LUND UNIVERSITY

Enchaining L2 vocabulary using Anki

Spaced repetition study

Emin Gaaya

Supervisor: Shinichiro Ishihara

Centre for Language and Literature, Lund University
MA in Language and Linguistics, Japanese
SPVR01 Language and Linguistics: Degree Project – Master's (Two Years) Thesis, 30 credits
May 2025

Abstract

Learners of Japanese studied one hundred fifty Japanese-English word pairs in a four-week treatment either using a spaced repetition schedule (SRS) called Anki or their own method. The current study examined an evidence-based technique proven to enhance long-term retention. Research shows that both equal spacing, for example, reviewing a word every four days, and expanding spacing, for example, reviewing a word after three, seven, and nine days, outperform studying without spacing, called massing. In recent years, SRS tools such as Anki, Quizlet, and Duolingo have attracted increasing attention in second-acquisition language research, yet most studies have only briefly explored their long-term benefits, leaving important questions about their role in vocabulary learning unanswered. The current study examined thirty-four respondents who were assigned to an Anki group (N = 17) or a control group (N = 17). The results showed that the Anki group scored higher than the control group on both the immediate posttest and a delayed posttest two weeks later, with statistical significance. The current study highlights a realistic outcome where life comes in the way, and one cannot realistically study every day. But if committed enough to your studies, one can show daily study which is efficient and easy to follow and show great results.

Keywords: Anki, Spaced repetition, CALL, Expanding spacing, self-study

List of figures

List of tables

Convention

Japanese vocabulary is written with Japanese hiragana, and the translation in English in perenties.

Abbreviations

SRS Spaced Repetition Scheduler / software

Acknowledgements

The current study wants to thank Shinichiro Ishihara for the helpful meetings that enabled me to write this whole study. I want to also thank my respondents for giving me six weeks of their time, and I apologize if I did not include the data even after helping me for so long. I also want to thank my family and my wonderful girlfriend for helping me and keeping me on track.

Table of Contents

1.0 Introduction

Achieving proficiency in a second language (L2) requires learners to remember vocabulary long-term. They must master thousands of words, from individual to multi-word expressions, to support compelling reading, listening, writing, and speaking (Nation, 2006). To achieve this, deliberate vocabulary practice, such as using flashcards (word cards), is one of the most effective strategies for establishing meaning from connections (Nakata, 2019; Webb et el., 2020). Flashcards are usually physical paper with a question written (L1 word) on one side and an answer (L2 word) written on the other (Mondria & Mondria-de Vries, 1994; Nation, 2001, p. 296). Which is a form of paired-associate learning and, when used correctly, help strengthen connections between two items and support rapid vocabulary acquisition (Thorndike, 1908; Fitzpatrick et el, 2008; Steinel et el., 2007; Nakata, 2011). However, even if flashcard practice is effective, retention will decay and needs to be reinforced (Nakata, 2019, p. 308). To achieve long-term retention, using flashcard studying with spaced retention is one of the most effective ways to practice new vocabulary. The current study discusses only relative spacing, which can be divided into two spacing methods. First is equal spacing, where reviews occur at consistent intervals; for example, reviewing material every four days and expanding spacing, where the intervals between reviews gradually increase, such as reviewing after three days, then seven days, and later after nine days (Nakata, 2011, p. 201). Both methods are optimal to increase long-term retention. However, equal spacing is perhaps not as realistic as learning a new language, as adding new words constantly and studying new vocabulary every fourth day is a workload no one can handle (Schuetze and Weimer-Stuckmann, 2010, 2011). For those reasons, expanding is

perhaps more optimal as it pushes old cards further away and lets new cards be added (Bjork, 1988, p. 399; Mondria & Mondria-de Vries, 1994). However, the researchers do not agree on what spacing method is the best, but they all agree that spacing is better than massing, a method in which a person studies everything at once instead of spacing between study seasons. For example, having one continues study season of three hours instead having three study seasons of one hour each spaced along the day. Spacing seems to always win against massing in this aspect.

The current study lets learners of Japanese use the flashcard program Anki, which is called a spaced repetition scheduler (SRS) because it incorporates spaced repetition in the program. Anki also uses the Leitner system, an old method previously used with physical flashcards. The principle is simple, when a flashcard is answered correctly, it will graduate to the next learning schedule set by the program, and when answered incorrectly it will go back to the schedule before. The program will ensure that spacing will never become too long for words that are too hard to remember and not show cards known to the learner too often. Physical flashcards are still a useful tool and should be used to this day. But they are bulky and take a lot of space. Digital spaced repetition software can be useful for future research and having more realistic research is necessary over controlled lab studies.

The gap in the research comes from the fact that laboratory studies often do not accurately reflect real-world data. This study addresses this issue by using a realistic approach in a six-week study involving 34 learners of Japanese. Participants used the spaced repetition scheduler Anki or their study own method in a control group to learn 150 English-Japanese word pairs. The effectiveness of their learning was measured through a pre-test and post-test format, with an additional delayed post-test to assess retention. Additionally, two interviews via google forms

used to collect data on motivation and their opinion of Anki. The aim is to present Anki as a valid software used in both education and self-study.

The study's structure begins with an introduction, followed by a literature review that summarizes previous research findings while primarily emphasizing the theoretical framework surrounding spaced repetition. This section also includes an analysis of prior studies on Anki. The third section presents the research question, while the fourth outlines the methods and materials used. The fifth section presents the results, followed by a discussion and an appendix.

2. Literature Review

Vocabulary knowledge is arguably the most essential component in language acquisition. While grammar plays an important role, communication is impossible without vocabulary (Schmitt, 2008). Numerous studies on vocabulary learning have demonstrated that retrieval practice significantly enhances retention (McDaniel & Fisher, 1991). Furthermore, when paired with spaced repetition, retention gains become even more pronounced (Cepeda et al., 2006). Flashcards are among the most common tools used for retrieval practice. In Japan, for instance, 60% of students report using flashcards to study vocabulary; however, many of them apply ineffective strategies (Nakata, 2011; Zung et el., 2022).

The aim of this literature review is to examine existing research on spaced repetition and retrieval practice, particularly as implemented in computer-based flashcard programs. Special attention will be given to the underlying theories supporting these methods, with a focus on demonstrating the benefits of such tools and analyzing the science behind Anki's spaced repetition system. I will begin by going over early research on spaced repetition, then explain studies on spacing, computer assisted language learning (CALL), flashcards.

2.1 Early research on spaced repetition

When we aim to learn something new, we often encounter moments when inevitable information slips from our memory. Forgetting is a natural aspect of being human, and despite our best efforts, it is something we cannot entirely avoid (Ellis, 1995; Hulstijn, 2001; Nation, 2001). However, many have experienced remembering something we had forgotten, experiencing

a stronger connection the second time but nonetheless forgetting it again the third time. The concept of forgetting can be abstract and complex to grasp, but research by Ebbinghaus in 1885 began to shed light on this topic. Research by Ebbinghaus in 1885, marks the beginning of systematic, empirical research on memory. His pioneering work laid the groundwork for understanding how we acquire, retain, and eventually forget information. One of his most significant contributions was the development of the "forgetting curve," a graphical representation that demonstrates the rapid decline of memory retention following the initial learning of new information (Ebbinghaus, 1885; Mondria & Mondria-de Vries, 1994).

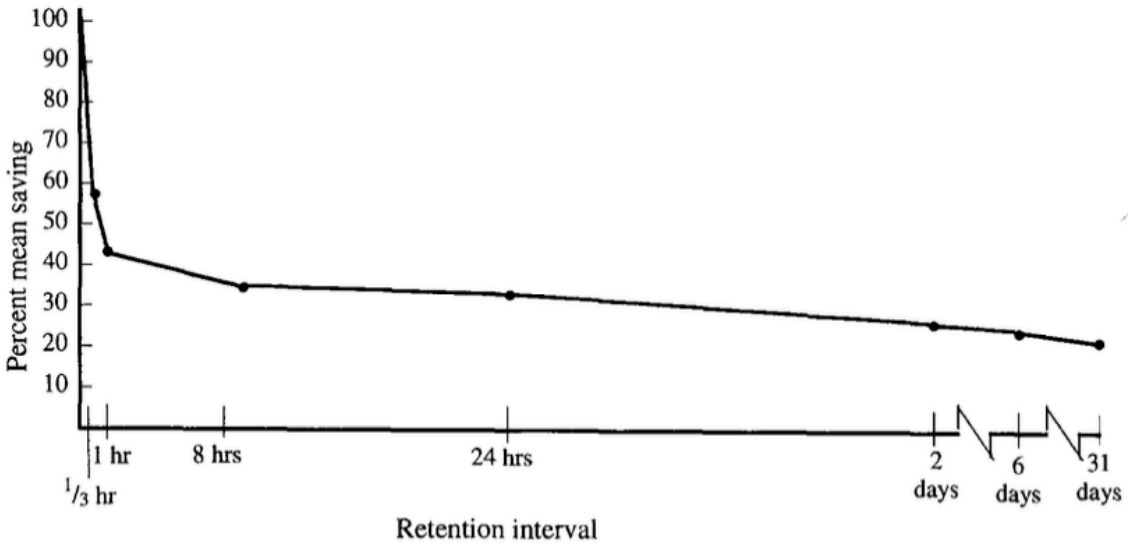


Figure 1 Forgetting curve (Ebbighhaus,1885; Mondria & Mondria-de Vries, 1994).

According to Ebbinghaus, forgetting is not a flaw in our learning system but an inherent part of the memory process. He noted that when we are on the verge of forgetting newly acquired information, reviewing it at that critical juncture can significantly aid retention (see figure 1). For

instance, when you learn new information, the curve shows that the initial forgetting is around 60-70% during the first 24 hours. However, if you review this information the following day, the memory trace is strengthened, and it remains accessible for a longer period before being forgotten again. As you continue to review the information, the rate at which it is forgotten decreases. This is because each subsequent review reinforces the memory further, causing the forgetting curve to flatten. A flatter curve indicates that the information takes longer to be forgotten, making it more likely to be retained in long-term memory. In other words, with each review session, you are effectively extending the period during which the memory remains accessible. Over time, the intervals between reviews can be lengthened, reflecting the fact that the reinforced memory is now more robust and resistant to decay. Ebbinghaus lays the empirical foundation for modern spaced repetition system and research on forgetting. The method most closely accosted with Ebbinghaus is using of nonwords and seeing how fast these words were forgotten (see Murre & Dros, 2015, for an updated replication and analysis of the forgetting curve).

Further expanding on this research is by Pimsleur1967 most famous for his language learning program. Paul Pimsleur sadly died at age 45 but already at age 36 further developed on how fast one should review information again to gain 100% recollection (see figure 2) following the same theory of the forgetting curve. Pimsleur landed on 60% and by reviewing one can expand spaces until the spaces become so big that they information do not show up anymore.

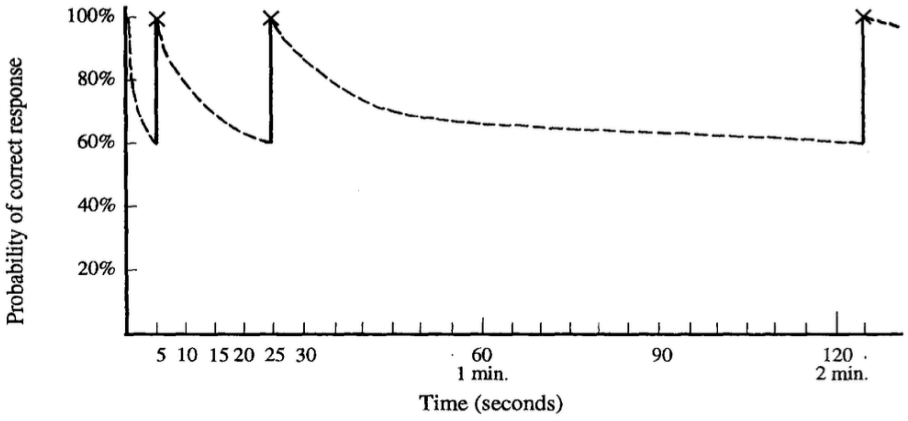


Figure 2 Forgetting curve (Ebbighhaus,1885; Mondria & Mondria-de Vries, 1994).

More research on long intervals has been studied by both Lado (1967) and Landauer and Bjork (1978), which started the push for Pimsleur expanding spacing. The reviews are seen to improve long-term retention, and according to Lado, even after each review, two- or three-times amount of the original interval is needed until the next review (Mondria & Mondria-de Vries, 1994, p. 50). However, one should only expand spacing between reviews if someone recalls the correct information, and there should be a system the learner can follow if they do not recall incorrectly.

One effective learning system is the Leitner system, developed by Sebastian Leitner in 1972. This flashcard organization method lets learners to schedule their review sessions using expanding spacing more efficiently. The Leitner system consists of a box with five compartments (see figure 3).

To begin using this method, learners must first create their flashcards. As Nakata (2011) describes, Flashcards (word cards) are usually a piece of paper with typically an L2 word on one

side and its meaning on the other, usually provided as a translation in the learner's first language (L1). However, these flashcards can be definitions and explanations. There could also be an equation of one side and the solution on the other (Polly et al., 2025).

In the first session, the learner starts with approximately 30 to 40 cards (Mondria & Mondria-de Vries, 1994, p. 52). They examine both sides of a flashcard that displays a word in their native language on one side and its translation on the other. The learner goes through all the cards until they repeat the first card. If they recall the card correctly, they move it to the second compartment. The learner continues studying the first compartment until they can recall all the cards accurately.

When learners correctly recall a word from the second compartment, they graduate it to the third one. However, if they recall it incorrectly, the card returns to the first compartment. The learner reviews and graduates known cards as compartments fill up while returning incorrectly recalled ones to the first compartment. When the final compartment is complete, the learner can discard the cards as understood or store them, with occasional reviews to ensure they retain the information.

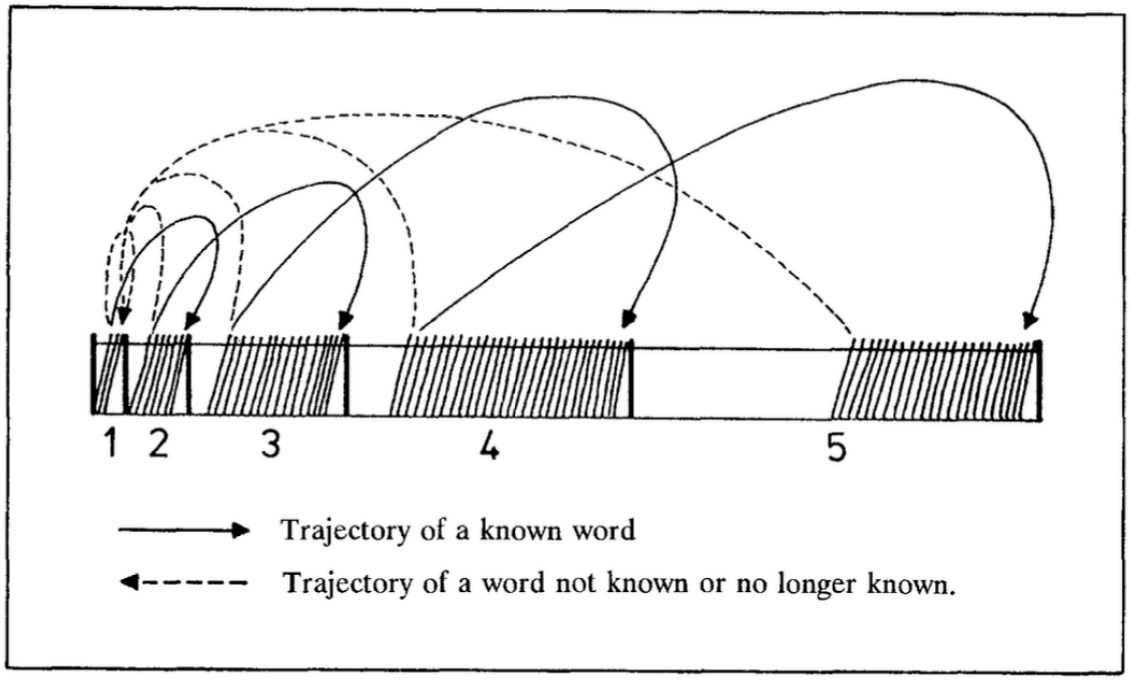


Figure 3 Leitner's hand computer, 1972; Mondria & Mondria-de Vries, 1994, p. 52).

Combining these two principles forms the basis of many spaced repetition software programs, such as Anki, SuperMemo, Memrise, and Quizlet (see Nakata, 2015, for a review). These digital tools also address some of the limitations of the physical method, such as the serial position effect and the physical method can be bulky and take up much space, especially depending on how many cards are added. A few studies on Leitner's hand computer (Leitner's learning box) have been conducted using the method more recently. First is Farhadi (2012), who tested the method compared to a control group on a vocabulary study with a pre-test and a post-test showing the of Leitner's hand computer outperformed to the control group. Another study, Whitmer et al (2022), implemented the method in a digital lab setting and saw the Leitner's hand computer to be more efficient in study time but did not show better retention.

2.2 Theoretical background

Previous studies have investigated different forms of spacing and its effect on retention. The current study investigates spaced repetition software Anki, which uses expanding spacing with the incorporation of Leitner's learning box to try to help learners remember longer with digital flashcards. The study takes early research on spaced repetition and creates a program around this method. However, spacing has shown great significance on retention compared with different methods.

Spacing can be divided into two forms: absolute spacing and relative spacing. Absolute spacing refers to the total amount of time between the first and last study session. For example, if someone studied a word three times with a 2-day interval between each session, the absolute spacing would be 6 days (Karpicke & Bauernschmidt, 2011; Sonbul et al.,2024).

In contrast, relative spacing concerns the pattern or distribution of intervals between study sessions. Relative spacing can be further categorized into two types: equal spacing and expanding spacing (Pyc & Rawson, 2007; Karpicke & Schmidt, 2011; Kang et al., 2014; Nakata, 2011). Equal spacing involves reviewing material at consistent intervals, such as every 4 days. In expanding spacing, the intervals between reviews increase over time, for example, reviewing a word after 3 days, then after 5 days, and then after 9 days (Karpicke & Bauernschmidt, 2011).

Another important factor influencing the effectiveness of spacing is the lag effect, which refers to the impact of the length of intervals between study sessions. For instance, if someone studies over a total period of 12 days with either equal intervals of 3 days (e.g., days 0, 3, 6, 9,

12) or expanding intervals like days 1, 4, and 7, the absolute spacing in both cases would still be 12 days, but the relative spacing differs. The lag effect shows that studying with longer spacing between sessions generally leads to better long-term memory retention, whereas studying with shorter intervals results in better short-term retention (Cepeda et al., 2006; Nakata, 2011; Karpicke & Bauernschmidt, 2011; Nakata et el, 2023). The phenomenon is called the spacing effect, which empirical research has proven leads to better long-term retention when one spaces out their study sessions over time, rather than studying every day with little to no gap between sessions, which is called massing (Feng et al., 2019; Yamagata, et al., 2023). Additionally, when one uses massing right before an exam that is a common term called cramming which is what massing is, but it does not necessarily mean a test is common. Massing is a very common approach which does not lead to better long-term retention but can lead to short-term gains which is why it is used tests often (Nakata, 2011). However, right after the tests all the information disappears fast unless it is reviewed again which will turn it into spacing, but people who use massing do not normally return to the literature and instead have one long season of studying. However, cramming is probably the most common way of cramming as pure massing is not that common (Cepeda et al., 2008; Kornell, 2009).

Numerous studied has yet to prove the best form of spacing between equal and expanding. Studies have shown results that equal spacing to be better at post-test scores (e.g. Pyc & Rawson, 2007; Storm et al., 2010). However, studies have also shown an advantage for expanding spacing (Vlach et al., 2014; Nakata, 2015; Kanayama & Kasahara, 2016). Although, studies on relative spacing and massing shows that spacing is always better than massing (e.g. Zulkiply, 2013;

Sonbul, 2024). Furthermore, studies on language learning have shown statistical advantage of spacing (e.g. Verkoeijen et el., 2008; Rogers, 2015; Kim & Webb, 2022; Yan & Zhou, 2023)

Nakata (2015) noted that second language (L2) vocabulary acquisition consistently shows that spaced repetition is more effective than massing (Cepeda et al., 2006; Nation, 2001). Within spaced repetition, there are two major scheduling approaches as mentioned before: equal spacing and expanding spacing. Although expanding spacing has received considerable empirical attention, results have been mixed, and no study has conclusively shown that expanding spacing is more beneficial than equal spacing in terms of test scores (Karpicke & Bauernschmidt, 2011; Pyc & Rawson, 2007). Some studies have found that direct or delayed feedback combined with equal spacing yields significant results (Logan & Balota, 2008; Storm et al., 2010). However, Nakata (2015) reported that expanding spacing produced slightly better results compared to equal spacing in a realistic second-language paired-associate learning setting that included productive recall and immediate feedback. Further research is needed to explore these findings, but the overall literature indicates that spaced repetition is superior to massing for L2 vocabulary acquisition.

Another study by Nakata & Suzuki (2019), which investigated the effects of massing compared to spaced repetition. The study focused on vocabulary retention by examining both semantically related and unrelated words. The sample consisted of 133 Japanese university students who had received at least six years of English language instruction. Respondents were divided into two groups. One group used a massing approach and studied semantically related or unrelated words consecutively, while the other group employed spaced repetition, with

repetitions of the same words distributed over time. The study materials comprised 48 low-frequency English words paired with their Japanese translations. The words were carefully matched for factors such as frequency, length, and familiarity. Half of the words were semantically related, for example, groups of animals or plants, while the other half were unrelated. Recall of the Japanese translations was assessed in both immediate and delayed post-tests to investigate long-term retention. The results demonstrated significant benefits from spaced repetition. In addition, semantically related words produced more recall errors than unrelated words, suggesting that semantic clustering may hinder retention.

Additional, research has explored various techniques within spacing. Studies have examined the benefits of repeated reading (Serrano & Huang, 2018), differences in testing formats (Nakata et al., 2021), the effects of spacing on vocabulary acquisition (Yan & Zhou, 2023), and comparisons between study materials such as flashcards and word lists (Nakata, 2008). Further investigations in classroom settings have also demonstrated the superiority of spaced repetition for learning (Rogers & Cheung, 2020). Nakata (2008), in prearticular showed that computer-based flashcards significant outperform both physical flashcards and using a word list. Spacing or spaced reptation as it is often called has been shown to be beneficial or learning information long-term. Taking this information, we might find ways to be able to study more with programs and software that aid the spacing process.

Vocabulary knowledge

The current study employs both productive and receptive recall tasks in its pre- and post-tests. According to Nation (2001), *productive knowledge* refers to the ability to actively use

vocabulary in tasks such as speaking and writing, while *receptive knowledge* involves recognizing and understanding vocabulary through listening and reading. Prior research has shown that receptive knowledge is generally easier to acquire than productive knowledge, as it relies on recognition and is closely tied to meanings already established in the learner's first language (L1). In contrast, productive knowledge requires greater retrieval effort, as learners must retrieval the second language (L2) word which require more effort (Yanagisawa, 2016).

Learning vocabulary with sound

The current study employs the Anki flashcard program to enhance vocabulary learning by incorporating audio. Although research on the use of sound in vocabulary acquisition remains limited (e.g., Chun & Plass, 1996; Nation, 2001; Plass & Jones, 2005; Teng, 2023;), and no prior work has examined user-created flashcards with multimedia elements. However, Teng's study provides a helpful insight on the benefits of multimedia input in language learning. In his experiment, respondents were assigned to one of four conditions: "definitions alone, definitions with detailed word information, descriptions and word information accompanied by audio or definitions, and word information accompanied by a video", and their retention was measured with a delayed post-test administered two weeks after the initial treatment. The findings showed that all conditions outperformed the ‘’definition’’ only condition. However, the study did not support sound to be the most significant in the results it was the second-best factor on retention on all the conditions in the study (p. 747). Teng (2023) mentions that dual coding theory may explain why retention was higher in the sound and video condition. This theory, introduced by Allan Paivio in 1986, proposes that verbal and non-verbal processing occur simultaneously,

which enhances working memory and ultimately leads to better long-term retention (Teng, p. 739).

2.3 Space repetition software (SRS) and flashcard programs

Numerous studies have compared the effectiveness of computer-assisted language learning programs with the incorporation of spaced repetition schedules, showing their benefits over traditional methods (Cakmak et al., 2021; Bower and Rutson-Griffiths, 2016; Yüksel et al., 2022). Traditional benefits can be defined as any method that do not include digital software. Such as, Pen and paper, physical flashcards, but also reading and taking notes.

Some studies also compare ways digital programs are better than others. With the popularity of various apps such as Duolingo and Quizlet, decided to compare alternative methods to see if one is better. For example, Larchen et al. (2020), investigated Quizlet and a virtual reality program to study 10 idioms. The study tested two spacing conditions 15 min and 1 week and then tested their benefits on a one-week retention post-test. The study concludes that virtual reality program showed better retention on both spacing conditions. Another study by Jia et al. (2023), also resulted on a three-week delayed post-test better long-term retention compared to Quizlet and paper a flashcard game. Another study also involving Quizlet by Bueno-Alastuey and Nemeth (2020), compared podcasts and Quizlet on receptive and productive knowledge but this study resulted in no significant difference between the two groups.

2.4 Evaluating Spaced Repetition Software Anki

2.4.1 Anki in the medical field

Kaitsu and Nakata (2025) proposed comprehensive criteria for evaluating computer-assisted language learning (CALL) software (2011) and mobile-assisted language learning (MALL) applications (2025). Both papers assess widely used programs such as Quizlet, SuperMemo, and iKnow!; however, Anki, the program used in the present study, was not evaluated. According to Kaitsu and Nakata (2025, p. 3), this omission was due to prior evaluations conducted by Koleini et al. (2024) and by Dunlosky and O'Brien (2020), though those evaluations addressed only the desktop version of Anki. To address this gap, we will apply Kaitsu and Nakata (2025) mobile criteria to evaluate both the mobile and desktop versions of Anki. We chose this approach because the desktop criteria Nakata established in 2011 are similar to those from 2025 and improved. Furthermore, unlike Dunlosky and O'Brien (2020), who focused on evidence-based learning strategies rather than language learning per se, Kaitsu and Nakata's frameworks are explicitly grounded in language-learning theory.

Dunlosky and O'Brien (2020) examined whether Anki and other spaced-repetition programs support effective learning strategies such as spaced retrieval and successive relearning (SR). A program that implements these strategies well can improve students' performance regardless of background or field of study (Dunlosky & O'Brien, 2020, p. 227). Their study employed a thirteen-criterion evaluation framework to assess the effectiveness of spaced-repetition software. Anki scored near-perfectly across all thirteen criteria, including the ability to add images, customize scheduling options, and accept typed answers. However, Dunlosky and O'Brien (2020) presented in their results that Anki lacks built-in support for features such as multiple-choice testing and automated external reminders (e.g., email or SMS notifications),

which must be enabled via add-ons or user configuration via code. For users without technical expertise, Anki's large community offers a wide range of pre-made decks (Set of flashcards) and third-party add-ons available on AnkiWeb. In sum, Dunlosky and O'Brien conclude that Anki effectively supports successive relearning and enhances learning outcomes, though the specific version evaluated is not clearly stated (presumably the desktop client, possibly alongside the web version). Future research should specify the exact Anki version under investigation, as functionality can differ substantially across platforms.

Koleini et al. (2024) explicitly evaluated the mobile (application) version of Anki. Their study assessed the acquisition of 100 technical vocabulary terms by 80 Iranian university students majoring in Psychology over a ten-week treatment period, comparing mobile-assisted digital flashcards (Anki) with traditional paper flashcards. Using the Vocabulary Knowledge Scale (VKS) as a pre-test, immediate post-test, and six-week delayed post-test, they analyzed results via a 2×3 mixed-design ANOVA. Respondents studied for fifteen minutes per day, Monday through Friday. Although Anki's built-in statistics tracked total study time and number of cards reviewed, these metrics were not reported. The digital-flashcard group significantly outperformed the paper-flashcard group on both immediate and delayed post-tests. However, the study relied solely on self-reported data, without objective usage metrics, satisfaction surveys, or qualitative interviews, commonly included in another medical-education research. For instance, Harris and Chiang (2022) and Jape et al. (2022) report high user satisfaction with Anki among medical students, and Wothe et al. (2023) found improved sleep quality in Anki users.

Goldman et al. (2024) conducted a systematic review of eight studies published between 2015 and 2022, demonstrating a positive correlation between Anki usage and significant

improvements in exam scores (e.g., Deng et al., 2015; Wothe et al., 2023; Lu et al., 2021; Gilbert et al., 2023; Levy et al., 2023) as well as a lower failure rate (2.8% vs. 10.94%; Cooper et al., 2023). Strauss, et el., (2019) similarly found that residents using Anki achieved a 92% pass rate, well above the national average of 67%. However, some studies (e.g., Sun et al., 2021; Levy et al., 2023; Cooper et al., 2023) report no significant score improvements compared to alternative methods. These discrepancies may stem from differences in Anki implementation, user engagement, or enjoyment levels.

Studies that focus on Anki in language learning are limited. However, since the current study studies this topic, reviewing some research that has used Anki to assess its effectiveness in language learning is essential. Indonesian second-grade students (Jaya, 2020) and adult EFL learners (Iravi & Malmir, 2023) made significant post-test gains following computer-based Anki practice, while college-level ESL students improved their vocabulary exam scores (Ozer et el, 2017). Additionally, Iranian learners who received Anki-based instruction retained more new words than students in traditional classes (Khoshsima & Khosravi, 2021), and Indonesian vocational students achieved significant Japanese-vocabulary acquisition using the AnkiDroid app (Nender et al., 2022). Furthermore, anki studies has shown after 10 weeks treatment of with a long-term retention of Spanish vocabulary compared to the non anki ground (Mujahidah et al., 2024), and university students attained higher end-of-semester test scores after Anki treatment (Hanson and brown, 2019). Based on results from spaced repetition software one can theories that Anki in future studies will show expected results similar to other studied on spaced repetition.

2.4.1 Evaluating the Flashcard App Anki

We evaluate Anki using the framework proposed by Kaitsu and Nakata (2025). Our primary focus is the desktop client, which offers the most extensive functionality. Particularly because of the third-party add-ons. Nevertheless, we also examine the mobile version of Anki, because respondents in the current study use both platforms. Even if the present study does not exploit every available feature, it is still important to assess Anki's overall suitability for language learning.

Kaitsu and Nakata (2025) identifies twenty-four criteria to assessing mobile flashcard software. While Nakata's earlier framework (2011) for flashcard software for the desktop covered the same aspects. The current revision refines the criteria and introduces new ones, with a more comprehensive tool consisting of seven criteria for flashcard creation and editing and seventeen for learning benefits (p. 12). Following Kaitsu and Nakata's scoring system Anki will be scored in the same way. A plus sign (+) indicates that a criterion is met and earns one point; a minus sign (–) indicates that the criterion is not met and earns zero points; and "N/A" is assigned to features that require coding and likewise earns zero points. Double plus (++) also earns one point but signals that the feature goes above and beyond the baseline expectation. For instance, criterion 16 (block size) receives a double plus when users, rather than the app, are free to set the block size themselves (p. 13).

According to Kaitsu and Nakata (2025, p. 5), a flashcard program should provide learners with a library of ready-made decks so that they can begin studying immediately. Although self-created flashcards have been shown to improve retention (Dodigovic, 2013; Lei & Reynolds, 2022), the software must still enable learners to create their own cards. It should also support

fully multilingual input (alphabetic and non-alphabetic) so that users can study any language, which is a clear advantage when writing in the L2 during learning to enhance acquisition (Gyllstad et al., 2023). Additionally, the software should accommodate multi-word items (e.g., idiomatic expressions), which contribute to greater fluency in the target language (Schmitt, 2023).

Kaitsu and Nakata (2025) also recommends features such as the ability to organise cards into sets for easier categorisation and to share those decks with other learners and instructors, thereby facilitating feedback and collaboration (Dunlosky & O'Brien, 2022). Finally, incorporating multimedia elements accords with dual-coding theory, which posits that pairing verbal and visual information strengthens retention (Paivio & Desrochers, 1980); empirical evidence likewise supports the use of images in vocabulary learning (Carpenter & Olson, 2012; Ramonda, 2022). According to this creation and editing criteria, Anki scores a perfect seven out of seven. Users benefit from an extensive community-curated collection of decks hosted on AnkiWeb, covering a wide range of languages (e.g., Japanese, French) and subjects (e.g., physics, music). The same advantages apply to the mobile apps.

The major difference between the desktop and mobile versions is that the mobile apps do not support user-created add-ons. Consequently, on mobile. Kaitsu and Nakata's last 17 criteria focussed only on learning and Anki mobile fails to meet 6 out of 17 of them. Anki desktop can add this function with the help of add-ons or coding. Kaitsu and Nakata's criteria 10 (receptive recognition), 12 (productive recognition), 14 (varied encounters and use), 18 (fluency development), 19 (automatic speech recognition), and 24 (motivational feedback). As discussed in the literature review, productive and receptive recall are more beneficial than recognition, but

an effective flashcard program should support all formats according to Kaitsu and Nakata. Criterion 14, varied encounters and use, refers to exposure to words in multiple contexts (e.g., several example sentences) to deepen lexical knowledge (Kaitsu and Nakata, 2025, p. 7). Fluency development (criterion 18) describes progression from beginner to communicative competence.

Furthermore, the mobile apps do meet criteria 11 (receptive recall) and 13 (productive recall), which lie at the core of Anki's functionality. Anki also satisfy criterion 17 (interference avoidance) by allowing users to study specific tagged cards or decks. Criterion 20 (adaptive sequencing) is met through Anki's spaced-repetition algorithm. Criterion 23 (formative feedback) is not provided automatically but can be added by users or generated through add-ons. Anki's direct feedback lets learners include as much information as they need for later review. Motivational feedback (criterion 24) is available via add-ons such as Anki Leaderboard, which introduces gamification elements that enhance motivation (p. 10).

Kaitsu and Nakata excludes criteria 21 (retirement) and 22 (expanding retrieval) from the overall score due to limited empirical evidence regarding their optimal implementation (p. 14). Applying his scoring to the mobile version of Anki, alone yields 15.5 points out of 46. This is one of the lowest scores if we compare it to the already evaluated programs evaluated by Kaitsu and Nakata. However, this outcome does not affect the current study, which employs only recall retrieval, a feature supported by all versions of Anki. When the desktop version with add-ons is considered, Anki's score rises to 45 out of 46. These results reflect Anki's flexibility: it imposes no limitations and enables users to create fully customized study environments. This openness has driven its widespread adoption in medical and language-learning contexts. Although Anki may initially seem challenging, requiring basic coding for full functionality. The large user

community, free add-ons, and vast library of pre-made decks make it as user-friendly as out-of-the-box applications such as Quizlet and Memrise.

3. Research Questions and Hypotheses

The current study hypothesizes that the SRS software Anki will lead to better long-term retention of 150 Japanese-English word pairs on the two week delayed post-test compared to a control group that could use any methods. However, the study does not predict that the immediate post-test results will show advantages for the Anki group compared to the control condition. The Anki group is expected to perform worse than the control group on the immediate post-test. Expanding spacing is effective for long-term retention but does not provide immediate retention gains for the short-term immediate post-test. Additionally, the Anki condition is also expected to have more positive feelings towards their method compared to the control condition. This could relate to more structure that perhaps is not easy to get in a self-study method.

The three research questions are as follows.

1. Is there a significant difference in vocabulary gains between digital flashcard using the spaced repetition software Anki compared to other control condition.

2. Is there a significant difference in vocabulary retention between digital flashcard using the spaced repetition software Anki compared to control condition.

3. What are the respondents perceptions and experiences regarding the effectiveness and usability of the Anki condition compared to the control condition.

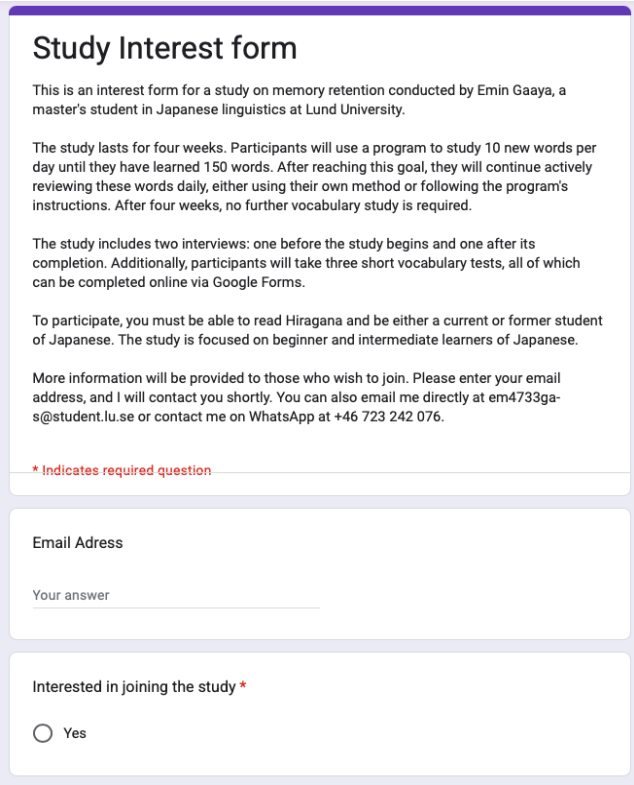
4.0 Methodology

This section begins by reviewing how we recruited and what criteria were used to select the respondents for the study. Next, we describe the materials and instruments, explicitly focusing on vocabulary selection and wordlist creation. We then explain the flashcard program, Anki. Finally, we provide a detailed procedure outline and explain how we gathered and analyzed the data.

4.1 Respondent

The recruitment phase took place during February and March 2025. The researcher emailed universities in Sweden, Norway, Denmark, and Finland that offer Japanese as a second language. Four digital Zoom information sessions held by the researcher for different classes, along with two in-person information sessions on campus to attract interested respondents.

The researcher distributed fliers having a QR code around campus, on social media, and online forums for Japanese learners, inviting them to express their interest by email or through a Google Form. One of these forums is the popular online website Reddit. Reddit has a forum that focuses solely on Anki, and many respondents who study Japanese gather there. The Google Form redirected interested people who clicked the link to the following page displayed in (Figure 5) and the last version of the filer displayed in (Figure 6). The QR code redirected interested people automatically, sending an email directly to the researchers' email showing their interest in the study.



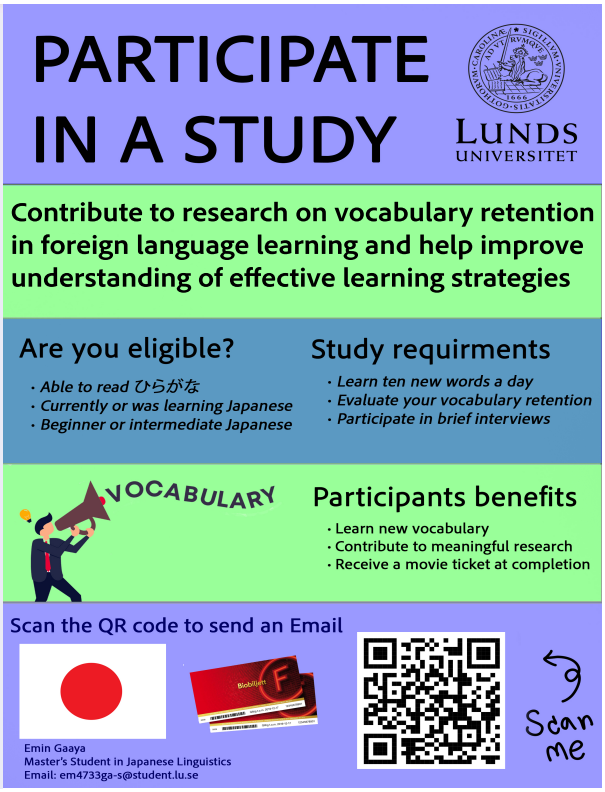


Figure 4 (Flier)

Figure 5 (Study interest form)

During the recruitment period, the researcher changed the inclusion and exclusion criteria. Initially, we set three inclusion criteria: (a) respondents should have little or no prior experience with Japanese, (b) they must be able to read hiragana, and (c) they should be first-semester university students studying Japanese. We later changed these criteria because they were too restrictive to recruit enough respondents for the study. We broadened criteria (a) and (c) to increase our ability to find more respondents. Since many university students often enter their

second semester, we realized that few true beginners could join. We expanded criterion (c) to include second-semester students and eventually adjusted it to include any active or prior learners of Japanese. We also broadened criterion (a) to allow anyone with limited to no experience, regardless of their Japanese level, with eligibility assessed by the pre-test. Criterion (b), the ability to read hiragana, stayed the only inclusion criterion not changed during the study.

Exclusion criteria included respondents who knew all the vocabulary on the pre-test, those who had Japanese as their first language, and anyone unable to meet the inclusion criteria. The current study would have been easier to conduct if students could receive class credit as an award for finishing the study or if teachers had been able to collaborate seen in earlier studies (e.g., Yüksel, 2020; Shahipanah et el., 2025; Hanson & Brown, 2019).

About 90 Japanese learners expressed interest in taking part and received a consent form that they could return digitally or in person. The consent form outlined the study's general guidelines and what we expected from them as respondents (Appendix A). We addressed any questions before they signed the consent form. However, we could not provide information about the groups they would join or the vocabulary items. We could only mention minor details about how we would conduct the study online using Google Forms. We also informed the respondents about their right to anonymity, and they could leave the study at any time without any restrictions. Additionally, each respondent needed access to a device (such as a smartphone or laptop) capable of running Anki and commit to studying 10 words daily.

Of the original 90 respondents who expressed interest in taking part in the study, only fifty-three completed the pre-test. The researcher asked three respondents to withdraw due to high pre-

test scores (see Table 4). The remaining fifty respondents received an email assigning them to one of two groups: Control or Anki. During the treatment phase, 10 respondents either stopped communicating or formally withdrew from the study via email. However, the researcher anticipated that some respondents would drop out before finishing the study because of the troubles with online studies (e.g., Hanson & Brown, 2019). Many respondents faced the same issue that they did not studying daily and asked if they should catch up by studying more. The researcher recommended that they continue studying as if nothing had happened. The researcher communicated with respondents once a week during the treatment phase. However, some respondents took several days to reply to first contacts. Future researchers should consider sending reminders through email or SMS and explore alternative communication methods to keep respondents engaged and ensure they follow the study procedures effectively with online respondents. In the end, forty respondents completed all tests. However, six respondents could not provide enough data, or there was a long gap between the final study session and the immediate post-test, leading to their exclusion from the analysis.

The final respondents (N = 34) that completed the whole study ranged in age from 20 to 40 years (M = 26.79, SD = 4.41) They revealed a diverse range of first languages (L1), including Swedish, Norwegian, English, Spanish, German, Luxembourgish, Croatian, Danish, and Russian. On average, respondents had studied Japanese for 3.6 years (SD = 3.02, range = 0.5–13) and had lived in Japan for about 0.37 years (SD = 0.57, range = 0–2.23). Around one-third of the respondents (n = 14) studied Japanese at a university, while 12 had taken classes in the past, and eight were entirely self-taught. The gender distribution included 21 males, 12 females, and one non-binary respondent. The respondents (n = 24) rated themselves between JLPT N4 and N3

levels, while the remaining 10 respondents fell within the range of beginner (N5) to early intermediate (N2). A few respondents (N = 6) also mentioned taking the JLPT, with one of the respondents taking the N2, four taking the N3, and one taking the N5. Most of the respondents came from Sweden (n = 19), followed by Norway (n = 3), Denmark (n = 2), and the USA (n = 2), with one respondent each from Spain, Luxembourg, Germany, Croatia, China, Canada, and Russia. Many respondents had prior experience with Anki (n = 22), often combining it with other platforms such as Quizlet (n = 12) and Memrise (n = 6). A few also used other tools like WaniKani, Duolingo, or Bunpro, while a minority of respondents reported no experience with any spaced repetition system (SRS) (n = 6).

The Anki and control groups included the same number of respondents, with 17 respondents in each group. The Anki group used only Anki, while the Own Method group could use any study method except Anki. The researchers provided the Anki group with brief information through an instruction sheet and a YouTube video showing how to make a new profile on Anki. We answered any other questions via email

In contrast, the Own Method group received an instruction sheet, a vocabulary list, and a study log. Researchers instructed this group that studying for even five minutes a day was sufficient. The consent form also said that respondents did not need to study for more than 30 minutes daily. Some respondents interpreted this guideline and chose to study for up to 30 minutes a day. Additionally, the Own Method group could skip words during their study sessions, while the Anki group had to review all the words, including those they already knew. However, this choice may be unfair, and if the Anki group also had this choice. However, knowing when

someone knows a word could be difficult and the Anki group only needed to answer correctly on known words a couple of times before they disappeared.

4.2 Material and instrument

4.2.1 Vocabulary selection process

The study used one hundred and fifty Japanese–English word pairs as target vocabulary. We selected the vocabulary according to various criteria to ensure a useful and representative list for language learners of Japanese. (1) Only vocabulary deemed useful for both intermediate and beginner learners of Japanese could be selected, and words found in three textbooks corresponding to all levels of the Japanese-Language Proficiency Test (JLPT) were deemed useful for the study. The JLPT is a standardized test used in Japan to evaluate the Japanese language proficiency of non-native speakers. The test consists of five levels, with N5 being the most basic and N1 the most difficult. (2) Only nouns were selected, simplifying what respondents needed to learn during the study. (3) The study included only words that respondents could write in hiragana and had kanji characters; it excluded all katakana words and words without kanji characters. (4) Most words had a single translation to minimize ambiguity, though this was not always perfectly achieved. Applying these four criteria, the initial selection included 500 words of verbs, adjectives, and nouns. Only nouns remained after applying criterion (2), excluding all adjectives and verbs. The remaining vocabulary was 300 nouns, including both hiragana-only and katakana words. Following criterion (3), all hiragana-only and katakana words were excluded

from the study. The remaining 150 vocabulary items were then selected from all five levels of the JLPT, as specified in criterion (1). These vocabulary items were divided according to JLPT levels, with adjustments to ensure a balanced list across the diverse levels. Seventy-five words from the third edition of Genki (Banno et al., 2020), a textbook commonly used by first-year Japanese students that covers JLPT levels N5 and N4. Fifty words from Tobira (Oka et al., 2009), which intermediate learners of Japanese typically use. The remaining 25 words are taken from the Shin Kanzen Master: JLPT N1 Preparation book (3A Network, 2011). The final vocabulary includes terms from all JLPT levels; this collection provides learners with valuable words to enhance their language journey.

4.2.2 Creation of the wordlist

When the word selection process ended, the researcher created a list containing all target vocabulary items used in the study. Each item included its translation and the corresponding Kanji characters. We organized the list adding ten vocabulary items per page to encourage respondents to study 10 new words per day, and to mimic Anki's default settings (20 cards per day). The list order also mirrored the Anki group order to mimic their experience. However, the respondents in the control group had no limit on what order they they could study the vocabulary, and they could study as many words as they preferred. To check study habits in control condition, respondents recorded their daily study time and the date at the top of the document (see appendix B). We collected these self-reported logs and analyzed as part of the study's data to compare study times between the two conditions with the post-test scores (e.g., Kornell, 2009; Mondria, 2003; Pyc & Rawson, 2007).

Additionally, we separated orthographic, or semantically related words so they did not appear on the same page in the control conditions word list, nor were they scheduled to be introduced on the same day in the Anki condition. We based the decision on findings by Nakata (2019), which suggest that related words can lead to greater interference and hinder retention compared to unrelated words. We intended to hinder interference caused by semantic or orthographic similarity (e.g, でんとう and でんせつ).

4.2.3 Creation of the flashcards in Anki

Anki is an open-source spaced repetition system (SRS) that supports long-term retention through digital flashcards incorporating spaced repetition. The term "Anki" means "memorization" in Japanese, and the developer developed Anki to be a language learning tool, which has expanded and become especially popular among medical students. Anki is available on four platforms: desktop, iOS (AnkiMobile), Android (AnkiDroid), and browser-based (AnkiWeb). All versions provide the same core functionality and allow users to synchronize their study progress across devices using a single user account. Respondents assigned to the Anki group could use or mix any of these versions in this study.

Creating flashcards in Anki, takes a small amount of time, depending on how much information one wants to add. The addition of add-ons makes the process even easier, creating flashcards automatically. The current study used two flashcards to practice productive and receptive recall. Productive recall flashcards require the respondents to recall the L2 word form the L1 word meaning (Happiness____). Receptive recall flashcards require the respondents to

recall the L1 word meaning from the L2 word form (うれしい____). The two flashcards also required the respondents to type the answer, an adaptation from an earlier study (Nakata, 2011) that is not part of the original Anki settings. Instead, we added code to gain this functionality. On the other hand, the respondents did not favor this approach as one needed to switch keyboard language to answer the flashcards. But, according to Nakata, the retrieval effort hypotheses states that adding more retrieval efforts such as typing the answer forces learners to engage with the precise orthographic form, which strengthens form-meaning connections to have the most returns on one's efforts (Pyc & Rawson, 2009; Nakata, 2011).

Anki's interface (see figure 6) is easy to use but has a learning curve for inexperienced users. Luckey, according to the first interview questionnaire mentioned in section 3.1. Most respondents (N = 22) had used Anki previously, which made teaching them how to use the Anki deck unproblematic. After the researchers randomized the respondents in each group, and after excluding respondents from the analysis, 11 respondents who had earlier experience with Anki before remained in the Anki group, and three had used an SRS. Only three respondents had no experience with any SRS but mentioned enjoyment using Anki:" I generally enjoyed using Anki." In the post-interview questionnaire.

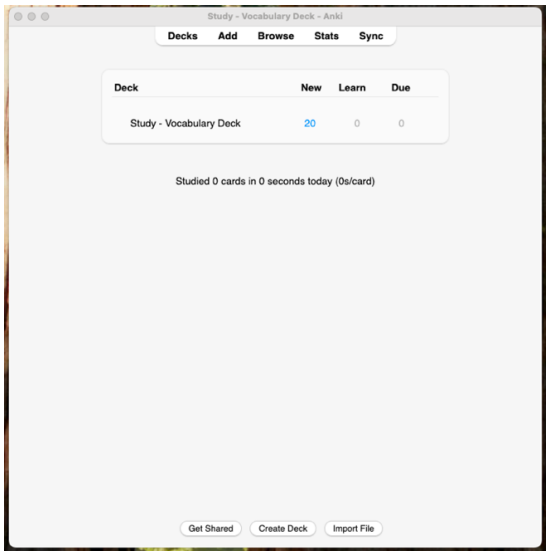
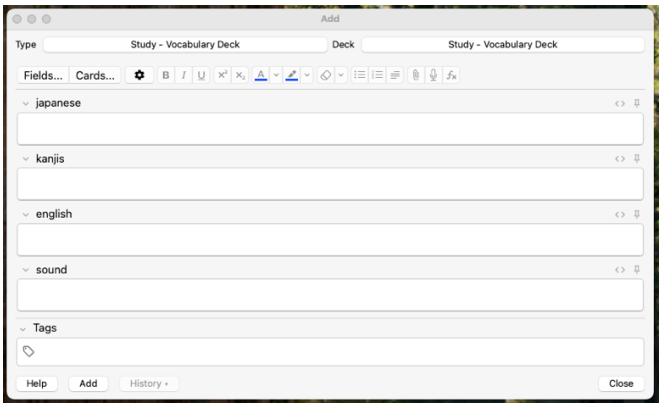


Figure 6 Landing page for the Anki desktop (Mac version)

To create flashcards for the study, we started by pressing the 'Add' button on the main interface. After pressing 'Add,' a screen will appear displayed in Figure 7. This screen is not the default setting; it holds the fields relevant to our current study. We can rename, add, reposition, and remove the fields by pressing the 'Fields' button, as seen in Figure 8.



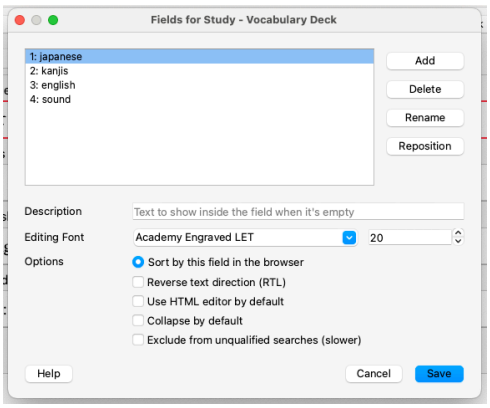


Figure 7 (flashcard creation screen (Right)

Figure 8) (Flashcard field editing screen) (Left)

For this study, the fields include 'Japanese,' which displays the Japanese word written in Hiragana. 'Kanji,' which displays the Japanese word displayed in Kanji characters. 'English,' which is the translation of the Japanese word, and 'Sound,' which is a recording of a native Japanese speaker from Tokyo pronouncing the word. The researcher recorded the sound using a mobile device in a small room. The recording was 15 minutes, which we cut down to 150 single sound files using the program Audacity. We then using a laptop added the sound into the 'sound' field of the flashcard. Figure 9 displays a completed flashcard with all fields filled out.

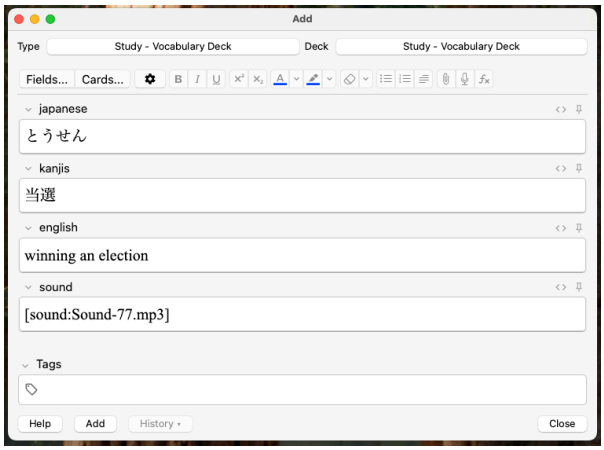


Figure 9 Completed flashcard

To add the typing feature to the flashcards, you can easily follow the instructions provided in the Anki Manual on Ankiweb. By pressing ‘Cards’ in the flashcard’s creation screen, you can incorporate the necessary code into the flashcards. This code enables respondents to answer the card by typing, as shown in Figure 10 for productive recall and figure 11 for receptive recall.

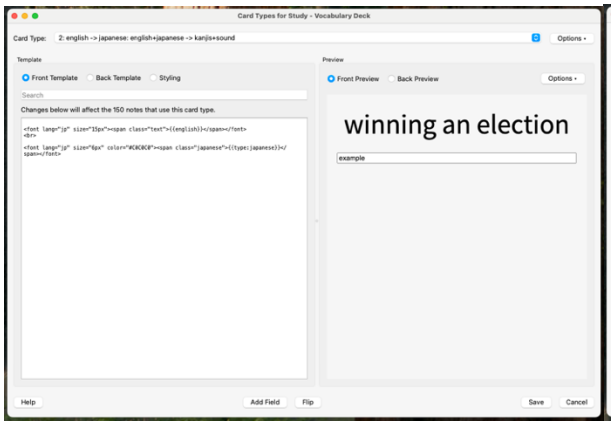


Figure 10 (productive recall flashcard)

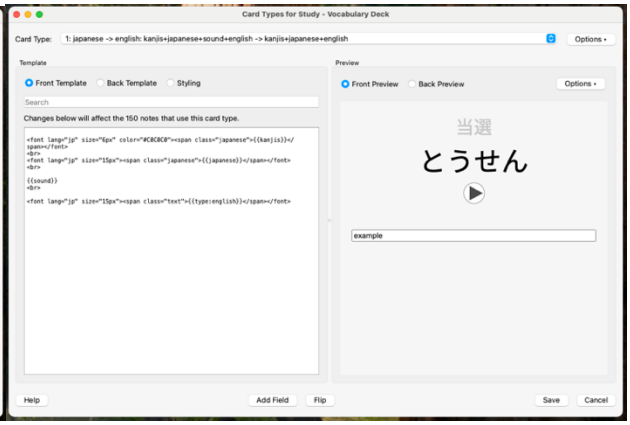


Figure 11 (Receptive recall flashcard)

Not knowing how to code can intimidate some users, but one can add fields to their flashcards by pressing ‘Add Field’ in the code editor screen, displayed in Figure 12. One can then add all the fields they want to show on the front of the flashcard. The researcher added ‘Kanjis’ at the top with half visibility, as we wanted the focus to be on the Hiragana. We then added ‘Japanese’ and then the ‘sound’ field. The last field is the word we wanted the respondent to type. We added so the respondents could answer by typing the answer in the flashcard by writing the code ‘type:’ before the field name, and Anki will do the rest.

{{**type:**english}}

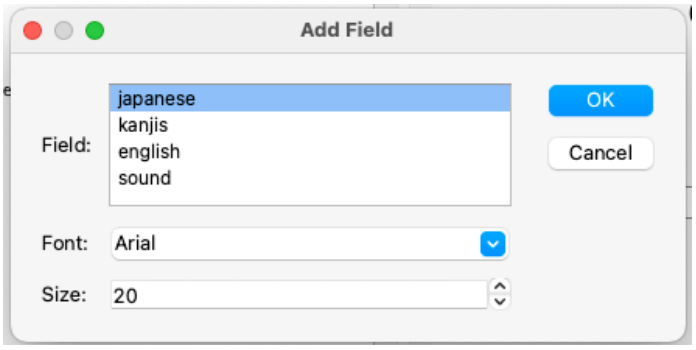


Figure 12 (Add fields to code screen)

We needed to complete the back template to finish the flashcard, as flashcards consist of two sides. To finish the back template, we must add the code {{FrontSide}}, and the flashcard is

complete. However, the current study required a different approach for the productive recall flashcard, so there is a distinction between the two codes. We changed the back template to exclude sound since the front template already plays it. Including the sound on the back would be redundant, so we removed it. One can see the back template code in Figure 13 for productive recall and Figure 14 for receptive recall. Another difference is the position of the Kanji, and we decided that the receptive recall flashcard looked better for it to be under the answer.

One must create a card type for each flashcard one wants for the vocabulary item. The current study had two card types and thus had two flashcards per vocabulary item with 300 cards. One can add a card type by pressing the 'options' button at the top right of the screen and then pressing the 'add card type' button. We must approve and then done. Figure 15 displays the directions of the two card types. 1: Productive and 2: Receptive recall.

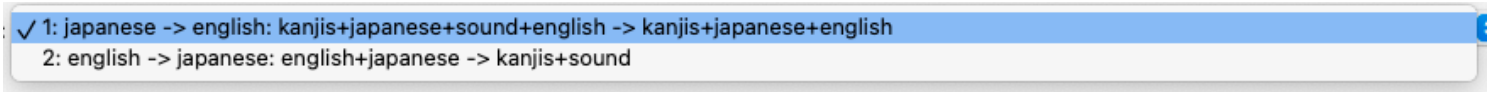


Figure 15 (Add note type menu)

4.2.4 How to use Anki during the treatment

The researcher sent the Anki deck with instructions on how to use Anki deck before the treatment begun. The respondents could then choose when they started studying the vocabulary. As mentioned previously they received two forms of cards to test productive and receptive recall.

Respondents received instructions on creating a new profile before importing the deck. Current Anki users have specific settings, and creating a new profile resets these settings to the

default Anki configuration. By setting this rule, we wanted to minimize the respondents' use of the wrong settings during the study. After creating their profiles, they imported the deck into Anki by either clicking and dragging it directly or double-clicking the Anki deck, which resulted in an automatic import as shown in Figure 16. When the import process succeeded, we can see all 150 new notes as demonstrated in the figure. Each note consisted of two card types for the current study, each standing for one flashcard for a specific word. We learned that one respondent encountered issues and could not import the deck. Currently, we are unsure of the cause. However, the respondents could use the deck on their Android phone via Akidroid, which became their primary device for the study.

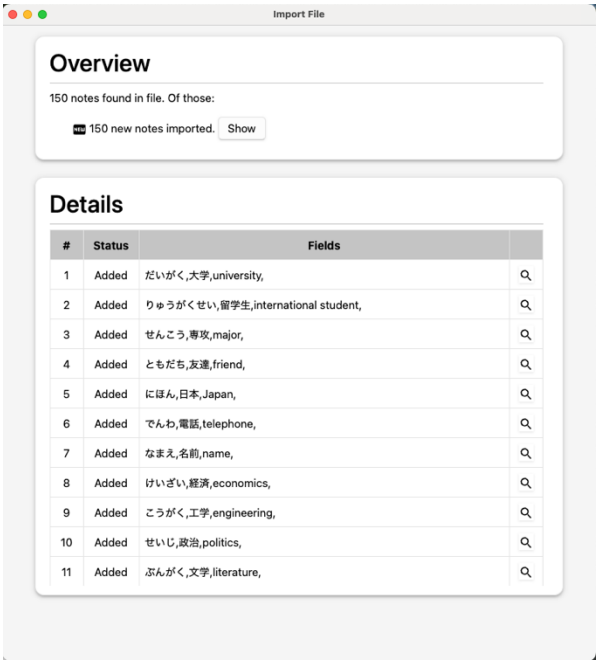


Figure 16 (Successful import of Anki deck)

Figure 16 shows the successful import of the Anki deck, confirming that Anki has the correct settings and that there are now 20 new cards to study (10 vocabulary items). However, three respondents reported 15, 10, and 5 new cards per day due to not creating a new profile. The respondents received instructions to change setting to the correct 20 new cards per day and then continued studying with the deck. Anki uses distinct colors to indicate the status of each card: new cards (blue) are ones you have not seen before; review cards (green) are those you have seen recently and learning cards (red) are cards seen before but answered incorrectly when trying to recall them.

The respondents started using the deck. As mentioned previously, each flashcard came in a fixed order set by the researcher before the study. Productive recall flashcards wanted the respondents to use the hiragana alphabet or click the "show answer" button to receive immediate feedback. The immediate feedback included the L2 word form in hiragana, Kanji characters of the item, a voice recording of the word, and correcting their spelling if they incorrectly typed the answer. Typing an incorrect answer did not automatically penalize the respondent; it only showed the incorrect spelling in red on the feedback screen (see figure 17).

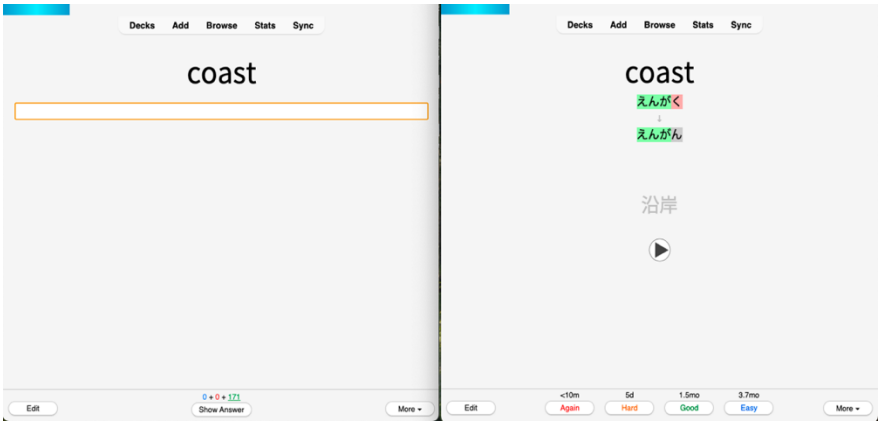


Figure 17 (Productive recall flashcards)

The receptive-recall flashcards displayed the Japanese word in hiragana, Kanji characters, and an audio recording. Respondents then had to produce the L1 meaning of the word by either typing it using the Roman alphabet or clicking the "show answer" button to receive immediate feedback with the L1 translation. We informed the respondents not to be overly concerned about the specific translation in the consent form. However, they answer for the post-test still used the translation from the treatment on the receptive recall test (see Figure 18).

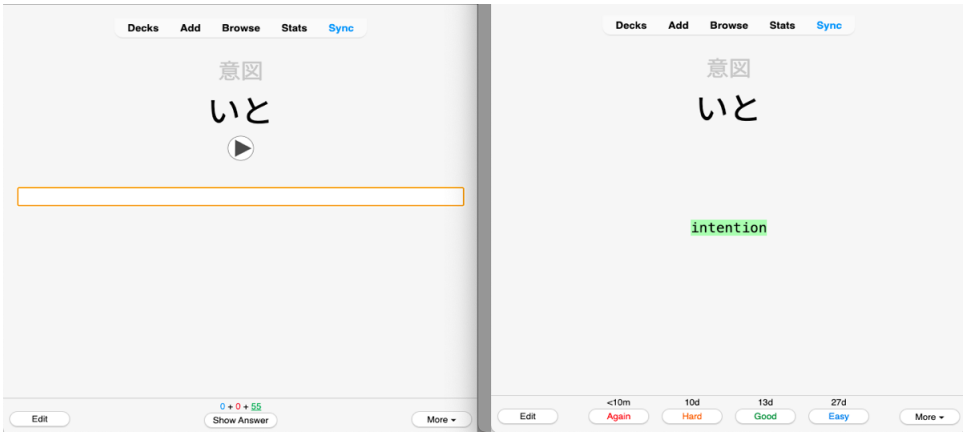


Figure 18 (Receptive recall flashcards)

4.2.5 Anki's algorithm

Anki uses an algorithm based on the SM-2 for supermemo. The algorithm adjusts based on user responses: Each response lets the user self-judge how well they recalled a card. Every card starts with an initial Ease Factor (EF) of 250% (Vermeer, 2017). The algorithm then adjusted the EF as follows. For pressing the 'Easy' button, the EF increases by 15 percent. For pressing the 'Good' button, the EF stays unchanged, and the interval progresses normally. For pressing 'Hard' button, the EF decreases by 15 percent, and the interval decreases. However, still treated as a correct response. Pressing the 'Again' button, resets the card to a short learning interval for

immediate reinforcement (10 min), and the ease factor is decreased by 20%. Each card has its own Ease Factor and its interval, ensuring that the most challenging card for the learner is reviewed more frequently than others.

Anki also uses the Leitner system with the combination of the algorithm to graduate cards that are correctly remembered. The initial step is the learning step of 1 min, 10 min, and 1 day, then it graduates cards, by pushing them to the next day. The card is now at the review stage, and the algorithm kicks in. If a user does not know a card, the algorithm pushes the card back into relearning and stays there until the card graduates from the relearning step again.

An example of a schedule of a difficult card of respondents during the treatment displayed in Table 1. One can also note that the ease factor does not decrease even if the user uses the 'Again' button more than once during the same season.

Review	Response	EF (Ease Factor)	Next Interval	Day
1	Again	250 %	1 min	Marsh 21
2	Good	250 %	10 min	Marsh 21
3	Good	250 %	1 day	Marsh 21
4	Good	250 %	4 days	Marsh 22
5	Good	250 %	10 days	March 26
6	Again	230 %	10 min	April 5
7	Again	230 %	10 min	April 5
8	Good	230 %	1 day	April 5
9	Good	230 %	3 days	April 6
10	Good	230 %	7 days	April 9

Table 1 (Easy Anki card)

In contrast, with a schedule of an easy card where the respondent does not have difficulty in learning, shown in Table 2. One can also take note that first day the ease factor will not change during the first session even if the user presses the ’Again’ button.

Review	Response	EF (Ease Factor)	Next Interval	Day
1	Again	250 %	1 min	Marsh 15
2	Good	250 %	10 min	Marsh 15
3	Again	250 %	1 min	Marsh 15
4	Good	250 %	10 min	Marsh 15
5	Good	250 %	1 day	Marsh 15
7	Easy	265 %	6 days	Marsh 16
8	Easy	280 %	20 days	Marsh 22
9	Good	280 %	2 months	April 11

Table 2 (Hard Anki card)

The algorithm calculates a flashcard by multiplying the previous interval by the ease factor. Additionally, cards that graduate receive a bonus of 1 day. Anki also applies a bonus that ranges from 1.2 to 1.3, depending on how quickly the user answers the cards. When we see that review three and four of Table G, the interval goes from 1 day to 4 days. The calculations are as follows. 1 day (previous interval) + 1 day (graduation bonus ’learning to review stage’) + (2.5 ease factor × 1.2 bonus) = 4 days. Leaving both 10 min and 1 day leads with a 1-day graduation bonus. However, the algorithm is not that simple and adds an extra factor to cards to minimize sequence effect cards that are always shown together are pushed aside via the algorithm.

4.3 Making an online study

The current study included three Google Forms that tested receptive and productive knowledge, as well as two Google Forms used for interviews.

Google form was selected as it is free and accessible to almost everybody. The google forms were created using the quiz format, calculating points automatically from an answer sheet. Each test included instructions on how to perform the test seen in figure 19.

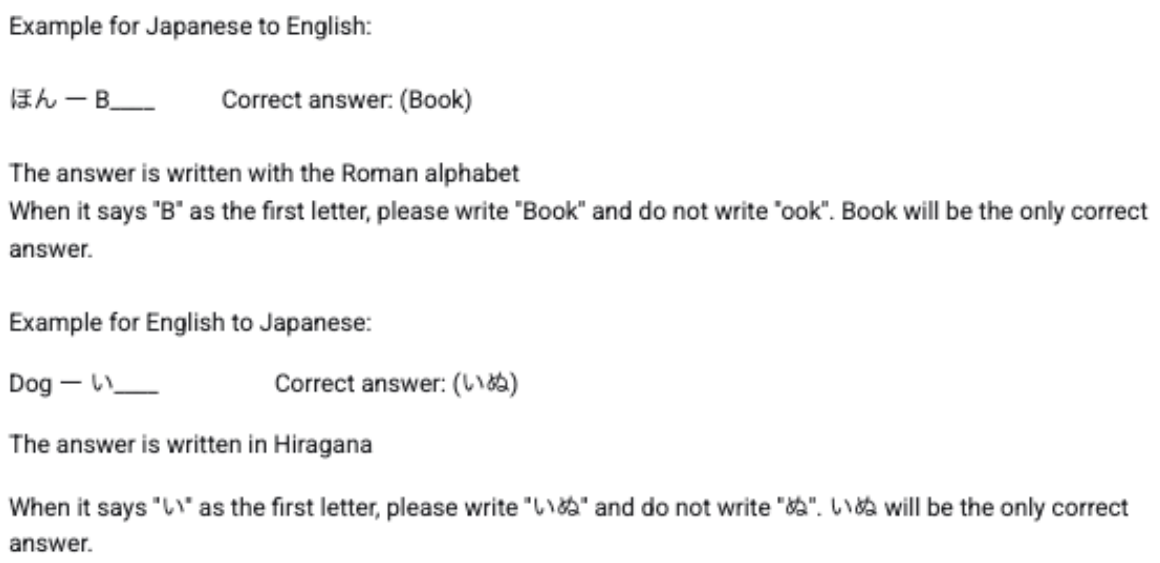


Figure 19 (test instructions)

All tests were tests had a randomized order to minimize any potential serial effect problems. A serial effect occurs when the order in which items are presented influences memory retrieval, with the previous item acting as a cue (Delaney et el., 2011). The tests varied in size, featuring groups of 30, 80, and 90 vocabulary words. The immediate post-test included 80 items

and 50 new vocabulary items. In comparison, the delayed post-test contained 90 items, featuring 40 original vocabulary items from the immediate post-test and 40 new words not previously encountered on the tests.

For the pre-test, the distribution goes as s follows: 20 words from Genki, five from Tobira, and five from Shin Kanzen Master. The pre-test consisted of both a form-recall and meaning-recall task to evaluate productive (15 vocabulary items) and receptive (15 vocabulary items) Japanese vocabulary knowledge. The researcher included initial-letter retrieval cues to assist respondents in identifying the correct target words, especially when synonyms were involved (Nakata, 2011, 2015).

The current study used retrieval cues on the pre-test to help them get the target vocabulary as they had no prior knowledge on what the target vocabulary were. Seen bellow is an example on how it should look.

Form recall: The task in form recall is to supply the L2 target word.

“Happiness” - し _ _ _ (target: しあわせ

Meaning recall: The task in meaning recall is to demonstrate a supply the meaning of the L2 word.

しあわせ - H _ _ _ _ _ _ (translation: Happiness)

However, the current study differs from prior implementations in two ways. First, because many of our test words have near-identical spellings to synonyms (e.g., both りえき and りじゅ

ん mean “profit”), providing the retrieval cue only on the first letter (e.g., “Profit – り _ _ _for り じゅん”) did not distinguish between synonyms, unlike in Nakata (2011, 2015). Instead, the retrieval cue should have been on the second letter (e.g., _ じ _ _) to use the retrieval cue effectually. Second, we did not display the total number of letters in each word with underscores and spaces (e.g., り _____), making it impossible for respondents to infer word length before recalling the target word.

The expected results of the pre-test depend on whether the respondents are beginner or intermediate. For example, we expected beginner students to have little prior knowledge of Japanese, so scoring zero or extremely low on the pre-test. At the same time, we also expected intermediate-level Japanese learners to gain a score of 20 or higher. We did not expect anyone to know any of the JLPT N1 vocabulary, while some might know some from Tobira. The score served solely as a baseline for later comparisons. The pre-test ended by thanking the respondents and informing them that the researcher would contact them shortly.

Additionally, all tests had the same ratio except for the delayed post-test, all tests maintained the same easy-to-complex vocabulary ratio based on the Japanese Language Proficiency Test (JLPT) levels. The delayed post-test included a higher proportion of complex vocabulary explicitly designed to assess retention, which could yield interesting results regarding learning outcomes.

4.4 Procedure

4.4.1 Setting

The entire study took place online. The respondents did all tests and interviews using Google Forms, and communication with most of the respondents occurred via SMS, WhatsApp, and email. There were no physical meetings with most of the respondents, and some remained completely anonymous, so the researcher does not know their identities. Respondent completed the treatment at home, during trips, and mostly late at night, reflecting a realistic view of how language learning might occur in everyday life.

4.4.2 Pilot study

The pre-test and pre-interview questionnaire also served as a pilot study to assess respondents' prior knowledge of the vocabulary items in the study. Additionally, the interview questions were refined based on feedback from the first respondents, improving the study's quality. We allowed some non-respondents to test the study before officerly released. Unfortunately, some mistakes still occurred in the final study. However, respondents informed the researcher that this did not negatively affect the study according to the post-test questionnaire of both conditions. For example, "I didn't notice them," and" No, though I did not catch the 'mistakes'?"

4.4.3 The procedure

Once all the respondents had finished all the pre-study preparations (consent form, questions), each respondent received an email with a google form link to the study's first phase. The study consisted of a Google Form with a pre-interview questionnaire that automatically redirected them to the pre-test upon completion; there was no time limit for these tasks. Each respondent took their time to finish the task, but the researcher sent a reminder email to anyone who had not completed it. After a respondent completed the pre-interview and pre-test, they advanced to phase two of the study. The researcher randomly assigned each respondent to one of the two conditions, the Anki group or the control group, also called the own method group. As mentioned previously, each respondent received the vocabulary through Anki or a word list. The researcher informed the respondents to email when they started to study the vocabulary, which indicated the start of the four-week treatment.

Treatment Anki (Phase 2)

The respondents studied 150 Japanese English word pairs using Anki in the Anki condition. The treatment lasted four weeks, with each respondent working on a single Anki deck and studying 10 new words daily. On the very first day, respondents Anki program introduced the first 10 new words, which amounted to twenty flashcards in total (ten “Japanese-to-English” cards and ten “English-to-Japanese” cards), and studied them until Anki displayed the “Congratulations! You have finished this deck for today” screen (see figure). Anki added 10 new words the

following day, and the first 10 words were due for review. Respondents began each session by reviewing any cards that Anki's spaced-repetition algorithm had scheduled. If the respondent had correctly remembered a card, Anki would have pushed it to be repeated further into the future. If the respondents had forgotten the card (Pushing the 'Again' button), it would reappear later in that session. The respondents repeated studying the cards until all card finished. The respondents had seen all 150 words after fifteen study sessions. Furthermore, the workload became smaller after each season until only the most difficult vocabulary only remained.

Respondents in the Anki condition were not allowed to study the vocabulary outside the Anki program. Once the daily study session ended. Their study time for that day was considered complete. This restriction did not prevent them from encountering or using the words in natural contexts (e.g., during conversation or reading). However, there were clear instructions from engaging in any active vocabulary study of the 150 vocabulary items outside of Anki. This included writing down the words and using handwritten lists or other materials to review the vocabulary outside of the scheduled Anki sessions. In a more realistic setting, Anki would be combined with some form of short-term retention method, such as massing, to prepare before a test or exam, but the Anki condition did not allow this choice.

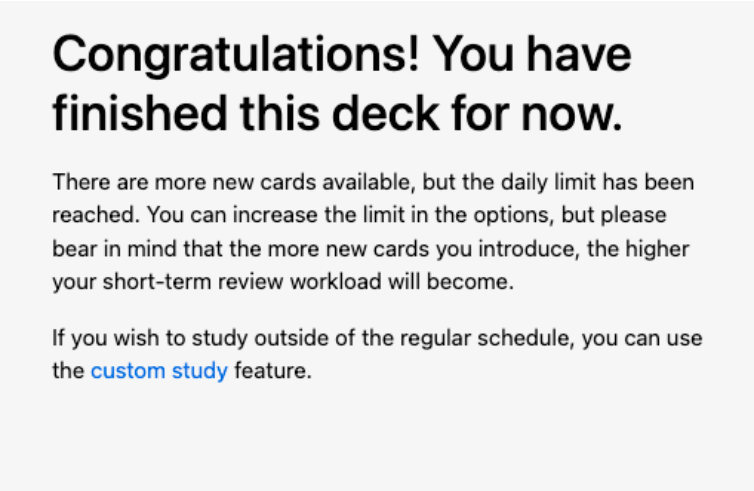


Figure 20 (Anki daily study completion screen)

Treatment control group (own method group) (Phase 2)

In the control condition, 150 Japanese–English word pairs were studied using a printable PDF or Word document. The treatment lasted four weeks, with each respondent working from a single word list and studying with no restrictions, except they were not allowed to use Anki. Respondents in the control condition could use preferred methods to learn the vocabulary. This included, for example, creating Quizlet flashcards, studying with friends, or writing the words out by hand. Unlike the Anki group, the control group followed no tightly controlled parameters or schedules. However, the consent form had instructed each respondent to study at least once daily throughout the four weeks and document each session using the study log.

Phase three of the study

At the end of the four-week treatment, the researcher contacted each respondent through their contact information and sent two google form links: one for the post-interview and one for

the immediate post-test. Each condition received separate interview questions, asking about their experience. Some questions were condition-specific, while others shared similarities across both groups. However, the post-test was identical for all respondents, regardless of condition. After the respondents completed the post-interview and immediate post-test, the researcher followed up with any individuals as needed. These follow-ups included clarifications regarding interview responses and questions about the immediate post-test. Once all the respondents had answered the questions, we informed them that they could not study the vocabulary for approximately two weeks. The respondents received no official date for the delayed post-test to keep respondents unaware of the exact timing; while giving them a general sense of the timeframe so they would be available to complete the test when contacted.

Phase four of the study

At the end of the two-week break, the researcher contacted each respondent again via their contact information and sent them two links: one to the delayed post-test Google. After completing the test, the respondents officially finished the study. The researcher thanked the respondents for their assistance over the six weeks and informed them they would receive a copy of the study once it was complete. If available, the researcher compensated the respondents for their contribution.

4.5 Data Collection Instrument

To address the research questions, the researcher employed both quantitative and qualitative instruments. Collecting quantitative data using three vocabulary tests administered through

Google Forms and gathered qualitative data through two interviews conducted via Google Forms. All data were safely stored on a google form using the university email. Data analyzed were on a laptop and an Ipad and delated after use.

4.5.1 Pre-interview questionnaire

The initial screen of the Google Form provided general information, encouraging respondents to answer truthfully, and that the questionnaire included open-ended questions, yes/no questions, and items rated on a five-point Likert scale: strongly disagree, Disagree, I do not know, agree, and Strongly Agree.

The first questions gathered demographic data, including gender, country, and age. The second section asked open-ended questions about the respondent's language background, including their L1, how long they have studied Japanese, length of stay in Japan, and how many times they have visited. Additionally, if they are or have been taking Japanese classes, what are their general thoughts about language learning, and if they liked studying Japanese? The section ended by asking them if they had taken the JLPT before and what levels they would rate themselves on.

The third section gathered information about the respondents' study habits, such as how many days a week they study Japanese, what methods they use to study Japanese, and what methods they use in and outside the classroom. Lastly, the section asked if they have ever used an SRS such as Anki, Quizlet, Supermemo, etc.

The final section of the questionnaire is an Attitude/Motivation Test Battery (AMTB) designed to assess the respondents' attitudes and motivation toward learning Japanese (see table 5). The test is a modified version based on three studies, two analyzing English and one analyzing Japanese, adapted from Gardner's (1985) research on motivation and attitudes (Okamura, 1990; Ushida, 2005; Effendi, 2022). The test battery is a five-point Likert scale: strongly disagree, Disagree, I do not know, agree, and Strongly Agree. The questionnaire ended with thanking them and providing a Google Form link to the pre-test.

Japanese - English vocabulary pre-test

The respondents received a link to both the pre-test and the interview. When they finished the interview the completed the pre-test as mentioned in making a study. The data were collected and put in a excel file. Each respondent's score had to be calculated one and one which took days.

Post-interview questionnaire

The two variations of the post-interview questionnaire began by congratulating the respondents for completing the four-week treatment and reminded them to answer all questions truthfully. The questionnaire included a combination of open-ended, yes/no, and multiple-choice questions. The researcher then reminded the respondents that the post-test would follow the questionnaire and that both Google Form links are in the email they have received.

The first section asked all respondents to upload data from their treatment period via the Google Form. The Anki group exported and uploaded their Anki deck with all study data intact, while the control group uploaded their study logs. The questionnaire then asked about their

general experiences during the four weeks, for example, if they liked the study method they used. Additional questions followed, asking when, where, and on what device they studied during the treatment.

Followed by questions on whether their regular academic work interfered with the treatment. The control group received further questions about their method, such as which words they prioritized during the treatment. This section concluded with questions about whether they would like to integrate their method into classroom settings and whether they considered it an effective tool for studying Japanese.

The second section focused on the use of sound. Since only the Anki group had sound in their flashcards, the questions are general impressions of sound, such as whether sound helps with retention, whether a native speaker's voice is preferable to AI-generated audio, and whether they found certain Japanese sounds perceptually difficult to distinguish.

The third and final section contained more reflective and specific questions about the respondents' study habits during the treatment. The first question asked how many words they thought they would remember after the upcoming two-week break and whether they had encountered any of those words outside the study context. We also asked whether they believed their Japanese proficiency had improved during the treatment and, if they preferred, a massed learning approach or spaced repetition. The section concluded by asking whether they found the study valuable, and whether the inclusion of Kanji interfered with their ability to learn hiragana. Lastly, the questionnaire ended by thanking the respondents for participating and providing the Google Form link for the immediate post-test.

Japanese - English vocabulary immediate post-test

Following the post-interview questionnaire, the respondents received the immediate post-test. This test is a copy of the pre-test with 50 additional vocabulary items compared to the pre-test, which only had 30. The immediate post-test tested as the pre-test both productive and receptive recall with each form having 40 items each. Respondents completed the immediate post-test on average the day after the treatment concluded (range = 0–3 days, where 0 = same-day completion and 3 = three days later). We expected average accuracy to be around 90% in both groups. However, the two conditions should differ in their preparation. The control group was free to review all 150 vocabulary items before testing, whereas the Anki condition could only study those items that Anki had scheduled for review.

Japanese-English vocabulary delayed post-test

The delayed post-test was administered approximately two weeks (15 days) after the immediate post-test. Respondents took on average, one day to complete the test (range: 0–4 days). The delayed post-test included only the most difficult vocabulary items from the treatment materials, comprising 40 words that appeared on the immediate post-test plus 40 additional items, for a total of 90 words.

Scoring of Pre-test and Post-test protocol

The pre-test and post-test responses were evaluated using two scoring methods: strict and sensitive. These methods were adapted from Nakata's (2013, pp. 50 - 55), study with some modifications to include scoring of receptive recall and because the languages taught in the original study were opposite (Learners of English instead of learners of Japanese).

In the strict scoring method, only perfectly spelled words in the productive recall test were respondent recalled Japanese target words were considered correct. In contrast, the sensitive scoring method employs what Nakata (2013), calls Lexical Production Scoring Protocol (LPSP; e.g., Barcroft & Rott, 2010; Deconinck et el, 2010), which assigns scores of 0.00, 0.25, 0.50, 0.75, or 1.00 based on how many letters in the response match the target word. ''A letter is counted as correct if it appears in the exact same position as in the target word'' (p. 51). However, if another word from the treatment is used as an answer it would be treated as incorrect even if the words are similar.

For the strict scoring of the receptive recall test, misspellings were treated as correct if the intended target word from the treatment was clear (e.g., hygine for hygiene), since the current study is not an English spelling test. Additionally, plural forms or gerund (e.g., speak(ing)) of the target word is treated as correct (p. 145). Finally, only the vocabulary from the treatment is considered as correct and any synonyms are treated as incorrect. In the sensitive scoring method however, synonyms found in a dictionary were also treated as correct and awarded 0.5 points. Allowing leaners that already have previous knowledge of the word or remembered a synonym

be awarded points for understanding the word. Below is an extraction of the scoring method by Nakata (2013), for the productive recall test with some modifications.

> ”1.00: all letters in the response are correct
>
> 0.75: 50% or more but less than 100% of the letters in the response are correct
>
> 0.50: 25% or more but less than 50% of the letters in the response are correct
>
> 0.25: at least one letter in the response is correct or 25% or more but less than
>
> 0.00: all other responses.”
>
> Nakata (2013, p. 51)

An example of sensitive LPSP scoring is when a respondent recalls the response けいさい for the target word けいざい (economy). This response receives a score of 0.75 in LPSP because 75% (3 out of 4 letters: けい_い) of the characters in the response are correct. Another example is when a respondent recall でんとう for the target word でんせつ (folklore). This response is treated as incorrect and does not receive a partial score of 0.50, even though 50% (2 out of 4 letters: でん_ _) of the characters are correct. This is because でんとう is a valid, answer for another vocabulary that is part of the treatment.

An example of sensitive scoring for the receptive recall test is when a respondent produces the response "main character" for しゅじんこう, when the target word is "protagonist." The

response receives a score of 0.50, because the respondent used a valid dictionary definition but did not show full recall of the exact target word from the treatment.

Analysis of the Data

The study used one-way Analysis of Variance (ANOVA) in Excel to determine whether learners using digital flashcards with Anki exhibited differences in vocabulary gains and retention compared to the control group. Pre- and post-tests were analyzed to assess the total words retained between the condtions. Separate ANOVAs were conducted for gain and retention scores, and effect sizes (η^2) were calculated in Excel to determine statistical significance.

In addition, the researcher calculated correlation coefficients in Excel to examine whether individual motivation levels were associated with final vocabulary test scores, specifically testing if more motivated learners achieved higher post-test performance regardless of the study condition. Lastly, semi-structured interviews were conducted to explore respondents' perceptions of Anki's usability and instructional value compared to the control condition.

5.0 Results and Findings

This section will show the Anki and Control condition data and the results. The first research question will try to answer whether the Anki condition has a larger effect on retention on a delayed post-test two weeks after a four-week treatment. The researcher first reviews a couple of conditions that could have caused this but then goes over the results of all the test results. The interview questions are then discussed, and the researcher shows if motivation can be a factor in post-test performance.

5.1 Study time

The current study did not limit any of the conditions on how much they could study, with one caveat: the Anki condition was not allowed to study more than what Anki allowed them to study. However, during the study period, respondents were allowed to review a card for as long as they wanted. In contrast, the control condition the respondents decided their own limitations.

In the post-interview questionnaire, the respondents in the control group were asked, "You were not limited in how much or how little you were able to study. Did you find limiting yourself from studying too much difficult, or was it the opposite?" One of the respondents answered; ”It wasn't difficult to limit the study time. I just studied until I finished everything I wanted to do that day." Another respondent who found it hard said, "Opposite, I had a hard time getting myself to sit down at the table and learn." Which was the most common response in the control condition. However, the data from study time show that the control condition studied more than the Anki condition. On average the Anki condition, used less time to study but studied more total days.

The control condition studied for an average of 23 days (SD = 7.6), and about 16.57 minutes (SD = 9.52) were spent studying daily. These findings are surprising, as the control condition experienced difficulties finding time to study yet still managed to study for around one hours more than the Anki condition (5h 35m vs 4h 20m). While the control condition had fewer study days, it compensated with longer study sessions, which suggests that massing occurred in the control condition. In contrast, the Anki condition had an average of 27 study days (SD = 1.6), with study sessions lasting about 9.2 minutes (SD 7.45) each day.

Study time for both immediate post-test scores and delayed post-test scores did not show statistical significance at p value of 0.05. Figure 10 and 11 illustrates the relationship between study time and scores on the post-test, indicating that increased study time did not result in significant gains. The Y-axis represents study time in minutes, while the X-axis displays the scores. It remains unclear why study time did not have an impact on post-test scores, even within the same condition. However, the Anki condition often landed on around 2 hours with which showing perhaps only studying 6 min per day could be enough if one uses Anki to study vocabulary daily. Even if it did not correlate looking over study time there were only one respondent in the Anki condition that studied for 16 hours according to Anki stats and still did not get a good score at the delayed post-score. Looking at the data from Anki stats we can see that the respondent studied a total of 29 days but had difficulties retaining the vocabulary. The data is not as clear in the control condition and the respondents six respondents almost studied 10 hours, but not all managed to achieve high scores.

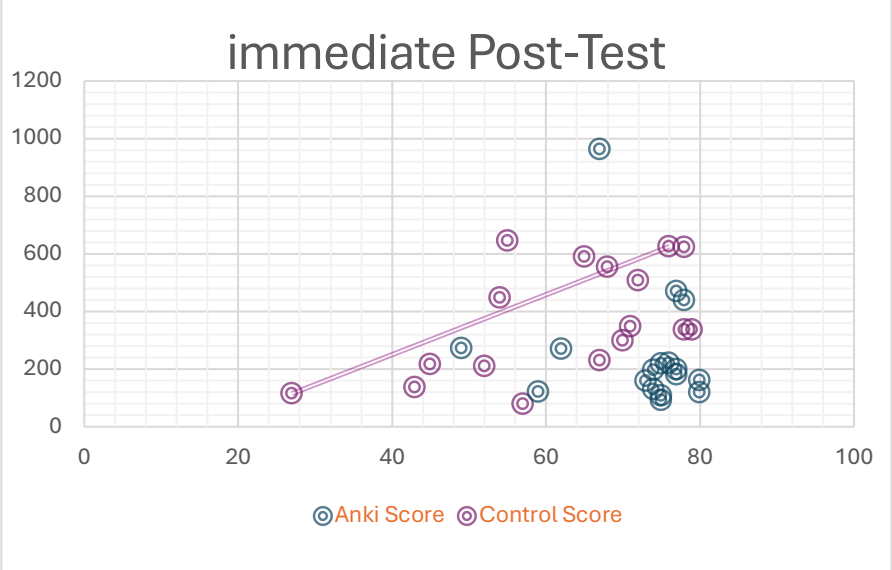


Figure 21 (Time vs score difference on immediate post-test)

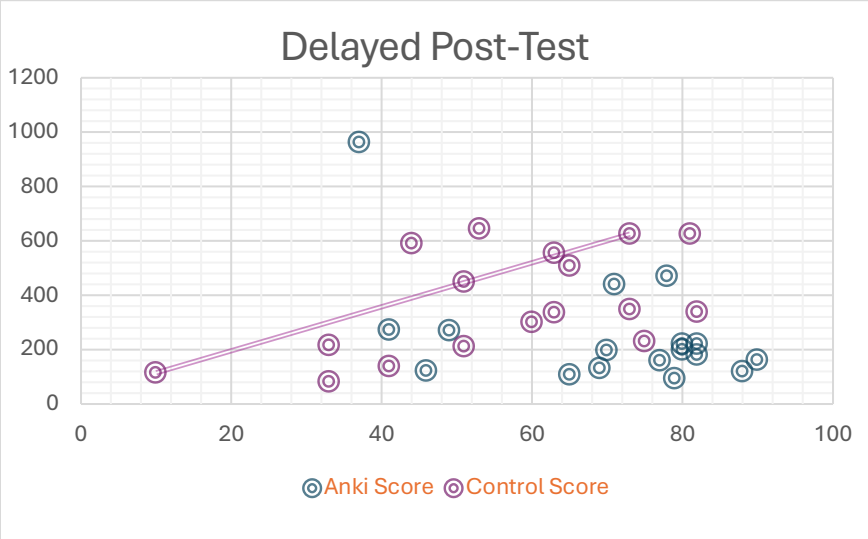


Figure 22 (Time vs score difference on Delayed post-test)

Interference errors

According to Nakata and Suzuki (2019), semantically related vocabulary tends to cause more interference errors than unrelated vocabulary. The current study aimed to excluded semantically related vocabulary from the word list to minimize this effect. However, despite these efforts, some related pairs still appeared. Interestingly, interference errors occurred at similar rates in the Anki and control conditions, showing no significant difference in frequency. Receiving the same results as Nakata and Suzuki (2019), however, the types of interference errors differed between the two conditions.

One of the most frequent types of interference in both conditions were related to similar orthography. Words that looked similar in writing but had different meanings were often mixed up. For example, respondents commonly mixed up でんとう (tradition) and でんせつ (folklore), as well as とうひょう (vote) and とうろん (debate). Additionally, Prior knowledge also contributed to errors. In some cases, words that resembled vocabulary previously learned outside the study caused interference. For example, せいさく (policy) was often confused with せいかく (personality), a word many respondents already knew.

Homonyms introduced another type of interference. The word しゅうかん can mean either “habit” or “week.” While only the “habit” meaning was part of the target vocabulary, many respondents already familiar with “week” from the Genki textbook, which was not included in the study.

Finally, synonym-related interference occurred in both conditions but was more common in the control condition. This often happened when respondents used a dictionary to look up meanings and remembered the dictionary translation rather than the one provided in the study. As a result, some respondents gave answers marked incorrect under the strict scoring method but acceptable under the sensitive scoring method. In some cases, this difference in scoring led to a variation of up to ten points.

The current study tried separate semantically related vocabulary not to be studied together but some interference still occurred. Removing all semantically related words is not a valid approach but a good question for future research. Furthermore, using uncommon words that all respondents might not is a common method to test retention but cannot be used if the study is aimed for a more realistic approach. The vocabulary must be useful for the respondents which can be hard to achieve because only the respondents themselves know what it useful to them.

Pre-test results

The data showed on a One-way ANOVA no significant difference in the pre-test results between the conditions with $f(1, 32) = 1{,}33$, $p = 0.99$, With a negative effect size of $\eta 2 < -0{,}03$. Each condition knew around the same amount of vocabulary from the pre-test, with the average score of 18,23 (3.8) out of 30 with a range of (8 - 23) for the control condition using the strict scoring method. Using the sensitive scoring method the results showed a score of 18.25 (3.7).

The Anki condition also scored 18,23 (4.42) out of 30, with a range of (6 – 25) for the strict scoring method and 18.76 (4.42) for the sensitive scoring. The respondents showed that they

knew a on average 60% of the vocabulary coming into the study. Knowing around 20 vocabulary items from the pre-test indicated that respondents were familiar with around 50% of the entire study vocabulary, which included 50% of the words from Genki (Banno et al., 2020). Therefore, the immediate post-test should show at least 40 points for all respondents who scored 20 or above on the pre-test. Additionally, the vocabulary items from the JLPT N1 Preparation book (3A Network, 2011) were unknown to all respondents. Thus, these will be the best indicator of retention in the two-week delayed post-test. Figure 4 shows the distribution between the conditions.

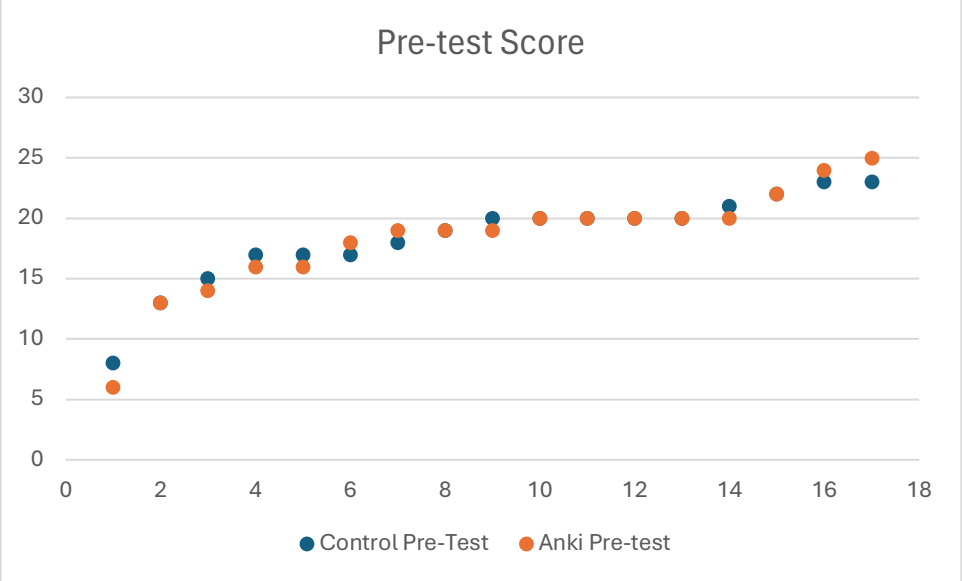


Figure 23 (Pre-test scores between the conditions)

Immediate post-test

On the immediate post-test, the scores showed a significance differed on a One-way ANOVA. The results of f (1, 32) = 6,0, p = 0.01, with a large effect size of η2 < 0,122. The researcher expected the results of both conditions to achieve around identical scores because of the equal pre-test scores and that the immediate post-test is right after the treatment. But it was expected for the Anki condition to score worse as they did not have the advantage of able to practice all vocabulary before the post-test as they could only study what Anki scheduled them to study. Which means that the study design to let the Anki condition first get an initial exposure of all vocabulary items and then lets them review these vocabulary items for around 15 more days resulted in higher post-test scores. The test effect could have played a role as 30 of the items where old items for both the pre-test and the immediate post-test but this was an advantage of both conditions.

It was also predicted that respondents who scored around 20 on the pre-test would score above 40 on the immediate post. Respondents who scored 20 points or above on the pre-test had a mean of 71 (SD 7.6) out of 80. The respondents with scores 20 or less scored 63 (SD 15.54) out of 80, displaying that knowing the vocabulary before the study did affect the immediate test scores. However, the variation between the two conditions that had respondents that scored 20 or above did not differ. Looking at the percentage between the groups on the post-test the Anki condition scores around 90% and the control condition around 77.72%. Further, looking at the difference between the conditions the control condition improved by around 17% after the treatment and the Anki condition around 29.50%. Doing a One-way ANOVA on the differences

between the conditions to see if the treatment had a statical significant difference we can get, f (1, 32) = 5,7, p = 0.02, with a large effect size of η2 < 0,122.

However, this only looks at scores using the strict scoring method. Looking also at the sensitive scoring, we see the difference become smaller but still statistically significant as we get f (1, 32) = 4,5, p = 0.04, with a medium effect size of η2 < 0,09. We can see that the Anki condition scores better than the control condition on both the sensitive and the strict scoring methods. However, it is unclear if they retained the vocabulary better than the control condition. The researcher decided to test the retention of ten of the hardest vocabulary items from the pre-test to see how well the treatment affected these vocabulary items on the immediate test. Five from the productive recall test and five from the receptive recall test. The vocabulary was counted as one point for retained when the respondent did not know the vocabulary before the study and zero if it was known before the study. They could, in total, collect ten points. On average, the control condition gained 5.17 (51%) of the vocabulary, and the Anki condition gained 7.8 (78%) with a statistically significant f (1, 32) = 8,8, p = 0.005, with a large effect size of η2 < 0,188.

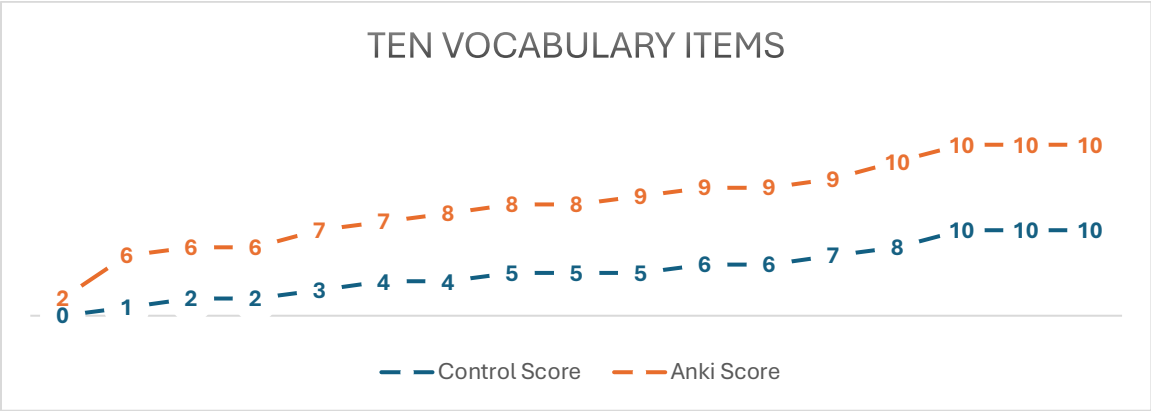


Figure 24 (Immediate test on difficult vocabulary items from the pre-test)

Delayed post-test

The delayed post-test revealed the most significant differences between the conditions. A one-way ANOVA showed the following results: for strict scoring, $F(1, 32) = 4.9$, $p = 0.03$, with an effect size of $\eta^2 = 0.103$; and for sensitive scoring, $F(1, 32) = 4.6$, $p = 0.03$, with an effect size of $\eta^2 = 0.083$. In Table 44, we can see the number of correct responses for both the immediate and delayed post-tests. Taking both productive and receptive scoring into account, the respondents scored similarly on both halves of the tests which do not agree with previous research that says that productive recall is more demanding than receptive recall (Mondria & Wiersma, 2004 ; Schneider et al., 2002).

Table 3 (Strict and Sensitive score of the productive and receptive results)

The control group scored an average of 55 points (SD = 18) out of 90, while the Anki condition scored 69 points (SD = 16) with strict scoring. Using a more sensitive scoring method, the control condition received an average of 66 points (SD = 18), and the Anki condition scored 77.2 points (SD = 15.6). When calculating the difference between the conditions, we found that the control condition forgot approximately 15.67% of the vocabulary, whereas the Anki condition forgot around 12.9%. This indicates that the Anki condition retained more information and achieved better results in less time than the control condition. Table 4 displays the overall scores

Number of Correct Responses of the Post-tests

		Immediate post-test				Delayed post-test			
		Productive		Receptive		Productive		Receptive	
Condition		Strict	Sensitive	Strict	Sensitive	Strict	Sensitive	Strict	Sensitive
Anki	M	36.2	36.8	36	37	35.2	37	34.4	34.34
(N = 17)	SD	5,1	4,8	3,6	3,5	7,8	6,5	9,5	9,5
Control	M	31,2	31,6	30	31,8	28.7	30	27	32
(N = 17)	SD	8,4	8,1	7	7	8,5	12	11	9.4

Note. The maximum score is is 40 on the immediate post-test and 45 on the delayed post-test. Strict = strict scoring; Sensitive = sensitive scoring (see).

and percentages of the post-test compared to the delayed post-test performance.

	Immediate Post-test				Delayed Post-test			
Scoring method	Strict		Sensitive		Strict		Sensitive	
Condition	Correct	%	Correct	%	Correct	%	Correct	%
Anki Condition (N = 17)	1228	90.2%	1255,5	92.3%	1184	77.3%	1236,25	80.80%
Control (N = 17)	1057	77.77%	1096.25	80.60%	951	62.1%	1018	66.5%

Table 4 Maximum Points: Immediate Post-test: 1360 points (17 x 80); Delayed Post-test: 1530 points (17 x 90)

Now that we have looked at the whole now let us look at the ten vocabulary items that we looked at the immediate post-test and how well they were retained between the immediate post-test and delayed post-test. First testing statistical significance with a One-way ANOVA shows that $f(1, 32) = 7.6$, $p = 0.01$, with an effect size of $\eta^2 = 0.76$. There is a statistically significant difference and looking at gains the Anki condition went from knowing 78% to knowing 65% of the vocabulary after two weeks. The control condition went from knowing 51% to knowing 35%. Showing that both conditions lost around the same in percentage but lost more overall. However, on average the Anki condition did score better and retained the information longer. But perhaps a larger sample size and a more controlled control group is needed. Again figure 11 and 55 shows how well each respondent remembered the vocabulary.

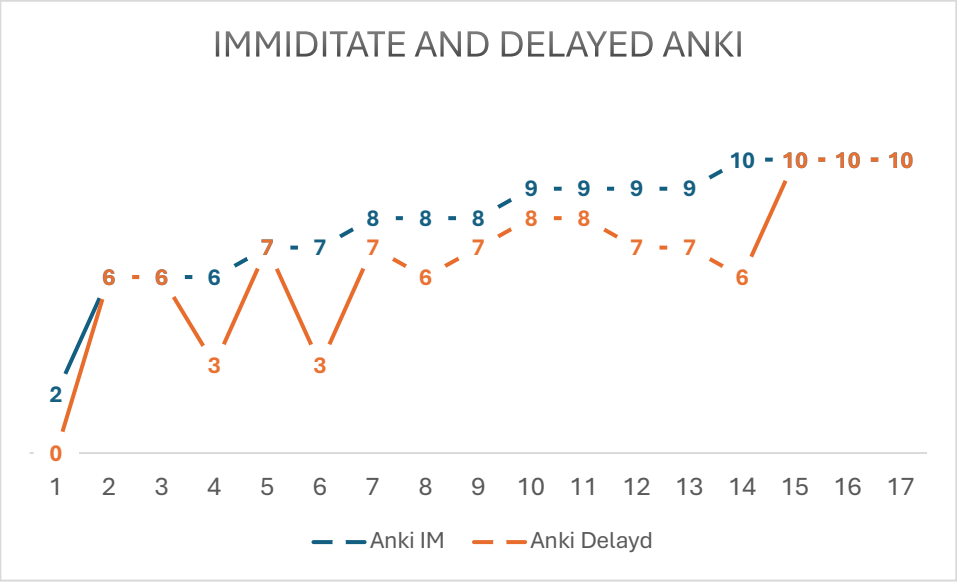


Figure 25 (Immediate and delayed Anki on ten vocabulary items (Anki))

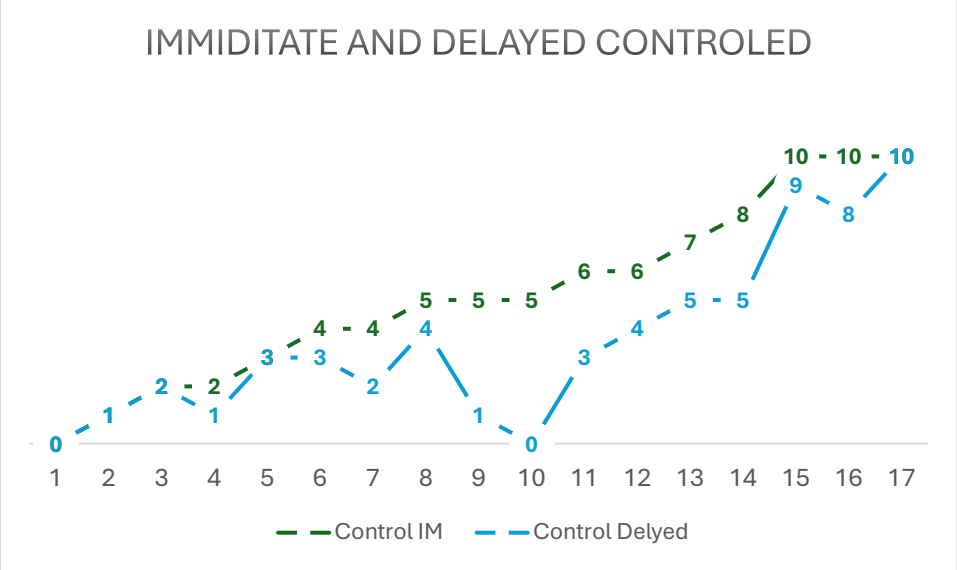


Figure 26 (Immediate and delayed Anki on ten vocabulary items (Control)

Motivation Test

At the end of the first interview questionnaire, we included a motivation test adapted from Okamura (1990) to assess learners' reasons for studying Japanese. We did not do a further attitude test as we only needed motivation from the respondents. The results were compiled in Table 20, which presents data from the conditions separately. All respondents strongly agreed that the primary reasons for learning Japanese included (Q) "to travel to Japan" and (C) "interest in the Japanese language," which aligned perfectly with language learners of Japanese in New Zealand in Okamura's study from 1990. The two groups continue to be similar but differ in the order of their reasons. The Anki group chose (S)" I would like to be able to read Japanese books, newspapers, or magazines." And the control group chose (A)" I am interested in Japanese

culture." This is followed by a point both groups agree on, which is (T)" I would like to have a better understanding of Japan and the Japanese people." Followed by (E)" Japanese proficiency is important to me because it will allow me to get to know various cultures and peoples." For both conditions.

In contrast, the reasons for studying Japanese that were most disagreed upon were (I) for both conditions:" I would like to get a university degree, and Japanese seemed to be the best way to get one." And (G)" I would like to contribute to tourism through my job." For the Anki conditions and (V)" I would like to work for a Japanese company." For the control conditions. This mimics the prior study by Okamura. Both conditions also agree on (P). I'm studying Japanese because it will help me get a good job. It should be noted that both conditions strongly depend on (N)" I only study Japanese when I have to for class."

		Anki Condition		Control Condition	
	AMTB Questions	Mean	S	Mean	S
A.	I am interested in Japanese culture.	4.47	0.8	4.76	0.44
B.	I study Japanese because I want to watch animes/dramas/movies in Japanese.	3.88	1.22	4.24	0.75
C.	I am interested in Japanese language.	4.82	0.39	4.88	0.33
D.	I think knowing Japanese will be useful in getting a good job.	2.76	1.48	3.47	1.37
E.	Japanese proficiency is important to me because it will allow me to get to know various cultures and peoples.	3.94	1.39	4.65	0.79
F.	I would like to catch up with, or be with, my friends who are also learning	3.47	1.74	4.0	1.12
G.	I would like to contribute to tourism through my job.	2.12	0.93	2.59	0.94
H.	I would like to help to establish better relations with Japanese people.	3.47	1.28	3.94	0.75
I.	I would like to get a University degree and Japanese seemed to be the best way to get one.	2.06	1.43	1.94	1.03
J.	I would like to get a job which requires Japanese language.	3.0	1.7	2.88	1.22
K.	I study Japanese as much as possible in my free time.	2.24	0.9	2.71	1.05
L.	I would like to improve my communication with Japanese friends or relatives.	3.24	1.6	3.76	1.25
M.	I would like to live in Japan some day.	3.24	1.52	3.29	1.31
N.	I only study Japanese when I have to for class.	1.65	1.11	1.88	0.6
O.	I would like to be able to teach Japanese in the future.	2.76	1.48	2.53	1.18
P.	I'm studying Japanese because it will help me to get a good job.	2.24	1.25	2.47	1.23
Q.	I would like to travel in Japan.	4.94	0.24	4.88	0.33
R.	Learning Japanese is one of the most important things for me right now.	2.94	1.3	3.41	1.12
S.	I would like to be able to read Japanese books, newspapers or magazines.	4.53	0.87	4.65	1.0
T.	I would like to have a better understanding of Japan and the Japanese people.	4.53	0.51	4.76	0.44
U.	No matter how much I study, Japanese is very difficult.	3.71	1.1	3.88	0.78
V.	I would like to work for a Japanese company.	2.47	1.01	2.47	1.01
W.	I'm studying Japanese because I would like to spend a longer period abroad.	3.88	1.27	4.12	1.05
	*On a five point scale ranging from 1 = ''Strongly disagree'' to 5 = ''Strongly agree''.				

Table 5 (Motivation between the conditions)

The table mimics the study by Okamura and correlates the data between each respondent of the 23 different motivation questions. What can be shown in the data is predictable, such as (P) '' I'm studying Japanese because it will help me to get a good job.'' and (J) ''I would like to get a job which requires Japanese language.'' Correlates well with (D):'' I think knowing Japanese will be useful in getting a good job'' because the respondents who want a job want these aspects. We can also see that (F) ''I would like to catch up with or be with my friends who are also learning'' correlates well with (L) ''I would like to improve my communication with Japanese friends or relatives.'' And (K) ''Learning Japanese is one of the most important things for me right now'' with (R) ''Learning Japanese is one of the most important things for me right now.''

However, several data points do not correlate well, such as (A) ''I am interested in Japanese culture'' and (I) ''I would like to get a university degree, and Japanese seemed to be the best way to get one.'' Respondents who want to know about Japanese culture do not necessarily want a degree. We can also see (N) ''I only study Japanese when I have to for class'' and (T) ''I would like to have a better understanding of Japan and the Japanese people'' do not correlate as one need to use Japanese outside the classroom if they want to achieve this goal.

The results and motivation data showed no coloration between the test results and the amount of motivation they had for studying Japanese. Respondents that scored low on motivation score still scored high on test score. Not all the respondents were current learners of Japanese as some were past learners of Japanese. This can have caused big differences in motivation even if they scored high on the tests. Respondents on the top of the motivation scale of almost five scored on average 68 points on the delayed post-test. Respondents who had the lowest motivation score had only one point less on average with 67 points.

	A	B	C	D	E	F	G	H	I	J	K	L	M	N	O	P	Q	R	S	T	U	V	W
A	1																						
B	0,26	1																					
C	0,14	0,52	1																				
D	0,24	0,32	0,38	1																			
E	0,35	-0,02	0,11	0,44	1																		
F	0,11	0,17	0,10	0,69	0,56	1																	
G	0,32	-0,18	-0,02	0,50	0,45	0,53	1																
H	0,10	0,24	0,28	0,54	0,61	0,53	0,29	1															
I	-0,34	0	0,21	0,42	-0,04	0,30	0,29	0,28	1														
J	0,14	0,19	0,33	0,84	0,37	0,58	0,50	0,54	0,52	1													
K	0,33	-0,03	0,11	0,61	0,35	0,40	0,40	0,16	0,30	0,44	1												
L	-0,08	0,19	0,26	0,57	0,54	0,70	0,22	0,46	0,27	0,43	0,40	1											
M	-0,08	0,07	0,26	0,69	0,41	0,52	0,52	0,56	0,46	0,84	0,30	0,44	1										
N	-0,11	-0,15	-0,02	-0,05	0,13	0,37	0,32	0,02	0,11	0,13	-0,04	0,17	0,15	1									
O	-0,13	0,31	0,40	0,59	0,34	0,62	0,44	0,46	0,45	0,65	0,13	0,59	0,66	0,26	1								
P	-0,02	0,28	0,33	0,78	0,20	0,44	0,38	0,41	0,58	0,72	0,36	0,43	0,65	0,00	0,55	1							
Q	-0,02	0,02	-0,13	0,24	0,08	0,37	0,12	-0,09	0,09	0,20	0,26	0,33	0,21	0,03	0,15	0,09	1						
R	0,20	0,19	0,34	0,73	0,37	0,52	0,39	0,44	0,28	0,57	0,76	0,55	0,45	0,01	0,43	0,61	0,13	1					
S	0,18	0,48	0,72	0,29	0,23	0,12	-0,14	0,27	0,03	0,25	0,25	0,43	0,13	-0,05	0,25	0,16	-0,14	0,36	1				
T	0,23	0,29	0,39	0,49	0,35	0,25	0,28	0,26	0,20	0,40	0,48	0,48	0,28	-0,20	0,22	0,27	0,20	0,47	0,54	1			
U	-0,18	0,17	-0,09	0,04	0,00	0,20	-0,12	-0,09	0,03	0,03	0,14	0,03	0,04	0,19	-0,01	0,04	0,15	0,11	-0,07	-0,03	1		
V	0,29	0,21	0,28	0,57	0,24	0,38	0,43	0,45	0,20	0,63	0,29	0,21	0,58	-0,04	0,25	0,53	0,04	0,51	0,28	0,36	-0,15	1	
W	0,44	0	0,07	0,31	0,40	0,29	0,53	0,07	0,06	0,40	0,42	0,16	0,43	0,29	0,12	0,17	0,46	0,19	0,06	0,32	0,11	0,40	1

Figure 27 (Correlation between the motivations)

Post interview questionnaire

Respondents expressed a positive view of Anki as a vocabulary-learning tool during the treatment. When asked, "*How did you find using Anki as a vocabulary learning tool?*" all respondents answered positively. For example, one respondent said, "*I think it's a good tool for learning vocabulary; it works well.*" Another stated, "*It is great! The spaced repetition system (SRS) is very effective."* Additionally, one respondent mentioned, "I found it very useful and effective! It seems Anki understands how people best remember vocabulary and has designed their app accordingly."

Many respondents were already familiar with Anki and one respondent answered; *"I've used Anki for 3 years, so I'm used to the app. I generally like it, especially when you actually make sure you do the deck every day to not get overloaded."* Even those with less experience found the application intuitive and effective. Several mentioned the satisfaction of daily progress and appreciated the structured nature of the reviews. In contrast, Respondents in the control condition, who used their own methods such as word lists or passive reading, often described their learning process as frustrating, inconsistent, or difficult to sustain.

Another question asked whether respondents ever felt frustrated when they couldn't remember a particular word. This was a common experience across both conditions. Some Respondents in the Anki condition reported feeling frustrated when forgetting a word. Only a few Respondents expressed frustration specifically with the Anki program, and these concerns were

typically linked to minor technical issues, such as bugs or syncing problems, rather than the learning process itself.

In contrast, Nearly all control group respondents expressed dissatisfaction with their method's effectiveness, and some indicated that they did not plan to continue using those methods after the study.

Overall, Respondents in the Anki condition reported a more enjoyable and engaging study experience. Multiple respondents described their study sessions with Anki as "I was impressed by all of the features. I especially liked that card could reappear, and that there was a limit to how much you could revise every day," "good, I like anki," or " Very smooth." In contrast some respondents in the control condition answered, '' Hard to say, id lean towards no'' and a respondent that use Anki outside the study said'' Not anymore because I use anki often these day, it's easier to use when your not home''

Furthermore, when asked if they would continue using Anki after the study, all respondents in the Anki condition responded with yes. When asked, "*Do you think Anki is an effective tool for people to learn Japanese?*" all respondents agreed that it is. Likewise, every respondent in the Anki group expressed interest in incorporating Anki into classroom settings, suggesting that they viewed it not only as effective for self-study but also as a valuable complement to formal instruction. In contrast the control condition said, '' *It might not be an ideal tool for the classroom but i think it could as a supplementary tool/method to language learning in general, particularly in terms of approaching new vocab.*''

6.0 Discussion and Conclusion

The answer to the first research question ‘’Is there a significant difference in vocabulary gains between respondents using digital flashcard via spaced repetition software Anki compared to other study methods,’’ is yes there is a difference between the Anki condition and the control condition on gains. On both the immediate post-test scores and the delayed post-test score the respondents in the Anki condition achieved higher scores overall. Anki uses expanding spacing and spacing as motioned in the literature review helps with long-term retention of information. The Anki condition scored better overall. This was despite the advantages the other conditions had of the ability to skip words and longer study times. The one advantage Anki had was sound and more structure. Which could have been a factor that sound caused the vocabulary to lead to more gain if we take dual code theory into account.

The second research question asked; ‘’Is there a significant difference in vocabulary retention between digital flashcard using the spaced repetition software Anki compared to other study methods.’’ There is a slight statistical significance on the delayed post-test scores that showed that showed that the Anki condition had better retention compared the control condition but only slightly. The results also show the data from the ten vocabulary items and on average only 3% difference between the conditions and the Anki condition loses more because it also gained more. However, if the study ran for longer there is a possibility that this lead would grow stronger and not weaker. Even if only a slight the Anki condition showed to have better post-test scores. One should take note that the respondents varied and some used Anki for the first time. If

this study were not online the respondents could have received more help, and a lot of problem would dimmish.

The third research question and the last one asked, is: "What are the respondents' perceptions and experiences regarding the effectiveness and usability of the Anki program compared to method's used in the control condition?" This question was designed to determine whether Anki is a viable program and demonstrate to educators that it is a user-friendly tool that can lead to improved test scores. The first research question asked yes it probably can do that but also the Anki program was well liked during the treatment. Every respondent in the Anki condition thought it could be helpful in the classroom. However, more research is needed to determine whether this is an option. Future research needs to assess Anki in a more structured way with both pre-test and post-test results. This also includes adding multimedia options to Anki and testing whether it can be integrated into the classroom efficiently. Past research has added it to the classroom, such as in the medical field. However, there is little to be known about spaced repetition software. Furthermore, further research needs to prove that equal spacing in a longitude study where flashcards are continually added is viable compared to expanding spacing.

6.1 Limitations of the current study

The current study indicated that people using Anki can achieve higher outcomes than other methods. However, the study has some limitations. One major issue was the lack of clear instructions for completing the treatment. The researcher received emails from respondents seeking clarification, even during the treatment, which negatively impacted the study. For

instance, some respondents had incorrect settings, while others followed rules inconsistent with the guidelines. The study ultimately needed more structure, which would have made the data easier to calculate and analyze. Another limitation is that it is an online study. The researcher's lack of observation and reminders caused some respondents to forget to do the treatment and many skipped days. Sometimes the respondents took a long time to answer, and the unclear communication caused study problems. For example, the post-test and delayed-post-test had on average couple of days difference between the respondents as the researcher had not created clear instructions to answer emails in due time.

Additionally, the interview questions were primarily random, lacked structure, and did not effectively follow previous research. However, free qualitative questions with detailed answers could have interesting data. The current study asked for too much from the respondents. This study could have achieved the same goal in two weeks instead of six weeks.

7.0 References

Banno, E., Ikeda, Y., Ohno, Y., Shinagawa, C., & Tokashiki, K. (2020).

"Genki 1: An integrated course in elementary Japanese 1" (3rd ed.). The Japan Times.

Bower, J. V., & Rutson-Griffiths, A. (2016). The relationship between the use of spaced repetition software with a TOEIC word list and TOEIC score gains. Computer Assisted Language Learning, 29(7), 1238–1248. https://doi.org/10.1080/09588221.2016.1222444

Bueno-Alastuey, M. C., & Nemeth, K. (2020). Quizlet and podcasts: effects on vocabulary acquisition. Computer Assisted Language Learning, 35(7), 1407–1436. https://doi.org/10.1080/09588221.2020.1802601

Barcroft, J., & Rott, S. (2010). Partial Word Form Learning in the Written Mode in L2 German and Spanish. Applied Linguistics, 31, 623-650. http://dx.doi.org/10.1093/applin/amq017

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), Metacognition: Knowing about knowing (pp. 185–205)., MIT Press.

Cakmak G. (2021). Evaluation of Scientific Quality of YouTube Video Content Related to Umbilical Hernia. Cureus, 13(4), e14675. https://doi.org/10.7759/cureus.14675

Carpenter, S. K., & Olson, K. M. (2012). Are pictures good for learning new vocabulary in a foreign language? Only if you think they are not. Journal of Experimental Psychology: Learning, Memory, and Cognition, 38(1), 92–101. https://doi.org/10.1037/a0024828

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. Psychological bulletin, 132(3), 354–380. https://doi.org/10.1037/0033-2909.132.3.354

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: a temporal ridgeline of optimal retention. Psychological science, 19(11), 1095–1102. https://doi.org/10.1111/j.1467-9280.2008.02209.x

Chun, D. M., & Plass, J. L. (1996). Effects of multimedia annotations on vocabulary acquisition. Modern Language Journal, 80(2), 183–198. https://doi.org/10.2307/328635

Cooper, S., Twardowski, N., Vogel, M., Perling, D., & Ryznar, R. (2023). The Effect of Spaced Repetition Learning Through Anki on Medical Board Exam Performance. International Journal of Medical Students, 11(4), 271–275. https://doi.org/10.5195/ijms.2023.1549

Deconinck, J., Boers, F., & Eyckmans, J. (2010). Helping learners engage with L2 words: the form-meaning fit. AILA REVIEW, (23), 95–114. https://doi.org/10.1075/aila.23.06dec

Deng, F., Gluckstein, J. A., & Larsen, D. P. (2015). Student-directed retrieval practice is a predictor of medical licensing examination performance. Perspectives on medical education, 4(6), 308–313. https://doi.org/10.1007/s40037-015-0220-x

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and testing effects: A deeply critical, lengthy, and at times discursive review of the literature. Psychology of Learning and Motivation, 53, 63–147.doi:10.1016/S0079-7421(10)53003-2

Dodigovic, M. (2013). Vocabulary learning with electronic flashcards: Teacher design vs. student design. Voices in Asia Journal, 1(1), 15–33

Dunlosky, J., & O'Brien, A. (2022). The power of successive relearning and how to implement it with fidelity using pencil and paper and web-based programs. Scholarship of Teaching and Learning in Psychology, 8(3), 225–235. https://doi.org/10.1037/stl0000233

Ebbinghaus, H. (1964). "Memory: A contribution to experimental psychology" (H. A. Ruger, C.E. Bussenius, & E. R. Hilgard, Trans.). Dover Publications. (Original work published 1885)

Effendi, P. P. (2022). Attitude-motivation of learning English as L2: Study case of Airlangga linguistics students. Loquen: English Studies Journal, 15(2), 115128. https://doi.org/10.32678/loquen.v15i2.7732

Ellis, N. C. (1995). The psychology of foreign language vocabulary acquisition: Implications for CALL. Computer Assisted Language Learning, 8, 103–128.doi:10.1080/0958822940080202

Farhadi, P. (2012). The effect of Leitner's learning box on the improvement of vocabulary teaching and learning (Case study: First year students at Parsabad Moghan Branch, Islamic Azad University, Parsabad Moghan, Iran). Academic Discourse: An International Journal

Feng, K., Zhao, X., Liu, J., Cai, Y., Ye, Z., Chen, C., & Xue, G. (2019). Spaced learning enhances episodic memory by increasing neural pattern similarity across repetitions. The Journal of Neuroscience, 39(27), 5351–5360. https://doi.org/10.1523/JNEUROSCI.2741-18.2019

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

Gardner, R. C. (1985). Social Psychology and Second Language Learning: The Role of Attitudes and Motivation. London: Edward Arnold. https://doi.org/10.1037/h0083787

Gilbert, M. M., Frommeyer, T. C., Brittain, G. V., & al. (2023). A cohort study assessing the impact of Anki as a spaced repetition tool on academic performance in medical school. "Medical Science Educator", 33, 955–962. https://doi.org/10.1007/s40670-023-01826-8

Goldman, M., Bryan, J., & Lucke-Wold, B. (2024). Evidence-based educational algorithm "Anki" for optimization of medical education. Journal of Biomed Research, 5(1), 1–7. https://doi.org/10.46439/biomedres.5.037

Gyllstad, H., Sundqvist, P., Sandlund, E., & Källkvist, M. (2023). Effects of Word Definitions on Meaning Recall: A Multisite Intervention in Language-Diverse Second Language English Classrooms. Language Learning, 73(2), 403-444. https://doi.org/10.1111/lang.12527

Hanson, A., & Brown, C. (2019). Enhancing L2 learning through a mobile-assisted spaced-repetition tool: An effective but bitter pill? "Computer Assisted Language Learning", 33(1), 1–23. https://doi.org/10.1080/09588221.2018.1552975

Harris, D. M., & Chiang, M. (2022). An Analysis of Anki Usage and Strategy of First-Year Medical Students in a Structure and Function Course. Cureus, 14(3), e23530. https://doi.org/10.7759/cureus.23530

Hulstijn, J. H. (2001). Intentional and incidental second language vocabulary learning: A reappraisal of elaboration, rehearsal, and automaticity. In P. Robinson (Ed.), Cognition and second language instruction (pp. 258–286). Cambridge University Press.

Iravi, Y., & Malmir, A. (2023). The effect of lexical tools and applications on L2 vocabulary learning: a case of English academic core words. Innovation in Language Learning and Teaching, 17(3), 636–649. https://doi.org/10.1080/17501229.2022.2102638

Jape, D., Zhou, J., & Bullock, S. (2022). A spaced-repetition approach to enhance medical student learning and engagement in medical pharmacology. BMC Medical Education, 22, 337. https://doi.org/10.1186/s12909-022-03324-8

Jia, W., Pack, A., Guan, Y., Zhang, L., & Zou, B. (2023). The influence of game-based learning media on academic English vocabulary learning in the EFL context. Computer Assisted Language Learning, 1–25. https://doi.org/10.1080/09588221.2023.2276800

Jaya, Ervan. (2020). Using Anki (A computer-based flashcard program) in improving student's vocabulary. 10.31219/osf.io/3j5kc.

Kanayama, K., & Kasahara, K. (2016). Is Expanding Spacing More Effective than Equal Spacing for L2 Vocabulary Learning? ARELE, 31, 1–15.

Kang, S. H., Lindsey, R. V., Mozer, M. C., & Pashler, H. (2014). Retrieval practice over the long term: should spacing be expanding or equal-interval?. Psychonomic bulletin & review, 21(6), 1544–1550. https://doi.org/10.3758/s13423-014-0636-z

Kaitsu, T., & Nakata, T. (2025). Analysis of smartphone-based flashcard apps for second language vocabulary acquisition. Computer Assisted Language Learning. Advance online publication. https://doi.org/10.1080/09588221.2025.2481396

Khoshsima, H. and Khosravi, M. (2021). Vocabulary Retention of EFL Learners through the Application of ANKI, WhatsApp and Traditional Method. Journal of Foreign Language Teaching and Translation Studies, 6(4), 77-98. doi: 10.22034/efl.2022.325424.1136

Koleini, N., Boroughani, T., Eslami, Z.R., and Xodabande, I. (2024). Exploring the impacts of mobile-assisted learning on university students' technical vocabulary knowledge, International Journal of Educational Research Open, Volume 7

Kornell, N. (2009). Optimizing learning using flashcards: Spacing is more effective than cramming. "Applied Cognitive Psychology", 23(9), 1297–1317. https://doi.org/10.1002/acp.1537

Kim, S.K. and Webb, S. (2022), The Effects of Spaced Practice on Second Language Learning: A Meta-Analysis. Language Learning, 72: 269-319. https://doi.org/10.1111/lang.12479

Logan, J. M., & Balota, D. A. (2008).Expanded vs. equal interval spaced retrieval practice: Exploring different schedules of spacing and retention interval in younger and older adults. "Aging, Neuropsychology, and Cognition", 15(2), 257–280. https://doi.org/10.1080/13825580701322171

Lado, R. (1964). Language teaching: A scientific approach.

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), Practical aspects of memory (pp. 625–632). Academic Press.

Lei, Y., & Reynolds, B. L. (2022). Learning English vocabulary from word cards: A research synthesis. Frontiers in Psychology, 13, 984211. https://doi.org/10.3389/psyg.2022.984211

Leitner, S. (1972) So lernt man lernen. Freiburg: Herder (16th impression 1991).

Larchen Costuchen, A., Darling, S., & Uytman, C. (2020). Augmented reality and visuospatial bootstrapping for second-language vocabulary recall. Innovation in Language Learning and Teaching, 15(4), 352–363. https://doi.org/10.1080/17501229.2020.1806848

Levy, J., Ely, K., Lagasca, G., Kausar, H., Patel, D., Andersen, S., Georges, C., & Simanton, E. (2023). Exploring Anki Usage Among First-Year Medical Students During an Anatomy &

Physiology Course: A Pilot Study. Journal of Medical Education and Curricular Development, 10. https://doi.org/10.1177/23821205231205389 (Original work published 2023)

Lu, M., Farhat, J. H., & Beck Dallaghan, G. L. (2021). Enhanced Learning and Retention of Medical Knowledge Using the Mobile Flash card Application Anki. Medical science educator, 31(6), 1975–1981. https://doi.org/10.1007/s40670-021-01386-9

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. Contemporary Educational Psychology, 16(2), 192–201. https://doi.org/10.1016/0361-476X(91)90037-L

Mondria, J.-A. (2003). The effects of inferring, verifying, and memorizing on the retention of L2 word meanings: An experimental comparison of the "Meaning-Inferred Method" and the "Meaning-Given Method." Studies in Second Language Acquisition, 25, 473–499. doi:10.1017/S0272263103000202

Mondria, J. A., & Mondria-de Vries, S. (1994). Efficiently memorizing words with the help of word cards and 'hand computer': Theory and applications. System, 22(1), 47-57. https://doi.org/10.1016/0346-251x(94)90039-6

Mondria, J. A., & Wiersma, B. (2004). Receptive, productive, and receptive + productive L2 vocabulary learning: What difference does it make? In P. Bogaards, & B. Laufer (Eds.), Vocabulary in a Second Language: Selection, Acquisition and Testing (pp. 79-100). John Benjamins Publishers. https://doi.org/10.1075/lllt.10.08mon

Mujahidah, Hasanah, N., Yusuf, M., Zulfah, & Fatmasyamsiar, A. A. (2024). The implementation of AnkiApp to improve students' vocabulary mastery. "Southeast Asia Language Teaching and Learning (SALTeL) Journal", 7(1), 9–18. https://doi.org/10.35307/saltel.v7i1.115

Murre, J. M., & Dros, J. (2015). Replication and Analysis of Ebbinghaus' Forgetting Curve. PloS one, 10(7), e0120644. https://doi.org/10.1371/journal.pone.0120644

Nation, P. (2001). "Learning vocabulary in another language". University Press. http://dx.doi.org/10.1017/CBO9781139524759

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

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. https://doi.org/10.1017/S0958344008000219

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. https://doi.org/10.1080/09588221.2010.520675

Nakata, T. (2013). Optimising second language vocabulary learning from flashcards.

Unpublished doctoral dissertation, the Victoria University of Wellington, New Zealand.

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. doi:10.1017/S0272263114000825

Nakata, T. (2019). Learning words with flash cards and word cards. In S. Webb (Ed.), "The Routledge handbook of vocabulary studies" (pp. 304–329). Routledge. https://doi.org/10.4324/9780429291581

Nakata, T., & Suzuki, Y. (2019).Effects of massing and spacing on the learning of semantically related and unrelated words. "Studies in Second Language Acquisition", 41, 287–311.

https://doi.org/10.1017/S0272263118000219

Nakata, T., Suzuki, Y. and He, X. (2023), Costs and Benefits of Spacing for Second Language Vocabulary Learning: Does Relearning Override the Positive and Negative Effects of Spacing?. Language Learning, 73: 799-834. https://doi.org/10.1111/lang.12553

Nender, A. J., Muntuuntu, M., & Rombepajung, P. (2022). Increasing students' vocabulary by using Anki-flashcard (A study conducted at SMP Berea Tondano). JoTELL: Journal of Teaching English, Linguistics, and Literature, 1(6), 707–719.

Okamura, Y. (1990). Motivation and attitudes in learning Japanese in New Zealand (Master's thesis, University of Canterbury). University of Canterbury. School of Educational Studies and Human Development. http://hdl.handle.net/10092/2132

Oka, M., Lawrence, N., Iwasaki, Y., Kondo, M., & Siegel, M. (2009). Tobira: Gateway to advanced Japanese: Learning through content and multimedia. Tokyo: Kurosio Publishers.

Ozer, E., Yağmur, Koçoğlu, & Zeynep, Z. (2017). The Use of Quizlet Flashcard Software and Its Effects on Vocabulary Learning. https://doi.org/10.1501/Dilder_0000000238

Paivio, A., & Desrochers, A. (1980). A dual-coding approach to bilingual memory. Canadian Journal of Psychology / Revue canadienne de psychologie, 34(4), 388–399. https://doi.org/10.1037/h0081101

Pimsleur, P. (1967), A Memory Schedule. The Modern Language Journal, 51: 73-75. https://doi.org/10.1111/j.1540-4781.1967.tb06700.x

Polly, D., Reinke, L. T., Colonnese, M. W., & Blackwelder, A. (2025). Examining differences between games and pictorial flashcards on multiplication basic fact fluency. The Journal of Educational Research, 118(2), 77–89. https://doi.org/10.1080/00220671.2024.2446889

Plass, J. L., & Jones, L. C. (2005). Multimedia Learning in Second Language Acquisition. In R. E. Mayer (Ed.), The Cambridge handbook of multimedia learning (pp. 467–488). Cambridge University Press. https://doi.org/10.1017/CBO9780511816819.030

Pyc, M. A., & Rawson, K. A. (2007). Examining the efficiency of schedules of distributed retrieval practice. Memory & Cognition, 35, 1917–1927. doi:10.3758/BF03192925

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? Journal of Memory and Language, 60, 437–447.doi:10.1016/j.jml.2009.01.004

Ramonda, K. (2022). A double-edged sword: Metaphor and metonymy through pictures for learning idioms. International Review of Applied Linguistics in Language Teaching, 60(3), 523–561. https://doi.org/10.1515/iral-2018-0336

Rogers, J., & Cheung, A. (2020). Input spacing and the learning of L2 vocabulary in a classroom context. "Language Teaching Research", 24(5), 616–641. https://doi.org/10.1177/1362168818805251

Schmitt, N. (2008). Review article: Instructed second language vocabulary learning. Language Teaching Research, 12, 329–363. doi:10.1177/1362168808089921

Schneider, V. I., Healy, A. F., & Bourne, L. E. (2002). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. Journal of Memory and Language, 46, 419–440. doi:/10.1006/jmla.2001.2813

Schmitt, N. (2023). Norbert Schmitt's essential bookshelf: Formulaic language. Language Teaching, 56(3), 420–431. https://doi.org/10.1017/S0261444822000039

Schuetze U., Weimer-Stuckmann G. (2010). Virtual vocabulary: Research and learning in lexical processing. CALICO Journal, 27, 517–528.

Schuetze U., Weimer-Stuckmann G. (2011). Retention in SLA Processing. CALICO Journal, 28, 460–472.

Serrano, R., & Huang, H.-Y. (2018). Learning vocabulary through assisted repeated reading: How much time should there be between repetitions of the same text? "TESOL Quarterly", 52, 971–994. https://doi.org/10.1002/tesq.445

Shahipanah, A., Khajavy, G. H., & Elahi Shirvan, M. (2025). The effect of textual and textual-pictorial glosses on incidental vocabulary learning in mobile-assisted listening. ReCALL, 37(1), 79–95. doi:10.1017/S0958344024000193

Sonbul, S., Macis, M., & Gyllstad, H. (2024). The effect of equal versus expanding spacing practice on the deliberate learning of L2 collocations. TESOL Quarterly. Advance online publication. https://doi.org/10.1002/tesq.3364

Steinel, M. P., Hulstijn, J. H., & Steinel, W. (2007). Second language idiom learning in a paired-associate paradigm: Effects of direction of learning, direction of testing, idiom imageability, and idiom transparency. Studies in Second Language Acquisition, 29, 449–484. doi:10.1017/S0272263107070271

Storm, B. C., Bjork, R. A., & Storm, J. C. (2010). Optimizing retrieval as a learning event: When and why expanding retrieval practice enhances long-term retention. "Memory & Cognition", 38(2), 244–253. https://doi.org/10.3758/MC.38.2.244

Strauss EJ, Markus DH, Kingery MT, Zuckerman J, Egol KA. Orthopaedic resident burnout is associated with poor in-training examination performance. J Bone Joint Surg Am. 2019 Oct 2;101(19):e102.

Teng, M. F. (2022). The effectiveness of multimedia input on vocabulary learning and retention. Innovation in Language Learning and Teaching, 17(3), 738–754. https://doi.org/10.1080/17501229.2022.2131791

Thorndike, E. L. (1908). Memory for paired associates. Psychological Review, 15,

122–138. doi:10.1037/h0073570

Ushida, Eiko. (2005). The Role of Students' Attitudes and Motivation in Second Language Learning in Online Language Courses. CALICO Journal. 23. 49-78. 10.1558/cj.v23i1.49-78.

Vermeer, A. (2017). "Anki essentials: The complete guide to remembering anything with Anki (Version 1.1)". Foggy Mountain Pass. https://foggymountainpass.com

Verkoeijen, P. P.J. L., Rikers, R. M. J. P., & Ozsoy, B. (2008). Distributed rereading can hurt the spacing effect in text memory. Applied Cognitive Psychology, 22(5), 685

695. https://doi.org/10.1002/acp.1388

Webb, S., Yanagisawa, A., & Uchihara, T. (2020). How effective are intentional vocabulary-learning activities? A meta-analysis. Modern Language Journal, 104(4), 715–

738. https://doi.org/10.1111/modl.12671

Whitmer, Daphne & Johnson, Cheryl & Marraffino, Matthew. (2022). Examining Two Adaptive Sequencing Approaches for Flashcard Learning: The Tradeoff Between Training Efficiency and Long-Term Retention. 10.1007/978-3-031-05887-5_10.

Vlach, H. A., Sandhofer, C. M., & Bjork, R. A. (2014). Equal spacing and expanding schedules in children's categorization and generalization. Journal of Experimental Child Psychology, 123, 129–137. https://doi.org/10.1016/j.jecp.2014.01.004

Yamagata, S., Nakata, T., & Rogers, J. (2023). Effects of distributed practice on the acquisition of verb-noun collocations. Studies in Second Language Acquisition, 45(2), 291–317. https://doi.org/10.1017/S0272263122000225

Yanagisawa, A. (2016). The effects of receptive and productive word retrieval practice on second language vocabulary learning. The Journal of the Chubu English Language Education Society, 30, 139–152. https://doi.org/10.20806/katejournal.30.0_139

Yan, T., & Zhou, D. (2023). The influence of the spacing effect on L2 vocabulary learning: A study on Chinese university students. "System", 115, 103049.

https://doi.org/10.1016/j.system.2023.103049

Yüksel, H. G., Mercanoğlu, H. G., & Yılmaz, M. B. (2020). Digital flashcards vs. wordlists for learning technical vocabulary. Computer Assisted Language Learning, 35(8), 2001–2017. https://doi.org/10.1080/09588221.2020.1854312

Zung, I., Imundo, M. N., & Pan, S. C. (2022). How do college students use digital flashcards during self-regulated learning? Memory, 30(8), 923–941. https://doi.org/10.1080/09658211.2022.2058553

Zulkiply, N. (2013). Effect of interleaving exemplars presented as auditory text on long-term retention in inductive learning. Procedia - Social and Behavioral Sciences, 97, 238–245. https://doi.org/10.1016/j.sbspro.2013.10.228

3A Network. (2011). Shin Kanzen Master: JLPT N3 Vocabulary Goi. Tokyo: 3A Corporation.

8.0 Appendix

Appendix A

Consent Form for Participation in Research on Memory

Emin Gaaya

I understand the Purpose of the Study

This research project aims to investigate memory retention of vocabulary in language learning. Participation in this study will contribute to a better understanding of how vocabulary acquisition can be improved. Furthermore, information on the study will be explained in a separate paper provided to the participants that are joining the study.

I understand what Participation Involves

If you agree to participate, you will be asked to:

1. Either use a program to study vocabulary or follow your study method.
2. Participants will take part in a pre-interview at the start of the study, followed by a vocabulary assessment test.
3. Participants will later complete a post-test, followed by an additional test later. No preparation is required for these tests.
4. An additional post-interview will be conducted to gather additional insights.
5. Participants using the program will be asked to share the data generated from their usage. This does not apply to those using their study method.

I understand that this is a Voluntary Participation

Your participation in this study is entirely voluntary. You are free to refuse to participate or withdraw from the study at any time without any negative consequences.

Consent form page 1

I understand that my participation will remain confidential

All data collected during this study will be anonymized and used solely for research purposes. Your identity will not be disclosed in any reports or publications. The data, including interview recordings and personal information, will be securely stored until the completion of the study.

Furthermore, excerpts from the interview may be quoted in [the thesis, presentations, published papers.].

I understand the Risks and Benefits

There are no significant risks associated with participating in this study. The benefits include contributing to academic research on memory and language learning, which may help improve educational tools and methods in the future. Additionally, participants will benefit from learning new vocabulary, supporting their language-learning journey.

I understand the Obligations

Study 10 new words each day until you have seen 150 words. After that, continue actively reviewing these words daily, using your method or following the program's instructions. After four weeks, you are no longer required to study the vocabulary. One must follow the instructions of the study to keep the research as valid as it can be. Be honest and do not use outside materials or assistance.

Consent form page 2

LUNDS UNIVERSITET

Consent

By signing below, you confirm that:

1. You have read and understood the information provided about this study.
2. You will also have the opportunity to ask questions about the study, within the limits of what the author is permitted to disclose.
3. You voluntarily agree to participate in the research as described.
4. You consent to the use of your data for the purposes of this thesis, with the understanding that it will be anonymized and deleted after the study is completed.
5. I agree to follow the study as instructed and be honest without cheating.

Name: …….

Email: …….

Telephone: …….

Signed: …….

Date: …….

Consent form page 3

Appendix B

Please document how much you study every day. This will be helpful when comparing the program with the self-study method.

Study Log Table

Day	Date	Time Studied (minutes)

Study log

Appendix C

だいがく 大学 university
りゅうがくせい 留学生 international student
せんこう 専攻 major
ともだち 友達 friend
にほん 日本 Japan
でんわ 電話 telephone
なまえ 名前 name
けいざい 経済 economics
こうがく 工学 engineering
せいじ 政治 politics

がっこう 学校 school
かいもの 買い物 shopping
いぬ 犬 dog
ねこ 猫 cat
ひと 人 person
こども 子供 child
しゃしん 写真 photograph
はな 花 flower
ごはん ご飯 meal
びょういん 病院 hospital

あね 姉 older sister
おとうと 弟 younger brother
きょうだい 兄弟 brothers and sisters
かいしゃ 会社 company
しょくどう 食堂 cafeteria
めがね 眼鏡 glasses
くるま 車 car
おなか お腹 stomach
はれ 晴れ sunny weather
あめ 雨 rain

しゅるい 種類 Type
でんりょく 電力 electric power
えいせい 衛生 hygiene

ぶんがく 文学 literature
れきし 歴史 history
いしゃ 医者 doctor
しゅふ 主婦 housewife
べんごし 弁護士 lawyer
いもうと 妹 younger sister
かさ 傘 umbrella
さいふ 財布 wallet
しんぶん 新聞 newspaper
とけい 時計 watch

まえ 前 in front (of)
たべもの 食べ物 food
のみもの 飲み物 drink
くだもの 果物 fruit
りょこう 旅行 travel
うみ 海 sea
しゅくだい 宿題 homework
たんじょうび 誕生日 birthday
かんじ 漢字 Chinese character
おかね お金 money

ゆき 雪 snow
きおん 気温 temperature
ふゆ 冬 winter
かいしゃいん 会社員 office worker
しごと 仕事 job
こうりつ 効率 efficiency
とうせん 当選 winning an election
しじ 支持 support
せいさく 政策 policy
ろんぶん 論文 Thesis

わるぐち 悪口 bad-mouthing
かんし 監視 Monitoring
かせき 化石 fossil

ぼうし 帽子 hat
ゆうびんきょく 郵便局 post office
ちゅうごくじん 中国人 Chinese
えいご 英語 English
えいが 映画 movie
おんがく 音楽 music
ざっし 雑誌 magazine
おちゃ お茶 green tea
みず 水 water
いえ 家 house

にもつ 荷物 baggage
でんき 電気 electricity
でんしゃ 電車 train
くに 国 country
かぞく 家族 family
おじいさん お爺さん grandfather
おばあさん お婆さん old lady
ちち 父 father
はは 母 mother
あに 兄 older brother

でんとう 伝統 tradition
かつやく 活躍 taking an active part
れんらく 連絡 contact
ちほう 地方 Region
けいこう 傾向 tendency
ごうかく 合格 passing an examination
せんぞ 先祖 Ancestor
たいさく 対策 countermeasure
えがお 笑顔 Smile
しゅうかん 習慣 habit

きおく 記憶 memory
こてい 固定 fixed (in place)
しゅうしょく 就職 getting a full-time job

むりょう 無料 free of charge
しんりん 森林 forest
はかい 破壊 Destruction
さばく 砂漠 desert
ちょきん 貯金 money saved up
えんがん 沿岸 coast
いびき 鼾 snoring

けつぎ 決議 decision
あいまい 曖昧 ambiguous
ふこう 不幸 misfortune
しゅじんこう 主人公 protagonist
ほうりつ 法律 law
しゅっけつ 出血 bleeding
じょうしき 常識 common sense

とうひょう 投票 vote
たいど 態度 attitude
とうろん 討論 debate
かんしゃ 感謝 gratitude
かっこ 括弧 parentheses
けが 怪我 injury
げんしょう 現象 phenomenon

げんりょう 原料 raw materials
はんとう 半島 peninsula
ちょくせつ 直接 directly
そうぞう 想像 imagination
かし 歌詞 song lyrics
きぼう 希望 hope
げんいん 原因 cause
でんせつ 伝説 folklore
にきび 面皰 Pimple
かいご 介護 Taking care of [Old People or Sick People]

へいき 兵器 weapon
しゅっさん 出産 childbirth
けいばつ 刑罰 (criminal) punishment
せきにん 責任 responsibility
れんたい 連帯 joint
りじゅん 利潤 profit
しょうめい 証明 proof
きゅうしゅう 吸収 absorption
いと 意図 intention
びじん 美人 beautiful woman

かくとく 獲得 Acquisition
れんぽう 連峰 mountain range
ぎょうじ 行事 event
ないよう 内容 content
がか 画家 painter
きろく 記録 record
めいし 名刺 business card
げんきん 現金 cash
はんざい 犯罪 crime
じゆう 自由 freedom

All 150 Japanese English word pairs from the study.