A Study Into the Effectiveness of WordEngine

—Does it Help Learners Transfer Vocabulary to Long-Term Memory?—

Stephen Clarke　Jaime Morrish

Abstract

WordEngine is a software that provides a digital, gamified version of bilingual flashcards. The developers claim that through repeated meetings with level-appropriate vocabulary items at spaced intervals through online quizzes, learners can fully acquire L2 words over the long term. This study examines the extent to which learners developed the ability to recognise and recall the target L2 words they had studied via the software and its effects on overall proficiency. The subjects were college English majors and spent between 45 to 60 minutes using the software on their smartphones per week. Test instruments were administered five and eight months after the subjects had correctly selected each word six consecutive times on WordEngine and measured active recognition and active recall (Laufer and Goldstein, 2004). Results showed high scores on the proprietary tests of active recognition, but subjects scored only around 33% on tests of active recall. After discontinuing WordEngine usage, average TOEIC and VELC scores in the same department as the test subjects were not negatively affected. These results question the efficacy of WordEngine as a vocabulary learning tool that promotes the significant long-term acquisition of both the active recognition and active recall types of vocabulary knowledge.

The ultimate goal of second language learning is to be able to use the new language in any given setting; however, to be able to do so, a learner must have the necessary tools. Wilkins (1972) succinctly writes that without grammar, very little meaning can be conveyed, but without vocabulary, no meaning can be conveyed. Indeed, a growing vocabulary in the language being studied (L2) supports the development of all four language skills. As Nation (2013, p. 45) notes, “Vocabulary is not an end in itself. A rich vocabulary makes the skills of listening, speaking, reading, and writing easier to perform.”

Moreover, Chukharev-Hudilainen and Klepikova (2017, p. 334) state that the importance of increasing students’ L2 vocabulary is a process that “cannot be overestimated.” Fully understanding vocabulary and knowing words in such a manner as both to recall them receptively and use them productively is a considerably long process that takes decades, even for native speakers (Nation, 2013). Furthermore, as our learners are not living in the target language country, the opportunities to absorb English through incidental learning are minimal. Due to this fact, studying L2 vocabulary is often a lengthy and laborious process, and the relatively short periods of time spent in EFL classrooms also increase the burden of this task. We,

therefore, felt it imperative to find new and innovative ways to increase the English vocabulary size of our students. In addition, we were looking for a L2 vocabulary study method that could be implemented outside class to maximise the time learners spent in contact with English, their peers and the teacher.

In the present study, we investigated WordEngine, which is a web-based software designed to be used on mobile devices such as smartphones, tablets and even personal computers. A significant motivation for choosing WordEngine was that its maker claims that vocabulary becomes “fully acquired” after users provide six consecutive, correct responses within 90 days and that long-term memory is then assured (Lexxica, n.d.). This is a bold and attractive claim, and we will return to this topic later in our discussion.

To use WordEngine, users must access the website through their browser on their chosen device; from there, the user then interacts with a digital flashcard system. However, before the flashcard study can start, first-time users are tasked with taking a short vocabulary test called V-Check. This diagnostic tool is used to measure the approximate lexical competence of the user. After completing the V-Check, WordEngine has obtained an approximation of how many words the student knows. With this completed, learners can then start studying new words and the order in which they appear is based on frequency, with high-frequency words coming first. In this way, WordEngine is tailored to the vocabulary which learners do or do not know, according to the V-Check (Cihi, 2018). Users can choose words from a “General English” programme or other types of vocabulary, such as TOEIC or TOEFL.

The process by which WordEngine aims to increase the user’s vocabulary is based on the flashcard method; the user is shown the word in Japanese, and then there are three choices in English (L2) from which to choose the correct translation. There is also a timer so that students do not have enough time to look up the word, which is an action that would obviously skew the results. One flashcard question can be seen in Figure 1, which shows a screenshot of what the user actually sees when using the software, including the timer bar, which reduces from full to empty over the course of ten seconds. When the correct word is chosen, the user hears an audio recording of the correct word.

Figure 1. A screenshot of a WordEngine flashcard question

Figure 2 shows a visualisation of a word's progress through the WordEngine spaced repetition system. Words are considered to move through a series of boxes, depending on the number of correct reviews they have obtained. When a word is entirely new to the learner, it starts on the far left, and after its first correct response, it moves into box one. After subsequent correct reviews, the word will gradually move to the right. If the learner chooses the wrong answer, that word is sent back to the beginning of the series of boxes and is regarded as a new word again. After a total of six consecutive correct responses or five correct reviews, this word is removed from the study area as WordEngine classifies it as being "fully acquired" (Lexxica, n.d.).

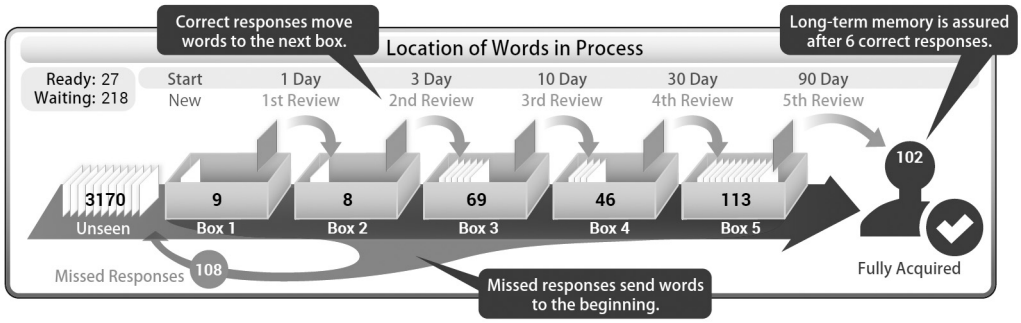

Figure 2. A visualisation of WordEngine's spaced repetition process (Lexxica, n.d.)

Literature Review

There are various methods to try and increase learners' L2 vocabulary, and recently the adoption of various digital methods has been explored (Li & Hafner, 2021). Gordania (2012) examined the use of digital corpora for this purpose, whereas Ranalli (2013) investigated web-based dictionaries. Stockwell (2007), on the other hand, advocated the design of smartphone applications with various vocabulary activities included and Nakata (2008, 2011) and Zhu, Fung, and Wang (2012) researched the possibilities of spaced repetition programmes in a digital flashcard-based approach. Online applications have been shown to help increase learners' motivation and engagement, thus developing the effectiveness of learners' language acquisition. Fageeh (2013) conducted a study on the effectiveness of how smartphone applications (apps) can enhance vocabulary learning and motivation to acquire vocabulary, and the results showed that students who used apps acquired more vocabulary than those who did not. Moreover, subjects who used WhatsApp, online dictionaries, and text messaging reporting increased motivation to learn vocabulary (Fageeh, 2013).

We will next discuss different features of vocabulary knowledge because we need to be clear about what this study is attempting to measure. The broad distinction between receptive and productive aspects of vocabulary knowledge is frequently made in the literature (e.g., Fitzpatrick, Al-Qarni & Meara, 2008; Lee & Muncie, 2006; Schmitt, 2014). Receptive knowledge can be described as the ability to understand words in reading and listening, whereas productive knowledge is the ability needed when producing words in

writing and speaking (Schmitt, 2010). Unsurprisingly, research has generally shown that learners can demonstrate more receptive than productive knowledge (e.g., Laufer & Paribakht, 1998). Nakata (2011) states that a common feature of flashcard software such as WordEngine is that it utilises receptive recognition. That is, the word is seen on the user's screen, and then the learner must choose the correct answer from a choice of several items. There is no productive element to the WordEngine software if one defines 'productive' as requiring writing or speaking. However, the user has to select the correct L2 word after being given a prompt in L1, and this arguably requires a somewhat different type of knowledge than would be employed if the prompt was in the L2 and the user had to provide the correct L1 translation. The depth of knowledge required for providing the L2 following an L1 prompt is undoubtedly more profound than the opposite type of translation. Moreover, the knowledge required to produce vocabulary when a learner is focused on meaning to create a spoken or written utterance differs from that required to provide translations of single words in the absence of any communicative context. Thus, the concepts of receptive and productive knowledge are problematic and may not sufficiently describe the kind of knowledge promoted by WordEngine usage or attempts to measure it.

Laufer and Goldstein (2004) hypothesised a hierarchy of different kinds of vocabulary knowledge with four levels, and subsequent investigation led them to claim that the hierarchy was valid at all levels of word frequency. The four levels rest on two distinctions, which they describe as follows:

> The first distinction implies that there is a difference in knowledge between people who can supply an L2 word and those who can only supply the meaning when the L2 word is presented to them. The second distinction implies that there is a difference in knowledge between those who can recall the form or the meaning of a word and those who cannot recall but can recognize the form or the meaning in a set of options. (Laufer & Goldstein, 2004, p. 406)

Laufer and Goldstein helpfully describe and illustrate the four degrees of vocabulary knowledge by referring to a bilingual test. The strongest level of knowledge is called "active recall," and it involves a learner being able to provide the L2 target item after being given a prompt that is the L1 translation of the target word (Laufer & Goldstein, 2004, p. 406). The second strongest is called "active recognition," and this involves a learner choosing the target L2 word from a series of options after being given the L1 prompt (Laufer & Goldstein, p. 406). The two weaker strengths of knowledge are called "passive recall" and "passive recognition" and mirror the above distinctions, except that they involve prompts that are in the L2 and require translation into the L1 (Laufer & Goldstein, p. 406). Therefore, the type of knowledge required to use the WordEngine software is active recognition. Laufer and Goldstein's (2004) types of vocabulary knowledge usefully differentiate between the knowledge required to answer questions in various types of vocabulary tests correctly, and they avoid many of the problems that are attached to the terms of receptive and productive knowledge. As a result, we will adopt their terminology in this paper.

Next, we turn to previous studies that have investigated WordEngine's efficacy and discuss their methodology and results. Three previous studies are relevant. Agawa, Black and Herriman (2011) compared increases in the average TOEIC scores of an English language programme over eleven years,

during the final year of which students had used WordEngine for eight weeks, studying TOEIC words. The increase during this final year was double the mean average but only the second-highest overall during the eleven years. Time on task (WordEngine usage) was compared with increases in individual subjects' TOEIC scores, but no correlation was found. The authors concluded that "for achieving success WordEngine will not, on its own, be a panacea" (Agawa, Black and Herriman, 2011, p. 197).

Phillips (2011) encouraged subjects to study the TOEIC program on WordEngine for 15 minutes per day over 11 weeks. TOEIC score increases on the post-test eight months following the pre-test generally correlated with time on task, but there were several limitations to the study which urge caution when judging the results. The reliability of the time-on-task data was questionable, and the subjects who used WordEngine for a greater time may have been more highly motivated than those who did not, meaning that WordEngine use alone may not have been the correlating factor.

McClean, Hogg and Rush (2013) attempted to measure the effect of WordEngine usage on the overall vocabulary size of their subjects. They used as their instrument a shortened version of the Vocabulary Size Test (VST), a multiple-choice test created by Nation and Beglar (2007). McClean, Hogg and Rush's (2013) shortened version comprised 80 items, with ten items from each thousand of the first eight-thousand word families found in the British National Corpus list. Thus, ten items represent the vocabulary knowledge of 1,000 words. Each test item contained a short non-defining sentence with a target word. Subjects had to select the item which could be substituted for the target word from four choices. Thus, translation from L1 was not required. Subjects in the two treatment groups used WordEngine for either one or two hours per week for 28 weeks (two semesters), and results showed that they outperformed a control group that did not use WordEngine. Although there were several possible confounding factors, McClean, Hogg and Rush (2013, p. 92) reported that "it is plausible that a significant portion of the increased gains in VST scores were due to the use of WordEngine." However, their results did not clearly show that two hours of WordEngine usage per week led to more significant gains than only one hour of usage. Overall, the study by McClean, Hogg and Rush (2013) lends some support to the efficacy of WordEngine in promoting vocabulary knowledge. However, its limitations include concerns over the sensitivity of the measuring instrument (using only ten items to represent 1,000 words), uncertainty over the level of vocabulary that subjects in the control group were exposed to and their time on task, as well as the fact that no delayed post-test was conducted.

From the description of the studies above, it is apparent that no study has yet measured long-term retention of the actual words studied via the WordEngine digital flashcard software. In addition, no study has employed translation of L1 words into L2, which is a feature of the flashcard study programme. Thus, we felt there was a gap in the literature for a study that measures the delayed effects of WordEngine, and which investigates active recognition of the actual words learned through the software. In addition, we were also curious about the extent to which WordEngine usage promoted the active recall of vocabulary learned through WordEngine. Webb (2005, in McClean, Hogg & Rush, 2013), for example, found that receptive learning resulted in both receptive and productive word knowledge and that receptive tasks were

superior to productive tasks when time on task was accounted for. Although we are not using this terminology in the present study, this finding illustrates that different kinds of word knowledge can be developed even when learners do not need to use the target words in speaking or writing. Finally, it would be worthwhile to use something other than the TOEIC test to measure overall proficiency, so we endeavoured to use another instrument to fully understand WordEngine's efficacy.

Research Questions

RQ1. To what extent does the use of WordEngine increase the active recognition of L2 vocabulary in the long term?
RQ2. To what extent does the use of WordEngine increase the active recall of L2 vocabulary in the long term?
RQ3. To what extent does the use of WordEngine improve measures of general English proficiency in this context?

Method

The study was conducted at a junior college in the Tokai region. The subjects were 66 female, second-year English major students. Time spent using WordEngine was the independent variable in the study, and usage was measured by the number of correct responses to the flashcard quizzes that the subjects had to complete per week. During their first year at the college, the subjects were required to obtain 400 correct responses per week and 350 in their second year. In the academic year 2019, following the graduation of the subjects, the number of required correct responses was again reduced to 210, and from April 2020, the use of WordEngine was discontinued altogether.

Materials and Procedure

Materials used in this study included four different types of tests to measure the dependent variables: long-term retention of the words learned via WordEngine and general English proficiency. It should be noted that although a total of 66 subjects took part in the study, not all of them could take every test of vocabulary retention due to being absent from class. Therefore, the number of subjects who took each test is described below.

The tests that measured active recognition of the words studied via WordEngine were downloaded from the software's administrator site, V-Admin. These tests are generated by a proprietary system, a detailed description of which is unavailable. The tests were tailored to each student because of individual differences in the words studied with the WordEngine software. The V-Admin website allows tests to be made on words with different numbers of correct responses, from one through to six. Another variable feature offered by the V-Admin site is for teachers to select the time at which words had entered a

particular stage in the learning process. In this study, word tests that were downloaded from the WordEngine system only contained words that had reached the “fully acquired” stage (they had received six consecutive correct responses, which is to say they had been successfully reviewed five times). The time at which they entered that stage varied from four months to eight months prior to the test day. Each test contained 30 multiple-choice questions with four items from which to choose. The test questions consisted of a Japanese word, and subjects were directed to choose the correct L2 translation from the four options.

The first tests administered were unaltered versions of the multiple-choice tests downloaded from the V-Admin website. The testing was conducted during class time in November 2018 (Time 1), and tests contained words that had entered the “fully acquired” box four months earlier in early July 2018. The purpose of this time gap was to enable the measurement of the long-term retention of the words. Testing took five to ten minutes, after which the subjects marked their own papers using the answer key provided by the V-Admin website. The authors later checked this marking. On this day, only 47 of the 66 subjects were present and able to take the test.

The second testing took place in December 2018 (Time 2), and the purpose this time was to measure the subjects’ active recall of the words they had studied via WordEngine. At Time 2, 57 of the 66 subjects were present and able to take the tests, which were based on a new set of multiple-choice tests downloaded from the V-Admin website. These tests were based on words that had entered the “fully acquired” stage in late July 2018, a gap of approximately five months. However, the four multiple-choice items had been completely removed this time, leaving only 30 Japanese words, which the subjects were instructed to translate into English. Testing again took up to approximately 10 minutes, and subjects checked their own answer sheets using the answer key they were given. When the authors later checked this marking, sometimes bonus points were added for correct translations of the test word that were not actually the word that the subject had studied via WordEngine. For example, a bonus point was awarded if a subject wrote ‘result’ as the translation of ‘結果’, even if the translation they had studied through WordEngine was ‘outcome’. It is unclear whether WordEngine use promoted the recollection of the close synonym, so test scores without these ‘bonus points’ were also recorded.

In January 2019 (Time 3), an additional set of L2 translation tests, similar to those administered at Time 2, was given to 60 subjects. The gap following the words being classified as “fully acquired” was eight months, the largest yet. Once class teachers had collected this test, the students were given the proprietary multiple-choice test (similar to that administered at Time 1) upon which the L2 translation tests were based. The purpose of the testing at Time 3 was, therefore, to measure both active recall and active recognition of the exact words that the subjects had studied via WordEngine, but with a longer time interval between learning and testing, to gain further data on the long-term nature of vocabulary retention.

In addition, one further procedure was conducted to analyse the results of the L2 translation tests of active recall conducted at Times 2 and 3. The results were analysed in such a way as to investigate whether the more proficient subjects in the department scored significantly higher, on average, than the less

proficient subjects. Using scores on the VELC proficiency test, subjects in the top 25% and bottom 25% of the cohort were separated, and the average scores of each group on the active recall test were calculated.

Increases in general English proficiency were measured in two ways: by comparing scores on the TOEIC test and the VELC proficiency test. The subjects were required to take the TOEIC test upon entering the college, at the end of their first year in February 2018 and at the end of their second year in February 2019. Subjects were required to take the VELC test at the end of the first, second and third semesters that they spent at the college. Average increases in the results of these tests were calculated to compare year groups. This made it possible to investigate if decreasing or discontinuing the use of WordEngine negatively affected the two measures of overall English proficiency among the different cohorts.

Results

The results of the tests of active recognition (multiple-choice) and active recall (L2 translation) of words learned via Word Engine can be found in Table 1. The maximum score on all the tests was 30. The scores in parentheses are those which only include the exact L2 translation learned via the WordEngine software (with no 'bonus points' included).

It can be seen that whereas the results for the multiple-choice tests were around 28 points out of 30, or 93%, the scores on the active recall tests were much lower, close to 11 points out of 30. If close synonyms are disallowed, then the scores dip below nine points out of 30. A noticeable feature of the results is that despite the variable time lags between the words being successfully reviewed a fifth time and the test day, there was little variation in the total scores.

Table 1. Results of Vocabulary Retention Tests

Time	Type of Test	Mean average of all subjects	Mean average for top 25% on VELC test	Mean average for bottom 25% on VELC test
Time 1 November 2018	Multiple-choice	27.98	–	–
Time 2 December 2018	L2 translation	10.60 (8.42)	10.83	9.2
Time 3 January 2019	L2 translation	10.80 (8.72)	12.0	10.0
Time 3 January 2019	Multiple-choice	28.0	–	–

The average increases in TOEIC scores during the period in which WordEngine was used at the college and when it was subsequently discontinued can be found in Table 2.

Table 2. TOEIC Score Increases

Student Year of Entry	Number of years of WordEngine Use	Increase after 1 year	Increase after 2 years
2017	2	91	120
2018	2	80	91
2019	1	91	188
2020	0	94	118

Figures for the 2019 cohort show the largest increase over two years, even though these students did not use WordEngine at all during their second year at the college. The increase of 188 points was double that of the previous year and around 50% higher than two years previously. The increase after one and two years for the 2020 cohort, which did not use WordEngine at all, was similar to previous cohorts.

Table 3 shows the average increase in VELC test scores over the same period.

Table 3. VELC Score Increases

Student Year of Entry	Number of years of WordEngine Use	Increase after 1 semester	Increase after 2 semesters	Increase after 3 semesters
2017	2	13	35	56
2018	2	25	37	39
2019	1	22	49	61
2020	0	30	33	53

The figures for the 2019 and 2020 cohorts do not seem to be significantly different from the previous years.

Discussion

In response to RQ1, data from the multiple-choice tests would appear to show that WordEngine usage led to a robust active recognition knowledge of the target words studied by the subjects. Scores of around 93% would seem to be a cause for satisfaction. In response to RQ2, the results show a much lower ability to actively recall the words studied with WordEngine. Scores of around 33% on the active recall tests would seem to be disappointing, given Lexxica's claim that using WordEngine leads to the "full acquisition" of vocabulary after five correct reviews. We believe that "full acquisition" entails the development of active recall as well as active recognition knowledge. Taken at face value, the results would appear to show that in response to RQ1 and RQ2, WordEngine usage promotes robust active recognition but not active recall. This is undeniably one possibility. However, in order to strongly make this claim, it would be necessary to have high confidence in the construct validity of the multiple-choice tests. However, we have to state that we have doubts about this due to a lack of transparency concerning the choice of distractors on these tests. These distractors may or may not have been words that subjects had

previously studied via the WordEngine software. If the distractors were indeed words that the students had studied at around the same time as the target words, then we could state confidently that the multiple-choice tests truly tested active recognition because subjects would have to genuinely know the correct translation of the L1 target item in order to answer correctly (if we discount lucky guessing). If the distractors had somehow been chosen at random, however, then when subjects came to choose an answer from the four multiple-choice items, they may simply have chosen the word that they remember having studied rather than the word that they could confidently match as a translation of the L1 target item. Such a scenario would seriously compromise the test's construct validity. It must also be noted that choosing distractors in this random way would serve to boost the scores on the multiple-choice tests, perhaps with the aim of providing customer satisfaction.

When constructing a multiple-choice test, the incorrect alternatives should be plausible, and as the name suggests, they should serve as distractors (Brame, 2013). The distractors in our subjects' multiple-choice tests are predominantly from the same 2800 frequency word list (NGSL, n.d.). However, we are still unsure if the students had previously studied the distractor words or if they were simply chosen at random from the new general service list (NGSL) or another vocabulary frequency list. We contacted WordEngine to clarify how the distractors are chosen. However, we did not receive a clear reply. Therefore, we cannot be sure that the multiple-choice tests are reliable measures of vocabulary knowledge, and our confidence in the effects of WordEngine as a promoter of active recognition is compromised.

In response to RQ3, we did not see an apparent influence of WordEngine usage on measures of overall proficiency, such as the TOEIC and VELC tests. In fact, average TOEIC scores increased more after students in the department had stopped using WordEngine. When making any conclusions about this particular result, it must be acknowledged that exceptional circumstances may explain the significant jump in average TOEIC scores in 2020. During the 2020 academic year, many classes were provided online rather than face-to-face. As a result, students may have had more time to study English or had fewer distractions, and therefore their TOEIC score increases may, to some extent, have been due to these conditions. Due to the unique nature of the circumstances, we will refrain from making claims from this particular result and would conservatively conclude that WordEngine usage did not seem to affect the measurements of overall English proficiency to any significant extent. This conclusion underscores our doubts about Lexxica's claim about the WordEngine software.

Our results did not align with previous studies such as Phillips (2011) and McClean, Hogg and Rush (2013), which showed generally positive results of WordEngine usage, although there were caveats. We feel, however, that our study was a more thorough attempt to judge the validity of Lexxica's claim of promoting "full acquisition" for two reasons. Firstly, we measured the recollection of the actual words studied by the subjects in terms of both active recognition and active recall. Secondly, we included a time lag into the design of our investigation and could therefore measure the extent to which words had entered long-term memory.

Conclusion

Between five and eight months following the completion of a spaced-repetition study programme using the WordEngine software, students could recall L2 translations for approximately one-third of the target words. Therefore, it is difficult to argue that the subjects had "fully acquired" even half of the words they had studied via WordEngine since full acquisition would entail active recall as well as active recognition knowledge. Thus, Lexxica's "full acquisition" claim must be seriously questioned. Moreover, it is uncertain whether WordEngine usage led to robust, long-term active recognition knowledge of vocabulary due to concerns over the design of the proprietary multiple-choice tests that were used as measuring tools. Therefore, it would be worthwhile for future studies to include self-made tests of active recognition knowledge made up only of words that students had studied around a specific time period, if this were feasible. Finally, comparison of the department's average TOEIC and VELC scores following the decrease in time on task and the subsequent discontinuation of its usage does not lead to the conclusion that WordEngine made a significant contribution to increasing students' overall proficiency.

Limitations

This study suffers from several limitations that need to be considered when evaluating our conclusions. First, our sample size was relatively small, and due to class absences on the three test days, we could not acquire data from precisely the same group of students each time. Second, the effects of COVID-19 may have increased the scores on the tests of overall proficiency in the academic year 2020. Thirdly, our use of the proprietary tests downloaded from V-Admin as instruments to measure active recognition did not lead to reliable results. In hindsight, it would have been better to create our own instruments, but this only became apparent after the research project had started. Moreover, it might not be feasible to create perfect instruments because the proprietary tests downloaded from V-Admin are the only way that researchers can find out what words subjects have been studying during a particular time period. Finally, tests that measured active recall may not have been a suitable measure for the type of learning promoted by WordEngine, and therefore WordEngine may want to include a productive element to their software in future.

Acknowledgement

Some of the data in this study previously appeared in the following paper:

Clarke, S. & Morrish, J. (2021). A critical review of WordEngine: Analysing its efficacy for studying vocabulary and transferring L2 words to long-term memory. *The JACET International Convention Proceedings: The JACET 60th Commemorative International Convention (Online, 2021)*, 143–144.

References

Agawa, G., Black, G., & Herriman, M. (2011). Effects of web-based vocabulary training for TOEIC. In A. Stewart (Ed.), *JALT2010 Conference Proceedings*. JALT.

Brame, G. (2013). Writing good multiple choice test questions. Retrieved from https://cft.vanderbilt.edu/guides-sub-pages/writing-good-multiple-choice-test-questions/

Cihi, G. (2018). V-Check lexical diagnostic assessment (v3.1). WordEngine. Retrieved from https://www.wordengine.jp/research/pdf/we30/V-Check_Lexical_Diagnostic_Assessment.pdf

Chukharev-Hudilainen, E & Klepikova, T. A. (2017). The effectiveness of computer-based spaced repetition in foreign language vocabulary instruction: a double-blind study. *Computer Assisted Language Instruction Consortium*, *33*(3), 334–354. https://doi.org/10.1558/cj.v33i3.26055

Fageeh, A. I. (2013). Effects of MALL applications on vocabulary acquisition and motivation. *Arab World English Journal*, *4*(4), 420–447.

Fitzpatrick, T., Al-Qarni, I., & Meara, P. (2008). Intensive vocabulary learning: a case study, *Language Learning Journal*, *36*(2), 239–248. https://doi.org/10.1080/09571730802390759

Gordania, Y. (2012). The effect of the integration of corpora in reading comprehension classrooms on English as a Foreign Language learners' vocabulary development. *Computer Assisted Language Learning*, *26*(5), 1–16.

Laufer, B., & Goldstein, Z. (2004). Testing vocabulary knowledge: Size, strength, and computer adaptiveness. *Language Learning*, *54*, 399–436.

Laufer & Paribakht, (1998). The relationship between passive and active vocabularies: Effects of language learning context. *Language Learning*, *48*(3), 365–391

Lee, S. H., & Muncie, J. (2006). From receptive to productive: Improving ESL learners' use of vocabulary in a postreading composition task, *TESOL Quarterly*, *40*(2), 295–320.

Lexxica. (n.d). Introducing WordEngine. Retrieved from https://www.wordengine.jp/research/pdf/we30/Introducing_WordEngine.pdf

Li, Y. & Hafner, C. A. (2021). Mobile-assisted vocabulary learning: Investigating receptive and productive vocabulary knowledge of Chinese EFL learners. *ReCALL FirstView*, 1–15. https://doi.org/10.1017/S0958344021000161

McClean, S., Hogg, N., & Rush, T. (2013). Vocabulary learning through an online computerised flashcard site. *The JALT Call Journal*, *9*, 79–98.

Nakata, T. (2008). English vocabulary learning with word lists, word cards and computers: Implications from cognitive psychology research for optimal spaced learning. *ReCALL*, *20*(1), 3–20.

Nakata, T. (2011). Computer-assisted second language vocabulary learning in a paired-associate paradigm: a critical investigation of flashcard software. *Computer Assisted Language Learning*, *24*(1), 17–38.

Nation, I. S. P. (2006). How large a vocabulary is needed for reading and listening? *Canadian Modern Language Review*, *63*(1), 59–82.

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

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

New General Service List Project (NGSL), (n.d.). Retrieved from http://www.newgeneralservicelist.org/

Phillips, J. (2011). An investigation into the effect of targeted vocabulary learning using a spaced repetition flashcard system on TOEIC scores. *Aoyamagakuin Joshi Tankidaigaku Bulletin*, *65*, 55–61.

Ranalli, J. (2013). Designing online strategy instruction for integrated vocabulary depth of knowledge and web-based dictionary skills. *CALICO Journal*, *30*(1), 16–43. doi:10.11139/cj.30.1.16-43.

Schmitt, N. (2010). *Researching vocabulary: A vocabulary research manual*. Palgrave Macmillan.

Schmitt, N. (2014). Size and depth of vocabulary knowledge: What the research shows. *Language Learning*, *64*(4), 913–953.

Stockwell, G. (2007). Vocabulary on the move: Investigating an intelligent mobile phone- based vocabulary tutor. *Computer Assisted Language Learning*, *20*(4), 365–383.

Webb, S. (2005). Receptive and productive vocabulary learning: The effects of reading and writing on word knowledge. *Studies in Second Language Acquisition*, *27*(1), 33–52.

Wilkins, D. A. (1972). *Linguistics in language teaching*. MIT Press.

Zhu, Y., Fung, A. S., & Wang, H. (2012). Memorisation effects of pronunciation and stroke order animation in digital flashcards. *CALICO Journal*, *29*(3), 563–577.

（受理日　2023年1月5日）