The Benefits of Spaced Repetition in Second Language Education

Brian Diep

12/22/2020

Keywords

Spaced-Repetition, Language Education, Learning

Abstract

Second language education has typically been confined to a traditional classroom experience with a teacher. However with the advent of the Internet and accessible technology, digital language education has been immensely successful in part due to its scale and flexibility. However, one of the major benefits of digitizing language education is that it is now possible to collect data in real-time and simultaneously use it to improve the learning experience for users. It is therefore important that we analyze the relationship between user-model interactions, aspects of language, and language acquisition to improve language acquisition for all. We approach this problem by using data gathered from the popular language education app Duolingo, we seek to model a user's language retention using logistic regression. Through this methodology we find that spaced-repetition is an efficient method for language acquisition.

Introduction

Learning a second language is a difficult task for most adults who have passed the critical stage of language acquisition as children. However, that has not stopped second language education from being a core part of education systems across the world. Unfortunately, these systems often leave much to be desired as they lack the flexibility and scope necessary to truly learn a language. Outside of the traditional system of language education, the widespread proliferation of technology such as the Internet, personal computing, and mobile devices have allowed other actors to break into the language education, in particular mobile apps. One of the greatest benefits that these apps have is that they can take user performance metrics and use them to guide decisions on teaching users.

In this report, we seek to explore this data driven approach using the free language education app Duolingo as a case study. The app has roughly 120 million users and is an example of one of the most popular of such language education apps available. We specifically focus on Duolingo's usage of spaced repetition as a method for teaching its users various grammar points. Spaced repetition is considered to be an efficient way to memorize new information and is broadly used in various teaching strategies (Tabibian, et al, 2019).

The remaining report will continue to further explore the data sets and methodologies used to gather the data, an analysis of the logistic model used, and results gathered from the fitted logistic regression model. Our findings suggest that spaced repetition is a useful tool that generally well models an individuals' retention of language skills, in particular grammar and vocabulary points. Furthermore, we discuss and investigate relationships between our predictor and response variables followed by a discussion of the strengths and weaknesses of the model, and further avenues of study.

Exploratory Data Analysis

The data used in this analysis was sourced from Duolingo containing almost 13 million sample interactions of users of the app. This provides a broad overview of interactions on the app and gives us a lot of the key metrics that are used to evaluate a user's learning. The Duolingo model for language education involves making lessons into small sections which allow for an individual to progress at their own pace. These lessons generally involve the user translating sample sentences, accompanied by vocabulary and grammar lessons as well as occasional listening and speaking tests administered through the microphone. Hidden from the user, the app records a user's performance at recalling specific words, referred to as lexemes, correctly and uses this information a spaced-repetition method to continually test users on information to improve overall retention of vocabulary and grammatical concepts.

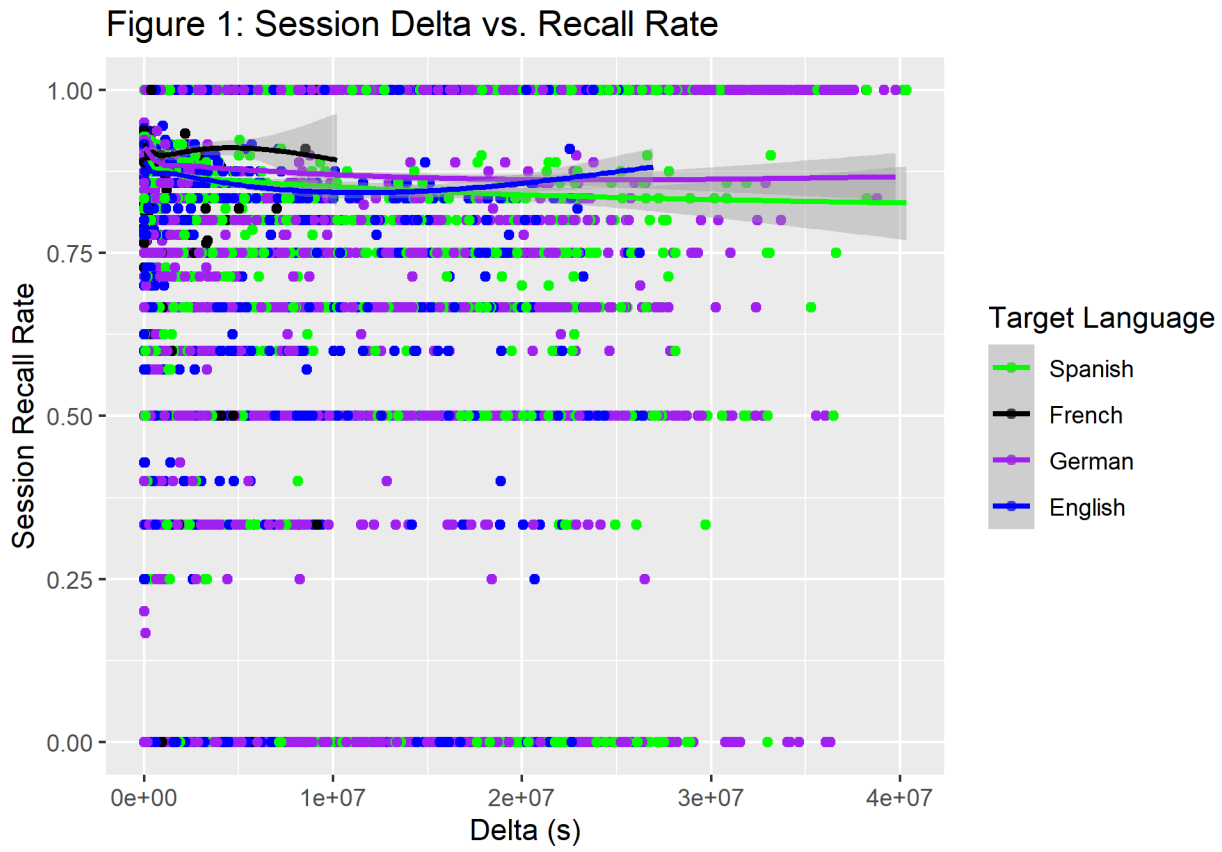
The population of interest for this dataset is second language learners who use apps. The population frame of this study is the users of the Duolingo app that were enrolled in specific courses. In particular, we restrict our analysis to speakers of English who are trying to learn Spanish, French, and German which are among the most popular languages on the app. For further analysis we also include a sample of speakers of Italian who are trying to learn English. Upon restricting the dataset in this way, we have observations from English users with a split of the observations between target languages Spanish (3.4 million samples), French (1.87 million samples), and German (1.45 million samples). The Italian to English samples comprise a further 424 000 samples. The private information of users was kept confidential and they are only referred to in the dataset according to a unique string of characters corresponding to their ID.

The data was gathered in a convenience sample of users on Duolingo since the company was able to access the pertinent information of all of its users and their progress through their course. Among all of the users enrolled in the courses that were selected to be included in the dataset, a random sample was drawn of which 115 222 users and their interactions were included in the dataset. This is an example of two-stage cluster sampling as we have broken up the app's userbase into different clusters (by course) and further sampled within. There is unlikely to be any major bias between the users learning any given language pair so this should not be of concern. Our use of simple random sampling in each course also eliminates any bias from sampling and we expect that our data is representative of the population.

The variables of interest in the given dataset include the probability of recall, delta, language being learned, the UI language, and the user's historical data with the same words/grammatical concepts.

The probability of recall is the ratio that a user correctly recalled a given lexeme during one session. We manipulate this variable such that we consider that a user has "learned" the lexeme correctly if they have a probability of recall of 1 for the given session. This is an admittedly narrow and strict definition of learning but it reduces the likelihood that we include instances where a user guesses the correct answer and better models a user's ability to retain an information if they can consistently recall it perfectly and are not only tipped off to the correct answer when they make a mistake during practice sessions.

Delta in this analysis is the time measured in seconds between the last time the user encountered this lexeme and the current learning session. An assumption is also made that the UI language of the app corresponds to the native language of the speaker as most users are likely to have their devices set to the language that they are most comfortable using. Historical user data such as the number of times the user has encountered the same lexeme and the number of times they have correctly recalled it in past lessons were also stored.



For this plot and the following plots, we have randomly sampled 10% of the data to simplify the graph while making salient the overall trends. The complete dataset is used for the analysis.

In Figure 1, when we plot the time between the last time an individual has seen a lexeme and their recall of that lexeme in their most recent session, we see a nice spread of observations from various values. We also see generally a slight trend that the user is less likely to correctly recall lexemes for which they were last reviewed a long time ago. This is consistent with our understanding that people are more likely to forget concepts that they have not seen recently.

Figure 2: Times Seen vs. Recall Rate

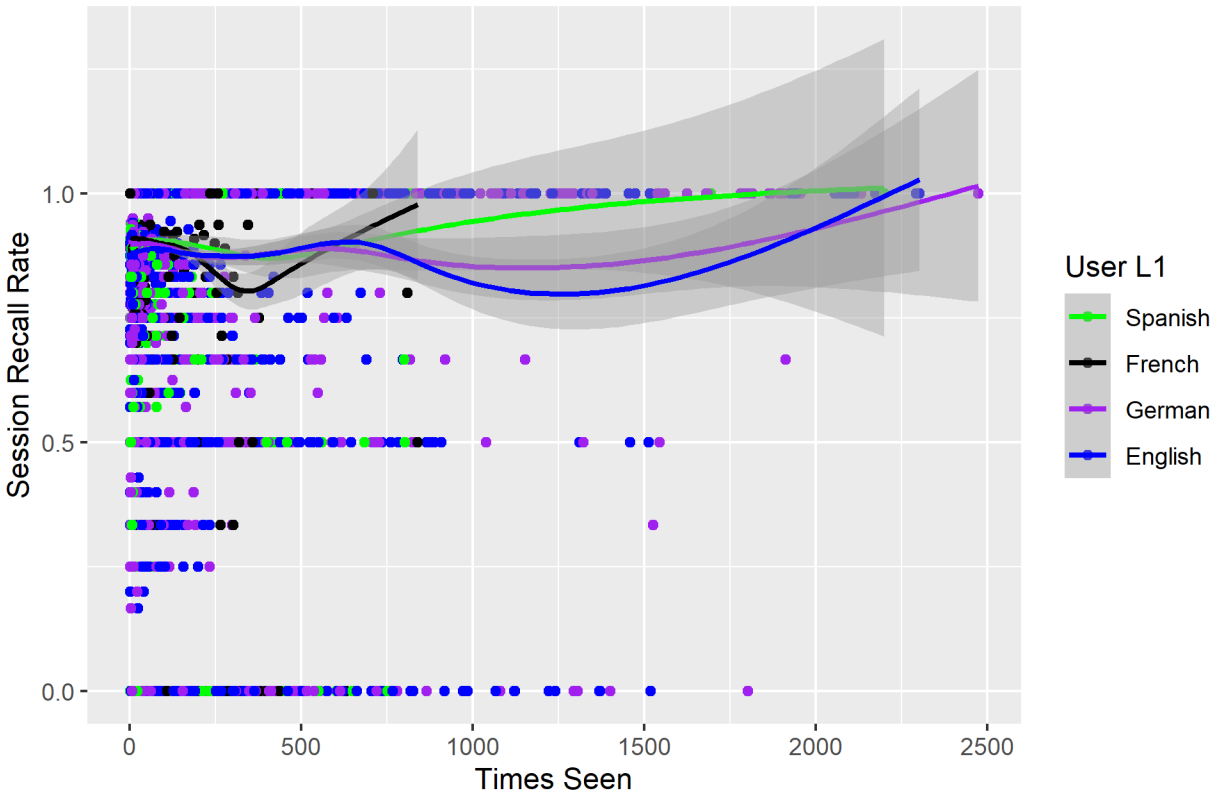
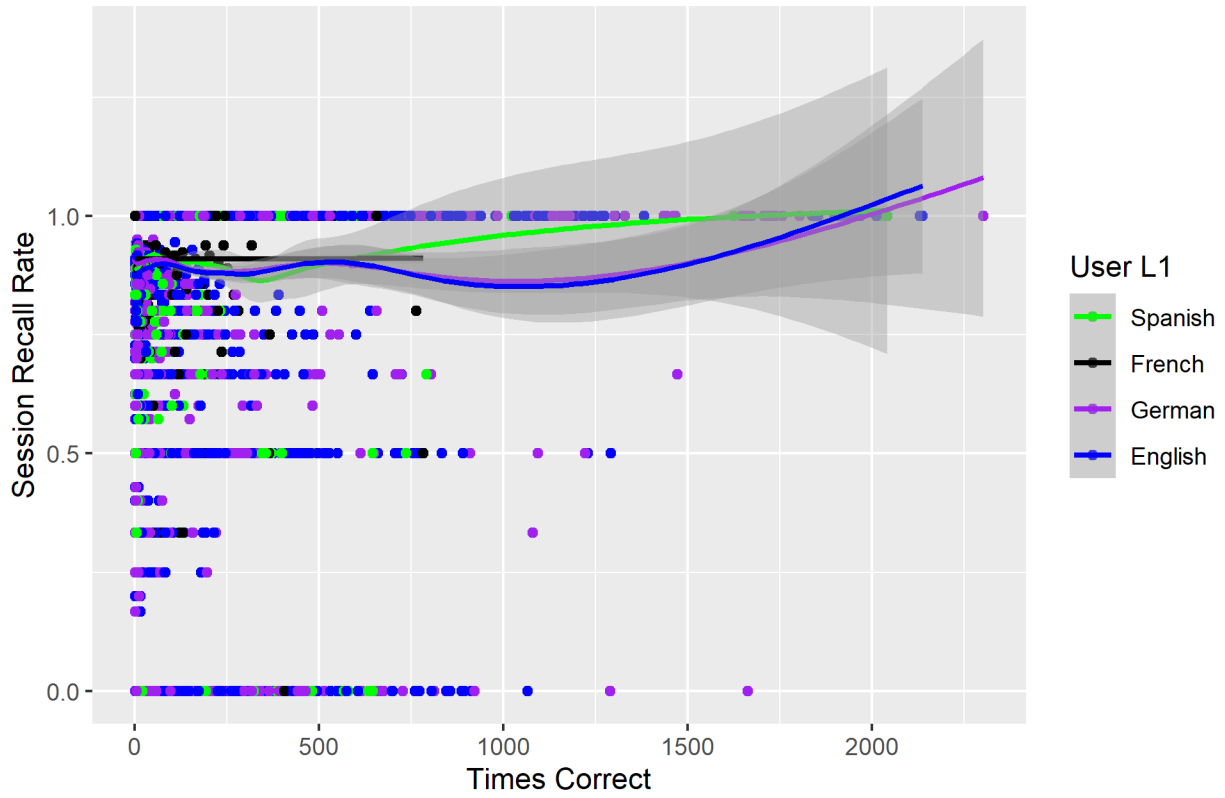


Figure 3: Times Correct vs. Recall Rate



In Figure 2, we plot the number of times that an individual has been exposed to a given lexeme against their recall of that lexeme in their most recent session. We see generally a trend that the user is more likely to recall recently introduced lexemes as well as those that have been introduced over 2000 times. There appears to be a gap in recall rates for lexemes that have been seen a middling number of times. A recency bias effect may explain why the most recently introduced lexemes have higher recall rates than those that have been seen multiple times. However, those that have been seen many more times than that are likely to have been committed to memory at that point and will have high recall rates as compared to in between when the user may not have learned the concept yet.

Figure 3 also shows us the times an individual has correctly identified a lexeme in an exercise as compared to their recall rate in their most recent study session. The dip in the middle is shared with Figure 2 but the effect is likely lessened as if a user is able to succesfully recall a term, it is a stronger indicator than if they have merely been exposed to it. Here we see a slight positive trend similar to in Figure 2. Users that have correctly identified a lexeme in the past are likely to continue to do so.

Across all three graphs, the graph of the mean values across all points follows roughly the same trends which suggests that any conclusiosn we can draw on one group will also be applicable to any others.

Model

Since recalling a particular lexeme is a binary variable, this immediately suggests that we should model it using a logistic model. Logistic models are used when trying to model the associations that exist between various independent variables and a binary categorical dependent variable. Mathematically in the context of our analysis, we can express our model as the following function of the various dependent variables.

$$Pr(lexeme\ retained) = logit^{-1}(\beta_0 + \beta_1 x_{delta} + \beta_2 x_{history\ seen} + \beta_3 x_{history\ correct})$$

This model is appropriate for our analysis since learning has been defined in such a way that it is a binary categorical variable and all the other dependent variables can be entered in as numerical values without further manipulation.

In our model, the features entered were delta, history seen, and history correct. Delta enters into our model as it represents the time in seconds between the last session. Intuitively it is likely that if the time between a user reviewing a lexeme is longer, the less likely an individual will remember that word. However, repeated revision as represented by the variables of history seen and history correct are also good indicators of whether a lexeme is retained since these are more likely to enter an individual's long-term memory. These variables refer to the total number of times the user has seen the given lexeme in all previous sessions and correct is the number of times they have responded correctly to a question involving the given lexeme. In particular, a user's history of correctly identifying the lexeme is the most obviously related to their ability to recall it again since it demonstrates that they have been able to recall it in the past.

We settle upon this final model after using backwards stepwise regression with AIC (Akaike's Information Criterion) to establish the most parsimonious model using the fewest and most significant predictors from our entire dataset. AIC acts as an indicator of how well our predictors are able to model the data and is used as a metric to minimize the number of predictors required to capture to model the relationship.

The model used was fit using the glm() function from the base R package.

For the purpose of this analysis, we also considered models such as a Bayesian model as well as a linear regression model, but these were eliminated based on their weaknesses in appropriately modelling our data.

Strictly linear models such as ordinary or multiple linear regression models were immediately excluded since they require that our be numerical which is clearly not the case for our binary variable.

Bayesian models were also excluded on the basis that they use informative priors to guide the regression. Having an informative prior is predicated on having information about the approximate distribution of the data that the model will be fit on. Since there are no clear trends to be found in our predictor variables nor any explained in the Duolingo dataset, our analysis would not assume any sort of informative prior.

Results

The overall model fit over all of the data samples results in the following logistic regression model:

$$(1)\ Pr(lexeme\ retained) = logit^{-1}(1.643 - (2.990 * 10^{-8})delta - 0.782 * history\ seen + 0.895 * history\ correct)$$

We can also find the following models (2), (3), and (4) which represent English to German, Spanish, and French respectively.

$$(2)\ Pr(lexeme\ retained) = logit^{-1}(1.671 - (3.375 * 10^{-8})delta - 0.135 * history\ seen + 0.155 * history\ correct)$$

$$(3)\ Pr(lexeme\ retained) = logit^{-1}(1.686 - (2.573 * 10^{-8})delta - 0.721 * history\ seen + 0.835 * history\ correct)$$

$$(4)\ Pr(lexeme\ retained) = logit^{-1}(1.499 - (2.573 * 10^{-8})delta - 0.721 * history\ seen + 0.835 * history\ correct)$$

For the model representing the learning of Italian to English speakers, we end up with a logistic regression model with the following coefficients.

$$(5)\ Pr(lexeme\ retained) = logit^{-1}(1.811 - (4.368 * 10^{-8})delta - 0.727 * history\ seen + 0.812 * history\ correct)$$

Discussion

After fitting the models, we find that all the models share key aspects which gives us insight into the power of certain predictors in our model. The values of each of the coefficients in the regression models all cluster within similar values which is evidence confirming our predictions from the exploratory data analysis where we expect there to be not a significant difference between any individual course.

With regards to interpreting the model coefficients, in general we find that delta is associated with an extremely small coefficient. This makes sense as for every additional unit of time between exposure to a lexeme (in this case, a second), there is unlikely to be a significant change. However this small delta will accumulate over long periods of time to cause a noticeable effect in the predicted retention of a given lexeme.

One finding of note is that history seen has a small negative coefficient whereas history correct always has a small positive coefficient who's magnitude is always slightly greater than history seen's coefficient. This makes sense as every time a user successfully recalls a lexeme, they are more likely to get it correct in the past.

Consequentially, for any fixed

$$Pr(lexeme\ retained) = p$$

, with every additional time a user correctly identifies a lexeme, it will require a longer delta for them to achieve the same value p. This implies that as a user continues to be exposed to a word (and demonstrates that they can correctly recall it), the app can lengthen the gap between reviewing the same topic. Note that the increased length of this gap does not grow linearly but actually grows rather exponentially as these coefficients are calculated under the logit function. This lines up with our intuitive understanding of learning and is backed up by research of spaced-repetition models (Tabibian, et al, 2019).

Examining further the interaction between delta, user historical performance, and the probability that a user has retained a lexeme. It is important for methods based on spaced-repetition as it provides a metric for which we can decide whether a user should review pertinent information. Language education programs should specifically be interested in identifying the threshold where the user is likely to be forgetting the word and is in need of revision. For example, with our models we can find when

$$Pr(lexeme\ retained)$$

dips below 0.5 as an estimate of when the user is beginning to forget a concept and the app can then include this concept in the user's next session.

This model actually allows us to find the optimal time for the app to review a lexeme as we can algebraically manipulate the logistic regression model equation to find the exact value of delta given.

$$log(\frac{p}{1-p}) = \beta_0 + \beta_1 x_{delta} + \beta_2 x_{history\ seen} + \beta_3 x_{history\ correct}$$

$$\frac{p}{1-p} = e^{\beta_0+\beta_1 x_{delta}+\beta_2 x_{history\ seen}+\beta_3 x_{history\ correct}}$$

$$p = \frac{e^{\beta_0+\beta_1 x_{delta}+\beta_2 x_{history\ seen}+\beta_3 x_{history\ correct}}}{1+e^{\beta_0+\beta_1 x_{delta}+\beta_2 x_{history\ seen}+\beta_3 x_{history\ correct}}}$$

$$p = \frac{1}{1+e^{-(\beta_0+\beta_1 x_{delta}+\beta_2 x_{history\ seen}+\beta_3 x_{history\ correct})}}$$

By setting the probability p to an arbitrary threshold value (0.5 as used in the actual Duolingo method of half-life regression) and inputting the user's history with that lexeme, we can use the above formula to find the correct value of delta we should wait before reviewing a concept.

In conclusion, we find that spaced repetition as in the style that Duolingo uses is an efficient strategy in determining what material a student needs to review in order to maximize the retention of concepts.

Limitations and Weaknesses

One major unavoidable limitation of our analysis is in the way we have operationalized language competency. Language is much more than being able to translate lexemes and the focus of Duolingo and similar language learning tools are on reading and writing skills, to the exclusion of key skills necessary for working fluency in a language such as listening and speaking (Husain, 2015). Thus, we should be cognizant of this limitation before making any broad conclusions about the efficacy of app based learning and traditional classroom based methods.

Another limitation is that Duolingo's course structure for each language are partially community sourced and may not all be of the same quality. Popular language pairs that have been supported for longer such as English for Spanish speakers or French for English speakers may be of higher quality than more obscure and/or newly supported languages.

Our analysis has also been limited to Germanic and Romance languages within the European sprachbund (a group of languages with shared features due to language contact and not genetic relationship). Existing bodies of research show that the ease of learning a new foreign language is heavily affected by any existing languages that the individual speaks (Broersma, et al. 2016). Although the general patterns of repeated exposure can be generalized to all language, the degree to which they affect an indivudal's ability to pick up a specific language will depend on how structurally similar the languages are.

Additionally, the platform of Duolingo may also lead to some level of self-reporting bias. Long term users of the app are likely to be more dedicated to learning a new language than those who have only recently or occasionally used the app. No distinction between these two groups is made in the dataset and we may find statistically significant differences between users who consistently use the platform as opposed to those who do not.

In this vein, although Duolingo provides a large amount of data and is a core resource for many language learners, it is not the only one out there and often times, those learning a language are consuming media and content in their target language outside of the app. Our analysis does not account for this and we cannot conclude on any interaction effects in the model as a result of this.

Future Work

Future work in understanding second language education should focus specifically upon different subfields such as grammar, syntax, or vocabulary. A working understanding of a language requires a speaker to combine all of these subfields. Furthermore, many language education programs also incorporate elements of listening and speaking, which is an avenue for further study. A comparison between whether learners retain new information better if they are given information through listening, reading, writing exercises, or a combination thereof would also be of note.

We can also take a comparison between learners of languages (in particular those outside of European languages) of vastly different origins such as English to Japanese or English to Arabic. These languages are generally known to be extremely difficult for monolingual L1 English speakers to learn due to cultural, phonological, and grammatical differences. Understanding which areas language learners find easier or more difficult when picking up these languages can allow us to find more universally applicable strategies for language education. This would also remove any bias inherent in learning languages similar to those that one already knows.

Our model provides an opportunity for Duolingo and other app based or self-guided language education programs to improve their materials and/or course structure by factoring in the strengths and weaknesses of their approaches found in this analysis.

Appendix

The code and the steps to find the dataset supporting this analysis can be found at the following link: https://github.com/brdiep113/duolingo-sta304-final

References

- Broersma, M. et al. (2016). Effects of the Native Language on the Learning of Fundamental Frequency in Second-Language Speech Segmentation. Frontiers in Psychology, 7(985), 10.3389/fpsyg.2016.00985
- Husain, Noushad (2015). Language and Language Skills. Maulana Azad National Urdu University.
- JJ Allaire et al. (2020). rmarkdown: Dynamic Documents for R. R package version 2.3. Retrieved from: https://rmarkdown.rstudio.com.
- R Core Team. (2020). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. Retrieved from: https://www.R-project.org/.
- Settles, B., & Meeder, B. (2016, August). 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) (pp. 1848-1858).
- Tabibian, B. et al. (2019). Enhancing human learning via spaced repetition optimization. Proceedings of the National Academy of Sciences, 116(10), 3988-3993.
- Wickham et al., (2020). dplyr: A Grammar of Data Manipulation. R package version 1.0.2. https://CRAN.R-project.org/package=dplyr
- Wickham et al., (2019). Welcome to the tidyverse. Journal of Source Software, 4(43), 1686, https://doi.org/10.21105/joss.01686
- Wickham, H. (2016). ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag. New York.