods to another country*, and 2) *a product that is sold to another country*. Simply recognising *export* as the noun form of the verb *to export* does not indicate which of these definitions is appropriate in the given context. In other words, when a verb is used as a noun, the aspects of meaning transferred are not always systematic. For example, one meaning of *twist* as a noun is the act of twisting something, while *fool* as a noun is meant to be interpreted as a person with certain characteristics. Why could it not describe a foolish act? When a similar word like *tease* can refer to an act of teasing someone, there seems to be no logical reason why *fool* as a noun cannot. As shown in the extract below, some participants had problems with this. All the interview extracts presented in this article were originally in Japanese and were translated into English by the interviewer.

Interviewer:	You seemed to have no problem translating "They fool me every time." Did you know all the words in this sentence?
MK1:	Yes.
Interviewer:	Okay, the next one "That fool did it again," you translated this sentence as "It was a lie again." Did you know the word *fool* as a noun?
MK1:	Honestly, I had no idea what this sentence might mean. As "fool" [verb] is like *damasu* (to trick someone by telling an untruth), I thought it might be 'it was a lie."

Another participant saw a gap in meaning transfer between the adjective and noun forms of *variable*.

Interviewer:	"They have variable prices…"
HS:	I knew *variable* as an adjective, but didn't know it as a noun.
Interviewer:	I see. You have translated the noun form as *henka* (changes).
HS:	As *variable* [adjective] means something is changing, I thought its noun might mean "changes."

Interviewer:	I see. One typical translation of this word is *hensu,* a factor that changes.
HS:	Oh, hmm, I couldn't guess that.

This participant had no problem seeing *variable* as something that has to do with changes, but could not identify what part of the changes were being referred to. Logically, "the state of being changed" and "something that causes a change" both seem plausible, and it is no surprise that "a factor that changes" was not always their first choice when learners have only learned this word as an adjective.

As shown above, the meaning relationship across different POS is not always as straightforward as it might appear, and learners must explore different possibilities that are logically plausible for them to reach the correct deduction. Such differences in the meaning relationship are sometimes evident in the translation of target items. Some POS pairs have very similar translations in Japanese with the simple addition of a grammatical marker (e.g., *kinou* and *kinou-suru* for the noun and verb forms of *function*, where *suru* indicates a verb), while other pairs are translated quite differently (e.g., *kawariyasui* and *hensu* for the adjective and noun forms of *variable*). For some word pairs that have similar translations, students who did not even know the word at all were able to provide the correct meaning in one POS when its definition was explained in the other.

SY:	I don't know "function."
Interviewer:	It means *kinou* (function) and is used as a noun. Can you tell what the other sentence might mean now?
SY:	Ah, *kore-wa yoku kinou-suru* (It functions well), then.
Interviewer:	That's right.

Such differing degrees of closeness in the translation pairs, combined with the complication of having multiple logically plausible meanings, seemed to influence students' performance. The more similar the translations, the easier it seemed for even the lowest-level learners to answer correctly.

In addition to the varying degrees of challenge that different word pairs posed, learners also had a great deal of variation in their ability to process meanings in different POS. The first thing we should note is that learners varied in their declarative knowledge of POS. Some participants actively used grammar knowledge and derived the meaning of words based on their awareness that the word was being used in another POS. NK was one such participant, and the case below is typical of the responses he provided throughout the interview.

Interviewer:	Did you know this word (*edit)*?
NK:	I only knew this word as a verb, but seeing it being used asa noun, I thought it would be a noun, so I translated it that way.
Interviewer:	How could you tell it was a noun?
NK:	It's following "an."
Interviewer:	I see. How about this word (*quote*)?

NK: I didn't know this word could also be a noun, but it is used as a noun here, right?

On the other hand, some participants seemed to have very limited awareness of POS itself (even for words they knew) and expressed difficulties when asked if they knew a word as a noun or a verb. The coding presented in columns six and seven of Table 2 is based on what the participants reported, although not all of them were sure about how they knew the POS of certain words. One participant (AS1) completely gave up answering such questions, while others hesitated and used a substantial amount of hedging such as "probably" and "I think" when describing their knowledge. Another participant (SW), who translated 75% of the sentences correctly and exhibited a good command of grammar in the follow-up interview, said she was not quite sure how she knew each word in terms of POS. She explained that she had some understanding of the core meaning of words, without specification of POS, and tried to find the best translation to fit the sentence both in terms of meaning and grammar. For instance, she associated the concept *henshu* with the word *edit,* which in her mind could both mean *to edit something* or *an action of editing something;* this is similar to us not being able to tell if the word *edit* is a noun or a verb simply by looking at it on its own. She then saw the word in a sentence "*We edit it well,*" from which she gathered that *edit* was being used as a verb, and correctly translated it into the verb form *henshu-suru,* with the verb marker *suru*. She added that although she claimed her previous knowledge of some words "as a noun" when she thought she had seen them being used as a noun, she did not think they were a part of her productive vocabulary and she was not sure whether she really knew their POS.

Not being able to recognise the grammatical features of a word on its own might not be an issue as long as we discuss learners' receptive knowledge. Learners like participant SW should be able to identify the words' grammatical features from the way they are used in a sentence. However, it can become a problem when a lack of awareness about POS is combined with weak parsing skills. Namely, some learners could not tell if a word was being used as a noun or a verb in a sentence, and therefore they stuck to the POS they knew.

Interviewer: You knew the word *function* as you translated "This has several functions" smoothly.
TK: Yes, it's *kinou* (function [noun]), right?
Interviewer: Alright, then what about "This functions very well"?
TK: *Sono kinou wa totemo yoi.* (This function is very good.)
Interviewer: Well, … what is the verb in this sentence?
TK: … Oh, there is no verb. So, is *function* a verb here?
Interviewer: That's right.
TK: Then, *kinou-suru* (function [verb])?
Interviewer: That's right.

This participant failed to see that the word *function* was being used as a verb in her first attempt and applied the noun-like translation *kinou* because that was how she remembered the word. When the interviewer drew her attention to

the grammatical structure, she could parse the sentence correctly, but until then, it did not occur to her that a word that she knew as a noun could also be used as a verb. What is interesting here is the fact that when it was pointed out to her that her translation gave no indication of a verb in the English sentence, she was able to immediately correct her answer. Therefore, the problem was not that she had no parsing skills. She would have been able to parse other sentences with the same structure, but it appears that she jumped to "this function" (*this* + noun) combination as she knew the word as a noun. It seems her parsing skills were not strong enough to alert her to the fact that such a reading would leave the sentence without a verb.

Here is another example of a participant who translated the prompt "I have one extra at home" as "I have an extra home."

Interviewer:	The next item … What's your understanding of this word?
SM:	*Extra* … it is like "there is a lot," like in "extra paper." That is how I know this word. So, I translated this sentence as "another home."
Interviewer:	Yes, the word *extra* means *yobun-na* (extra, additional).
SM:	Yes, yes, *yobun-na*.
Interviewer:	That's right. The meaning is right. In the second sentence, "We did some extra things," it fits well. Your translation is good.

But this one, "I have one extra at home," this is not about a rich person owning a second house.

SM:	Really? (Laughs)
Interviewer:	It says "one extra," which means *extra* is used as a …… noun.
SM:	Huh?
Interviewer:	What do you think this might mean?
SM:	……. Does that mean there is one more person?
Interviewer:	Well, probably there is one more of something. For example, when you suddenly need an umbrella, "Ah, I have another one at home" or something like that. It can be about anything, you have one that you don't need and you can lend it to someone, or something like that.
SM:	Oh, I see. Yeah, "at home," not "extra home." Now I see it.

As we can see from these examples, learners need to be able to detect the correct POS before they can consider what a word might mean in another POS. In some cases, it is possible that by remembering a translation with grammatical markers (e.g., a verb marker *-suru* or an adjective marker *-na*), learners with weak parsing skills try to put the words together based on the grammatical category they associate the word with, making it even harder to parse a sentence. Namely, unless a learner has sufficient parsing skills to recognise that an adjectival form does not fit the context, a learner that understands *extra* as *yobun-na* (adjective)

might be more likely to interpret "extra at home" as "extra home" (*yobun-na ie)*. In other words, they may disregard the "at" before "home."

In addition to the variation in parsing skills, there were also differences in the level of flexibility when exploring meanings. As explained earlier, some words have a straightforward connection between the meanings in the two POS, while others have a slightly less obvious connection in the eyes of learners. Some participants seemed more capable than others to handle such connections by exploring various possible interpretations in a flexible manner. For example, for *export*, some participants could deduce the meaning correctly even if they had not seen the word being used to mean "something to be exported."

Interviewer:	You translated "They buy our most important export." without difficulty…
HS:	I initially did not know what the noun form of *export* meant, but as it follows *buy*, it must be something people buy, so "things to be exported"? Was I right?

Not all participants had such good guessing skills, and some struggled more. Participant SY associated the word *export* with a translation word *yushutsu* (the act of exporting something), which she transformed to a verb form *yushutsu-suru,* when seeing it used as a verb. However, she could not translate the same word when it was used as a noun, as she did not see how the translation word she knew (*yushutsu*) would fit the sentence.

Interviewer:	You translated the sentence "They export many different things" correctly. Did you know about the word *export*?
SY:	Yes.
Interviewer:	And you did not provide any translation for "They buy our most important export."
SY:	I could tell that *export* was used as a noun, but I didn't know what "to buy *yushutsu*" could mean.

Had the prompt been "Export is important to the economy of our country," this participant might have been able to come up with a good translation, as she could tell that the word *export* was being used as a noun with a familiar meaning sense in this sentence. However, the shift to "things to be exported" from "the act of exporting something" seemed to be difficult for her to handle. It seems that the first translation she came up with, *yushutsu,* further hindered her processing and blocked her ability to comprehend the sentence. With more flexibility to explore different possibilities, thinking about what is involved in transactions described as *exports* and what could be bought, she might have reached the right answer.

4 Discussion and Conclusions

This study was a follow-up of Stoeckel et al.'s (2020) work, which showed that some learners fail to recognise the meaning of a word form even when they can recognise the same word in another POS. As was the case in their study, some

participants missed an item in one POS even if they knew the word itself in the other POS. Interviews with participants revealed the factors that are involved in applying the knowledge in one POS to another, and what problems learners might face in decoding the meaning of words.

To apply previous knowledge of a word in one POS to another, learners need to take the following steps:

1. Recognise the word form.
2. Note that the word is used in a different POS than the one they are familiar with.
3. Identify the new POS of the word.
4. Deduce the meaning based on context and the meaning of the word in the known POS.

L1 or advanced L2 speakers of English naturally go through these steps and do not recognise many obstacles. However, as has been shown in the examples from the interviews, some learners lack the parsing skills to notice or figure out the POS (Steps 2 and 3). In some cases, the translation they are familiar with in one POS contains an L1 grammatical marker (i.e., verb marker *-suru* or adjective marker *-na*), which could possibly hinder accurate parsing. It seems that reliable parsing skills are required to make adjustments to the previously acquired knowledge that contradicts the sentence being presented to them.

When comprehension broke down at steps 2 or 3, it could be argued that a lack of grammatical knowledge, not lexical understanding, was the primary reason learners had difficulty demonstrating an understanding of target words that were known in another POS. However, this belies the fact that in such cases, the correct translation of the sentence containing the target word in the known POS was a clear demonstration of the grammatical knowledge needed: the two sentences were grammatically similar. Moreover, it is reasonable to conclude from the interview results that, in many cases, if learners had been familiar with the target word in each POS, they could have correctly parsed both sentences with no additional global grammatical knowledge. Some participants could also correct their parsing once subtle hints were provided by the interviewer. This suggests that the problem is a poor understanding of some grammatical aspects of the tested words, rather than weak global grammatical knowledge. As Nation (2001) suggested in his list of components of vocabulary knowledge, understanding how a word behaves grammatically is a part of lexical knowledge.

There were also cases where participants accurately captured the grammatical structure of a sentence, but still could not correctly interpret the target word. This is because the relationships of meaning across different POS might not be as straightforward as they seem. For some words, determining the meaning (Step 4) is almost automatic after detecting the POS of the word. However, there are also cases where learners need to choose between multiple logically possible meanings (e.g., the noun form of *export*), and discrepancies between the meanings in different POS and distance in the translation words can also pose challenges to differing degrees (e.g., *kinou/kinou-suru* vs. *kawariyasui/hensu*).

Despite attempts made to select words with similar meanings, there was inevitably some variation in the semantic distance between the two POS tested. In some cases, the meaning of the target words in the test materials reflected the literal, core senses of the words, rather than a more frequent but specialised sense (e.g., *extra* as a noun is primarily used to refer to extras in movies, while the prompt in this study suggested an extra thing). Controlling for such variation may not be entirely possible, owing to the unique characteristics of each word pair, but as we state above, differences in relationships of meaning seem to contribute to the difficulties that some of our participants faced. To some extent, this also demonstrates the challenge of using flemma-based groupings in vocabulary tests to determine learners' knowledge.

Due to the limited sampling of the target items, we must concede that this study might not have captured the entirety of the issue. First, this study limited target words to those that supposedly had similar meanings in the two POS. It was rather surprising how much difficulty some, although not all, participants faced in identifying the different meanings. The interviews revealed that detecting meaning might require an active and flexible search of multiple possibilities, even when the meanings seem to be closely related to each other across different POS. In the broader English lexicon, there are words that have multiple POS with considerably different meanings (e.g., *rate, project*). The risk of flemma-based word counts leading to overestimations of lexical knowledge may therefore be even more substantial than what the current study suggests.

Second, the decision to avoid loanwords inevitably distorted sampling. In selecting the items, 900 out of the most frequent 3,000 flemmas were identified as having more than one POS. After eliminating the words that were considered to be part of the Japanese vocabulary (i.e., loanwords), nearly 40% of those words were discarded. Although these items did not qualify for the research purpose in this study, they are also an important part of English vocabulary that Japanese learners know in some way. We do not know how excluding them could have affected the results. If such items had been included, participants might have taken advantage of their knowledge in their first language, or the differences in the usage of the loanwords from the original English might have created an extra challenge. How the elimination of loanwords has influenced the results is unknown, but we need to acknowledge the fact that there is a large missing piece in this context.

There is also a limitation in the presentation of the target words. All the items were presented either in their base forms or base forms inflected with plural *-s* or third person singular *-s*. To obtain a better understanding of the utility of the flemma versus lemma, further research might also need to include other inflectional forms (e.g., past tense *-ed* or present participle *-ing*). In addition, in the current study, the contexts were kept to a minimum to measure the knowledge of target words rather than participants' skills at guessing the meaning of unknown words based on the context. However, there is a chance that this decision added to the difficulty in deducing the meaning of words known in one POS. If the context were richer, some participants might have performed better.

We should also concede some limitations in the coding procedure. The responses from the participants were coded by a person with previous experience

in a study using the same experiment materials as the current study. Additionally, the interview data served to clarify the presence or lack of understanding of the target items. Even so, it would have been better if we had had another coder so that we could assess the reliability of the judgements.

Despite the small number of participants and potential issues with target word sampling, the current interview study provides some warnings regarding the use of flemmas in vocabulary testing. Whichever word counting units we apply, there are two assumptions that underlie vocabulary test construction and use. First, when a learner correctly answers a test item assessing knowledge of a word form, it is assumed that they also know most or all the other word forms in the lexical unit. Second, it is assumed that this knowledge is employable in specific, typically real-world, uses of the language. Accordingly, with a flemma-based test of written receptive vocabulary knowledge like the one used in this study, it is assumed that a correct response corresponds with the ability to receptively understand all flemma constituents when reading.

Stoeckel et al. (2020) have already shown that this assumption may not be viable in all cases, and the interviews in this study provide further insights into why this is the case. There is more complexity than we often assume in the relationship between the meanings of a single word form in two different POS. Examples of this complexity include differences in how the two POS are translated into the L1 as well as the presence of multiple possible meanings. Such complexity might be overcome if the learners have advanced parsing skills as well as the ability to explore different meanings flexibly; however, some participants in this study were not yet at that level.

As the proficiency levels of the participants in this study were within the range typically found in Japanese universities, the flemma appears to be too encompassing a word counting unit for written receptive vocabulary tests with this population. Regarding pedagogical word lists, those based on the flemma or word family may help Japanese learners to see the relationship among morphologically related word forms, but such resources need to overtly list the various lemmas that make up the larger lexical units in the list.

References

Bauer, L., & Nation, P. (1993). Word families. *International Journal of Lexicography, 6*, 253–279. https://doi.org/10.1093/ijl/6.4.253

Gardner, D. (2007). Validating the construct of word in applied corpus-based vocabulary research: A critical survey. *Applied Linguistics, 28*, 241–265. https://doi.org/10.1093/applin/amm010

Iwaizumi, E., & Webb, S. (2021). Measuring L1 and L2 productive derivational knowledge: How many derivatives can L1 and L2 learners with differing vocabulary levels produce? *TESOL Quarterly*. Advance online publication. https://doi.org/10.1002/tesq.3035

Laufer, B., Webb, S., Kim, K. S., & Yohanan, B. (2021). How well do learners know derived words in a second language? The effect of proficiency, word

frequency and type of affix. *ITL—International Journal of Applied Linguistics.* Advance online publication. https://doi.org/10.1075/itl.20020.lau

Leow, R. P., & Morgan-Short, K. (2004). To think aloud or not to think aloud: The issue of reactivity in SLA research methodology. *Studies in Second Language Acquisition, 26,* 35–57. https://doi.org/10.10170S0272263104261022

McLean, S. (2018). Evidence for the adoption of the flemma as an appropriate word counting unit. *Applied Linguistics, 39*, 823–845. https://doi.org/10.1093/applin/amw050

Mochizuki, M., & Aizawa, K. (2000). An affix acquisition order for EFL learners: An exploratory study. *System, 28*, 291–304. https://doi.org/10.1016/S0346-251X(00)00013-0

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.

Nation, P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*, 9–13. https://jalt-publications.org/tlt/issues/2007-07_31.7

Pinchbeck, G. G. (2014, March 22–25). *Lexical frequency profiling of a large sample of Canadian high school diploma exam expository writing: L1 and L2 academic English* [Roundtable presentation]. American Association for Applied Linguistics Annual Conference.

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. https://doi.org/10.1177/026553220101800103

Stoeckel, T., Ishii, T., & Bennett, P. (2020). Is the lemma more appropriate than the flemma as a word counting unit? *Applied Linguistics, 41*(4), 601–606. https://doi.org/10.1093/applin/amy059

Zipf, G. K. (1945). The meaning-frequency relationship of words. *The Journal of General Psychology, 33,* 251–256. https://doi.org/10.1080/00221309.1945.10544509

Vocabulary Learning and Instruction
Volume 10, Issue 1, December 2021
http://vli-journal.org

Predicting Japanese Primary Schoolchildren's English Vocabulary Knowledge from a Sound-Meaning Recognition Test

Kohei Kanayama
Hokkaido University of Education

Abstract

The purposes of this study were to: (1) build an appropriate model for predicting primary schoolchildren's English vocabulary knowledge; (2) examine whether the developed model applies to new data; and (3) discuss how to apply the model to L2 vocabulary instruction. More specifically, the study asked third- and fourth-grade public primary school students to take a sound-meaning recognition test. All the study's target words were *katakana* English and on the first 1,000-level of word frequency according to JACET (Japan Association of College English Teachers) 8000 and SVL (Standard Vocabulary List) 12000; however, different numbers of word phonemes were included. The collected data were divided into training and testing data. To build the predictive model from the training data, the study used a generalized linear mixed model, which revealed that: (1) third-graders' scores were as high as those of fourth-graders, and (2) the more phonemes a word has, the lower the test score. From the model, the study developed a regression formula that made it possible to predict how many primary schoolchildren knew the meaning of a certain word when they listened to its pronunciation. It also found that the predicted percentages from the training data correlated moderately with the actual test scores from the testing data. Therefore, the model achieved moderately high quality for predicting new data.

Keywords: Predictive model, L2 vocabulary knowledge, Machine learning

1 Introduction

In April 2020, English education became a formal subject in Japanese primary schools. Fifth- and sixth-grade children learn English as a formal subject, and third- and fourth-grade children are introduced to it through a compulsory but not formal subject called *gaikokugo katsudo* [foreign language activity]. According to the *Course of Study* (MEXT, 2017), the Japanese primary school English curriculum, teachers in primary schools introduce 600–700 words over 4 years (third to sixth grade) to facilitate English communication. They also teach collocations and high-frequency phrases or idioms such as "get up," "look at," "excuse me," and "I'm sorry." Despite the large number of English words to be learned, children study English for only 157.5 hours over 4 years. Third- and fourth-graders learn English only once a week, while fifth- and sixth-graders learn

Please cite this article as: Kanayama, K. (2021). Predicting Japanese primary schoolchildren's English vocabulary knowledge from a sound-meaning recognition test. *Vocabulary Learning and Instruction, 10*(1), 16–29. https://doi.org/10.7820/vli.v10.1.kanayama

twice a week. A lesson lasts 45 minutes, and the academic year has 35 weeks; therefore, the 4 years of English study unfold over 210 lessons for 157.5 hours. In other words, teachers have very limited instructional time compared to the number of English words to be taught. Thus, for effective English vocabulary instruction within a limited amount of time, distinguishing between words that children know and do not know is important for teachers to concentrate on unfamiliar words. In addition, teachers should spend time turning the words already known by many students into productive words to be used in actual English communication.

However, how should teachers distinguish which words children are familiar with? One of the purposes of this study is to develop a suitable model to predict Japanese primary schoolchildren's English vocabulary knowledge. For example, how many children recognise the meaning of "apple" when they listen to its English sound (/æpl/)? The features of the word are key to this question. Previous studies have found that the shorter a word is, the easier it is to remember (Willis & Ohashi, 2012); therefore, it is predictable that "apple" is easier to remember than "strawberry." Moreover, previous studies have revealed that Japanese learners can recognise English words that exist in Japanese as *katakana* English better than those that do not (Kasahara et al., 2012; Willis & Ohashi, 2012; Yoshimura, 2009). Therefore, many primary schoolchildren know "apple," which has the feature of *katakana* English [アップル], but do not know "pear," which does not. The next section introduces some previous studies that conducted English vocabulary knowledge tests on Japanese learners of English and investigated word features that affect the ease of vocabulary learning.

2 Literature Review

2.1 Effects of Katakana English on the Ease of Remembering for Japanese Learners of English

Some studies have conducted English vocabulary knowledge tests to examine what words are easy for Japanese learners of English to learn and remember. Yoshimura (2009) conducted a sound-meaning recognition test for first- to sixth-grade students. The test asked the participants to choose an appropriate answer from four options (e.g., *chairo* [brown], *midoriiro* [green], *murasaki* [purple], *shiranai* [I do not know]) when they heard the target word (e.g., /pəːrpl/). Findings revealed that even among first- and second-grade children, some *katakana* English words (e.g., "house," "fruit," "milk," "eleven," "purple") had high correct rates (over 90%).

The fact that words with the features of *katakana* English are helpful for Japanese learners of English to learn is also supported by other studies (Kasahara et al., 2012; Willis & Ohashi, 2012). Willis and Ohashi (2012) examined what English words are easy for Japanese university students to learn and retain in the long term, that is, words with the features of *katakana* English, high-frequency words, and words with fewer phonemes. These researchers emphasised that *katakana* English was the most influential.

The Japanese use many *katakana* English words (e.g., ピアノ(*piano*) [piano], パンダ (*panda*) [panda], オレンジ (*orenji*) [orange]). In fact, Daulton (2008) reported that about half of the 3,000 high-frequency English words are often used as *katakana* English, spread widely across newspapers, magazines, and television shows (Shibuya, 2012). Moreover, Hoshino and Shimizu (2019) analysed the textbooks *Let's Try!* (MEXT, 2018a) and *We Can!* (MEXT, 2018b), which are both widely used in English lessons (foreign language activities) in Japanese primary schools. Their findings revealed that *katakana* English accounts for about 70% and 50% of textbooks' vocabulary for reading and listening, respectively. These facts imply that many children in Japan already know some *katakana* English even before they start learning English in school.

2.2 Various Types of English Vocabulary Knowledge Tests for Primary Schoolchildren in Japan

A few previous studies have conducted English vocabulary knowledge tests on Japanese children. Kasahara et al. (2012) conducted three types of English vocabulary knowledge tests for fifth- and sixth-grade children: (1) a sound-meaning recognition test in which examinees chose the correct picture from four options (e.g., "wallet," "plastic bottle," "cashbox," "bag") when they heard the target word pronounced (e.g., /bæg/); (2) a sound-spelling recognition test in which examinees chose the correct spelling from four options (e.g., "giraffe," "laugh," "picture," "zebra") when they heard the target word pronounced (e.g., /dʒəræf/); and (3) a spelling-meaning recognition test in which examinees chose the correct picture from four options (e.g., "hat," "scarf," "glasses," "dress") when they saw the target word's spelling (e.g., "dress"). Among all three tests, the correct answer rate was significantly higher for the sound-meaning recognition test (92.8%) than for the sound-spelling recognition test (84.6%), and the spelling-meaning recognition test (84.7%).

Moreover, Koshiba (2015) conducted three types of tests for third- to sixth-graders: (1) a sound-meaning recall test in which children wrote meanings of the words after listening to their pronunciation; (2) a meaning-sound recall test in which children pronounced the word when they looked at its illustration; and (3) a spelling-meaning recall test in which children wrote the meanings of the words when they saw their spelling. Among the four grades, the sound-meaning recall score was the highest.

In general, previous studies have shown two main results: first, previous studies have indicated some word features that make it easier for Japanese learners to remember English words. Yoshimura (2009) indicated that test scores for *katakana* English were higher than those for non-*katakana* English. On the other hand, word frequency level and word length are key for the ease of vocabulary learning (Willis & Ohashi, 2012).

Second, most studies have conducted a sound-meaning test, in which students performed the best. Conceivably, this is because the *Course of Study*'s main goals for primary school English education is not to help children develop a written language, but to help them develop basic communication skills and provide

many opportunities to interact with their classmates, homeroom teachers, and assistant language teachers (MEXT, 2017). Therefore, children's enhancement of sound-meaning connections is crucial for familiarising them with basic English oral communication. This is consistent with Barcroft (2020), who argued that "all language begins with input" (p. 489). This is why most of the previous studies used the sound-meaning test and found it to yield more significant results than other types of tests, such as spelling-meaning and sound-spelling tests.

2.3 Previous Studies' Limitations and This Study's Purposes

The studies conducted previously however, had several limitations. First, the participants in Willis and Ohashi (2012) were not primary schoolchildren. They investigated variables that made it easy to remember words. However, their survey was conducted with Japanese university students, and their results have not yet been confirmed among children. Second, their study used vocabulary lists derived from the British National Corpus (BNC), which includes 100 million written and spoken British English words. Even though their study found high-frequency words easier to remember, the BNC is not designed for Japanese learners of English; thus, the target words used in this study were selected from JACET 8000 (The Japan Association of College English Teachers, 2016) and SVL 12000 (ALC, 2001), both developed for Japanese learners of English.

Third, previous studies did not discuss how their results could be applied to actual L2 vocabulary teaching in primary schools. Although previous studies conducted vocabulary knowledge tests for primary schoolchildren in Japan, their purpose was to compare the mean scores between the different types of tests.

Considering the limitations of the previous studies, the present study is original in that it aims to: (1) conduct a sound-meaning recognition test and build a predictive model from training data and (2) examine whether the developed model is applicable to new data. In short, this study is based on a framework for supervised machine learning. Training data or input data were collected for statistical analysis and for building a predictive model. On the other hand, testing data or new data are used to determine whether the model developed from the training data can fit into new data.

Relative to the first aim mentioned earlier, third- and fourth-graders took a sound-meaning recognition test, in which they listened to a target word's pronunciation and chose the appropriate picture from four options. As English beginners are not prepared for demanding vocabulary tests, such as a sound-spelling recall test, and because it would be ideal if English beginners could acquire words for spoken communication, this study conducted a sound-meaning recognition test.

Furthermore, this study aimed to develop a model that primary school English teachers can use. Two main types of variables that affect test scores are: (1) learners and (2) words. In the case of learners, even though it is possible to adopt such variables as learners' grades or lengths of English learning experience, this study adopted learners' grades as the first variable. The length of English learning experience is known to be an important variable that affects scores in the sound-meaning recognition test. Ojima et al. (2011) reported a clear relationship

between second-language proficiency and total study time; in other words, the earlier children begin studying English, the longer their experience of second-language learning. However, the reality is that English teachers have a variety of students with different needs to be considered, and English learning experience cannot be used in the predictive model because it depends heavily on individual children's experiences. Therefore, this study adopted a learner's grade as the first variable.

As for the variable of words, previous studies have revealed features that ease vocabulary learning, such as word length, word frequency, and *katakana* English. In addition, other features include L2 vocabulary words' pronunciation (Willis & Ohashi, 2012), imaginability (Karakas, 2020), and familiarity (Hoshino & Shimizu, 2019; Saji & Saeki, 2013). However, this study adopted word length as its second variable. Even if placing many variables in the model makes it more accurate, it can become too complex to use. For example, to measure word familiarity, Nishide and Mizumoto (2009) asked Japanese university students to score a certain word's familiarity on a 5-point scale, with a score of 5 meaning the word is completely familiar. For example, the word "apple" received an average score of 4.91 out of 5. Therefore, if this study had constructed its predictive model by adopting such a variable, users would always be required to check each target word's familiarity value and input the value into the regression formula, which is a time-consuming process. The significance of this study lies in its construction of an accurate predictive model from a small number of variables: children's grades and word length.

On the other hand, as detailed in the second aim of the study, the constructed model's quality for predicting new data still requires confirmation. The model made it possible to calculate the predicted percentages of children correctly identifying a certain target word in the sound-meaning recognition test by placing variables in the regression formula. Therefore, this study compared the predicted percentages from training data with actual percentages from the testing data to investigate whether the newly developed model can be reliable in predicting new data. The research question is as follows: can a model that predicts Japanese primary schoolchildren's English vocabulary knowledge be applied to new data?

3 Study

3.1 Participants and Target Words

A total of 144 third- and 155 fourth-graders (n = 299) participated in this study. The experiment was conducted in July 2019. The third-graders had been learning English for 3 months in 10 lessons since April 2019. In contrast, the fourth-graders had been learning English for 15 months in 30 lessons since April 2018. All participants learned from *Let's Try!* (MEXT, 2018a), a textbook widely used in Japan.

This study's 25 target words were all nouns with the feature of *katakana* English chosen from *Let's Try!* (MEXT, 2018a), and all within the first 1,000 words ranked by level of word frequency according to JACET 8000 and SVL 12000 (see Table 1 below). JACET 8000 contains eight levels of 1,000 words for Japanese

Table 1. Target Words Used in this Study

Words	Letters	Syllables	Phonemes	Words	Letters	Syllables	Phonemes
Apple	5	2	3	Bag	3	1	3
Ball	4	1	3	Bed	3	1	3
Book	4	1	3	Bus	3	1	3
Cake	4	1	4	Card	4	1	4
Cat	3	1	3	Dinner	6	2	5
Dog	3	1	3	Egg	3	1	2
Flower	6	2	6	House	5	1	4
Library	7	3	8	Milk	4	1	4
School	6	1	4	Science	7	2	6
Shirt	5	1	4	Star	4	1	4
Table	5	2	5	Telephone	9	3	8
Tree	4	1	3	Rain	4	1	4
Room	4	1	3				

learners of English. Similarly, SVL 12000 includes 12 levels of 1,000 words. New Word Level Checker (https://nwlc.pythonanywhere.com/) was used to determine the frequency level for both JACET 8000 and SVL 12000. In addition, this study used two *katakana* English dictionaries published by Sanseido Henshujo (2019) and Horiuchi (2013) to ascertain that the target words were *katakana* English. Based on Allen (2019) and Shibuya (2012), this study defines *katakana* English as: (1) Japanese words that share Japanese form and meaning with English words; (2) words that have been borrowed into Japanese; (3) words that are often used in daily life in Japan; and (4) words that are not Japanese–English words, such as カステラ (castella) [castella cake] and エアコン (aircon) [air-conditioner], which cannot be understood by English speakers.

3.2 Procedure and Data Analysis

This study conducted a sound-meaning recognition test, which required third- and fourth-grade Japanese primary school students to listen to an English word's pronunciation (e.g., /æpl/) and then choose the appropriate picture from four options (see Appendix I). During the experiment, the participants took the sound-meaning recognition test, where they listened to a native English speaker's pronunciation using *GlobalvoiceEnglish3*. They were able to listen to an English sound twice per question. It took approximately 10 minutes to complete the experiment.

After the experiment, the collected data were divided into training and testing data. For training data, 70% of the collected data were chosen at random, and the remaining 30% were chosen as testing data, based on Katada and Fukuzawa (2019). Training data were used for statistical analysis, and testing data were used to calculate the mean score of the sound-meaning recognition test. Among the 144 third-graders, 100 participants' data were categorised into training data and 44 into testing data. Similarly, among the 155 fourth-graders, 109 participants' data were categorised into training data and 46 into testing data.

The study employed a generalized linear mixed model (GLMM), which combines a generalized linear model (GLM) and a linear mixed model (LMM). The GLM includes some regression models, and this study adopted the logistic regression model (LRM), which can be used when the outcome or dependent variable is binomial data. Since participants chose an appropriate picture from four options, they were either correct or incorrect, and thus produced binomial data. The GLMM can also be used to develop a regression formula that enables primary schoolteachers to predict the difficulty of students learning a certain word. Moreover, the GLMM considers fixed and random effects.

Table 2 shows how the collected data were entered into an Excel file. In the case column, there are 7,475 cases (Case 1 to Case 7,475) in total because 299 participants took the test that included 25 target words (299 × 25). The column labelled "Participant" reflects children's sample number from 1 to 299. In the "Grade" column, third-graders were coded as 0, and fourth-graders were coded as 1. Similarly, in the "Data" column, training data were coded as 0, and testing data were coded as 1. In the "Letter," "Syllable," and "Phoneme" columns, the numbers depended directly on each target word's features. Moreover, in the "Correct" column, 1 represented a correct answer, and 0 an incorrect answer.

For example, Case 1 means that a third-grade child (Participant No. 1) was assigned to training data and correctly chose "apple"—a target word with five letters, two syllables, and three phonemes. Similarly, Case 8 means that a fourth-grade child (Participant No. 4) was assigned to testing data but could not choose the correct picture of "egg," which has three letters, a syllable, and two phonemes.

The GLMM was conducted to investigate whether the training and testing data were equally divided. Correct was considered a dependent variable, and grade and data as independent variables. Finally, participant was considered to be a random effect.

The GLMM was also conducted to build the most accurate predictive model from the training data. Two independent variables, grade and length of words (i.e., letter, syllable, and phoneme) represented the fixed effect, and participant the random effect. The most appropriate model was based on Akaike's information criterion (AIC).

Indeed, the regression formula constructed from training data made it possible to predict the percentages of children who knew certain words depending

Table 2. Examples of Data Input

Case	Participant	Grade	Data	Letter	Syllable	Phoneme	Word	Correct
1	1	0	0	5	2	3	Apple	1
2	1	0	0	3	1	3	Bag	1
3	2	0	1	4	1	4	Cake	0
4	2	0	1	6	2	5	Dinner	1
5	3	1	0	5	1	4	House	0
6	3	1	0	9	3	8	Telephone	1
7	4	1	1	5	1	4	Shirt	1
8	4	1	1	3	1	2	Egg	0

on their grade and on the length of words. A correlation analysis was conducted between the predicted percentage from the training data and the actual percentage from the testing data to examine the research question.

4 Results

Results of the sound-meaning recognition test (*Max* = 25) from the training data showed that third-graders had a mean of 21.5 (standard deviation [*SD*] = 8.68), and fourth graders 21.9 (*SD* = 8.25). On the other hand, results from the testing data showed that third-graders had mean of 21.0 (*SD* = 9.25) and fourth-graders of 21.7 (*SD* = 8.48). Table 3 shows the results of fixed effects and presents the following: estimated coefficient (*Estimated*), odds ratio (*OR*), standard error (*SE*), *z*-value, *p*-value, and a 95% confidence interval (95% *CI*). Table 3 shows no significant effect of either grade or data, indicating that the two datasets were equally divided.

Models 1, 2, and 3 were developed to ascertain the ideal model. Grade and letter were adopted in Model 1, grade and syllable in Model 2, and grade and phoneme in Model 3. Table 4 shows the three models and their AIC values. The results revealed that Model 3 was the most appropriate because the AIC values of this Model were the lowest among the three models.

Table 5 shows Model 3's fixed effects and presents an estimated coefficient (*Estimated*), *OR*, *SE*, *z*-value, *p*-value, and a 95% confidential interval (95% *CI*). The Table shows no significant effect of grade (*Estimate* = 0.26, *SE* = 0.25, $z = 1.03$, $p = 0.305$), but a significant effect of phoneme (*Estimate* = −0.56, *SE* = 0.04, $z = -12.89$, $p < 0.001$). This means: (1) there is no difference in test scores between

Table 3. Results of the Sound-Meaning Recognition Test

Fixed effect	*Estimated*	*OR*	*SE*	*z*-value	*p*-value	95% CI
Intercept	2.23	9.34	0.15	14.42	<0.001	[1.93, 2.54]
Grade	0.18	1.20	0.19	0.97	0.33	[−0.18, 0.55]
Data	−0.13	0.88	0.20	−0.65	0.52	[−0.53, 0.27]

OR, odds ratio; SE, standard error; CI, confidence interval.

Table 4. The Value of AIC Affecting the Best Model

	Model 1	Model 2	Model 3
Variables	*Estimated* (*SE*)	*Estimated* (*SE*)	*Estimated* (*SE*)
Intercept	2.82 (0.27) ***	3.94 (0.25) ***	4.88 (0.30) ***
Grade	0.14 (0.23)	0.24 (0.24)	0.26 (0.25)
Letter	−0.12 (0.04) **		
Syllable		−1.08 (0.09) ***	
Phoneme			−0.56 (0.04) ***
AIC	1855.7	1734.6	1667.7

***$p < 0.001$, **$p < 0.01$.
SE, standard error; AIC, Akaike's information criterion.

Table 5. Fixed Effect from GLMM (Model 3)

Fixed effect	*Estimated*	*OR*	*SE*	*z*-value	*p*-value	95% CI
Intercept	4.88	131.78	0.30	16.45	<0.001	[4.30, 5.46]
Grade	0.26	1.29	0.25	1.03	0.305	[−0.23, 0.75]
Phoneme	−0.56	0.57	0.04	−12.89	<0.001	[−0.65, −0.48]

OR, odds ratio; SE, standard error; CI, confidence interval.

third- and fourth-graders; and (2) since the estimated coefficient is negative, the more phonemes a word has, the lower the test score.

Since GLMM is part of LRM, the OR can illustrate a certain variable's effectiveness. The estimated coefficients in Table 5 illustrate the logarithmic ORs. When values change into ORs, their meanings become clearer. The value of the estimated coefficient of phoneme was −0.56, which is equal to the 0.57 OR (= exp [−0.56]). This means that the correct rates of words with four phonemes (e.g., "cake") were 0.57 ORs higher than words with three phonemes (e.g., "apple"). The same was true in the comparison of five-phoneme (e.g., "dinner") and four-phoneme words (e.g., "card"). Furthermore, correct rates of words with six phonemes (e.g., "flower") were 0.33 ORs higher than those of words with four phonemes (= exp [−0.56 × 2]).

The following regression formula made it possible for primary schoolteachers to predict the ease or difficulty of teaching certain words:

$$\text{Regression formula} = 4.88 + 0.26 \times \text{grade (0 or 1)} - 0.56 \times \text{phoneme}$$

To predict whether third-graders knew the *katakana* English word "apple," the number of phonemes (3) and the participant's grade (0) were included in the regression formula. Then, the value of the estimated coefficient (3.20) was calculated as equal to 24.53 odds (= exp [3.20]). Odds signifies the value attained when the probability of success is divided by that of unsuccess (success percentage/unsuccess percentage). In this case, success and unsuccess imply that learners can answer correctly and incorrectly, respectively. Odds can be used as shown in the following formula:

$$\text{Successful Percentage} = \frac{\text{odds}}{1 + \text{odds}}$$

When odds of 24.53 were placed in the formula, the percentage of success was 0.961. This suggests that when third-graders listen to the word "apple" (/æpl/) being pronounced, the probability that they know its meaning is 96.1%. In this way, teachers can predict the probability of known words based on the word's number of phonemes and students' grades. Table 6 displays the predicted percentages from GLMM (Model 3) obtained from the training data and the actual percentages obtained from the testing data.

Spearman's rank method was applied to examine the correlation between the predicted percentages from the training data and the actual percentages from the testing data. The two types showed moderate correlation in third-graders

Table 6. Relationship between Predicted Percentages and Actual Percentages (%)

Data Types	Grade	Apple	Bag	Ball	Bed	Book	Bus	Cake
Training Data	3rd	96.1	96.1	96.1	96.1	96.1	96.1	93.3
	4th	97.0	97.0	97.0	97.0	97.0	97.0	94.5
Testing Data	3rd	100	95.5	72.7	100	86.4	100	100
	4th	100	88.9	88.9	100	94.4	100	100
Data Types	Grade	Card	Cat	Dinner	Dog	Egg	Flower	House
Training Data	3rd	93.3	96.1	88.9	96.1	97.7	82.1	93.3
	4th	94.5	97.0	91.2	97.0	98.2	85.6	94.5
Testing Data	3rd	86.4	100	72.7	100	90.9	77.3	100
	4th	94.4	94.4	92.9	100	100	72.2	100
Data Types	Grade	Library	Milk	Rain	Room	School	Science	Shirt
Training Data	3rd	59.9	93.3	93.3	96.1	93.3	82.1	93.3
	4th	65.9	94.5	94.5	97.0	94.5	85.6	94.5
Testing Data	3rd	40.9	63.6	77.3	86.4	54.5	72.7	68.2
	4th	46.4	83.3	89.3	89.3	50.0	67.9	72.2
Data Types	Grade	Star	Table	Telephone	Tree			
Training Data	3rd	93.3	88.9	59.9	96.1			
	4th	94.5	91.2	65.9	97.0			
Testing Data	3rd	95.5	95.5	72.7	90.9			
	4th	94.4	96.4	78.6	85.7			

Table 7. Predicted Successful Percentages of Katakana English Words with Different Phonemes

Grade	Phoneme	*PSP*	Examples	Grade	Phoneme	*PSP*	Examples
3rd	3	96.1%	Pen	3rd	6	82.1%	Orange
4th	3	97.0%	Red	4th	6	85.6%	Mountain
3rd	4	93.3%	Heart	3rd	7	72.3%	Baseball
4th	4	94.8%	Horse	4th	7	77.2%	Hospital
3rd	5	88.9%	Salad	3rd	8	59.9%	Library
4th	5	91.2%	River	4th	8	65.9%	Telephone

PSP, predicted successful percentage.

($\rho = 0.53$ [large effect size], 95% $CI = [0.18, 0.77]$, $p = 0.006$) and in fourth-graders ($\rho = 0.54$ [large effect size], 95% $CI = [0.18, 0.77]$, $p = 0.005$), meaning that a word predicted to be easy tends to have a higher actual score and vice versa. Thus, Model 3 was moderately higher in terms of predicting Japanese primary school-children's knowledge of English vocabulary.

Table 7 shows the predicted successful percentage (*PSP*) based on grade and phoneme, and some examples from the textbook, *Let's Try!* (MEXT, 2018a), respectively. The use of the PSP is discussed in the next section.

5 Discussion

As Table 4 reveals, Model 3 was the most appropriate for predicting new data, and the grade and phoneme variables were placed into it. Of letter, syllable,

or phoneme, the most suitable variable was phoneme, which was attributed to the type of test. The study adopted the sound-meaning recognition test, in which children listened to an English word pronounced (e.g., /æpl/) and chose the appropriate picture (of an apple) from four options. The participants did not look at the target word's spelling; in other words, the phoneme variable, related to sound information, was key to the sound-meaning recognition test.

Moreover, the syllable variable is not flexible in representing the features of target words. As Table 1 shows, the range of syllables was limited from one to three. On the other hand, word phonemes ranged from two to eight, which is why Model 3 was flexible enough to predict new data.

As Table 5 shows, the grade variable showed no significant difference: third-graders' scores were as high as those of fourth-graders. This result observed a ceiling effect, in that the total mean scores and SD exceeded the maximum score (Kanayama, 2020). The third-graders' total average score (M = 21.5) and the standard deviation (SD = 8.68) exceeded the maximum score (Max = 25). The same was true for fourth-graders (M = 21.9, SD = 8.28). Indeed, the ceiling effect prevents researchers from measuring the knowledge acquired by learners (Kanayama, 2020). This likely occurs because the target words were all nouns with features of *katakana* English and because they were from the first 1,000 levels of JACET 8000 and SVL 12000. In short, participants knew most of the target words well, enabling them to achieve high scores and leading to a ceiling effect for both grades.

Answering this study's research question of whether or not a model that predicts Japanese primary schoolchildren's English vocabulary knowledge can be applied to new data, Spearman's rank method found a moderate correlation between predicted percentages from training data and actual percentages from testing data, as shown in Table 6. This revealed a general tendency of schoolchildren to know "easy" words—that is, words with fewer phonemes—but not to be as familiar with "difficult" words—for example, words with more phonemes.

One purpose of this study was to provide pedagogical implications for teaching English vocabulary in Japanese primary schools. As mentioned above, the regression formula was found to be applicable. For instance, when third- and fourth-graders listened to the pronunciation of *katakana* English words and chose their meanings, the regression formula made it possible for primary schoolteachers to predict the ease or difficulty of teaching certain words. As Table 7 shows, it is predicted that most learners knew the meaning of *katakana* English words with three to four phonemes because *PSP* was over 90% for both grades. Therefore, teachers can spend less time teaching such words. Instead, they should encourage children to use the words they already know, enabling them to be used productively. In contrast, third-graders' rates for *katakana* English words with eight phonemes were lower than 60%, indicating that teachers should spend more time introducing such words.

6 Conclusion

In this study, third- and fourth-graders took the sound-meaning recognition test, and GLMM from training data revealed no significant effect of grade, but a

significant effect of phoneme. From the fixed effect of GLMM, the most appropriate model was constructed to predict the scores of the sound-meaning recognition test from the training data. The model was found to be moderately reliable because of the moderate correlation between the training and testing data. Thus, the results of this study are significant because the developed model can predict primary schoolchildren's English vocabulary knowledge as long as the tested words are nouns from *katakana* English and high-frequency.

This study has some limitations. First, because participants were limited to third- and fourth-graders, whether the predictive model will apply to fifth- and sixth-graders remains unknown. Second, the sound-meaning recognition test was the only one administered; therefore, the model could not predict other types of English vocabulary knowledge. Further studies should employ, for instance, a sound-meaning recall test, a meaning-sound test, a spelling-sound test, and a spelling-meaning test. Third, the largest limitation is that target words were limited to *katakana* English nouns from the first 1,000 frequency levels of both JACET 8000 and SVL 12000; thus, the model cannot predict all words. For instance, the *katakana* English word "peach" cannot be predicted because it is not in the first 1,000 level but in the second 1,000 frequency level of both JACET 8000 and SVL 12000. Finally, a limited number of learners participated in the study. Considering that a large amount of data would make the regression formula more accurate and reliable, future studies should gather more data. Considering these limitations into account, future studies should ask various grades to take various vocabulary knowledge tests that include both non-*katakana* and *katakana* English with different word frequency levels.

References

ALC. (2001). *Standard vocabulary list 12000*. https://www.alc.co.jp/eng/vocab/svl/list.html

Allen, D. (2019). An overview and synthesis of research on English loanwords in Japanese. *Vocabulary Learning and Instruction*, *8*(2), 8–25. https://doi.org/10.7820/vli.v08.2.Allen

Barcroft, J. (2020). Key issues in teaching single words. In S. Webb (Ed.), *The Routledge handbook of vocabulary studies* (pp. 479–492). Routledge.

Daulton, F. E. (2008). *Japan's built-in lexicon of English-based loanwords.* Multilingual Matters.

Horiuchi, K. (2013). *Gendaiyougo no kisochishiki katakana gairaigo ryakugo jiten* [*Basic knowledge of modern term: Dictionary of Katakana*] (5th ed.). Jiyukokuminsha.

Hoshino, Y., & Shimizu, H. (2019). Katakana words in English textbooks used in Japanese primary schools—A comparative study of word senses between Japanese and English. *The Japan Association of English Teaching in Elementary School, 19*, 117–129. https://doi.org/10.20597/jesjournal.19.01_117

Kanayama, K. (2020). Is expanding spacing more effective than equal spacing for L2 vocabulary learning? *Annual Review of English Language Education in Japan, 31*, 1–16. https://doi.org/10.20581/arele.31.0_1

Karakas, A. (2020). Exploring psycholinguistic difficulties in learning foreign language vocabulary in light of the English language. *i-manager's Journal of Educational Psychology*, *13*(3), 36–45. https://doi.org/10.26634/jpsy.13.3.16546

Kasahara, K., Machida, N., Osada, E., Takanashi, T., & Yoshizawa, S. (2012). Vocabulary knowledge of 5th and 6th graders at elementary school: Connections between sound, meaning, and spelling. *The Japan Association of English Teaching in Elementary School*, *12,* 90–101. https://doi.org/10.20597/jesjournal.12.0_90

Katada, Y., & Fukuzawa, S. (2019). *Chokkan de wakaru Excel de kikaigakushu* [*Machine learning using Excel by intuition*]. Impress.

Koshiba, W. (2015). Comparison of the vocabulary knowledge of third and six grade students—Through meaning, phonology, and orthography. *The Japan Association for the Study of Teaching English to Children, 34,* 39–52.

MEXT (2017). *Shogakkou gakushu shido yoryo kaisetsu gaikokugo katsudo gaikokugo hen* [Study of course guideline for primary school English version]. Kairyudo.

MEXT (2018a). *Let's try!* Tokyoshoseki.

MEXT (2018b). *We can!* Tokyoshoseki.

Nishide, K., & Mizumoto, A. (2009). An attempt to measure the familiarity level of 8000 English words. *Tsuru University Graduate School Review, 13,* 57–92. http://trail.tsuru.ac.jp/dspace/bitstream/trair/180/1/Y013057.pdf

Ojima, S., Matsuba-Kurita, H., Nakamura, N., Hoshino, T., & Hagiwara, H. (2011). Age and amount of exposure to a foreign language during childhood: Behavioral and ERP data on the semantic comprehension of spoken English by Japanese children. *Neuroscience Research*, *70*, 197–205. https://doi.org/10.1016/j.neures.2011.01.018

Saji, R., & Saeki, Y. (2013). Word familiarity and English pronunciation ability of Japanese 4th graders. *The Japan Association of English Teaching in Elementary School*, *13*, 150–162.

Sanseido Henshujo [Sanseido Editing Center]. (Ed.). (2019). *Miyasui katakana shingo jisho* [*Easy-to-read dictionary of katakana*] (3rd ed.) Sanseido.

Shibuya, T. (2012). *Shougakkou yonensei no katakana eigo no imi no rikai: Eigo bogo washa no hatsuon suru eitango no rikai* [*Recognition of English loanwords by fourth year pupils: Understanding English words pronounced by native speakers*]. *The Japan Association of English Teaching in Elementary School, 12,* 44–56. https://doi.org/10.20597/jesjournal.12.0_44

The Japan Association of College English Teachers. (2016). *The new JACET 8000 basic words*. Kirihara.

Willis, M., & Ohashi, Y. (2012). A model of L2 vocabulary learning and retention. *The Language Learning Journal*, *40*(1), 125–137. https://doi.org/10.1080/09571736.2012.658232

Yoshimura, H. (2009). Koritsu shogakko jido no goichisiki ni kansuru chousa no jissenhoukoku [Practical report on examining vocabulary knowledge of children in public primary school]. *The Japan Association of English Teaching in Elementary School, 9,* 23–30. https://doi.org/10.20597/jesjnl.9.0_23

Appendix I: Sample of Sound-Meaning Recognition Test

Listen carefully and circle the picture of the words you hear.

Question 1 (/æpl/)

Question 2 (/bæg/)

Question 3 (/m1lk/)

Note. All of the original instructions were written in Japanese.

Vocabulary Learning and Instruction
Volume 10, Issue 1, December 2021
http://vli-journal.org

Longitudinal Measurement of the Vocabulary Size of Japanese Junior High School Students: Developing a Vocabulary Size Test for Beginner Learners

Tsuyoshi Sato
Hirosaki University Faculty of Education

Abstract

As vocabulary plays an important part in L2 learning at the beginner level, investigation into vocabulary learning constitutes an essential part of second language acquisition research (Koda, 1989; Laufer, 1992; Meara, 1992). Assuming the case of Japanese, English education can provide a meaningful example. In addition, noting that precise measurements of the English vocabulary size of Japanese junior high school students have not been attained (Katagiri, 2000), the present study examines the growth of vocabulary size of Japanese junior high school students across their 3 years of English learning. For this purpose, a vocabulary size test (VST) was specially developed with a word list compiled from Japanese junior high school English textbooks based on the frequency and range of their vocabulary items (Sato, 2016). A total of 159 junior high school students participated in the research. The vocabulary size of the participants was measured five times over the period of 3 academic years at intervals of 6 months. Analysis of their vocabulary growth by frequency revealed a distinctive pattern for each level, yielding implications for more effective vocabulary teaching for learners at the beginner level in light of the characteristics of their vocabulary growth.

Keywords: vocabulary size, test development, Japanese junior high school students

1 Introduction

Extensive evidence suggests that vocabulary plays a vital role in L2 learning, showing a significant correlation with learner proficiency (Arnaud, 1992; Koda, 1989; Laufer, 1992; Meara, 1992). However ample research indicates that L2 learner vocabulary size is much smaller than those expected in the previous studies (Olmos, 2009; Pathan et al., 2019; Tang, 2007). Furthermore, most of these studies have focused on a "one-shot" measurement of vocabulary size at a given specific time, assessing participants' attainment of certain vocabulary levels at the end of a course or academic year. In other words, there is still considerable ambiguity with regard to the pattern of vocabulary growth longitudinally from the beginning stage. Specifically, little research has evaluated how many English words learners know at the beginning of a certain course and how their vocabulary

Please cite this article as: Sato, T. (2021). Longitudinal measurement of the vocabulary size of japanese junior high school students: developing a vocabulary size test for beginner learners. *Vocabulary Learning and Instruction, 10*(1), 30–48. https://doi.org/10.7820/vli.v10.1.sato

size develops throughout their academic career. Research on the process of learner vocabulary growth at the beginning stage likely holds significant implications for more effective vocabulary teaching at this important stage (Nation, 2006; van Zeeland & Schmitt, 2012).

In this regard, English education in Japan may become a suitable context which will lead to a test case in investigating beginner learners' vocabulary knowledge. Japan is now undergoing a dramatic shift in English education from teaching the language mainly through grammar-translation methods to those intended to foster communicative competence. Various changes have been made, such as introducing English as a required subject into the elementary school curriculums since 2020, expecting teachers to teach English mainly using English. Among the changes, an increase in the amount of vocabulary to be taught is notable. According to the new *Course of Study* by the Ministry of Education, Culture, Sports, Science and Technology (MEXT, 2018), a total of 4,000 to 5,000 words should be taught over 10 years of English education from elementary school to senior high school. This denotes that 600 to 700 words must be introduced in elementary school, with an additional 1,600 to 1,800 words in junior high school, and with an additional 1,800 to 2,500 in high school. It is crucial to periodically assess students' achievements in meeting these standards of vocabulary size in order to examine the validity of these standards and determine the necessary changes. However, little research has been conducted with Japanese junior high school students as beginner level English learners. Thus, assessment of the degree to which students have met the mandated standards in terms of vocabulary is mostly based on the experience, even intuition, of the teachers in charge of the lessons. Consequently, vocabulary size and growth of Japanese junior high school students have been overlooked and it remains a question to resolve.

2 Literature Review

It is generally accepted that vocabulary knowledge plays a crucial role in L2 learning. It is assumed that the more vocabulary a learner knows, the better his or her understanding and ease of communication becomes in the target language. There is also considerable evidence that the learner vocabulary size is strongly related to his or her proficiency in the language. Indeed, copious research has been conducted to measure vocabulary size, and has shown that it strongly correlates with several aspects of language proficiency, both in L1 and in L2 (Arnaud, 1992; Koda, 1989; Laufer, 1992; Meara, 1992; Pathan et al., 2019; Tang, 2007). Consequently, a variety of vocabulary size tests (VSTs) have been developed to measure learners' vocabulary proficiency (Aizawa, 1998; Koizumi, 2003; Meara, 1992; Meara & Buxton, 1987; Mochizuki, 1998; Nation, 1990, 2001; Nation & Beglar, 2007; Sato, 2003; Schmitt et al., 2001). Using these tests, plentiful research on vocabulary size has been conducted.

As Nation (2001) suggested, the number of words needed to use the language is utilitarian information to help decide how much vocabulary needs to be learned. In the case of second languages, it is generally accepted that 3,000

to 4,000 word families are needed for intermediate learners (Nation, 2006; van Zeeland & Schmitt, 2012). This is based on the findings identifying this as the vocabulary size necessary for 95% coverage of various written texts such as novels, newspapers, and movies (Nation, 2006). Moreover, Nation suggested this level as the primary standard for the vocabulary size that English learners need in order to use English independently. van Zeeland and Schmitt (2012) investigated the relationship between lexical coverage and comprehension of passages in listening tasks for 36 native and 40 non-native speakers. Based on their findings, they concluded that 98% is a good coverage target. However, they suggested that for less stringent comprehension rates, 90 to 95% should be acceptable with spoken passages. Based on this, van Zeeland and Schmitt (2012) concluded that language users need to know from a low of 2,000 to 3,000 word families (for 90 and 95% coverage) to a high of 6,000 to 7,000 families for adequate to ideal listening comprehension (for 98% coverage). In conclusion, various research has suggested that for L2 learners, it is appropriate to set the vocabulary learning goal for beginner learners to increase the vocabulary size up to 2,000 to 3,000 word families.

However, an increasing number of studies which have measured L2 learners' vocabulary size (Alonso, 2013; Gallego & Agustin Llach, 2009; Laufer, 1998; Qian, 2002), have concluded that the actual measure of their vocabulary size is generally much smaller than the requirements expected or proposed in those studies which induced the estimated vocabulary size based on the comprehension of the passages with different coverage rates. Pathan et al. (2019) investigated the receptive vocabulary growth of 123 first-year post-graduate university students in Pakistan. The vocabulary size of the students was measured at two different points and was estimated to be 3,500 word families and 4,090 word families at the beginning and end of the course, respectively. The VST developed by Nation and Beglar (2007) was adopted in this study. Pathan et al. concluded by expressing concern that the overall vocabulary size of the participants was much smaller than the proposed requirement of Nation's (2006) 95% coverage. The research showed that even in Pakistan, where English is an official language and people are generally considered to have high English proficiency, learner vocabulary size is not satisfactory. Olmos (2009) measured the vocabulary size of final-year students in a high school in Murcia, Spain, using Nation's (1990) Vocabulary Levels Test (VLT). A total of 49 students participated in Survey 1, measuring 1,000- and 2,000-word levels, and 38 students in Survey 2, measuring 3,000- and 5,000-word levels. The participants were considered to possess the vocabulary knowledge of a particular level if they scored above 88% of the established standard. The results showed that very few of the participants passed each level of the test: only three passed the 1,000-word level test, only one the 2,000-word level test, and not even a single participant passed either the 3,000- or 5,000-word level test. This suggests that the Spanish students, who spent 8–10 years studying English as a foreign language (EFL), were not able to reach the level necessary to be an efficient speaker of English. Tang (2007) investigated the vocabulary size of a total of 499 primary schools and secondary school EFL students in Hong Kong, also using VST, finding that the participants had a small vocabulary size and impoverished vocabulary knowledge. The primary school and secondary school students possessed a vocabulary size of about 925 and 2,891 word families, respectively. This suggests that Hong Kong

students still need help with basic English words even after learning English as a compulsory subject for 12,500 hours over 9 years.

Turning to Japanese learners, Kosuge (2003) measured the vocabulary size of Japanese junior high school students using a lemma-based VST for Japanese Learners of English (Mochizuki, 1998). The research concluded that at the end of their school year, the average vocabulary size of seventh-grade students was 1,185 lemmas and that of eighth-grade students was 1,859 lemmas. Katagiri (2000) researched the vocabulary size of 401 senior high school students using both written and listening versions of the same test. The results revealed that the vocabulary size of the participants ranged from 2,719 to 4,371 lemmas for the written version, and from 2,343 to 4,113 lemmas for the listening version. However, these findings need to be interpreted with caution because the VSTs adopted in these studies were based on lemma-based vocabulary lists, which are known to yield larger estimations than those based on word-family-based lists. According to Nation (2016), the approximate ratios for the 3,000 highest-frequency word families are 1: 3: 6 for word families, lemmas, and word types, respectively. Therefore, it is assumed that the estimated vocabulary size obtained from the studies above is around three times larger than would have been obtained by counting word families.

A serious limitation however with the above studies is that most of them adopt the multiple-choice format vocabulary size or levels tests such as VLT (Nation, 1990). As substantial research criticises, one of the major shortcomings of multiple-choice tests (MCT) is that they overestimate the learner vocabulary size because of the guessing effects (Gyllstad et al., 2015; Stewart, 2014; Stoeckel et al., 2020). Stoeckel et al. (2020) examined three studies which compared the score difference between meaning-recognition and meaning-recall tests, and found that the examinees answered correctly 13.9–62.9% of the meaning-recognition items that tested words unknown at the meaning-recall level of knowledge. Stewart (2014) revealed that the score overestimation for a four-choice vocabulary test could be as high as 25% for most ability levels of learners. Therefore, for precise measurement of the learner vocabulary size, much research suggests that it is preferable to avoid multiple-choice meaning-recognition tests and to adopt meaning-recall tests instead.

As discussed above, most studies of learner vocabulary size have generally focused on setting standards mainly for learners with a certain amount of learning experience using a cross-sectional study framework. This approach offers the possibility of substantial and valuable data. However, little research has sought to determine learner vocabulary size at the point when learners begin studying English and clarify the patterns of growth in their vocabulary size. Enhancing our understanding of how the vocabulary size of L2 learners at the beginning level develops to achieve the word level of 2,000 to 3,000 is significant as an initial step not only to improve effective curriculum development, but also to develop systematic vocabulary instructions for English learners.

2.1 Purposes of the Present Study

As stated above, the vocabulary size necessary for L2 learners to get 95% coverage of various written or spoken texts has been identified as 2,000 to 3,000

word families, constituting an appropriate first goal of vocabulary learning (Nation, 2001, 2013; Nation & Webb, 2011). However, neither the vocabulary levels of Japanese junior high school students at the beginning of each school year nor the pattern of development—how their vocabulary size increases—have been identified. In order to describe this longitudinal development, the present study set the following two objectives:

1. To determine the vocabulary size of Japanese junior high school students at the beginning and end of each school year.
2. To determine the longitudinal pattern of vocabulary size growth for Japanese junior high school students across 3 academic years.

3 Method

3.1 Participants

A total of 159 junior high school students in Japan participated in this study. The data from 10 participants who did not take part in one or more sessions were excluded from the analysis. The following analyses therefore were based on the remaining 149 participants. All participants are from a public junior high school attached to a national university in a rural area. It has five classes for each grade, thus it is regarded as a middle-size school. They had studied English from the first year of junior high school using MEXT authorised junior high school textbooks; at the time of the final measurement, they had studied English in Japan for 3 years as a required subject.

3.2 Materials

The experiment used the four-option multiple-choice VST, which was developed from the lemma-based vocabulary list compiled from all authorised junior high school English textbooks (seven publishers/textbooks for each grade) used in Japan (Sato, 2016). The vocabulary items are listed and ordered according to the indices attained from the following formula:

$$\text{Ranking criterion of the list} = \frac{range \times frequency}{grade\ level\ at\ which\ a\ word\ first\ appears\ in\ textbook}$$

The underlying concept of this formula is that in addition to frequency and range, learning period for any specific vocabulary item also has a strong effect on recognition of the word for Japanese learners. In other words, higher frequency and wider range words, those which students encounter earlier in their school years and have a longer period of exposure, are assumed to be learned better. Katagiri's (2000) data suggested that less frequent words seem, not to be necessarily, more likely to impair comprehension for Japanese EFL learners. The results of previous research (Sato, 2016) revealed that the frequency or range alone does not correspond to Japanese learners' vocabulary size. The

results demonstrate that some high-frequency words with wide ranges that are considered easier for English learners in general are in fact not so for Japanese junior high school students, and vice versa for low frequency and low range words. This is primarily because English in Japan is taught in the context of EFL, where students obtain most of their explicit vocabulary knowledge from school education. Based on this, Sato (2016) suggested that however high the frequency of a word is, it is rarely taken up by Japanese junior high school students until it is taught in English class. Therefore, Sato hypothesised that the learning period should be added as the third crucial factor along with frequency and range in developing word lists and VSTs for Japanese learners. For example, the word *language* first appears in the middle of the first year and appears a total of 817 times across all six authorised textbooks used in Japan. Its ranking criteria index is thus 817 × 6 ÷ 1.5 = 3,264. On the other hand, the word *energy* appears 840 times in five authorised textbooks, and first appears at the beginning of the second year. The ranking criteria index is thus 840 × 5 ÷ 2 = 2,100. Therefore, according to the index, *language* is listed higher than *energy*, even though its frequency is lower.

The VST used consists of four levels: Level 1—500-word level, Level 2—1,000-word level, Level 3—1,500-word level, and Level 4—2,000-word level. Each level comprises 30 test items; as each level includes 500 words, each test word represents about 15 words at that level, as can be seen in the Appendix. The VST adopts a Computer-Based Testing (CBT) system programmed to ran- domly extract the test items from the word list described above, which thus serves as an item bank. Using the CBT test format, it is expected to avoid memory effects from being used repeatedly with the same learners over a short span of time. The pilot test was conducted by Sato (2016) to confirm the equiv- alency in difficulties across the forms randomly generated by the computer programme. One-way AVOVAs were conducted for three different forms, and results show that there were no significant differences in means of raw scores across the forms (see Sato, 2016, 2017, for more detail). Consequently, in the present study, the vocabulary size of the participants was measured five times using this VST, but each time they took the test, the test items were different, even at the same level.

3.3 Procedure

The study examines the longitudinal vocabulary growth of Japanese junior high school students across their 3-year junior high school English education. Using the VST described in section 3.2 above, vocabulary size of the students was measured five times over three academic years: in July and December in the first and second years and July in the third year. One level, which comprises 30 test items, was administered per session, allowing students sufficient time to finish the whole test (approximately 15 minutes per session). The VST consists of four levels, consequently, the participants had four test sessions within each test period. Considering the burden on the participants, each test session was administered at the beginning of every English class. Thus, 4 days were needed for each test period.

3.4 Data Analysis

To examine the longitudinal vocabulary size growth of junior high school students, the vocabulary size was measured five times over the period of three academic years at intervals of 6 months, in July and December. A two-way repeated-measures ANOVA was conducted, considering the test time (five times over the 3 years) and levels of vocabulary size (500-, 1,000-, 1,500-, and 2,000-word levels) as independent variables and the test score as a dependent variable.

4. Results and Discussion

4.1 Vocabulary Size of Japanese Junior High School Students

The descriptive statistics of the test results of each word level are shown in Table 1. The vocabulary size was estimated by multiplying the ratio of correct answers by 2,000. Therefore, the estimated mean vocabulary size of the participants was: 1,295 in July of the first year; 1,420 in December of the first year; 1,570 in July of the second year; 1,655 in December of the second year; and 1,721 in July of the third year. Thus, the participants' vocabulary grew by an average of 425 words across the 3 academic years. It should be noted that the dispersion of the participants' vocabulary size remained relatively stable over the 3 years, while the mean vocabulary sizes increased.

However, as discussed above, the vocabulary sizes obtained from the present study can be overestimated compared with the actual vocabulary knowledge of the participants because of the MCT formats (Gyllstad et al., 2015; Stewart, 2014;

Table 1. Descriptive Statistics for the Vocabulary Size Test (*N* = 149)

		500 WL	1,000 WL	1,500 WL	2,000 WL
1st-J	Mean	23.62	21.77	18.94	13.36
	SD	4.42	4.75	4.83	4.21
	95% CI	[22.91, 24.34]	[21.00, 22.53]	[18.15, 19.72]	[12.68, 14.04]
1st-D	Mean	26.09	21.96	20.64	16.54
	SD	4.90	4.59	3.75	4.66
	95% CI	[25.29, 26.88]	[21.21, 22.70]	[20.03, 21.25]	[15.78, 17.29]
2nd-J	Mean	28.88	25.15	21.73	18.42
	SD	2.59	4.06	4.02	4.35
	95% CI	[28.46, 29.30]	[24.50, 25.81]	[21.08, 22.38]	[17.71, 19.13]
2nd-D	Mean	29.23	26.15	23.58	20.32
	SD	2.35	3.60	4.23	3.78
	95% CI	[28.85, 29.61]	[25.57, 26.74]	[22.90, 24.27]	[19.71, 20.93]
3rd-J	Mean	29.62	28.11	24.72	20.79
	SD	1.34	2.25	3.98	4.13
	95% CI	[29.40, 29.84]	[27.74, 28.47]	[24.08, 25.37]	[20.12, 21.46]

Note. The maximum score for each level of the VST (Sato, 2016) was 30. WL = world level, SD = standard deviation, CI = confidence interval, *1st-J* = July in the first year, *1st-D* = December in the first year, *2nd-J* = July in the second year, *2nd-D* = December in the second year, *3rd-J* = July in the third year.

Stoeckel et al., 2020). One possible way to eliminate the effect of overestimation as a result of guessing is adopting "correction-for-guessing formula" (Frary, 1988; Rowley & Traub, 1977). This formula estimates adjusted scores based on the number of questions which students got wrong on the test. The formula is: corrected score = # correct – (#incorrect/#options–1). For example, suppose a participant took the four-option multiple-choice VST adopted in this study and happened to get 105 correct and 15 incorrect items, then, the participant's score would be decreased from 105 to 100: 105 – (15/3) = 100. Consequently, it is plausible to estimate the mean vocabulary size of the participants somewhere between 1,060 and 1,295 in July of the first year, 1,227 and 1,420 in December of the first year, 1,426 and 1,570 in July of the second year, 1,540 and 1,655 in December of the second year, and 1,628 and 1,721 in July of the third year as described in Figure 1.

According to the *Course of Study* (MEXT, 2008), Japanese junior high school students are supposed to learn 1,200 words. Compared to this standard, the estimated vocabulary sizes of the participants of the present study were quite large, even in July of the first year at the beginning of their formal study of English. It is notable that the average size of their receptive vocabulary was beyond the 1,200 word level in December of the first year even when strict estimation considering the effect of overestimation because of guessing was used. Therefore, it is probable that the participants understand almost all the 1,200 words, that is, by reading, even at the very beginning of their English study. From the viewpoint of vocabulary knowledge, one of the greatest concerns of some English teachers in Japan is that students may have difficulty reading passages without Japanese verbatim translation, as is suggested by the *Course of Study* (MEXT, 2018). The findings contribute to minimising the concern in this regard and teachers should be encouraged to try to conduct reading lessons without word-for-word translation.

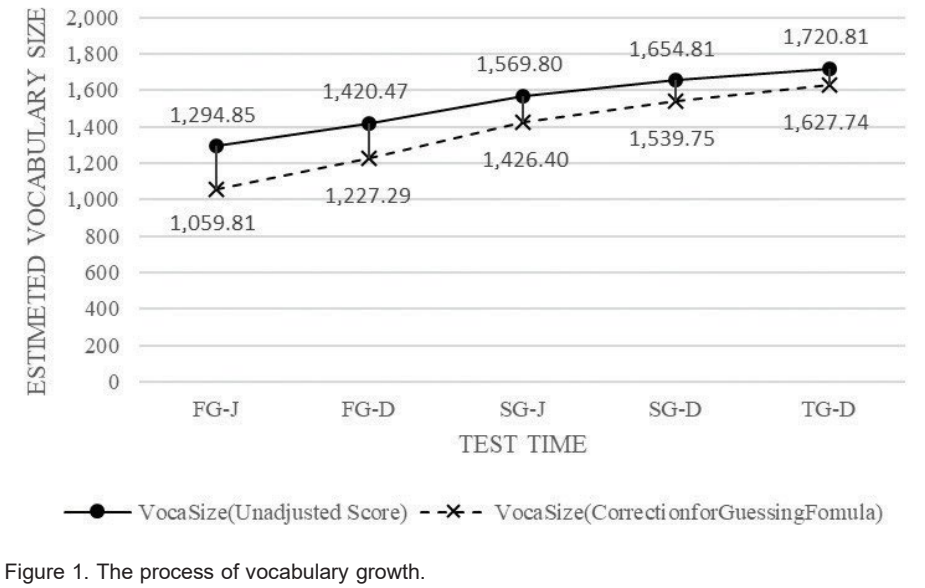


Figure 1. The process of vocabulary growth.

These unexpected results—that the participants' vocabulary size is as high as, if not higher than, the levels set by MEXT—may be partially explained by the abundance of English loan words experienced in students' daily lives. For students exposed to English loan words, it is not so difficult to either understand or infer the meaning of English words in receptive ways (Nation and Webb, 2011). Daulton (2008) investigated which and how many of the first 3,000 words of the British National Corpus list occur in Japanese and reported that over half are used in Japanese as loan words in one form or other. Under the current *Course of Study* (MEXT, 2018), Japanese students start studying English as a required subject from junior high school, and consequently, most of the English words introduced at this stage are high-frequency basic words. Indeed, Morinaga et al. (1978, 1979) suggested that Japanese authorised English textbooks include a substantial amount of loan words. As Nation and Webb (2011) suggested, learners' knowledge of loan words is a "tremendous resource for learning English" (p. 56). Therefore, many teachers assume that their students do know something of English for a variety of reasons. For example, English activities in elementary schools in Japan are prevalent, some students learn English from a young age at cram schools. This raises the need for further research investigating what and how many English loan words found in Japanese appear in the authorised English textbooks.

The results further suggest that it is necessary to revise the vocabulary test to measure larger vocabulary sizes (2,000 words) because of the revision of the *Course of Study* (MEXT, 2018). The number of vocabulary items to be taught in junior high school is to be increased up to 1,600 to 1,800 words, in addition to the 600 to 700 words to be taught in elementary school in the new English curriculum implemented from 2020 (MEXT, 2018). Therefore, students under the current *Course of Study* are expected to learn 2,200 to 2,500 words in total by the end of junior high school. Thus, it would be valuable to develop a VST capable of measuring vocabulary sizes up to 3,000 or 4,000 words.

Lastly, further studies to measure the productive vocabulary size of Japanese junior high school students would help teachers to make their vocabulary teaching more effective. It is problematic that the English *Course of Study* provided by the Japanese government fails to indicate whether the 1,600 to 1,800 words specified are receptive, productive, or both. It states only that "1,600 to 1,800 words should be taught in junior high school English lessons (MEXT, 2018, p. 33)." This seems not to be specific enough to show the goal of vocabulary teaching for teachers. As teachers might interpret this as requiring them to teach all these words as productive vocabulary, some may be overwhelmed by the large number of vocabulary to teach, while others might assign much harder work than necessary. Developing separate receptive and productive vocabulary lists, specifying the words to be learned by junior high school students based on empirical research rather than intuition, will contribute substantially to more effective vocabulary teaching and learning.

4.2. The Patterns of Vocabulary Growth by the Word Level

Table 2 shows the results of a two-way ANOVA, comparing the mean scores and standard deviations (SD) between the test times and the four levels of the

Table 2. Results of the Two-Way ANOVA for the Vocabulary Size Test

Sources	*SS*	*Df*	*MS*	*F*	*p*	η^2
Test Levels	37319.58	3	12439.86	796.04	<0.001	0.81
Test Time	16204.01	4	4051.00	259.23	<0.001	0.32
Interaction	868.20	12	72.35	4.63	<0.001	0.02
Error	46256.27	2,960	15.62			
Total	100648.06	2,979				

Note. *SS* = sum of squares, *df* = degrees of freedom, *MS* = mean sum of square, *F* = F ratio.

Table 3. Results of the Post Hoc Comparison (Level and Test Time)

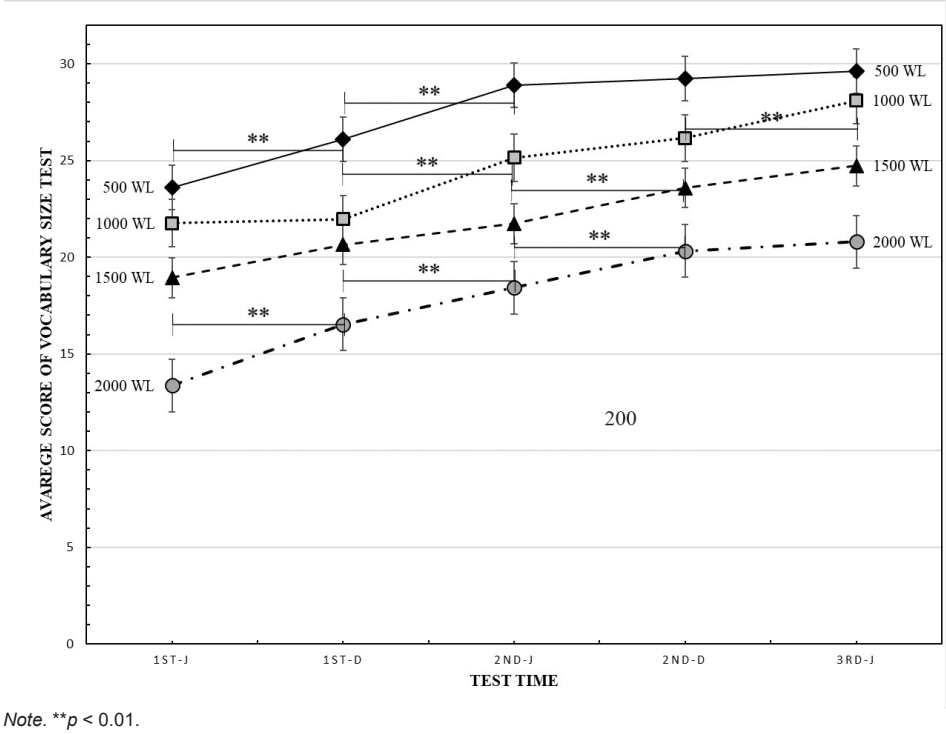
Level	Test Time		Mean Difference	*p*	Cohen's d
500 WL	1st-J	1st-D	2.46	<0.001	0.20
	1st-D	2nd-J	2.79	<0.001	0.22
	2nd-J	2nd-D	0.35	0.94	0.03
	2nd-D	3rd-D	0.40	0.91	0.03
1,000 WL	1st-J	1st-D	0.19	0.99	0.02
	1st-D	2nd-J	3.19	<0.001	0.26
	2nd-J	2nd-D	1.00	0.19	0.08
	2nd-D	3rd-D	1.95	<0.001	0.16
1,500 WL	1st-J	1st-D	1.70	0.03	0.14
	1st-D	2nd-J	1.09	0.68	0.09
	2nd-J	2nd-D	1.85	<0.001	0.15
	2nd-D	3rd-D	1.14	0.09	0.09
2,000 WL	1st-J	1st-D	3.17	<0.001	0.25
	1st-D	2nd-J	1.89	<0.001	0.30
	2nd-J	2nd-D	1.90	<0.001	0.15
	2nd-D	3rd-D	0.47	0.84	0.04

Note. WL = word level, *1st-J* = July in the first year, *1st-D* = December in the first year, *2nd-J* = July in the second year, *2nd-D* = December in the second year, *3rd-D* = December in the third year.

VST considering the test time (five times over the 3 years) and levels of vocabulary size (500-, 1,000-, 1,500-, and 2,000-word levels) as independent variables and the test score as a dependent variable. The results reveal a significant interaction between the two factors of the level and test time (F (12, 2960) = 4.63, $p < 0.001$, $\eta^2 = 0.02$ [small effect size, Field, 2005]). In addition, there were significant differences between the main effect at the four levels (F (3, 2960) = 796.04, $p < 0.001$, $\eta^2 = 0.81$ [large effect size]) and the test times (F (4, 2960) = 259.23, $p < 0.001$, $\eta^2 = 0.32$ [large effect size]).

To examine the pattern of vocabulary size growth in detail, the main effects were analysed using post hoc analysis. The results of Tukey's multiple comparison indicate that there were significant differences between the four levels (500-word level, > 1,000-word level, > 1,500-word level, > 2,000-word level), showing that each level measures the participants' vocabulary knowledge distinctively. This confirms the validity of the four-level construction of the VST (Sato, 2016).

The results for the simple main effect of the test time reveal the distinctive patterns of vocabulary growth for each level, as shown in Table 3 and Figure 2.



Note. $^{**}p < 0.01$.

Figure 2. The distribution of estimated vocabulary size.

The vocabulary size did not always increase linearly. Instead, a distinctive pattern is revealed for each level. For the 500-word level, the participants showed a significant increase until July of the second year and reached a plateau after that, indicating a ceiling effect. These results can be interpreted to suggest that the participants had acquired the vocabulary of the 500-word level by July of the second year. For the 1,000-word level, on the other hand, there were significant increases observed from December of the first year through July of the second year, and from December of the second year through July of the third year, the 7th to 8th grade levels. For the 1,500-word level, a significant increase was observed from July through December of the second year. Finally, for the 2,000-word level, the participants' vocabulary grew at each point until December of the second year.

These distinctive patterns of learners' vocabulary growth are explained by the characteristics of the words in each level and the contents taught at each period. For example, most basic and frequently used words, such as *do*, *have*, *like*, *can*, *what*, *how*, *good*, *many*, *time*, *Japan*, and *English* are included in the 500-word level. These words are used when discussing general topics rather than being highly content or topic specific. It is important to note that the vocabulary knowledge needed to express all tenses—past, present, and future—are included in the 500-word level, such as the base forms of important basic verbs and their past tense forms (*did*, *went*, *had*, and *came*), and auxiliary verbs (*will*, *may*, etc.). This is consistent with the structure of the syllabus of Japanese English textbooks,

where the verb tenses are introduced in the order of present, past, and future by July of the second year. In other words, Japanese junior high school students are expected to acquire basic sentence patterns and tenses of English using high-frequency words included in the 500-word level, which are used commonly in most interactions. Therefore, it is crucial for Japanese students to be able to understand the vocabulary included in the 500-word level by the end of the second year, which provides a fundamental base for their future English learning. This substantiates the suggestion made by Kosuge (2003) that loss of motivation and onset of difficulties with studying English begins to appear in the middle of the second year. Indeed, Japanese teachers of English often refer to the second year as "the crossroads of English learning," marking a significant period for students. Therefore, measuring student vocabulary growth at this point as a formative assessment may contribute to appropriate adjustments to classes and provision of remedial instruction if necessary.

The vocabulary of the 1,000 and 1,500-word levels, on the other hand, consists of words related to more specific topics or contents, for example, Japanese culture or tradition (*shrine*, *temple*, *cherry blossom*, *pray*, and *ancient*), or social issues (*environment*, *fossil fuel*, *volcano*, *wheelchair*, and *volunteer*). Once students learn the fundamental structures of English together with the vocabulary at the 500-word level, they are exposed to less frequently used words to express more specific contents. Vocabulary growth at the 1,000 and 1,500-word levels beginning after December of the first year reflects this delayed exposure to these words in the textbooks.

Finally, words of 2,000-word level are characterised as "topic-specific vocabulary reflective passage content," that is, vocabulary directly related to the topics dealt with in the textbook chapters. These words are neither high frequency words in general English use, nor is this vocabulary borrowed into Japanese as loan words. These words are necessary for the specific topics or stories of the textbook, thus they appear only in the relevant passages in English textbooks. For example, words such as *lullaby*, *sugarcane*, and *thatch* appear in a story about Okinawa in World War II, and *polar bear*, *aurora*, and *iceberg* appear in a biography of a photographer working in the Arctic. Knowledge of these words grows continuously from the first year to the second year because learners rarely have knowledge of the vocabulary of this level from the beginning. The mean score for this level is 13.36 out of 30, indicating room for development. However, the present study failed to account for the reason why the vocabulary knowledge of this level develops at each time point, even though these words appear only in limited lessons or one particular story. Further research is needed to seek possible explanations.

While intuition leads English teachers in Japan to generally assume that vocabulary size increases linearly from high-frequency to low-frequency words, the present study reveals that this is not the case. The results show a distinctive pattern according to each level and/or the characteristics of the words. Although future studies on this topic are needed to verify or generalise these findings considering that this pattern would change if a different set of textbooks were used at this school, the present research offers invaluable findings not only for vocabulary teaching but also for preparing appropriate reading materials for Japanese

students by controlling the vocabulary level to match the target learners based on the data obtained by measuring their vocabulary size.

5. Conclusion

The present study primarily aimed to investigate the pattern of vocabulary growth of beginning English learners. The findings suggest that the vocabulary size of junior high school students in this context increases differently across word frequency levels. This implies the following regarding the growth in vocabulary size of junior high school students.

It is clearly evident that the learners already have a fairly large vocabulary even in July (fourth month) of the first year at the beginning of their formal study of English, estimated at around 1,060 and 1,295 words. Learner vocabulary thereafter develops, showing a distinctive pattern for each level. The 500-word level increases significantly until July of the second year, at which point it plateaus. These are words to express general topics or contents that students should surely acquire as soon as possible. The 1,000-word level vocabulary size increases from December (ninth month) of the first year to July of the second year, and from December of the second year to July of the third year. For the 1,500-word level, a significant increase occurs from July through December of the second year. These words mainly relate to limited or specific topics. The 2,000-word level, including textbook topic-related words, develops consistently until December of the second year. The observed patterns of vocabulary development can be explained by the relationship between the characteristics of vocabulary items included in each level and the topics or contents of the textbooks at a time.

This research provides a new and effective approach for vocabulary instruction at the beginning level for Japanese English learners. The findings of the present study also constitute a valuable resource both for material development and setting guidelines for material selections appropriate to the levels of learners. They should guide instructors in controlling the vocabulary during the preparation of reading materials that will be helpful to students based on the data obtained by measuring their vocabulary size. These results should be confirmed by further research with junior high school students in different contexts in a case study approach considering the possibility that vocabulary of every learner does not develop precisely as outlined above.

The limitations of this study should also be noted. First, the vocabulary sizes of the third-year students are probably larger than the coverage of the VST adopted in the present study (2,000 words). In addition, for the fact that the vocabulary size to be taught is increased in the new *Course of Study* (MEXT, 2018), the need is to revise the vocabulary test to measure larger vocabulary sizes. Second, the current study failed to examine either individual test takers or test items. Therefore, the degree to which test takers and words conform to or deviate from these patterns remains unclear. Research comparing item measure and person measure indices and investigating how well the difficulty levels actually correlate with the adjusted frequency values utilising Rasch Analysis is already underway. Third, as the participants are from a junior high school attached to a national university,

they likely have larger vocabulary size than students from public schools. In that sense, application of this research method targeting students with various backgrounds should be fruitful in extending its implications. Finally, it is probable that the participants' vocabulary size was overestimated for adopting multiple-choice VSTs because of the guessing effects (Gyllstad et al., 2015; Stewart, 2014; Stoeckel et al., 2020). Therefore, it is urgent to develop alternative test formats which estimate learner vocabulary size precisely with more confidence. Stewart (2014) and Stoeckel et al. (2020) suggested that it can be achieved with adapting item response theory (IRT) or computer adaptive tests. Fourth, as the present study only focused on learners' receptive vocabulary size, a further study should examine the pattern of their vocabulary growth from a different aspect, such as productive vocabulary learning or depth of vocabulary knowledge.

Acknowledgements

I would like to thank the English teachers of the junior high school for cooperating in the 3-year longitudinal research, without whose help this would never have been possible. I would also like to thank Editage (www.editage.com) for English language editing.

References

Aizawa, K. (1998). Developing a vocabulary size test for Japanese EFL learners. *ARELE: Annual Review of English Language Education in Japan*, *9*, 75–85. https://doi.org/10.20581/arele.9.0_75

Alonso, A. C. (2013). Receptive vocabulary size of Spanish secondary EFL learners. *Revista de Lingüísticay Lenguas Aplicadas*, *8*, 66–75. https://doi.org/10.4995/rlyla.2013.1180

Arnaud, P. J. (1992). Objective lexical and grammatical characteristics of L2 written compositions and the validity of separate-component tests. In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 133–145). Macmillan.

Daulton, F. (2008). *Japan's built-in lexicon of English based loanwords*. Multilingual Matters.

Field, A. (2005). *Discovering statistics using SPSS* (2nd ed.). Sage Publications.

Frary, R. B. (1988). Formula scoring of multiple-choice tests (correction for guessing). *Educational Measurement: Issues and Practice, 7*(2), 33–38. https://doi.org/10.1111/j.1745-3992.1988.tb00434.x

Gallego, M. T., & Agustin Llach, M. P. (2009). Exploring the increase of receptive vocabulary knowledge in the foreign language: A longitudinal study. *International Journal of English Studies*, *9*, 113–133. http://revistas.um.es/ijes/article/viewFile/90681/87481

Gyllstad, H., Vilkaitė, L., & Schmitt, N. (2015). Assessing vocabulary size through multiple-choice formats: Issues with guessing and sampling rates.

ITL-International Journal of Applied Linguistics, *166*(2), 278–306. https://doi.org/10.1075/itl.166.2.04gyl

Katagiri, K. (2000). Quick and rough estimates of general English ability using Mochizuki's vocabulary size test for Japanese EFL learners. *The Japan Language Testing Association Journal, 3*, 83–99. https://doi.org/10.20622/jlta.3.0_83

Koda, K. (1989). The effects of transferred vocabulary knowledge on the development of L2 reading proficiency. *Foreign Language Annals*, *22*, 529–540. https://doi.org/10.1111/j.1944-9720.1989.tb02780.x

Koizumi, R. (2003). A productive vocabulary knowledge test for novice Japanese learners of English: Validity and its scoring methods. *JABAET Journal*, *7*, 23–52. http://www7b.biglobe.ne.jp/~koizumi/Productive_VKT_JABAET.pdf

Kosuge, A. (2003). Goi-size-test kara mita goi no shutoku [Vocabulary development from the perspectives of a vocabulary size test]. In H. Ota, A. Kosuge, K. Kanatani, & S. Hidai (Eds.), *Eigoryoku ha donoyouni shite nobite ikuka: Chugakuseino eigo shutoku kateiwo ou* [*How English proficiency develops: Examining the process of English acquisition of junior high school students*] (pp. 81–115). Taishukan.

Laufer, B. (1992). How much lexis is necessary for reading comprehension? In P. J. L. Arnaud & H. Béjoint (Eds.), *Vocabulary and applied linguistics* (pp. 126–132). Macmillan.

Laufer, B. (1998). The development of passive and active vocabulary in a second language: Same or different? *Applied Linguistics*, *19*, 255–271. https://doi.org/10.1093/applin/19.2.255

Meara, P. (1992). *EFL vocabulary tests*. Centre for Applied Language Studies, University College Swansea.

Meara, P., & Buxton, B. (1987). An alternative to multiple choice vocabulary tests. *Language Testing*, *4*, 142–151. https://doi.org/101177/026553228700400202

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2008). *Chugakko gakushu shido yoryo gaikokugo* [*The course of study for foreign languages in junior high schools*]. https://www.mext.go.jp/a_menu/shotou/new-cs/youryou/index.htm

Ministry of Education, Culture, Sports, Science and Technology (MEXT). (2018). *Chugakko gakushu shido yoryo gaikokugo* [*The course of study for foreign languages in junior high schools*]. https://www.mext.go.jp/a_menu/shotou/new-cs/1384661.htm

Mochizuki, M. (1998). Nihonjin eigo gakushusha no tameno goi size test [A vocabulary size test for Japanese learners of English]. *IRLT Bulletin*, *12*, 27–53. http://ci.nii.ac.jp/naid/40004654197

Morinaga, M., Ishidoya, S., Hagino, H., Osa, K., Yamaguch, M., & Yasuhara, S. (1978). *Gariargo to eigo kyoiku: Syogakko kyoukasho ni arawareta gairaigo risuto* [*Loan words and English education: The list of loan words in the textbooks of elementary school students*]. Tokyoto chugakko eigo kyouiku kenkyukai.

Morinaga, M., Ishidoya, S,. Hagino, H., Osa, K., Yamaguch, M., & Yasuhara, S. (1979). *Gariargo to eigo kyoiku: Chugakko kyoukasho ni arawareta gairaigo risuto* [*Loan words and English education: The list of loan words in the textbooks of junior high school students*]. Tokyoto chugakko eigo kyouiku kenkyukai.

Nation, I. S. P. (1990). *Testing and learning vocabulary.* Heinle & Heinle.

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*, 59–82. https://doi.org/10.3138/cmlr.63.1.59

Nation, I. S. P. (2013). *What should every ESL teacher know*? Compass.

Nation, I. S. P. (2016). *Making and using word lists for language learning and testing*. John Benjamins Publishing Company. https://doi.org/10.1075/z.208

Nation, I. S. P., & Beglar, D. (2007). A vocabulary size test. *The Language Teacher, 31*(7), 9–13.

Nation, I. S. P., & Webb, S. (2011). *Researching and analyzing vocabulary*. Heinle Cengage Learning.

Olmos, C. (2009). An assessment of the vocabulary knowledge of students in the final year of secondary education: Is their vocabulary extensive Enough? *International Journal of English Studies, 9*(3), 73–90. https://revistas.um.es/ijes/article/view/99531

Pathan, Z. H., Ismail, S., & Soomro, N. H. (2019). An investigation into receptive vocabulary growth and its predictability for reading development of university students in a semester course. *International Journal of Instruction*, *12*(1), 797–808. https://doi.org/10.29333/iji.2019.12151a

Qian, D. D. (2002). Investigation of the relationship between knowledge and academic reading performance: An assessment perspective. *Language Learning*, *52*, 513–536. https://doi.org/10.1111/1467-9922.00193

Rowley, G. L., & Traub, R. E. (1977). Formula scoring, number-right scoring, and test-taking strategy. *Journal of Educational Measurement*, *14*(1), 15–22. https://doi.org/10.1111/j.1745-3984.1977.tb00024.x

Sato, R. (2003). Developing the vocabulary size test for junior high school and high school students according with the new course of study. *STEP BULLETIN*, *15*, 25–37. https://www.eiken.or.jp/center_for_research/pdf/bulletin_archives/vol_15.pdf

Sato, T. (2016). Developing an English vocabulary size test for Japanese junior high school students: Examination of validity and reliability. *TELES Journal, 36,* 107–118.

Sato, T. (2017). Measuring vocabulary size of Japanese junior high school students: Cross-sectional and longitudinal viewpoints. *JLTA Journal, 20*, 90–106. https://doi.org/10.20622/jltajournal.20.0_90

Schmitt, N., Schmitt, D., & Clapham, C. (2001). Developing and exploring the behavior of two new versions of the vocabulary levels test. *Language Testing*, *18*, 55–88. https://doi.org/10.1177/026553220101800103

Stewart, J. (2014). Do multiple-choice options inflate estimates of vocabulary size on the VST?. *Language Assessment Quarterly*, *11*(3), 271–282. https://doi.org/10.1080/15434303.2014.922977

Stoeckel, T., McLean, S. & Nation, P. (2020). Limitations of size and levels tests of written receptive vocabulary knowledge. *Studies in Second Language Acquisition*, *43,* 181–203. https://doi.org/ 10.1017/S027226312000025X

Tang, E. (2007). An exploratory study of the English vocabulary size of Hong Kong primary and junior secondary school students. *The Journal of Asian TEFL*, *4*(1), 125–144.

van Zeeland, H., & Schmitt, N. (2012). Lexical coverage in L1 and L2 listening comprehension: The same or different from reading comprehension? *Applied Linguistics*, *34*, 457–479. https://doi.org/10.1093/applin/ams074

Appendix

Sample of the Vocabulary Size Test for the 500-Word Level

中学生のための語彙サイズテスト (500 word Level)

ID 名前

1〜30の単語の意味として最もふさわしいものをa〜dの中から選び記号を○で囲みなさい。

1. now
a. （一緒に）〜しましょう、〜しよう
b. 今
c. 〜の後に
d. 〜です。〜である。〜にある、〜にいる

2. your
a. あなた（たち）の
b. そこへ（で・に）
c. 〜を食べる
d. 〜について、およそ、約

3. our
a. 〜を使う
b. 私の
c. 私たちの
d. 水

4. park
a. 公園
b. しばしば
c. 〜を試す、試み
d. 面白い、興味深い

5. word
a. 一緒に
b. (写真）を撮った、〜を買った、連れていった
c. 冬
d. ことば、語

6. speak
a. 長い、長く
b. 外へ
c. 幸せな
d. 話す

Sample of the Vocabulary Size Test for the 1,000-Word Level

中学生のための語彙サイズテスト (1000 word Level)

ID 名前

1〜30の単語の意味として最もふさわしいものをa〜dの中から選び記号を○で囲みなさい。

1. enough
a. 〜しましょうか、〜しませんか
b. 行った、催した
c. 十分な（に）
d. ピザ

2. die
a. 忙しい
b. 声
c. 筆、ブラシ
d. 死ぬ

3. such
a. 地球
b. 他に（の）
c. そのような
d. すべて

4. news
a. 外で、屋外で
b. 静かな
c. 知らせ、ニュース
d. 音楽家

5. bridge
a. 事務所、会社
b. 橋（はし）
c. 木
d. おばあちゃん

6. cup
a. 〜の間で（に）
b. 城
c. 茶わん、カップ
d. 地震

Sample of the Vocabulary Size Test for the 1,500-Word Level

中学生のための語彙サイズテスト (1500 word Level)

ID 名前

1〜30の単語の意味として最もふさわしいものをa〜dの中から選び記号を○で囲みなさい。

1. professional
a. 大学
b. 〜を怖がらせる、おびえさす
c. 目の不自由な
d. 職業(上)の、本職（の）

2. shock
a. 奉仕（ほうし）、世話
b. 〜を作り出す、生み出す
c. 黒板
d. 衝撃、振動

3. address
a. （うちけしの文で）もう（〜ない）
b. （洋服の）ボタン
c. 住所
d. お母さん、ママ

4. square
a. ある量
b. 四角の、正方形
c. 安い
d. 植物、〜を植える

5. act
a. 行進、行進する
b. 行動する
c. 祖先、先祖
d. とても、本当に

6. medal
a. 神
b. 毒（どく）
c. 〜の上／の／で
d. 薬

Sample of the Vocabulary Size Test for the 2,000-Word Level

中学生のための語彙サイズテスト (2000 word Level)

ID 名前

1〜30の単語の意味として最もふさわしいものをa〜dの中から選び記号を○で囲みなさい。

1. import
a. 目が覚めた
b. 子守歌
c. 〜を輸入（ゆにゅう）する
d. 個人的な

2. relay
a. ぱっと発する光、閃光（せんこう）
b. 〜を稼ぐ（かせぐ）、〜を得る
c. ひざ
d. リレー競争

3. profession
a. 苦い
b. 〜を賞賛（しょうさん）、崇拝（すうはい）する
c. 職業、専門職
d. ナス

4. final
a. 郵便箱、ポスト
b. 景色、風景
c. （ほうきやブラシなどで）掃く（はく）
d. 決勝戦

5. tug-of-war
a. 輪、鳴る
b. 地下鉄
c. 綱引き
d. 残念なこと

6. acid rain
a. わらぶき
b. 縄（なわ）、綱（つな）
c. 酸性雨
d. うなずく

Vocabulary Learning and Instruction
Volume 10, Issue 1, December 2021
http://vli-journal.org

"That Sounds About Right"—Lexical Bundle Naturalness Intuitions in Japanese Learners of English

Julian Northbrook and Kathy Conklin
School of English, University of Nottingham

Abstract

The current study examines the perceived naturalness of lexical bundles learned from early-stage teaching materials in Japanese learners of English. Naturalness ratings of 24 native English speakers and 23 non-native speakers with Japanese as their first language were compared in relation to corpus derived frequencies from the British National Corpus (BNC) and a corpus of Japanese secondary-school English textbooks—the "Junior High School English Textbook Corpus" (JHSETC, Northbrook & Conklin, 2018). The rating scores of both groups were significantly predicted by lexical bundle frequency. However, the groups were sensitive to different metrics; the performance of native speakers was best predicted by the BNC, and that of the Japanese speakers by the JHSETC. This is taken as evidence that learner intuitions are affected by the input they receive from teaching materials, and that these intuitions may stay with learners long term.

Keywords: naturalness intuitions, Japanese learners of English, lexical bundles

1 Introduction

Anyone who has learned a second language has likely had the experience of being told something they said or wrote is, *"Okay, but not how a native speaker would say it."* In Hymes' (1972) taxonomy of situational competence, this relates to the fourth parameter, namely, that language should be *performed.* In other words, language needs to be "done" in the real world, not simply be possible, well formed (feasible) or appropriate (the first three parameters). Arguably, this is one of the greatest challenges language learners face: not distinguishing between grammatical and non-grammatical utterances but distinguishing between *possible* utterances and *likely utterances.* Early work in the area of formulaic language highlighted this fact and proposed a new way of looking at language—not as simply a set of syntactic rules and words, as generative grammar approaches had proposed, but as a memory-based system using lexical elements (Bolinger, 1975). Sinclair (1991) referred to this as a system that considers both an open choice principle and the *idiom* principle. The "open choice" principle refers to speakers' ability to construct unique and creative utterances using a system of rules and connected items. However, Sinclair argued that this is not the way in which we primarily produce language. Rather, we default to the *idiom principle*—that is, where possible we

Please cite this article as: Northbrook, J. & Conklin, K. (2021). "That sounds about right"—Lexical bundle naturalness intuitions in Japanese learners of English. *Vocabulary Learning and Instruction, 10*(1), 49–63. https://doi.org/10.7820/vli.v10.1.northbrook.conklin

use pre-constructed formulaic—or, "performed"—language that is entrenched in long-term memory (see, for example, Conklin & Schmitt, 2012).

Following this, Pawley and Syder (1983) coined the terms "nativelike fluency" and "nativelike selection," arguing that simply exercising the creative potential of syntactic rules is not sufficient to sound "nativelike" when speaking a language. Although native speakers could say *"Pleasant morning."* or, *"Could you aid me in this task?"* they tend not to do so. This is not because these phrases are ungrammatical but rather because there are more common, socially institutionalised phrases available (e.g., *"Good morning."* and *"Could you help me with this?"*). Most people would likely agree that these more common alternatives sound more natural, are understood more readily and are preferred by native speakers. At the same time, there are utterances such as, *"Thanks very much."* that are highly frequent expressions and commonly used by native speakers, but which are, in fact, ungrammatical (the expression being an amalgamation of *"Thank you very much"* and the archaic *"many thanks to you"* and, therefore, does not conform to typical grammatical conventions). We must conclude, then, that communicative competence in a language is not simply the ability to produce grammatically correct utterances. It also encompasses the ability to select frequent and conventional patterns that are *likely* in the language (see Biber et al., 1999; Cowie, 1998; Ellis, 2001; Howarth, 1998; Skehan, 1998, among others, for further discussion), and therefore, whether learners are able to do this is an important issue in English as a Foreign Language (EFL) and Second Language Acquisition. Before we consider studies that have looked at learner intuitions of formulaic sequences, it is first necessary to define the concept of formulaicity in the context of the current study.

1.1 Formulaic Sequences in Second Language Learning

There are many different types of formulaic language: idioms: *kick the bucket*, collocations: *plastic surgery*, binomials: *bride and groom*, phrasal verbs: *rely on*, lexical bundles: *I don't want to*, to name but a few. The current study focuses on lexical bundles, which are non-idiomatic sequences of words that are defined purely based on their frequency of occurrence (Biber, 2009; Biber et al., 1999; Stubbs & Barth, 2003) and can be thought of as extended collocations (Biber et al., 1999). At the early stages of second language acquisition, learners are thought to use memorised formulaic language, which allows them to express themselves above and beyond their current abilities (Myles et al., 1998, 1999; Nattinger & DeCarrucio, 1992; Wray, 2002). It is not necessary, for example, to understand that the phrase "*That sounds great*" is made up of a pronoun acting as a noun phrase, which is modified by a verb phrase consisting of a third-person singular verb in the present tense combined with an adjective—or indeed, to worry that the aforementioned, *"Thanks very much"* is, technically speaking, ungrammatical. Learners can simply use them as is. As proficiency advances, however, learners tend to construct much of their speech using rules rather than lexicalised routines (Foster, 2001; Nesselhauf, 2003; Skehan, 1998), perhaps believing that high-proficiency language requires originality. This is mistaken, however, because studies show that the use of formulaic language, far from indicating that a learner is a "beginner," is, in

fact, a mark of nativelikeness and, therefore, proficiency (e.g., Boers et al., 2006; Dickinson, 2012).

1.2 Exploring Naturalness in Second Language Learners

Dickinson (2012) trained students on specific formulaic sequences and found that those who practised and actively used them gave presentations that were more engaging and easier to understand than those who did not. In a similar study, Taguchi (2007) found that instructing and drilling formulaic sequences with learners of Japanese as a second language led to a higher sensitivity to discourse features and was useful for developing learners' oral ability. Wray and Fitzpatrick (2008) similarly concluded that by acquiring formulaic sequences that are common in conversations, learners not only increase their fluency and naturalness in the language but also increase their confidence in using the language. In another study, Boers et al. (2006) trained two groups of students for 22 hours each—one using a traditional words-and-rules approach; the other was trained to use formulaic sequences. Subsequently, two judges rated the group that had formulaic training as having higher proficiency.

Despite these findings, very little has been said about how well non-native speakers are able to judge the naturalness of sequences they encounter. Does instruction help learners to differentiate between utterances such as *can you aid me with this task?* and *can you help me with this?* Which, in turn, would allow them to better assess the formulaicity of their own speech and "select" un-nativelike expressions? In one of the few studies looking specifically at learners' ability to judge formulaic sequences, Edmonds (2013) administered an online naturalness judgement task to French natives and two groups of learners of French living in France (half long stay and half short stay). All groups judged conventional expressions as being more natural and did so faster than control phrases. Importantly, this included the short-stay French learners, which, as Edmonds concludes, indicates that even learners who have little relative exposure to the target language are "well on their way to nativelike selection" (p. 95). In two studies that looked at adjective-noun collocation intuitions, Siyanova and Schmitt (2008) and Siyanova-Chanturia and Spina (2015) asked native and non-native speakers to judge the frequency of collocations, which we take here to be related to their naturalness. Similar to Edmonds (2013), Siyanova-Chanturia and Spina (2015) found that both native and non-native speaker intuitions correlated strongly with corpus frequency. This was, however, only for high-frequency items. Neither group were able to judge medium- or low-frequency pairs, although the native speakers were able to accurately judge very low-frequency items. Siyanova and Schmitt (2008) had somewhat similar findings. They showed that native speaker intuitions correlated strongly with BNC frequency scores. However, the non-native group displayed poor collocational intuitions in comparison with the native speakers.

Overall, it seems that while learners may be able to accurately judge formulaic sequences, their intuitions may not be entirely accurate or in tune with authentic language use. An explanation for this may be found in usage-based approaches to second language acquisition, which put a premium on the linguistic input that learners receive. In such an approach, experience of and exposure

to language result in high-frequency, repetitive sequences of words being stored in long-term memory (Barlow & Kemmer, 2000; Bybee, 1998, 2006, 2013; Ellis, 2002a, 2002b, 2003), which would influence learners' intuitions regarding formulaic sequences. Thus, the question arises: is it that the second language learners cited above are not sensitive to collocation, or is it that their intuitions are simply *different*? That is, do they treat formulaic language as compositional and not as chunks, or is it simply that they have, through exposure to "un-nativelike" English (i.e., "textbook English") developed un-nativelike intuitions?

In order to answer this question, it is necessary to consider the input English learners receive. In the case of EFL students, this is, as previously stated, primarily classroom instruction and materials (Meunier, 2012; Nunan, 1991). Studies focusing on secondary school textbooks have found the dialogues in them to be inauthentic (Nguyen & Ishitobi, 2012; Northbrook & Conklin, 2018; Römer, 2004, 2005, see also Gilmore, 2004), in that they do not accurately represent the situations students need to be prepared for. Looking at formulaic language more specifically, Northbrook and Conklin (2018) examined lexical bundles in English textbooks for Japanese secondary school students and found them to occur frequently, although they were qualitatively different to those found in naturally produced English. Northbrook and Conklin (2019) found that junior high school students had a processing advantage on a phrasal judgement task for lexical bundles from their textbooks but not for matched controls. It appears that the learners were exposed to un-nativelike formulaic language and had an advantage for it, irrespective of native-speaker frequency distributions. Crucially, the lexical bundles tested in Northbrook and Conklin (2019) were selected *because* they occurred in the textbook series that the participants had used during their 3 years of junior high school, and the students were tested while in their final year. Subsequent exposure to English might help students to overcome their initial exposure and become more nativelike.

The current research is a small-scale study that investigates whether adult Japanese learners of English intuitions are based on the frequency distributions of the classroom English from their junior high school textbooks or whether they have become more nativelike. To address this, we asked adult native speakers of English and adult non-native, Japanese learners of English to rate the naturalness of sentences containing lexical bundles in an offline judgement task and examine whether their ratings were related to a native speaker corpus (the British National Corpus, BNC) or the Junior High School English Textbook Corpus (JHSETC).

2 Methodology

2.1 Participants

Participants for this study included 24 native speakers from the University of Nottingham and 23 native Japanese speakers. The native speakers participated for course credit. Initially, we had planned to recruit the Japanese English learners on-site and run the experiment in a lab setting, replicating Edmonds's (2013) study. However, due to the difficulty of recruiting appropriate participants in the area, we decided to modify the research design and run the experiment online

(see Procedures below). The Japanese speakers were all studying English as a second language and were recruited from an online email newsletter associated with a self-study e-learning platform targeted at intermediate English learners. We sent an email to a segment of the newsletter that represented the best fit for the study (Japanese English learners living and studying in Japan) asking for volunteers to participate. The group were, therefore, a sample of convenience. The Japanese participants were asked to provide their background information, as summarised in Table 1. In addition to what can be seen in the table, we asked the participants to indicate their profession: five were university students, and the remaining were in full-time employment (10 working in an office, three as engineers, one as a lawyer, one as a medical officer and one as a waitress). All of the Japanese participants reported that they lived in Japan, had never lived in an English-speaking country and had attended a typical Japanese-medium school.

2.2 Materials

We created 17 passages consisting of a line of text to provide context, followed by a short dialogue. One portion of each dialogue contained either a high-frequency lexical bundle or a low-frequency control phrase matched for meaning (see below for lexical bundle and control selection criteria). There were two versions of each passage (a total of 34 possible items). Participants saw either version A or B, but not both, so each participant judged 17 items. Lexical bundles are ideal for looking at questions of naturalness, because unlike idioms, they are not easily identifiable when embedded into a sentence. For example, items such as, *"James' aunt kicked the bucket / died last week*," might make it clear that the focus is on idioms and potentially influence participants' responses. In contrast items like, "*Really, I don't know why/ I'm not sure why I agreed to this*," should make the focus on formulaic language less apparent. Each item included a context line and dialogue that served to contextualise and mask the lexical bundle. In the following example, everything other than the underlined portion is supporting the context:

> Context: Andy is going on a double date with his friends.
>
> Lexical bundle: "Really, I don't know why I agreed to this," he grumbled to his friends Simon and Alice. "You know this isn't my sort of thing."
>
> Or:
>
> Control: "Really, I'm not sure why I agreed to this," he grumbled to his friends Simon and Alice. "You know this isn't my sort of thing."

Table 1. Summary of Non-native Speakers' Age and Self-rating of English Proficiency for Participants

	Age	Age of first contact	Speaking	Listening	Reading	Overall proficiency
Mean	31.15	12.55	3.23	3.66	3.29	3.36
SD	6.38	0.91	1.06	1.16	0.99	0.95

Note: *Speaking, Listening and Reading are self-rated on a seven-point scale (1 = very weak, 7 = very strong). Overall proficiency is calculated as the mean of these three categories.*

In lexical bundle studies, controls are normally made by switching a single word for another matched for frequency. For example, "*in the middle of the*" (1,513 occurrences in the BNC) might become "*in the front of the*" (47 occurrences in the BNC). This was not possible here because the lexical bundles and controls needed to have the same meaning so that they could be used in the same contextual sentences. Therefore, we first selected 17 lexical bundles with a minimum frequency of over 3 per million occurrences in the BNC. We then selected synonymous expressions, ensuring that their frequency was as low as possible in comparison with the lexical bundles. For example, the only difference between the two versions in the example above is the lexical bundle, "*I don't know why*" (frequency = 651 in the BNC) and "*I'm not sure why*" (15 occurrences in the BNC). Thus, any difference between the two can be attributed to the presence or absence of a lexical bundle.

2.3 Procedure

The passages were shown to participants in an offline naturalness judgement task. The participants saw each passage individually in the form of a survey and were asked to rate the underlined portion for how natural they thought it sounded on a seven-point scale (7 = very natural and 1 = very unnatural). The items were counterbalanced so that participants saw an equal number of lexical bundles and control items, but never both versions of the same item. It should be noted that the decision to use a seven-point scale was arbitrary and other scales would be possible, which will be taken up in the discussion of limitations of the current study.

2.4 Analysis and Results

Initially, we planned to model the data using the seven-point naturalness rating scale in its raw form; however, in order to use a simpler form of analysis, more suited to the (small) sample size (discussed below), we made the decision to collapse the scale into a binary, "High" or "Low" metric where a score of 5, 6 or 7 constituted a high level of naturalness and a score of 1, 2 or 3 constituted low naturalness. We removed scores of 4 from the analysis as they represent a middle ground, and therefore, likely represent a "do not know" decision (this resulted in a loss of 11% of the data from the Japanese speakers and 9% from the native speaker group. Table 2 provides a summary of responces.

A chi-square analysis reveals that the native English speaker group was significantly more likely to give items a "high" rating, irrespective of lexical bundlehood

Table 2. Summary of Responses

		High	Low
Japanese	Lexical bundle	114 (29%)	62 (16%)
	Control	81 (21%)	91 (23%)
Native	Lexical bundle	154 (38%)	34 (8%)
	Control	115 (28%)	70 (17%)

Note: *Percentages are calculated from the total number of responses across both conditions by participant group.*

(lexical bundles: $X^2(1) = 76.59$, $p \leq 0.001$, controls: $X^2(1) = 10.95$, $p \leq 0.001$), and that the Japanese participants were more likely to give lexical bundles a high rating ($X^2(1) = 15.36$, $p \leq 0.001$); however, there was no significant difference in the number of items they rated as high or low ($X^2(1) = 0.58$, $p = 0.45$). This indicates that the Japanese English learners are sensitive to higher frequency lexical bundles; however, in this case we would also expect them to rate control items significantly lower, which we would also expect from the native speaker group, as discussed below.

The data were further analysed using generalised linear mixed effects models (GLMER) with the package lme4 (Bates et al., 2015). The data were modelled with the collapsed naturalness metric as the response variable with condition, phrasal frequency and sequence length (calculated as the number of characters in a sequence) as fixed effects. For the Japanese learners, we also included age and age of first acquisition. Because we are primarily interested in differences in how the native speakers rate items when compared with Japanese English learners, we needed two separate phrasal frequency metrics tailored to each of the two groups. As a "native" metric we used values from the BNC. For the Japanese speakers, selecting a representative frequency metric is difficult as we know little about the relative exposure to English of each participant. What we do know, however, is that each participant learned English at school from the age of 13 years using Ministry of Education approved textbooks. Frequency scores from a corpus of these textbooks—the JHSETC (Northbrook & Conklin, 2018)—proved a robust indicator of online processing in a task with junior high school students (Northbrook & Conklin, 2019). Therefore, we used the JHSETC as an indication of participants' exposure to lexical bundles (and the controls). In order to account for idiosyncratic differences between participants (see Winter, 2014), participants were included as a random-effect factor in the model. Some predictors were correlated, and so were residualised by fitting linear models and creating a new metric consisting of the residual (i.e., information not accounted for by correlating factors, Baayen, 2008). For example, BNC phrasal frequency was correlated with JHSETC frequency, so BNC phrasal frequency was fitted as the response variable in a model with JHSETC frequency as a predictor. Other correlated factors were fitted in a similar manner. The resulting residuals all correlated significantly with their related variables (<0.001): BNC phrasal frequency ($r = 0.72$) and length ($r = 0.77$). All other continuous predictors were centred to avoid having a change in slope that might correlate with a change in intercept (see Baayen, 2008).

For both native and non-native speakers, a simple model was fit with residual JHSETC frequency, and with subject as a random effect. Complexity was then built into this model by adding new predictors in a step-by-step manner. These were also tested as adjustments for the participant random effect (as recommended by Barr et al. 2013). Each iteration of the model was tested against the previous best fit to see whether the change was significant (likelihood ratio test, $p \leq 0.05$). This process was continued until all predictors and interactions had been tested as both fixed and random effects. Table 3 summarises the best fit for both models.

Whether an item was classified as a lexical bundle or a control phrase made no difference for either native ($X^2(1) = 2.09$, $p = 0.15$) or non-native speakers ($X^2(1) = 0.29$, $p = 0.59$) indicating that lexical bundlehood cannot be reduced to

Table 3. Summary of Fixed and Random Effects for the Best-fit Model

	Native speakers					Japanese speakers				
Predictors	*b*	*CI*	*SE*	X^2	*p*	*b*	*CI*	*SE*	X^2	*p*
(Intercept)	5.37	2.22–18.54	3.47	-	0.001 **	1.19	0.00–1.69	0.04	-	0.083
BNC	1.7	1.08–1.56	0.12	8.25	0.004 **	1.13	0.90–1.25	0.09	0.49	0.48
JHSETC	0.81	0.57–1.23	0.16	0.79	0.38	1.91	1.23–2.33	0.28	10.69	0.001 **
Length	1.55	1.07–1.30	0.06	10.71	0.001 **	1.2	0.98–1.17	0.05	2.41	0.12
LB type	0.44	0.15–1.33	0.25	2.09	0.15	1.3	0.53–3.23	0.6	0.29	0.59
Age of FA	-	-	-	-	-	1.22	0.89–1.73	0.21	1.24	0.26
Age	-	-	-	-	-	1.25	0.99–1.08	0.02	1.61	0.21
Subject		Variance[1]					Variance			
		0.24					0.06			

Native Speaker Model formula: Naturalness Binary ~ Residual BNC Frequency + Residual JHSETC Frequency + Residual Length + Lexical Bundle Type + (1 | Subject). Participants = 24, Observations = 373.
Japanese Speaker Model formula: Naturalness Binary ~ Residual BNC Frequency + Residual JHSETC Frequency + Residual Length + Lexical Bundle Type + Age of First Acquisition + Age + (1 | subject). Participants = 23, Observations = 348.

a simple binary metric. However, there were significant effects based on lexical bundle frequency. The data was modelled with both native-speaker frequency values and frequency values from the JHSETC. Native speaker ratings were predicted best by the BNC phrasal frequency, and there was no significant effect of the JHSETC values (X^2 (1) = 0.79, p = 0.38). Japanese speaker ratings, however, were predicted best by the JHSETC phrasal frequency but not the BNC values (X^2 (1) = 0.49, p = 0.48). This is an important finding and will be taken up in the Discussion section. Phrasal length was a significant predictor for the native speakers, but not for the Japanese learners (X^2 (1) = 2.41, p = 0.12). Neither age (X^2 (1) = 1.61, p = 0.21) nor age of first acquisition (X^2 (1) = 1.24, p = 0.26) were significant factors for the Japanese speakers. None of the random effect adjustments improved either of the models.

3 Discussion

Although this study did not find any difference between lexical bundles and non-lexical bundles *(I don't know why* vs. *I'm not sure why*), within the category of "lexical bundle" native and non-native speakers display some interesting differences. The lexical bundles that native and non-native speakers rate as more or less natural are not the same. The native speaker group rate items that are of high frequency in the BNC as more natural (although the effect size for this is relatively small; discussed below), and there is no JHSETC frequency effect. Native English speakers have not been exposed to Japanese English textbooks, and therefore, we would not expect frequency in the JHSETC to be a significant predictor of their behaviour, which was, indeed, the case. In contrast, the Japanese learners of English are sensitive to the frequency of occurrence in their (early) English textbooks (JHSETC); that is, items with a higher frequency in the JHSETC are judged

as more natural and those with a lower frequency, less so. BNC frequency metrics did not predict the Japanese speakers' naturalness ratings.

As mentioned briefly in the Introduction section, the detailed analysis of the lexical bundles in the JHSETC from Northbrook and Conklin (2018) demonstrated that they are quantitatively and qualitatively different from those in a native-speaker corpus. The current results provide some additional evidence that the lexical bundles in the JHSETC do not represent native-speaker English; if they did, the native speakers would rate ones occurring more frequently in the JHSETC as more natural. Notably, the Japanese speakers were exposed to lan- guage from the JHSETC, on average, 15 years before the current study. The fact that we see an effect of the JHSETC, a corpus that reflects the learners' earliest experience with English, demonstrates the importance of teaching materials tar- geted at beginners. In addition, other studies focusing on secondary school text- books have found the dialogues in them to be inauthentic (Nguyen & Ishitobi, 2012; Römer, 2004, 2005), in that they do not accurately represent the situations students need to be prepared for. Based on the findings of the current study, not only is this problematic because we are not preparing our students for English use in the real world, but arguably presenting inauthentic, "unnatural" English to stu- dents is potentially also damaging in the long term as evidenced by their intuitions about the naturalness of English lexical bundles.

Native speakers rated items significantly higher than non-native speakers, irrespective of whether the item was a lexical bundle or matched control phrase. Given this, it is not surprising that lexical bundlehood was not a significant factor in the model, at least for the native speakers. However, rather than showing evidence of a binary distinction (lexical bundle vs. control), native speakers' intuitions about the naturalness of the phrases was related to their frequency in a native-speaker corpus (BNC), which is consistent with the previous study findings (e.g., Siyanova & Schmitt, 2008). It is important to note, however, that for the native speakers, the effect size for fixed effects was very small in relation to the intercept, indicating that such a conclusion should be treated with caution. This was not the case for the Japanese learners, where the intercept was not significant in the model and JHSETC frequency was the best predictor of their rating behaviour.

It is important to note that the current study was relatively small in scale and so included fewer observations than can be considered best practice for mixed effects modelling. Brysbaert and Stevens (2018) stated that a properly powered experiment with repeated measures should have at least 1,600 observations per condition. Clearly, this is considerably more than the 373 observations that this study has (for the native speaker group; the Japanese group included 348 observations). We do not believe this to be a problem given the relative simplicity of the models, however. Furthermore, we do not believe that this detracts from the core finding of the study, particularly regarding the Japanese speaker group. Results from the current study indicate that the participants are influenced by the materials they have studied, which we would expect from a usage-based approach and are in line with previous studies (discussed below). However, in light of the small effect size, notably regarding the native speaker group, subsequent research certainly needs to include a greater number of participants and experimental stimuli. Doing so would allow us to not only replicate the current findings but also build

more complexity into the statistical models, and thus, account for a wider range of potential factors that may influence intuitions of naturalness.

A further limitation of the current study is the use of the seven-point scale to rate naturalness. As mentioned previously, the original intention was to analyse the data using the raw ratings from this scale. However, doing so required a more complex form of analysis that was not appropriate given the (small) number of observations. This meant that we had to make a decision on how to use the data, which almost certainly affected the results. It is not clear whether a score of "3," for example, is really different to a score of "4" (which we took as a middle ground, and therefore, removed from the analysis). Ideally, we would have only kept the bottom and top end of the scale, removing everything between a score of 3 and 5; however, doing so would have resulted in considerable data loss. In terms of further research, this scale certainly needs further validation. In a future iteration of this study, we will either follow Edmonds (2013) and use a simple, "High," "Low" or "Can't Decide" scale, or else use a longer nine-point scale, which would have a clearer low, middle and high end.

Finally, the current study is limited in that it only focuses on a single group of non-native speakers: Japanese English learners. We chose this population because it continues on from and adds evidence to findings from previous studies in the author's research (Northbrook & Conklin, 2018, 2019). Furthermore, corpora such as the JHSETC are essential for generating phrasal frequency metrics tailored to a specific group's experience of English. In order to test the role of early input in other populations, similar corpora would be needed (which will be the topic of future research, specifically with Taiwanese secondary-school learners of English, who use very similar government-mandated textbooks to Japanese secondary-school students). However, the current research provides an indication of the lasting impact of the language materials that young learners are exposed to. In the intervening time between exposure to the textbook language and the current study, the learners will have been exposed to English in other ways (including their current English study in the e-learning platform from which they were recruited). Crucially, this additional exposure was not reflected by naturalness ratings that were in line with the native speaker ones and related to a native speaker corpus. It may be that frequency metrics from another source, say, for example, the Corpus of Contemporary American English (COCA) (Davies, 2008) or the SUBTLEXus (Brysbaert & New, 2009), may better predict Japanese speakers' rating scores because the Japanese learners may have had more exposure to American English, or indeed scores taken directly from materials used in the e-learning environment. This warrants further research, particularly to address whether very recent exposure to materials optimised with high-frequency lexical bundles (i.e., ones that could be embedded in their e-learning platform) would "overwrite" the (un-nativelike) lexical bundles that learners' have entrenched in long-term memory. This is the topic of a subsequent study by the authors (Northbrook et al., 2021).

Overall, as with Northbrook and Conklin's (2019) previous research, the results from this study align well with usage-based accounts of language acquisition, which maintain that language develops in reaction to use and exposure (Barlow & Kemmer, 2000; Bybee, 2006; Ellis et al., 2013; Tyler, 2010). Northbrook and Conklin (2019)

demonstrated a clear processing advantage on a phrasal judgement task by Japanese junior high school students for the lexical bundles that they had encountered in their textbooks. The students were sensitive to the phrasal frequency of items embedded in their early textbooks, and responded to them faster and more accurately. Such findings indicate that exposure to language underpins entrenchment in memory—even from the very earliest stages of learning. That frequency information from these textbooks predicts the naturalness ratings of *adult* Japanese learners of English suggests that the processing advantages they developed early on may be retained over time, and even in some cases—for example where there is a large discrepancy between the frequency distribution of an item in the textbooks and in naturally-occurring language—influence their judgement and prevent them from making (fully) "nativelike" evaluations of lexical bundles. These findings once again highlight the need for carefully considering the input students are given, even at the earliest stages of learning, as they may have a far-ranging influence.

4 Conclusion

This study, while small in scale, demonstrates that language learners may develop intuitions of English based on the actual English they have learned and have been exposed to, which supports a usage-based view of language acquisition. If the input learners receive is not nativelike, their *intuitions* will not be nativelike. Crucially, once formed, these intuitions may become entrenched in memory and stay with learners long term; the current findings of this study indicate an influence of English materials that the learners encountered, on average, 15 years previously. Notably, these findings highlight the need for carefully considering the input students are given and for textbook designers to create contents that reference native speaker English, rather than simply creating examples to demonstrate grammar points without consideration of the usage of the examples in which they are embedded.

Note

1. Variance (ICC) indicates between-person differences with a score ranging from 0 to 1 where a low score indicates a higher level of between-participant differences. There is a higher level of variance in rating behaviour in the Japanese participants than the native speaker group. This is expected given that the Japanese speakers were a sample of convenience and the native speakers from a single student population, as well as because we expect second language learners to have a more varied experience of English than monolingual speakers for whom English is a first language.

References

Baayen, R. H. (2008). *Analyzing linguistic data: A practical introduction to Statistics using R*. Cambridge University Press. https://doi.org/10.1017/CBO9780511801686

Barlow, M., & Kemmer, S. (2000). *Usage based models of language*. CSLI Publications.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language, 68*(3), 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1–48. https://doi.org/10.18637/jss.v067.i01

Biber, D. (2009). A corpus-driven approach to formulaic language in English: Multi-word patterns in speech and writing. *International Journal of Corpus Linguistics, 14*(3), 275–311. https://doi.org/10.1075/ijcl.14.3.08bib

Biber, D., Johansson, S., Leech, G., Conrad, S., & Finegan, E. (1999). *Longman grammar of spoken and written English*. Pearson Education Limited.

Boers, F., Eyckmans, J., Kappel, J., Stengers, H., & Demecheleer, M. (2006). Formulaic sequences and perceived oral proficiency: Putting a lexical approach to the test. *Language Teaching Research, 10*(3), 245–261. https://doi.org/10.1191/1362168806lr195oa

Bolinger, D. (1975). Meaning and memory. *Forum Linguisticum, 1*, 2–14.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*(4), 977–990. https://doi.org/10.3758/BRM.41.4.977

Brysbaert, M., & Stevens, M. (2018). Power analysis and effect size in mixed effects models: A tutorial. *Journal of Cognition*, *1*(1), 9. https://doi.org/10.5334/joc.10

Bybee, J. (1998). The emergent lexicon. *Proceedings of the Chicago Linguistic Society 34*, 421–455.

Bybee, J. (2006). From usage to grammar: The mind's response to repetition. *Language, 82*(4), 711–733. https://doi.org/10.1353/lan.2006.0186

Bybee, J. (2013). Usage-based theory and exemplar representations of constructions. In T. Hoffman & G. Trousdale (Eds.), *The Oxford handbook of construction grammar* (pp. 49–69). Oxford University Press. https://doi.org/10.1093/oxfordhb/9780195396683.001.0001

Conklin, K., & Schmitt, N. (2012). The processing of formulaic language. *Annual Review of Applied Linguistics, 32*, 45–61. https://doi.org/10.1017/S0267190512000074

Cowie, A. P. (1998). Phraselogical dictionaries: Some east-west comparisons. In A. P. Cowie (Ed.), *Phrasology: Theory, analysis and applications* (pp. 209–228). Oxford University Press. https://doi.org/10.1353/lan.2000.0046

Davies, M. (2008). The Corpus of Contemporary American English (COCA): 560 million words, 1990-present. https://corpus.byu.edu/coca/

Dickinson, P. (2012). Improving second language academic presentations with formulaic sequences. *Bulletin of Niigata University of International and Information Studies Department of Information Culture, 15*, 25–36.

Edmonds, A. (2013). Conventional expressions. *Studies in Second Language Acquisition, 36*(1), 69–99. https://doi.org/10.1017/S0272263113000557

Ellis, N. C. (2001). Memory for language. In P. J. Robinson (Ed.), *Cognition and second language acquisition* (pp. 33–68). Cambridge University Press. https://doi.org/10.1017/CBO9781139524780

Ellis, N. C. (2002a). Frequency effects in language processing: A review with implications for theories of implicit and explicit language acquisition. *Studies in Second Language Acquisition, 24*(2), 143–188. https://doi.org/10.1017/S0272263102002024

Ellis, N. C. (2002b). Reflections on frequency effects in language processing. *Studies in Second Language Acquisition, 24*(2), 297–339. https://doi.org/10.1017/S0272263102002140

Ellis, N. C. (2003). Constructions, chunking, and connectionism: The emergence of second language structure. In C. J. Doughty & M. H. Long (Eds.), *Handbook of second language acquisition* (pp. 63–103). Blackwell. https://doi.org/10.1002/9780470756492

Ellis, N. C., O'Donnell, M. B., & Römer, U. (2013). Usage-based language: Investigating the latent structures that underpin acquisition. *Language Learning, 63*(1), 25–51. https://doi.org/10.1111/j.1467-9922.2012.00736.x

Foster, P. (2001). Rules and routines: A consideration of their role in the task-based language production of native and non-native speakers. In M. Bygate, P. Skehan, & M. Swain (Eds.), *Researching pedagogic tasks: Second language learning, teaching and testing* (pp. 75–93). Longman.

Gilmore, A. (2004). A comparison of textbook and authentic interactions. *English Language Teaching Journal, 59*(4), 363–374. https://doi.org/10.1093/elt/58.4.363

Howarth, P. (1998). The phraseology of learners' academic writing. In A. P. Cowie (Ed.), *Phrasology: Theory, analysis and applications* (pp. 161–186). Oxford University Press.

Hymes, D. (1972). On communicative competence. In J. B. Pride & J. Holmes (Eds.), *Sociolinguistics* (pp. 269–293). Penguin.

Meunier, F. (2012). Formulaic language and language teaching. *Annual Review of Applied Linguistics, 32*(2), 111–129. https://doi.org/10.1017/S0267190512000128

Myles, F., Hooper, J., & Mitchell, R. (1998). Rote or rule? Exploring the role of formulaic language in classroom foreign language learning. *Language Learning, 48*(3), 323–363. https://doi.org/10.1111/0023-8333.00045

Myles, F., Mitchell, R., & Hooper, J. (1999). Interrogative chunks in French L2: A basis for creative construction? *Studies in Second Language Acquisition, 21*(1), 40–80. https://doi.org/10.1017/S0272263199001023

Nattinger, J. R., & DeCarrucio, J. S. (1992). *Lexical phrases and language teaching*. Oxford University Press. https://doi.org/10.1017/S0272263100013000

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics, 24*(2), 223–242. https://doi.org/10.1093/applin/24.2.223

Nguyen, H. T., & Ishitobi, N. (2012). Ordering fast food: Service encounters in real-life interaction and in textbook dialogs. *Japan Association for Language Teaching Journal, 34*(2), 151–186. https://doi.org/10.37546/JALTJJ34.2-2

Northbrook, J., Allen, D., & Conklin, K. (2021). "Did you see that?"—The role of repetition and enhancement on lexical bundle processing in English learning materials. *Applied Linguistics.* https://doi.org/10.1093/applin/amab063

Northbrook, J., & Conklin, K. (2018). "What are you talking about?" An analysis of lexical bundles in Japanese junior high school textbooks. *International Journal of Corpus Linguistics, 23*(3), 311–334. https://doi.org/10.1075/ijcl.16024.nor

Northbrook, J., & Conklin, K. (2019). Is what you put in what you get out?—Textbook-derived lexical bundle processing in beginner English learners. *Applied Linguistics, 40*(5), 816–833. https://doi.org/10.1093/applin/amy027

Nunan, D. (1991). *Language teaching methodology: A textbook for teachers*. Prentice Hall.

Pawley, A., & Syder, F. H. (1983). Two puzzles for linguistic theory: Nativelike selection and nativelike fluency. In J. C. Richards (Ed.), *Language and communication* (pp. 191–225). Longman.

Römer, U. (2004). Comparing real and ideal language learner input: The use of an EFL textbook corpus in corpus linguistics and language teaching. In G. Aston, S. Bernardini, & D. Stewart (Eds.), *Corpora and language learners* (pp. 151–168). John Benjamins. https://doi.org/10.1075/scl.17

Römer, U. (2005). *Progressives, patterns, pedagogy: A corpus-driven approach to English progressive forms, functions, contexts and didactics*. John Benjamins Publishing Company. https://doi.org/10.1075/scl.18

Sinclair, J. (1991). *Corpus, concordance and collocation*. Oxford University Press.

Siyanova, A., & Schmitt, N. (2008). L2 learner production and processing of collocation: A multi-study perspective. *The Canadian Modern Language Review, 64*(3), 429–458. https://doi.org/10.3138/cmlr.64.3.429

Siyanova-Chanturia, A., & Spina, S. (2015). Investigation of native speaker and second language learner intuition of collocation frequency. *Language Learning, 65*(3), 533–562. https://doi.org/10.1111/lang.12125

Skehan, P. (1998). *A cognitive approach to language learning*. Oxford University Press.

Stubbs, M., & Barth, I. (2003). Using recurrent phrases as text-type discriminators: A quantitative method and some findings. *Functions of Language, 10*(1), 61–104. https://doi.org/10.1075/fol.10.1.04stu

Taguchi, N. (2007). Chunk learning and the development of spoken discourse in a Japanese as a foreign language classroom. *Language Teaching Research, 11*(4), 433–457. https://doi.org/10.1177/1362168807080962

Tyler, A. (2010). Usage-based approaches to language and their applications to second language learning. *Annual Review of Applied Linguistics, 30*, 270–291. https://doi.org/10.1017/S0267190510000140

Winter, B. (2014). *A very basic tutorial for performing linear mixed effects analyses (Tutorial 2)*. http://www.bodowinter.com/tutorial/bw_LME_tutorial2.pdf

Wray, A. (2002). *Formulaic language and the lexicon*. Cambridge University Press. https://doi.org/10.1017/CBO9780511519772

Wray, A., & Fitzpatrick, T. (2008). Why can't you just leave it alone? Deviations from memorized language as a gauge of nativelike competence. In F. Meunier & S. Granger (Eds.), *Phraseology in foreign language learning and teaching* (pp. 123–147). John Benjamins. https://doi.org/10.1075/z.138

Vocabulary Learning and Instruction
Volume 10, Issue 1, December 2021
http://vli-journal.org

WordUp: An App that Teaches English Words through Extensive Exposure to Authentic Materials

Musa Nushi[a], Alireza Aghaei[a], and Maryam Roshanbin[b]
[a]*Shahid Beheshti University;* [b]*Iran University of Science and Technology*

Abstract

This paper reviews WordUp, a mobile application which fosters English vocabulary learning through exposure to new words in authentic and engaging contexts such as excerpts of movies, songs, and news programme. The samples of use are introduced after the definition of the target words have been provided in both the learners' first language and English. In this review, the advantage of extensive exposure to authentic use of the target words will be discussed. Moreover, some of the app's shortcomings such as its unsuitability for beginner learners and insufficient tasks to improve the users' production of the learned vocabulary will be highlighted. WordUp is recommended for learners above level A1 The Common European Framework of Reference for Languages (CEFR) who wish to learn the appropriate application of the learned material through authentic input.

Keywords: Applications, mobile learning, second language, technology, vocabulary, WordUp

1 Introduction

Developing the depth and breadth of learners' vocabulary has always been high on second language (L2) teachers' agenda (Nushi & Jenabzadeh, 2016). This concern is understandable given the importance of vocabulary in language learning. McCarthy (1990, p. 8) noted that "no matter how well the student learns grammar, no matter how successfully the sounds of L2 are mastered, without words to express a wide range of meanings, communication in an L2 just cannot happen in any meaningful way." Lewis (1993) goes even further, contending that "lexis is the core or heart of language" (p. 89). Learning vocabulary in another language is not without its own challenges and numerous studies have been conducted to facilitate this process, mainly through equipping learners with various vocabulary learning and retention strategies (Fan, 2020; Mizumoto & Takeuchi, 2009; Moir & Nation, 2002; Taka, 2008). Such an extensive scholarly enterprise seems warranted given the fact that acquiring vocabulary knowledge is a multi-faceted, experience-based process. Acknowledging that fact, Zhang (2020, p. 1) stated that "lexical knowledge is a multi-faceted and complex construct with multiple cognitive and linguistic skills involved." Nation (2013) argued that knowing a word includes knowing its *form*, *meaning* and *use*. More specifically, form refers to knowledge of the written and spoken representation of a word. Meaning entails

Please cite this article as: Nushi, M., Aghaei, A., & Roshanbin, M. (2021). Wordup: An app that teaches English words through extensive exposure to authentic materials *Vocabulary Learning and Instruction, 10*(1), 64–72. https://doi.org/10.7820/vli.v10.1.nushi

knowledge of the denotative (i.e., literal or conceptual) and connotative (i.e., figurative or associative) meanings of a word. Finally, use refers to the understanding of how a word should be used in linguistic and social contexts. Schmitt (2000) added that words can have different meanings in different contexts and lexical instruction needs to be comprehensive.

Teaching vocabulary, in comparison to other components of language instruction, has not received the attention it merits (Block & Fitzgerald, 2006; Cunningham, 2009). Research has also demonstrated that conventional vocabulary activities such as looking up the words and memorising their definition do not effectively help vocabulary learning (Cunningham, 2009). Advancements in technology, however, have provided L2 teachers and learners with many opportunities which can effectively increase the students' vocabulary knowledge (Lin & Lin, 2019). Yet, there is still a need for more research regarding the use of these technological innovations to extend and enhance vocabulary teaching and learning practices. One of the ways that technology can help vocabulary instruction is through exposing learners to large amounts of authentic reading materials. Nunan (1988, p. 99) defined authentic materials as those "which have been produced for purposes other than to teach language," In other words, authentic materials involve the natural use of spoken or written language for communication in the native speaker context. Although researchers have not reached a consensus regarding the use of authentic materials (e.g., YouTube videos, TV series and programmes, songs, movies, books, magazines, podcasts) for L2 teaching and learning purposes (see Gilmore, 2007 for a review), there are many studies showing that such materials can prove beneficial to L2 learners by not only exposing them to real-life examples of language use and thus preparing them for real world communication, but also raising their motivation to take part in classroom activities (Namaziandost et al., 2021; Peacock, 1997).

The present article reviews a mobile application that aims to teach leaners new English words and improve their vocabulary in the language through a variety of authentic examples of spoken English for each word from movie clips, songs and quotes. The review explores the possibilities the app can offer for learning English words, and provides a number of suggestions that the app designers can use to improve the app and make learning vocabulary a more fruitful experience for learners.

2 Description

WordUp can be downloaded from http://www.wordupapp.co, which redirects the users to App Store for iOS devices, Google Play for Android devices, and Microsoft for Windows devices. This application starts by asking the learners about their age, approximate language proficiency level, and preferred accent (Figure 1a–c).

Moreover, it gives the learners a placement test which indicates which position on the 25,000-word list best fits their vocabulary knowledge (Figure 2a). After they have taken the test and set a reminder for their daily vocabulary exercise (Figure 2b), every day they will take a similar test to discover more words, among the 1,000 words appropriate for their level, to learn. According to their usefulness,

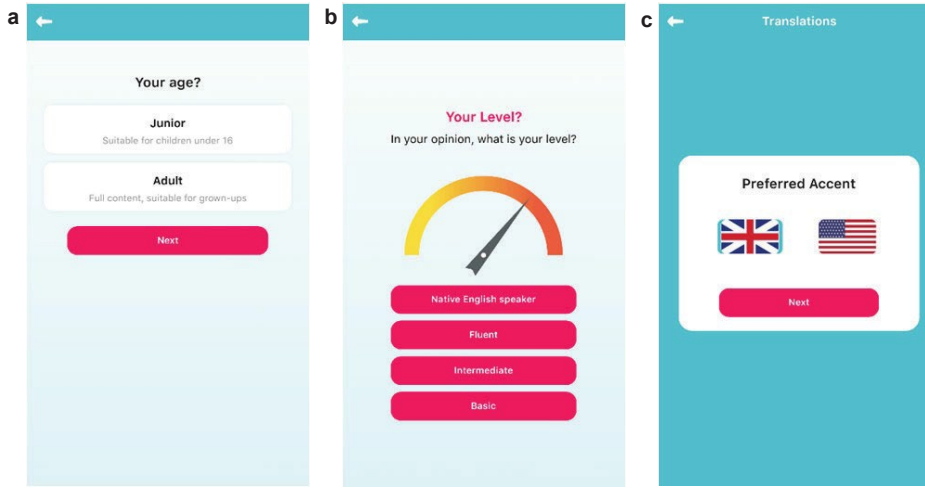


Figure 1. (a–c) The First Three Stages WordUp Users Go Through to Adjust the Application to Their Needs.

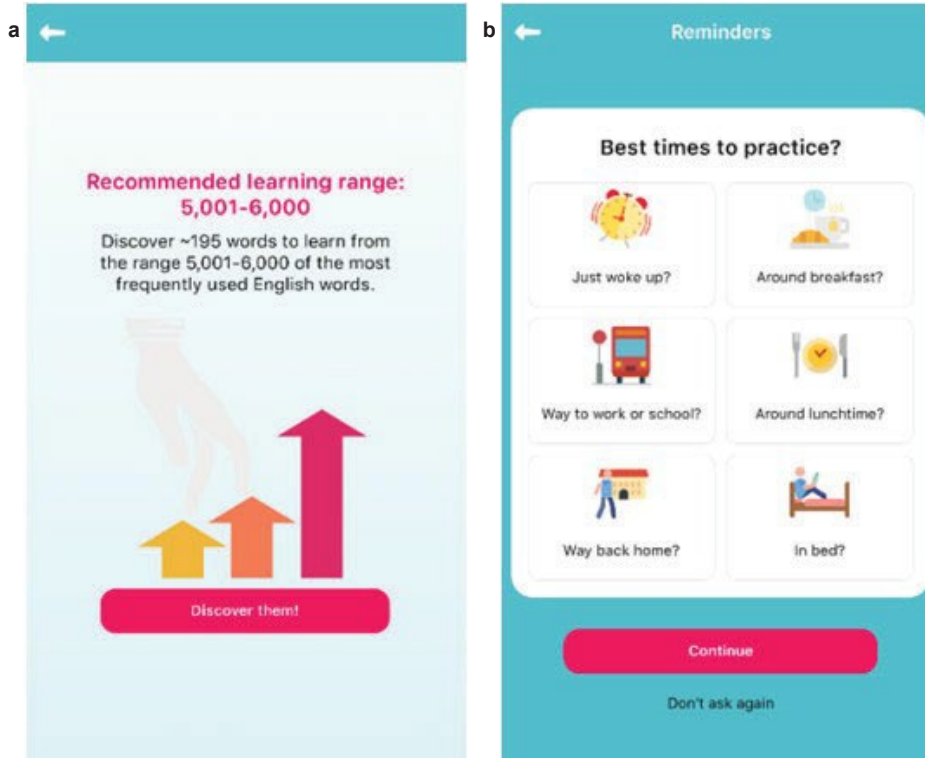


Figure 2. (a, b) Determining the Users' Appropriate Level and Reminding Schedule.

the words are categorised into 25 groups of 1,000 words, ranked from the first most useful and frequently used thousand words to the 25th thousand one. Facing a new word, the users will see three options that WordUp offers (Figure 3a). They can confirm that they know the word, confirm that they do not know the word and would like to learn it, or ask the application to test them if they are not sure about the word's meaning. When the learners start learning a word, a dictionary entry containing the word's definition, pronunciation, part of speech, first language (L1) equivalence, and some examples appear (Figure 3b).

What makes WordUp special, however, is the extensive list of authentic examples which are found using search engines and artificial intelligence technologies. This list is shown after the dictionary entry and the users are able to find and include quotations, excerpts of TV series, movies, news programmes and songs in these lists (Figure 4). The app developers claim that they "use a combination of well-grounded sources such as Oxford" to create their own list (WordUp Team, personal communication, September 7, 2021). This extensive list of various authentic examples will help the learners engage in the process of vocabulary learning and find opportunities to enjoy learning new words while they read, watch, and listen to interesting texts and other forms of media. The online corpora can also help learners become independent learners (see Giampieri, 2019;

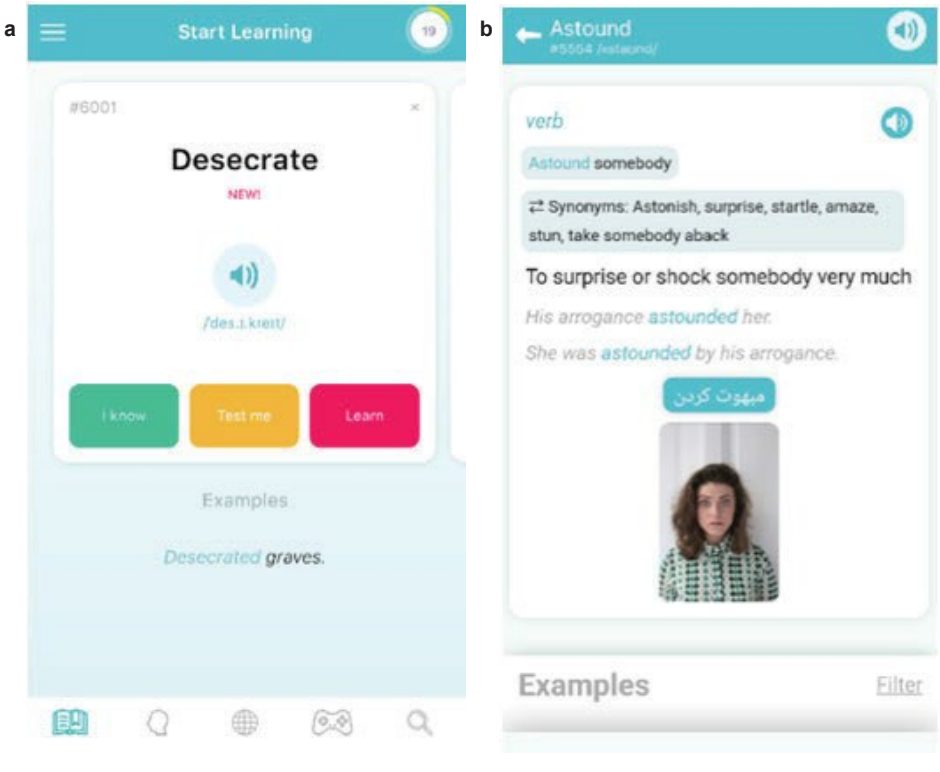


Figure 3. (a, b) Three Options Offered to the Users When They Face a Word and the Display of Each Word's Dictionary Entry.

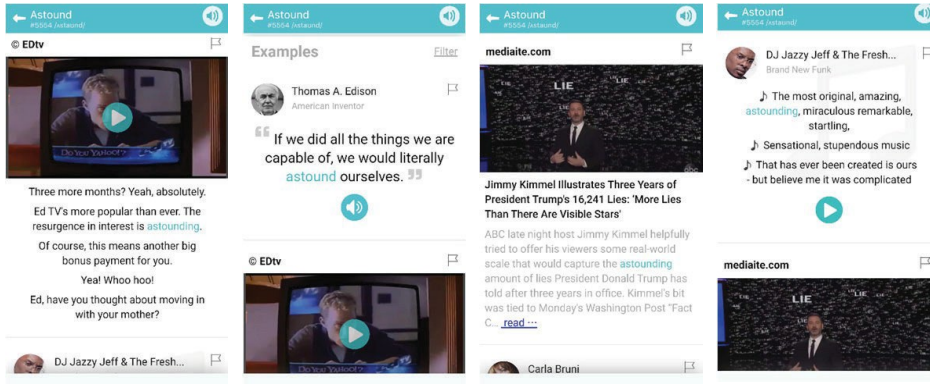


Figure 4. Authentic Examples of the Words Used in Several Contexts.

Mills, 2000; and Römer, 2011 for the use of Web-based corpora for second language teaching and learning). It also seems that WordUp uses spaced repetition to make sure users remember words in the long-term. Spaced repetition is an effective vocabulary learning technique that involves reviewing and recalling previously learned material at gradually increasing intervals until the information is satisfactorily learned (Nakata, 2015).

After the users have worked on the word and its authentic use in different contexts, WordUp will ask them the meaning of that word the next day. Provided that they answer correctly, it will ask them whether they remember that word after 3 days, a week, a month, 3 months, and a year. Whenever the users forget the word during the process, the reminding sequence will start all over again.

Although users can access many options freely, WordUp has some in-app purchase options that give students access to even more. There are two types of accounts which can be purchased, namely "essential" and "pro." The essential account, costing $11.99 per year or $39 for an unlimited period of time, cancels the advertisements (ads) which frequently interrupt the lessons. The pro subscription, however, gives users access to several more options. Costing $59 per a year and $99 for a lifetime, the pro account not only removes all the ads, but also enables access to vocabulary games, categorised words based on their topics, exam-specific words, idioms and phrases, and limitless examples.

3 Evaluation

WordUp has the potential to make a great contribution to the students' vocabulary learning and introduces an interesting and mesmerising way to the users to enhance their vocabulary size without the need of teachers and classrooms; hence, one can conclude that WordUp is capable of fostering learner autonomy as well. However, this is no more than a claim and research should indicate its correctness. Further research can also aim at answering these questions: Does WordUp result in enhancing the lexical knowledge of students at all levels? Do learners make use of all the options provided by WordUp or suffice to the L1 equivalences? How well

can the students use the learned words in their productive skills? How can teachers take advantage of such technology to students' benefit? What are students and teachers' attitude toward using this app? Does such technology align with flipped classroom principals which will lead into the new classroom methodologies such as the Dogme ELT approach (Thornbury, 2000)?

WordUp has a significant affordance for English as a foreign language (EFL) speakers all over the world since it provides the learners with not only the English definition of the words and authentic instances of their use in English, but also the words' translation in the students' L1. WordUp supports a wide range of languages as the students' L1; thus, learners with lower levels of language proficiency can benefit from the equivalences of the vocabulary in their L1, if needed. Another positive feature of this app is the variety in the authentic examples it presents, which turns learning into an amusing experience for learners with different tastes and learning styles.

Although the advantages in using WordUp abound, there are some drawbacks to this app as well. It was mentioned earlier that WordUp supplies the learners with the translation of the new words. The L1 equivalences, however, seem to be borrowed from online dictionaries, and are not level-appropriate for learners with different levels of language proficiency. Moreover, some words have more than one part of speech and this app does not provide the users with L1 equivalences for more than one part of speech in most cases. The same problem occurs for words with different meanings. Besides, WordUp mentions idioms in which the words are used, while no L1 equivalence for those idioms are provided. Another noteworthy criticism is against the way WordUp chooses examples to be presented to the app users; thus, there are some issues that need to be addressed by the app developers to make the examples more efficient. Firstly, some quotations in which the words are mentioned are difficult to make sense of without the context in which they were originally used. Furthermore, some quotations are in the form of dependent clauses and the rest of the utterance seems to be missing, which makes understanding the examples quite difficult for the students. Secondly, the excerpts containing the to-be-learned word might include vocabularies that are unlikely to exist in the learners' lexical knowledge; hence, the users might find it difficult to comprehend the texts and ignore those examples. This lack of curation for the language level required to understand the examples is ubiquitous in WordUp. In addition, every word has a meaning sense, regarding its connotation and denotation, which is of great importance when it comes to learning and using that word. WordUp seems not to have taken this fact into account and the examples provided after each word do not specify the meaning sense of the words. This lack or precaution has led to the absence of sufficient examples for different meanings and parts of speech of one word. Finally, since WordUp uses online sources to find examples and instances of the words, users who do not have constant network connection will be deprived of exposure to the authentic use of the words.

As mentioned earlier, one of the most significant features of WordUp is supplying the learners with authentic language. Literature ascertains that exposure to authentic language input is highly advantageous for language learners (Brown, 2014). Moreover, the principles of Communicative Language Teaching highlight the role of authentic material in language learning (Chambers, 1997). The appealing examples

of authentic use of the newly learned words in video clips, songs, TV shows, etc. can also make vocabulary learning an entertaining and enjoyable process. Hence, the original content provided in WordUp can promote vocabulary learning. Additionally, WordUp is programmed to repeatedly expose students to the words they have learned. The frequency of this repetition is determined by the number of times that the users answer correctly to questions regarding the meaning of the previously learned vocabulary. Nation (2001) stated that the repetition of words through the reminding process is proven to facilitate users' memorisation of the words up to a great extent. Unfortunately, this application does not provide learners with activities in which the learners can use the words and make sentences or do tasks using the words. In other words, we assume there is a need for more practices and exercises through which the users can utilise and work with the learned material.

In order to make a better judgement regarding the efficiency of WordUp, a group of EFL learners, including five at elementary level and seven at intermediate level, were asked to use this application over a 2-week period and report their opinion about their experience to the researchers. In return for their participation, the researchers provided the learners with free instruction on their language problems during the 2 weeks. Seven students (three elementary and six intermediate) commented that WordUp can successfully engage learners in the process of vocabulary learning because of the authentic and in some cases "entertaining" examples. Furthermore, almost all of the learners believed that the dictionary entries presented in the app made it easier for them to use the app. On that issue, one of the learners (intermediate) stated:

> I do not like working with apps that do not provide me with the meaning of new words because then I have to leave the app and look up a word in some other dictionary before I can proceed their vocabulary lesson.

Nonetheless, some students, three of the elementary level students and one at intermediate level, found the examples difficult to understand since they contained words which were not familiar for the users.

4 Conclusion

WordUp intends to foster its users' English vocabulary learning through extensive exposure to new words being used in authentic contexts. This application places the learners at an appropriate level and presents them with some new words from its 25,000 of most frequently used English words, which are grouped to the first most useful thousand words, to the 25th mostly used thousand words. The most significant features of this application are its thorough list of authentic examples and efficient method of reminding the words to the learners. However, more engaging practices and exercises which gives the users opportunities to apply the words in sentences uttered by themselves could help with the learning process and avoid mere storing of the words as the learners' passive knowledge (Ur, 2012). In addition, a topic-based categorisation of the words could enhance the efficiency of this application. All in all, WordUp can be a useful tool for English learners at all levels above A1 (CEFR) to work on their vocabulary knowledge outside of the class and get familiar with different uses of the words in various authentic

contexts. In order to maximise the effect of this application, teachers can introduce further practices on using the words by the students through their productive skills. In the absence of a teacher, the least the learners can do is writing sentences and texts containing the words they acquire on WordUp every day.

References

Block, C. C., & Fitzgerald, J. J. (2006). The new look for research-based best practice. In C. C. Block & J. N. Mangieri (Eds.), *The vocabulary-enriched classroom: Practices for improving the reading performance of all students in grades 3 and up* (pp. 8–17). Scholastic.

Brown, H. D. (2014). *Principles of language learning and teaching* (6th ed.). Pearson Education.

Chambers, F. (1997). What do we mean by fluency? *System, 25*(4), 535–544. https://doi.org/10.1016/S0346-251X(97)00046-8

Cunningham, P. M. (2009). *What really matters in vocabulary: Research-based practices across the curriculum.* Pearson.

Fan, N. (2020). Strategy use in second language vocabulary learning and its relationships with the breadth and depth of vocabulary knowledge: A structural equation modeling study. *Frontiers in Psychology, 11*(752), 1–14. https://doi.org/10.3389/fpsyg.2020.00752

Giampieri, P. (2019). The web as corpus in ESL classes: A case study. *International Journal of Language Studies, 13*(2), 91–108.

Gilmore, A. (2007). Authentic materials and authenticity in foreign language learning. *Language Teaching, 40*(2), 97–118. https://doi.org/10.1017/S0261444807004144

Lewis, M. (1993). *The lexical approach: The state of ELT and a way forward.* Language Teaching Publications.

Lin, J. J., & Lin, H. (2019). Mobile-assisted ESL/EFL vocabulary learning: A systematic review and meta-analysis. *Computer Assisted Language Learning*, *32*(8), 878–919. https://doi.org/10.1080/09588221.2018.1541359

Mills, D. (2000). Web-based technology as a resource for form-focused language learning. *TESOL Quarterly*, *34*(3), 603–615. https://doi.org/10.2307/3587749

Mizumoto, A., & Takeuchi, O. (2009). Examining the effectiveness of explicit instruction of vocabulary learning strategies with Japanese EFL university students. *Language Teaching Research*, *13*(4), 425–449. https://doi.org/10.1177/1362168809341511

McCarthy, M. (1990). *Language teaching: A scheme for teacher education*. Oxford University Press.

Moir, J., & Nation, I. S. P. (2002). Learners' use of strategies for effective vocabulary learning. *Prospect Journal*, *17*(1), 15–35.

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*(4), 677–711. https://doi.org/10.1017/S0272263114000825

Namaziandost, E., Razmi, M. H., Ahmad Tilwani, S., & Pourhosein Gilakjani, A. (2021). The impact of authentic materials on reading comprehension, motivation, and anxiety among Iranian male EFL learners. *Reading & Writing Quarterly, 38*, 1–38. https://doi.org/10.1080/10573569.2021.1892001

Nation, I. S. P. (2001). *Learning vocabulary in another language*. Cambridge University Press. https://doi.org/10.1017/CBO9781139524759

Nation, I. S. P. (2013). *Learning vocabulary in another language* (2nd ed.). Cambridge University Press.

Nunan, D. (1988). *The learner-centered curriculum*. Cambridge University Press.

Nushi, M., & Jenabzadeh, H. (2016). Teaching and learning academic vocabulary. *California Linguistic Notes*, *40*(2), 51–70.

Peacock, M. (1997). The effect of authentic materials on the motivation of EFL learners. *ELT Journal, 51*(2), 144–156. https://doi.org/10.1093/elt/51.2.144

Römer, U. (2011). Corpus research applications in second language teaching. *Annual Review of Applied Linguistics, 31*, 205–225. https://doi.org/10.1017/S0267190511000055

Schmitt, N. (2000). *Vocabulary in language teaching.* Cambridge University Press.

Taka, V. (2008). *Vocabulary learning strategies and foreign language acquisition*. Multilingual Matters. https://doi.org/10.21832/9781847690401

Thornbury, S. (2000). A dogma for EFL. *IATEFL Issues, 153*(2), 59–67.

Ur, P. (2012). *A course in English language teaching*. Cambridge University Press.

Zhang, H. (2020). Testing reciprocity between lexical knowledge and reading comprehension among Chinese children: A cross-lagged panel analysis. *European Journal of Psychology of Education*, *35*(2), 1–20. https://doi.org/10.1007/s10212-020-00462-9