A Computational Approach to Language Learning: Examining the Efficacy of Spaced Repetition

CSC-475: Seminar in Computer Science
Fall 2023

Jace Rettig
jace.c.rettig@gmail.com
Department of Computer Science
Furman University
jace.rettig@furman.edu

Abstract

When learning languages, it is important to choose an effective method for learning. Spaced repetition is a newer method focused on having the learner review a piece of information just before they are about to forget it. With each review, the length of time between reviews gets larger. Several digital language learning applications implement spaced repetition, such as Duolingo and Anki. This paper uses data from Duolingo's Half-Life Regression repository and the author's personal Anki review statistics to base its conclusions. Data from Duolingo suggests that spaced repetition works equally well across languages and parts of speech. In contrast, the Anki data suggests that spaced repetition's efficacy is inconsistent for reviews with an interval of 3 months or greater.

1 Introduction

In the modern age of globalization, there is an increasing need for bilingual speakers in the work force. While in other countries it is common to speak more than one language, this is seen far less in America. In fact, its been estimated that more than half of the world's population is bilingual, and yet only about 20% of Americans can speak more than one language [1]. In an effort to raise this percentage, it is important to identify the most effective way to learn a new language. Traditionally foreign languages have been taught in a classroom setting, but the question as to whether this method is actually the most effective way to learn a language, and if there exists other methods that are more effective, should be raised. While research has been performed on the half-life of knowledge and spaced repetition [1], these methods have never been analyzed against and proven to be more effective than classroom language courses. Working towards identifying which language learning models are objectively most effective, this project aims to quantitatively analyze the effectiveness of spaced repetition in language learning applications. This work promotes further research and data collection for language learning and spaced repetition, working towards a more rigorous study to find the best model for language learning. In identifying this model, it could be be generalized and applied to any field of study, reducing the time it takes one to learn new concepts. It is also important to understand how to maximize learning productivity so that we can not only learn more efficiently, but more effectively overall. In learning new things quicker than before, we can better spend our time and efforts elsewhere.

The first dataset came from a study on the Half-Life Regression Model for language learning in 2016 [1], which contains roughly 13 million learning traces from the language learning app Duolingo. The second dataset is the author's personal review data from Anki, a free open source and highly

[1] https://www.cambridge.org/core/books/life-as-a-bilingual/extent-of-bilingualism/93D0020186A5EF999F299F7265C03854

customizable flashcard application that also incorporates spaced repetition into its review structure. The review log contains roughly 92,000 review entries of 5,000 cards from over two years of daily reviews. A basic data analysis of these datasets was performed using Pandas in a Jupyter Notebook, investigating whether spaced repetition is effective overall, and if so if it is more or less effective for a subset of languages or parts of speech. To visualize the data, libraries for Matplot, Seaborn, and Bokeh were used. The initial results from Duolingo's dataset suggested that spaced repetition is equally effective across languages and parts of speech. For the anki dataset, spaced repetition appeared to work well when the interval between reviews is less than 75 days. Beyond this threshold the effectiveness of spaced repetition appeared less predictable.

2 Related Work

When students learn a foreign language in school, it typically occurs in a classroom setting. While there are many different strategies for teaching a foreign language in the classroom, there are two primary models that have dominated in history. The original model was known as the Grammar Translation Method (GTM), which as it's name suggests, focuses on grammar and translating from the learner's native language to the target language. Students prioritize learning new vocabulary and grammatical structures to improve their reading and writing skills, and grammar is learned deductively in reference to one's native language [2]. Some of the disadvantages to this approach is that students learn to translate from their native language rather than understand the target language's grammatical structures implicitly. Additionally, there is little emphasis on using language to communicate, resulting in a lack of fluency among learners. Frustrated by these drawbacks, scholars in the 1970's began devising new language learning models that would prioritize developing learner's listening and speaking skills, and focus on competency of language and communication rather than pure knowledge and grammatical constructs [3]. Communicative Language Teaching (CLT) was one of these models methods that has grown in popularity and is used in many foreign language classrooms today. As the name suggests, it's main philosophy is that communicative competence should be the goal of language learning. In a study that documented the views and practices of CLT, teachers describe the method as emphasizing communication in the target language, relying heavily on speaking and listening skills, involving little direct grammar teaching, and using time-consuming activities![4]. The communicative approach also puts the focus on the learner, as the communicative needs of the learner provides the framework for the program aiming for functional competence [5]. It is important to note that while overall teachers and students hold positive attitudes towards CLT as a language learning strategy, some students may experience anxiety in the classroom environment. It is therefore the teacher's role to make a less threatening atmosphere, encourage the students to succeed, and build student's confidence to maximize the effectiveness of CLT[6].

In recent years, technological advances have paved the way for new language learning strategies that make it easier than ever to learn foreign languages. The general term for using technology to facilitate language learning is Mobile Assisted Language Learning (MALL), which can be described as 'any technology that can be used while walking around' [7]. The convenience and accessibility of modern smart phones have led to large organizations and governments producing apps that encourage language learning anywhere anytime. Education publishers have also started pushing towards online learning tools and digital textbooks to save on physical distribution costs [8]. In addition to using smart phones to assist language learning, some teachers have begun including mobile-sourced materials in the classroom as a tool to enhance students learning. In particular, combining CLT with mobile sourced materials has shown to make language learning more engaging and self-driven by learners[9]. One learning technique that is frequently used in MALL applications is spaced repetition, which originated in the 19th century as a strategy to optimize the long-term retention of information. Cognitive psychology research in memory and learning shows that a successful recall from memory yields superior retention to a mere presentation of the target item, and successfully recalling an item from memory after a delay is more effective than immediately after we learn it [10]. Combining these two principles, it is most efficient to review items just when they are about to be forgotten [11] While traditionally this method of review is done using flashcards, it is much more convenient or even feasible to have a computer calculate the next optimal review date for each item in a digital deck of cards. Researchers at Duolingo have created a trainable spaced repetition model based on the half-life regression of knowledge which combines the psycho-linguistic model of human memory with modern machine learning techniques to generalize two popular algorithms used in language

learning technology [1]. By combining spaced repetition with MALL and language learning apps, language learning on the go becomes much more accessible to the average person.

3 Data

The first dataset contains 12.9 million learning traces from Duolingo users collected over a 2 week period. These traces were used in a study in 2016 on the Half-Life Regression Model that is implemented in Duolingo's mobile app [1]. This dataset is publicly available through Dataverse or from the project's GitHub page. Table 1 demonstrates the nature of the data collected.

Interface Language	Learning Language	Lexeme String	History Seen	History Correct	Session Seen	Session Correct	P_Recall
English	Portuguese	come/comer <vblex>	21	20	1	1	1.0
Italian	English	newspaper/ newspaper<n>	42	39	4	3	0.75
English	German	frau/frau <n><f>	6	5	2	1	0.5
English	Italian	mie/mio<det> <f><pl>	7	7	3	2	0.67
English	Spanish	hermana/hermano <n><f>	12	10	1	1	1.0
English	French	et/et <cnjcoo>	38	27	2	1	0.5

Table 1: Example table of learning traces (Simplified). Each entry contains related information about the language and word being studied

The first two columns describe the language the user is studying, as well as the language of their interface. The interface language is assumed to be the user's native language. Each entry is based on a specific word that was presented to the user for the first time or for review. This word and its related grammatical information is stored in the Lexeme String column. History Seen is the total number of times the user has been exposed to the current word, and History Correct is how many times they have correctly identified this word. Similarly, Session Seen is the total number of times the user has been exposed to the word in the current lesson, and Session Correct is how many times they have correctly identified this word in the current lesson. P_recall refers to the accuracy of recall for the current *session*, rather than the total history.

Additional columns not shown in Table 1 included an anonymous user ID, a lexeme ID for the word being studied, the timestamp of the current practice session and the time since the last practice session. These columns were not directly analyzed during this study. To facilitate data analysis, a new part of speech field was created for each entry by filtering out the part of speech tag from the Lexeme String.

It is important to recognize some of the limitations of this dataset. This dataset only represents a 2-week period of learning traces for users, potentially limiting the viability of making accurate long-term predictions. There is also a lack of language diversity in the dataset. In 2016, Duolingo only supported and tracked the learning of five languages: English, Spanish, French, Italian, and Portuguese. The distribution of these languages is also unbalanced, heavily favoring English and Spanish learners as seen in Figure 1. Last, it is possible there are users whose native language does not match the language of their interface.

The second dataset comes from the author's use of Anki over the past 3 years where they created and studied over 5,000 cards when learning Japanese and Korean. The dataset contains the statistics for the 92,034 reviews that were performed through daily use of Anki. Table 2 demonstrates the nature of the review log.

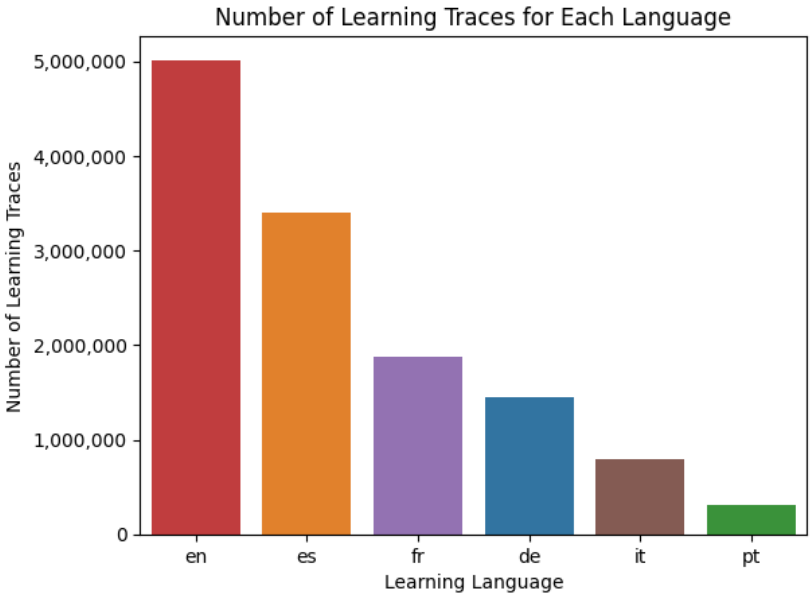

Figure 1: Distribution of languages being studied through Duolingo. There is a heavy bias towards English and Spanish traces.

Time of Review	Card ID	Ease of Recall	Interval for Next Review
1566452085029	1528564711188	2	1
1566452447881	1528565827041	2	2
1566626988674	1528564661613	3	5
1596122830309	1444182700694	3	62
1626313041376	1603083802365	4	108

Table 2: Sample anki review log. Card ID and Time of Review are stored as UNIX epoch timestamps

Time of Review represents the time the review was conducted in UNIX epoch time. Card ID is the unique identifier for a card, generated from the UNIX epoch time when the card was created. Anki reviews are self-graded, allowing users to rate how easily they could recall the term. The Ease of Recall represents the self-reported difficulty of the review, where 1 is review again, 2 is hard, 3 is normal, and 4 is easy. Interval for Next Review is the the amount of time before the next review will occur based on the spaced repetition and the self reported difficulty. When a new card is created or a card is updated, its interval is set to -600 by default. Several other columns were omitted due to not being used directly for analysis. These included the previous interval for review, the ease factor used to generate the next interval, the amount of time spent on the question and answer sides of the card before selecting a difficulty level, and the type of the card, varying from learning, review, relearning, or cram cards.

This dataset is also not without limitations. One issue is when the content of a card is updated, its card ID changes and it will be treated as a new card while maintaining the correct interval for review. This creates an issue for card analysis, only allowing complete tracking of cards that have never been updated. Additionally, due to the nature of self-reporting difficulty and this dataset representing a single person's experience with anki, there will be many unknown biases in the data associated with their use of the application. Any trends in the data can serve as a point of reference to guide more rigorous studies and future research, but cannot be used to draw definite conclusions on their own.

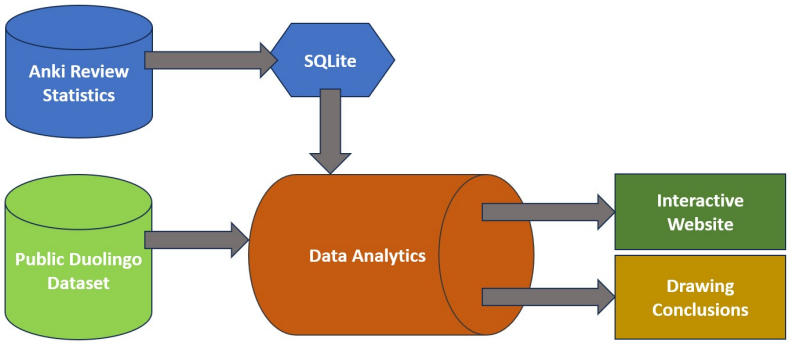

Figure 2: Pipeline of the project

4 Methods

Figure 2 gives an overview of the methods used to computationally analyze the effectiveness of spaced repetition during this project. During the data collection stage, Duolingo's dataset and related code from their study was found publicly available on GitHub. To provide an individual perspective on spaced repetition's effectiveness, the author's performance using Anki was analyzed. While there was an overview of the Anki statistics displayed natively through the application, the dataset the statistics were generated from was not directly accessible. To retrieve the dataset, SQLite was used to extract and convert the data to a CSV file for further processing. Pandas in a Jupyter Notebook file was used for the data analysis. As the author was unfamiliar with the field of data analytics, they relied heavily on their professor, ChatGPT, and GitHub Copilot for guidance on what to analyze and how to write the associated code. The data visualizations representing the findings on spaced repetition's effectiveness were made with a combination of code from Matplot and Seaborn libraries. To increase the audience of the findings, The Bokeh library was used to create interactive graphs and generate the HTML for an interactive web page. The website, Anki data set, and data analysis code will be publicly available on this project's GitHub page to support future work on this topic.

5 Results

5.1 Duolingo

Figure 3 shows a positive relationship between the number of times a word has been seen and the number of times a word was correctly recalled in a session, supporting spaced repetition's effectiveness within lessons and short periods of time. Figure 4 is a side by side comparison of spaced repetition's effectiveness between sessions and over time. Session shows a linear trend similar to Figure 3. History also shows a positive growth just on a much larger scale as users continue to use Duolingo and review previously learned material. While spaced repetition appears to be effective overall, its effectiveness across languages and parts of speech was also considered.

5.1.1 Across Languages

Investigating whether spaced repetition was more effective for learning certain languages over others, Figure 5 illustrates the average recall of words for each languages. Allowing for some minor deviations in the data due differences in sample size, all languages showed a similar increase in accuracy for each repetition within the session studied.

5.1.2 Across Parts of Speech

Looking at spaced repetition's effectiveness across parts of speech, Figure 6 illustrates the average recall of words based on their respective part of speech. Similar to the Figure 5, all parts of speech show similar increase in accuracy for each repetition within the session studied.

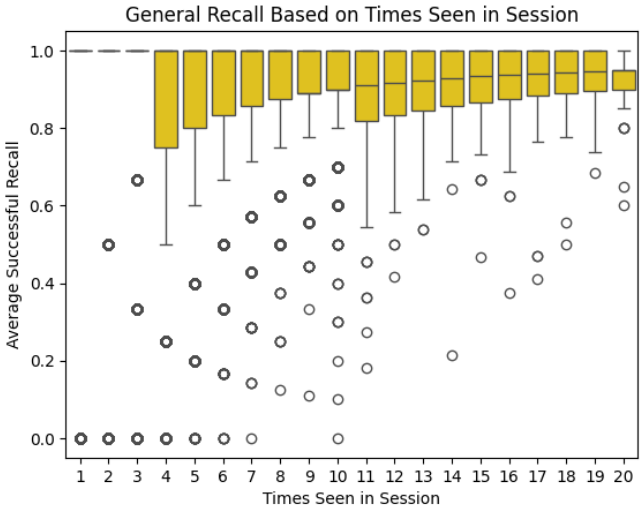

Figure 3: Correct recall based on times seen in the current session for all languages. The more times a word has been seen, the more likely the user will recall it correctly

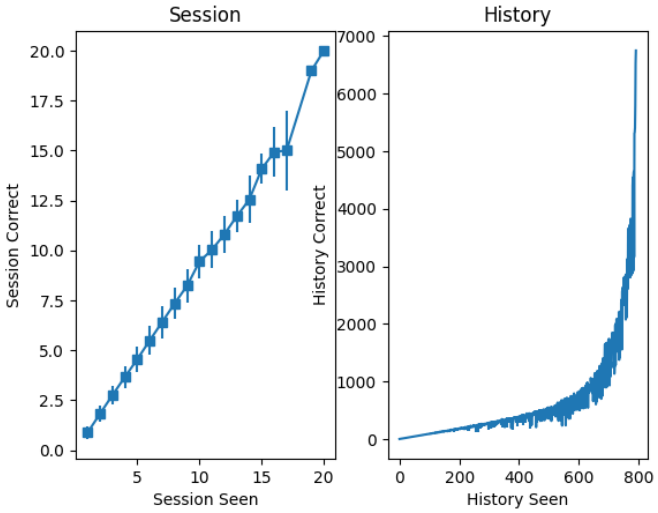

Figure 4: General trends for Duolingo learning with spaced repetition within and across sessions. History also shows a positive growth that is at a much larger scale

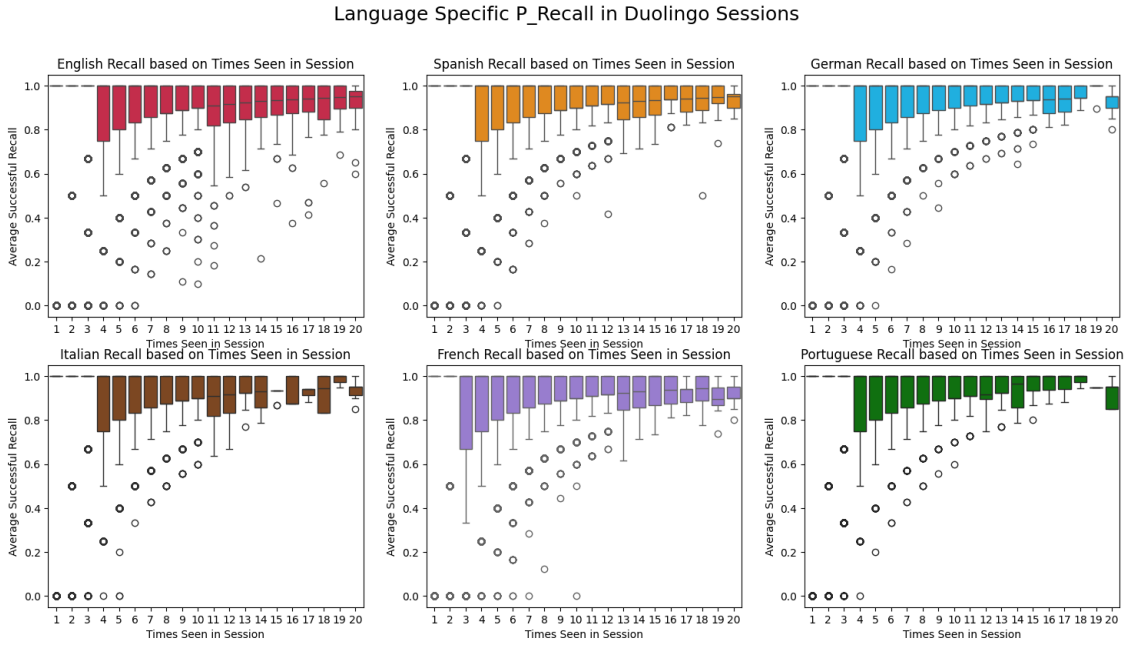

Figure 5: P_Recall trends across languages. The general trend of improved recall appears consistent regardless of the language

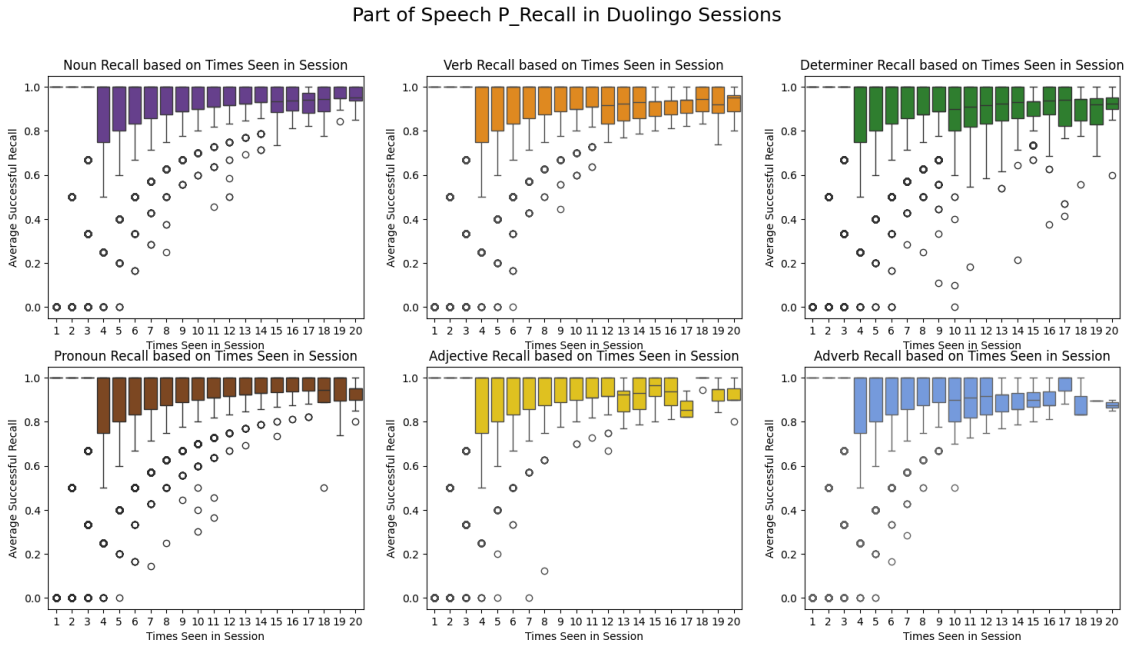

Figure 6: P_Recall trends across the most common parts of speech. The general trend of improved recall appears consistent regardless of the part of speech.

5.2 Anki

Figure 7 shows the overall learning progress in the Anki deck, divided into seven groups. The average learning curve for each group was calculated and can be seen on the left graph, while the right graph shows the learning curve for each individual card colored by the group it falls into. Orange and blue words were very easy to recall after only a few repetitions and quickly reached large review intervals. Green and red words took longer to consistently recall but then began reaching larger review intervals. In comparison, purple, brown, and pink words were more difficult to recall and were more frequently scheduled for review. Lines that go below zero represent cards that were forgotten and had their intervals reset to the default of -600 to be reviewed again in the same day. When looking at the median interval for the next Anki review as shown in Figure 8, spaced repetition seems to be consistently effective until review intervals greater than 75 days. Additionally, after reviewing a card more than 30 or 40 times, it is less predictable whether the word's recall will be effective or not.

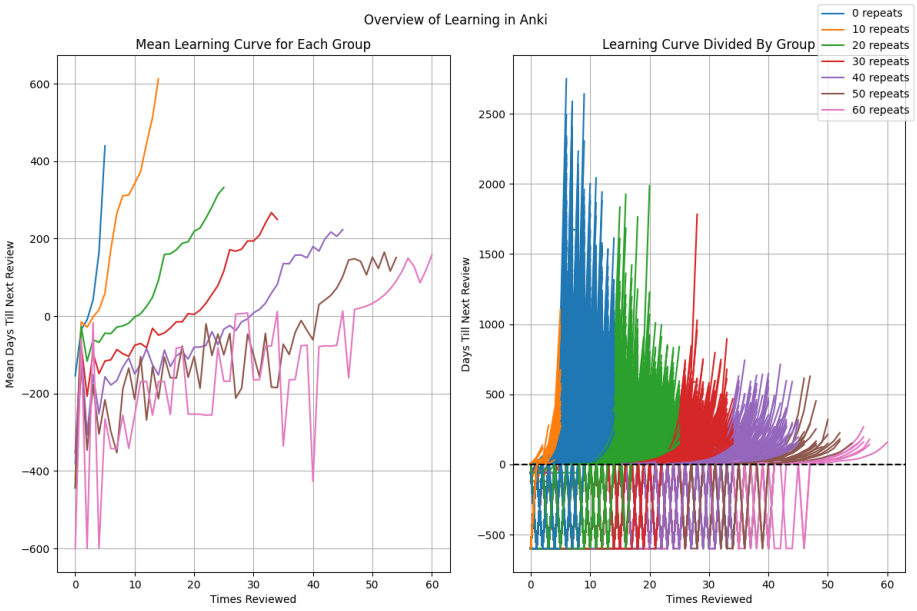

Figure 7: Anki learning progress for each card divided into seven groups. The average of each group can be seen on the left, with orange and red cards being easy to learn while brown and pink were very difficult to learn

6 Discussion

Looking at the results, the public data for Duolingo's spaced repetition seems promising. Regardless of the language or part of speech, spaced repetition appears to work equally across the board. As the original study and dataset from Duolingo was from 2016, it is natural to assume the algorithms used by the app have only improved over time. Despite the compelling data, some studies suggest that while Duolingo is great as a supplementary study method, the app's effectiveness for language learning mastery may be overstated when used in isolation [12].

Looking at the anki results, spaced repetition seems to be consistently effective for review intervals less than 75 days. After 75 days however, the effectiveness is much more unpredictable, with some words being easy to recall and others comparatively more difficult. Comparing Duolingo side by side with the Anki, there is a clear discrepancy in the results. The most simple explanation for the disparity in the results is that author struggles to consistently recall words with review intervals longer than 75 days. In this case, rather than spaced repetition, their own recall ability is inconsistent, and changes to the algorithm may needed to identify the more difficult cards and change the review interval accordingly.

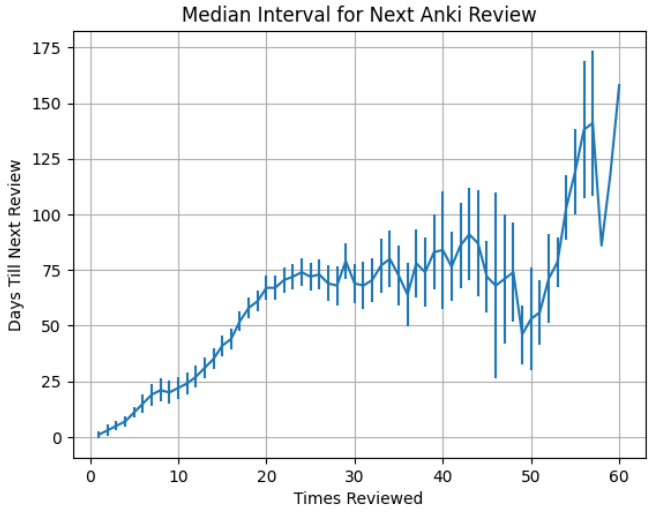

Figure 8: Median interval for next review with error bars. Spaced repetition is consistently effective for intervals less than 75 days or cards that have been reviewed less than 30 times

There is also the possibility that the public data released by Duolingo does not give an accurate representation of spaced repetition. As Duolingo is a private company with their own objectives, they might not want to publicly release data that shows any faults with their application and paints a bad image. In this case, one may ask what data Duolingo should publicly release that would paint a true picture about spaced repetition. A similar but updated version of this dataset in 2023 would be interesting to investigate, as Duolingo has since created new language courses, added features, and improved algorithms for the application.

7 Future Work

This research serves as a light introduction to the efficacy of Spaced Repetition in language learning apps, and should be used as a foundation to build off of for further research. Some limitations to this project include only analyzing one Anki deck, leading to potential bias and invalid conclusions. Due to some attributes not being included in the data analysis process, it is also possible these attributes had a significant impact that was not considered. To build upon this research, it would be interesting to have a larger sample of Anki Decks and analyze trends across them. Comparing variations of the Spaced Repetition algorithm across learning language applications outside of Anki and Duolingo may lead to interesting results. Finally, comparing the effectiveness of spaced repetition to traditional classroom language learning could be of interest. As for difficulties with these future endeavors, data acquisition may prove difficult to for datasets and anki decks. In Anki's case, each person's statistics are stored locally on their computer rather than online. Gathering student performance in classroom may be challenging at a large scale and present some issues of confidentiality. These examples are some of the many ways this research can be continued in the future. Overall, to extract more conclusive results on this topic, future work must expand upon this project through a more comprehensive study.

References

[1] Burr Settles and Brendan Meeder. A trainable spaced repetition model for language learning. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: long papers)*, pages 1848–1858, 2016.

[2] Muhammad Natsir and Dedi Sanjaya. Grammar translation method (gtm) versus communicative language teaching (clt); a review of literature. *International Journal of Education and Literacy Studies*, 2(1):58–62, 2014.

[3] Merlissa Elpedes Suemith. The communicative language teaching approach: Theory and practice. *Magister Scientiae*, (30):1–9, 2011.

[4] Kazuyoshi Sato and Robert C. Kleinsasser. Communicative language teaching (clt): Practical understandings. *The Modern Language Journal*, 83(4):494 – 517, 1999.

[5] SANDRA J. SAVIGNON. *Communicative Language Teaching: Linguistic Theory and Classroom Practice*, pages 1–28. Yale University Press, 2002.

[6] Kun-huei Wu. The relationship between language learners' anxiety and learning strategy in the clt classrooms. *International Education Studies*, 3(1):174 – 191, 2010.

[7] Olga Viberg and Åke Grönlund. Mobile assisted language learning: A literature review. In *11th World Conference on Mobile and Contextual Learning*, 2012.

[8] A Kolbuszewska. Adaptive learning in elt. *Humanising Language Teaching*, 17(6), 2015.

[9] Rupert Walsh. Teaching communicatively in a classroom with mobile-sourced materials. *Advances in Language and Literary Studies*, 12(5):23 – 29, 2021.

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

[11] Evgeny Chukharev-Hudilainen and Tatiana A. Klepikova. The effectiveness of computer-based spaced repetition in foreign language vocabulary instruction: A double-blind study. *CALICO Journal*, 33(3):334 – 354, 2016.

[12] Shawn Loewen, Dustin Crowther, Daniel R. Isbell, Kathy Minhye Kim, Jeffrey Maloney, Zachary F. Miller, and Hima Rawal. Mobile-assisted language learning: A duolingo case study. *ReCALL*, 31(3):293–311, 2019.