Identification of Behavioral Variables for Efficient Representation of Difficulty in Vocabulary Learning Systems

Parisa Supitayakul *, Zeynep Yücel *
Akito Monden *, Pattara Leelaprute †

Abstract

This study focuses on foreign language vocabulary learning in computerized medium and seeks behavioral variables that best reflect the difficulty level of learning material. In this respect, we employ a spaced repetition flashcard software and display English vocabulary belonging to various word classes as well as difficulty levels to a set of participants and determine the most efficient variables in representation of task difficulty and estimation of required effort. Based on an analysis of activity logs, we propose a set of behavioral variables, which have a potential relation to task difficulty. Subsequently, we examine the correlation between these behavioral variables and task difficulty levels. Our results indicate that variables at deck level have a stronger relation to difficulty than those at card level. In addition, when the correlation of the proposed variables with the difficulty level is contrasted to users' self assessment of difficulty, the proposed variables are found to be more reliable indicators of difficulty. Such variables are promising for improving the performance in user-adaptation, scheduling of learning tasks, and estimation of learning effort, motivation and engagement among others.

Keywords: E-learning, behavioral variable, adaptation, effort estimation, engagement.

1 Objectives and Related Work

Recently, computer based tutoring systems are have become common learning tool at various levels of education as well as in voluntary self-motivated pursuit of knowledge (i.e. lifelong learning)[1]. The rapid diffusion of such systems into education is suggested to be due to a series of reasons including their low cost, ease of access, and diverse range of study materials [1, 2, 3].

In addition to these, one of the most important advantages is considered to be user-adaptative properties, and specifically providing of personalized support through various interventions (e.g. adjustment of pace, introduction of incongruity, arousing of curiosity) [4,

---

* Department of Computer Science, Okayama University, Okayama, Japan
† Department of Computer Engineering, Faculty of Engineering, Kasetsart University, Bangkok, Thailand

5]. However, incorporation of such stimuli initially requires assessment of the amount of learning effort or user's motivational state.

The estimation of learning effort can be utilized for several purposes. For instance, if the required effort is very high, i.e. the study material is too challenging for the user, it can be considered to adjust it to a more reasonable degree for avoiding frustration. Similarly, if the effort is very low, i.e. the material is too trivial, it can be considered to adjust it to a higher level to avoid boredom and monotony. In addition to the adjustment of material difficulty, estimated effort can be incorporated with estimation of user's state of engagement (i.e. focused, distracted, fatigued etc.)[6].

Specifically, the learning platform (e.g. computer, smart phone, tablet) offers a convenient medium for observing various user behaviors summarized in activity logs [7]. In particular, several computer-recorded values such as frequency of logins, number and frequency of responses/views, time spent online etc. emerge as convenient attributes for estimation of learning effort, workable on almost any platform [8, 9].

Activity logs are an integrated part of almost any e-learning system independent of the specifics of the learning subject (e.g. foreign language, mathematics), target user profile (e.g. students, off-curriculum learners), or teaching method (e.g. tests, quiz), etc. [10, 11]. In addition, they are easy to record from a large number of users, and they capture an abundant variety of information enabling detection of regularities or deviations within numerous users. To date, various behavioral variables are derived from computer recordings, namely log files, such as task duration, response/decision time, answer correctness etc. [11], which bear the potential of observing certain actions at intermediate stages of the learning session, enabling a continuous monitoring of users' motivational state [12].

Numerous studies examine these behavioral variables at several difficulty levels. However, they often disregard any inherent variations on learning material. For instance Ma et al. study the relation between reaction time and difficulty regarding a textbook irrespective of the required skills for different subjects [13]. Similarly, Beck detects disengagement using response time to cloze type questions, where users need to identify deleted words in a sentence, irrespective of the word class or grammatical function [14]. In addition, although analysis of activity logs has been employed in detection of motivational state of users since a long time, most studies consider markers of a single particular task [7, 15, 16]. In that respect, this study distinguishes itself by investigating the effect of varying content types.

Moreover, this study investigates the relation between several behavioral variables and difficulty levels across content types. Any possible behavioral variation with respect to content type enables fine-tuning of the interventions or adaptations (e.g. optimization of task scheduling) or improving detection of disengagement [17, 12]. In that respect, it offers an additional and more profound evaluation of efficacy in addition to the conventional approaches relying merely on final test accuracy [18].

Specifically, we consider the task of memorization (of vocabulary), which is a substantial part of foreign language education. We prepare several tasks involving vocabulary from different word classes and with different difficulty levels. We then present them through a spaced repetition flashcard software to several participants and record their activity logs [19]. From activity logs, we derive certain behavioral variables and investigate the effect of content type on those variables. Particularly, recent research on neural processing mechanisms relating different word classes provide promising results and indicate a potential of fine-grained adaptation of the learning software in relation to such background information [20].

2 Experiments

In order to promote the expansion of behavioral spectrum of learners, we designed several diverse tasks. In particular, the tasks vary in (i) their content type and (ii) difficulty stage (i.e. required learning effort). The tasks are presented through a free and open-source e-learning platform, which is modified to register particular user actions. The registered actions are then employed in representing user behavior. A total of 6 users (henceforth, participants) carry out the tasks on the e-learning platform and their activity logs are collected[1]. In particular, the participant set is constituted by 4 females and 2 males, all mother tongue Japanese speakers, with an age of 43 ± 7.1. In what follows, we provide further details on the tasks, e-learning platform, and activity logs.

2.1 Experiment tasks

The tasks target English language vocabulary learning. In terms of *content type*, we consider different *sorts of vocabulary*, or lexically speaking, *parts of speech*, as (i) concrete noun, (ii) abstract noun and (iii) verb. Moreover, for each content type, we build three *sets of vocabulary* (henceforth, decks) at difficulty levels of (i) easy, (ii) medium or (iii) hard, denoted by E, M and H, respectively.

In particular, this *objective level of difficulty* is achieved by selecting the vocabulary from different sets of study materials of a standardized English test (i.e. STEP Eiken) at different stages (i.e. proficiency levels). Moreover, we represent it by an integer $v \in [1,3]$, where $v = 1$ corresponds to E, and $v = 2$ and $v = 3$ correspond to M and H, respectively. We adjusted the composition of the decks using the method of [21] such that E is performed effortlessly, while M is in fair correspondence with users' skills, and H is involves words least familiar to users, and in that, requires significantly higher effort to remember. We confirmed using rank bi-serial correlation that the assigned difficulty levels of the cards are *uniform* within each deck, *diverse* between different decks and *comparable* across content types [22].

2.2 E-learning platform

The aforementioned tasks are delivered through a spaced repetition flashcard software "Anki" [23, 24]. In analogy to the physical flashcards, a virtual flashcard has two *sides* as *front* and *back* (See Figure 1[2]). Here, front side involves the *query* (i.e. a word in Japanese), whose correspondence the user needs to remember in English. The correct correspondence, i.e. *answer* can be seen by *flipping* the card and disclosing the information on the back side.

The general course of tasks is initialized by receiving a query. Subsequently, the user tries to remember the answer to this query and then "flips" the card to (i) learn or confirm the answer and (ii) evaluate his/her confidence in remembering this answer. The evaluation consists of choosing one out of the three options of "Again", "Good" or "Easy". We consider this evaluation as the *subjective (i.e. perceived) level of difficulty* of that query, and represent it with $v' \in [1,3]$. In particular, choosing "Easy" is denoted by $v' = 1$, whereas "Good" and "Again" are denoted by $v' = 2$ and $v' = 3$, respectively. Note that, all words in

---

[1]The participants are informed in a clear manner about the nature and method of the research, volunteered to participate in the experiments, and gave their permission for data recording. The participants reported to have an approximately same skill degree in English (corresponding to high school graduate level).

[2]For readability, the interface is displayed in English, but in experimentation a Japanese interface is used.

Figure 1: (a) Front and (b) back sides of a sample card.

the same deck have the same objective level of difficulty v, irrespective of the time point in the experiment, whereas the subjective level of difficulty v' is expected to decrease in time, as the participant practiced it (multiple viewings of the cards). Upon evaluating difficulty, the succeeding word appears, until the user exhausts all the cards in the deck. Although there is no time limit on how long can the query be viewed or the answer can be reviewed, we let the users study a deck for at most 15 minutes.

2.3   Activity logs

The software registers user activity into a log file in terms of a set of markers. First of all, several identification information about each individual card is registered, namely d and i, which are both 13-digit integer codes representing deck ID and card ID, respectively. In addition, we consider 3 kinds of markers as t_p, t_e, t_f representing time course of the experiment. In particular, t_p denotes the instant, at which the query is displayed. In addition, t_f and t_e represent the instants, at which the user flips the card, and the instant he/she evaluates the difficulty of that vocabulary[3]. Finally, we register the subjective level of difficulty v', reflecting users' confidence in recalling a particular vocabulary as explained in Section 2.2.

3   Methodology

Based on the activity logs mentioned in Section 2.3, we propose a set of variables reflecting users' behavioral patterns. We then investigate the correlation of those behavioral variables with varying difficulty levels across different content types. In what follows, we provide the explicit definitions of our behavioral variables, and the investigation of the relation between the proposed variables and difficulty levels.

3.1   Behavioral variables

We process activity logs and derive several markers, which have the potential of indicating how much strain a card or a deck puts on a user (to accomplish the goal of memorizing). Specifically, we consider the following 10 behavioral variables.

Three behavioral variables are derived from number of card displays (or reviews) as n_c, n_{avg} and n_{tot}. Here, n_c denotes the total number of displays of a card. Namely, for a card i, which is displayed n_i times, $n_c[i] = n_i$. Since n_c is defined on a card-basis, it can

---

[3]All the variables relating time, i.e. t_p, t_e, t_f, are registered in standard Unix time in milliseconds.

potentially be subject to large variation, particularly if the difficulty is not uniform within the deck. Therefore, we define two other variables as n_{avg} and n_{tot}, which are expected to be less prone to the effect of irregularities. Namely, n_{avg} is average number of displays for all -displayed- cards in a deck, whereas n_{tot} is the total number of card displays from a deck within the time limit (of 15 minutes).

$$n_{\text{avg}} = \frac{\sum_{i \in d} n_i}{N}, \qquad (1) \qquad\qquad n_{\text{tot}} = \sum_{i \in d'} n_i, \qquad (2)$$

where d' is the set of cards that the user studied from deck d and N is the number of cards in deck d. Since n_{avg} and n_{tot} are defined on a deck-basis, they are expected not to be affected by minor irregularities of difficulty within the deck.

In addition, if the deck is trivial for the participant, he/she is likely to finish studying all N cards within the allowed time window. However, if the deck is challenging, the user may consume the time limit without exhausting the cards (but studying only the cards in d'). The introduction of the two behavioral variables of n_{avg} and n_{tot} allows us to better address this latter situation. Namely, while both variables are expected to grow with deck difficulty, only the growth in the latter one is affected by the ability of the user to finish the deck.

In addition to the set of variables derived from number of card displays, we derive another set of variables from the time stamps. For the card i, let $t_q[i]$ represent the average duration that the user viewed the front side. Moreover, let $t_f[i, j]$ represent the time card i was flipped on its jth viewing, and let $t_p[i, j]$ stand for the time of prompt for the same viewing. Then $t_q[i]$ can simply be defined by taking the average over all n_i viewings,

$$t_q[i] = \frac{\sum_{j=1}^{n_i} t_f[i, j] - t_p[i, j]}{n_i} \qquad (3)$$

In addition, the average duration spent on back side of card i and the average duration spent on card i can be expressed as,

$$t_a[i] = \frac{\sum_{j=1}^{n_i} t_e[i, j] - t_f[i, j]}{n_i} \qquad (4) \qquad\qquad t_c[i] = t_q[i] + t_a[i] \qquad (5)$$

Moreover, we account for rate of front (query) viewings to back (answer) viewings r_{qa},

$$r_{qa}[i] = \frac{\sum_{j=1}^{n_i} t_q[i, j]/t_a[i, j]}{n_i}. \qquad (6)$$

Note that $t_q[i], t_a[i], t_c[i]$ and $r_{qa}[i]$ relate a single card i. In order to represent user's reactions against the entire deck, we define $\bar{t}_q[d], \bar{t}_a[d], \bar{t}_c[d]$, which are averages of the corresponding variables for all cards in deck d. Let $\bar{t}_q[d]$ be the average duration, which the user spent on the front sides of all cards belonging to deck d,

$$\bar{t}_q[d] = \frac{\sum_{i \in d'} t_q[i]}{N'}, \qquad (7)$$

where N' is the number of cards in d'. Using a similar logic, we define three other variables for deck d as $\bar{t}_a[d], \bar{t}_c[d]$ and $\bar{r}_{qa}[d]$.

As we examine the distribution of behavioral variables, we see that there are several outliers. Since these values are quite few in quantity, we regard them not to represent the "normal" behavior and filter them out by preserving the data below a threshold which bounds the lower 95% of all data. After this sort of preprocessing, the probability density functions of the above-mentioned behavioral variables regarding all users, content types and difficulty levels are approximated as in Figure 2.
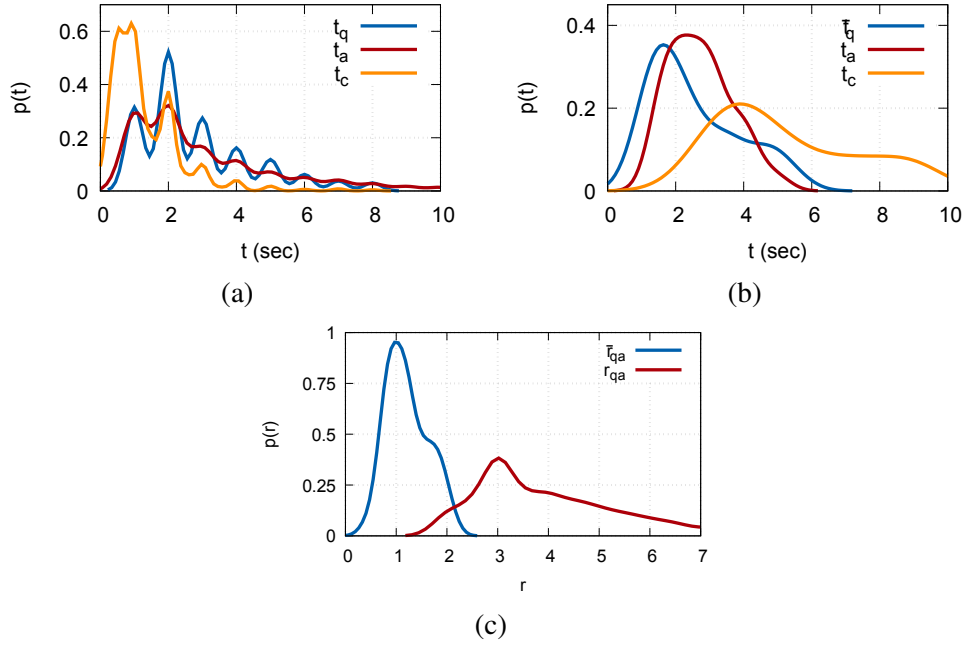
Figure 2: Density estimations of distributions of behavioral variables.

4 Results and discussion

In order to determine the variables that best reflect the effect of task difficulty on user behavior, we investigate the quality and extent of the relation between the difficulty levels (of the cards or decks) and the variables proposed in Section 3.1. To that end, we employ polyserial correlation, which defines the correlation between a quantitative variable and an ordinal variable [25]. In our specific case, numerical variables are the behavioral variables defined in Section 3.1, whereas ordinal variables are objective levels of difficulty v.

4.1 Correlation of behavioral variables and difficulty levels

The polyserial correlation values illustrated in Table 1 represent the relation between ten behavioral variables and three objective levels of difficulty. In addition, in order to provide a baseline, we consider assigned difficulty by the user v' as a pseudo-variable, and present its correlation with the objective difficulty v.

Examining the efficacy of each behavioral variable in representing the task difficulty, we see that n_{tot} and n_{avg} have higher correlation than n_c. This observation is in line with expectations, since these variables are less exposed to variations of difficulty between cards but relate the entire deck.

This effect can also be noticed when card level variables (i.e. t_q, t_a, t_c and r_{qa}) are contrasted to the deck level variables (i.e. $\bar{t}_q$, $\bar{t}_a$, $\bar{t}_c$ and $\bar{r}_{qa}$). Namely, averaging the variables across viewings of a card, does not help in eliminating the variations between cards. However, as we represent behavioral patterns addressing an entire deck, i.e. in terms of $\bar{t}_q$, $\bar{t}_a$, $\bar{t}_c$ and $\bar{r}_{qa}$, we notice that certain variables achieve a high correlation with deck difficulty v, in particular, $\bar{t}_a$. It is not surprising that $\bar{t}_a$ has a higher correlation than $\bar{t}_q$, since the front side involves only the prompt and in that does not require a significant effort, whereas the back

Table 1: Polyserial correlation between behavioral variables and difficulty levels for varying content types.

Behavioral variable	Abstract noun	Concrete noun	Verb
n_c	0.64	0.65	0.64
n_{avg}	0.90	0.95	0.90
n_{tot}	0.85	0.90	0.83
t_q	0.18	0.24	0.29
t_a	0.48	0.41	0.42
t_c	0.44	0.40	0.46
r_{qa}	0.28	0.19	0.24
$\bar{t}_c$	0.84	0.79	0.82
$\bar{t}_q$	0.56	0.60	0.70
$\bar{t}_a$	0.93	0.88	0.80
$\bar{r}_{qa}$	0.85	0.85	0.75
v'	0.76	0.68	0.72

side is viewed for longer since it presents the essential matter to memorize. Obviously, this is particularly true for the *first time interaction* (i.e. studying a deck of cards for the first time), as is the case in our experiments.

A very interesting observation from Table 1 is that users' subjective assessment of difficulty v' has a lower correlation to objective level of difficulty v than some of the proposed behavioral variables (e.g. $\bar{t}_a$, $\bar{t}_c$). Therefore, we can claim that those behavioral variables provide a more accurate estimation of effort than users' own estimation.

Another interesting point of Table 1 is that in terms of t_a and t_q, verbs present slightly different characteristics in comparison to abstract and concrete nouns. Moreover, the discrimination seems to be in opposite directions, i.e. t_q relating verbs is slightly larger than those relating abstract or concrete nouns ($0.70 > 0.56, 0.60$). On the contrary, t_a is slightly smaller for verbs (i.e. $0.80 < 0.93, 0.88$). The potential reason could be that actions are represented or associated in a different manner than namings and demand different neural processing [20]. We find these inferences drawn from the variables computed at the card level particularly interesting, since they have the potential of monitoring intermediate stages of learning sessions in addition to the entire course of learning [26].

4.2   Discussion on effect of stimulus length

According to literature, various potential markers have been proposed to be in relation to vocabulary difficulty, where one particular marker is word length. In particular, word length can affect the time, that is necessary for reading the query or the answer. Namely, it may take longer time for the participants to react to longer queries or answers than shorter ones. Therefore, we need to confirm that the correlation between the proposed variables is due to the level of difficulty, and not due to the -possibly- longer lengths of more difficult words.

In order to eliminate any doubts of bias due to word length, we compute word length of decks with various levels of difficulty and confirm that there is no significant relation

between length and difficulty. Tables 2-(a), (b), (c) illustrate length of the prompt and the answer in terms of characters and strokes, respectively.

It is noteworthy to mention that the characters on the front sides of the cards are ideograms (i.e. Chinese characters). Therefore, in order to quantify their *complexity*, we use both the number of ideograms and the number of strokes. In any case, we can see that there is no significant relation between word length (represented in number of ideograms or strokes) and level of difficulty. In other words, the length of an arbitrary word from one deck, is highly likely to lie withing a single standard deviation of another particular deck. This assures that any behavioral distinction is due to *vocabulary difficulty* and not due to a higher number of characters or strokes.

Table 2: Statistics of (a) query length in number of Chinese characters, (b) query length in number of strokes, (c) answer length in number of roman letters.

(a)			
	Abstract	Concrete	Verb
Easy	1.90 ±0.60	2.13 ±0.62	2.50 ±0.73
Medium	2.03 ±0.41	1.96 ±0.55	2.96 ±0.92
Hard	2.50 ±0.68	2.53 ±0.97	3.9 ±0.80
(b)			
	Abstract	Concrete	Verb
Easy	16.80 ±6.78	17.06 ±6.67	12.70 ±4.78
Medium	17.63 ±5.83	21.06 ±8.17	16.00 ±6.29
Hard	22.90 ±7.07	22.26 ±6.75	21.33 ±6.90
(c)			
	Abstract	Concrete	Verb
Easy	6.23 ±1.83	5.50 ±1.79	4.73 ±1.55
Medium	8.16 ±2.19	7.06 ±2.49	5.23 ±1.43
Hard	8.56 ±1.94	7.13 ±1.90	7.3 ±1.39

5   Conclusions

This study focuses on vocabulary learning task on digital platforms and seeks for variables that best reflect the effect of task difficulty on user behavior. To that end, vocabulary from various word classes and at different levels of difficulty are presented to learners through a spaced repetition flashcard software. Using activity logs, a set of behavioral variables are derived and the effect of task difficulty on these variables is investigated. It is observed that behavioral variables derived at the deck level present higher correlation with difficulty than those computed at card level. It is also noteworthy that the correlation of these variables are higher than the self assessment of the user regarding difficulty level, which indicates to the potential benefits of the behavioral variables. In addition, variables collected at card level show similar characteristics for abstract and concrete nouns, whereas verbs are somewhat distinguished by higher $\bar{t}_q$ and lower $\bar{t}_a$ values.

Acknowledgment

References

[1] A. Beinicke and T. Bipp, "Evaluating training outcomes in corporate e-learning and classroom training," *Vocations and Learning*, p. 501–528, 2018.

[2] E. O'Donnell, S. Lawless, M. Sharp, and V. P. Wade, "A review of personalised e-learning: Towards supporting learner diversity," *International Journal of Distance Education Technologies*, vol. 13, no. 1, pp. 22–47, 2015.

[3] G. M. Piskurich, "Online learning: E-learning. Fast, cheap, and good," *Performance Improvement*, vol. 45, no. 1, pp. 18–24, 2006.

[4] M. A. A. Dewan, M. Murshed, and F. Lin, "Engagement detection in online learning: a review," *Smart Learning Environments*, vol. 6, no. 1, p. 1, 2019.

[5] J. Keller and K. Suzuki, "Learner motivation and e-learning design: A multinationally validated process," *Journal of Educational Multimedia and Hypermedia* , vol. 29, no. 3, pp. 229–239, 2004.

[6] J. A. Fredricks, P. C. Blumenfeld, and A. H. Paris, "School engagement: Potential of the concept, state of the evidence," *Review of Educational Research*, vol. 74, no. 1, pp. 59–109, 2004.

[7] M. Cocea and S. Weibelzahl, "Can log files analysis estimate learners' level of motivation?," in *LWA*, pp. 1–4, 2006.

[8] C. R. Henrie, L. R. Halverson, and C. R. Graham, "Measuring student engagement in technology-mediated learning," *Computers & Education*, vol. 90, pp. 36–53, 2015.

[9] G. Ben-Zadok, M. Leiba, and R. Nachmias, "Drills, Games or Tests? Evaluating Students' Motivation in Different Online Learning Activities, Using Log File Analysis," *Interdisciplinary Journal of E-Learning and Learning Objects*, vol. 7, no. 1, pp. 235–248, 2011.

[10] L. Juhaňák, J. Zounek, and L. Rohlíková, "Using process mining to analyze students' quiz-taking behavior patterns in a learning management system," *Computers in Human Behavior*, vol. 92, pp. 496–506, 2019.

[11] A. Hershkovitz and R. Nachmias, "Learning about online learning processes and students' motivation through web usage mining," *Interdisciplinary Journal of E-Learning and Learning Objects*, vol. 5, no. 1, pp. 197–214, 2009.

[12] S. Chen, A. Lippert, G. Shi, Y. Fang, and A. C. Graesser, "Disengagement detection within an intelligent tutoring system," in *Proc. International Conference on Intelligent Tutoring Systems*, p. 127, 2018.

[13] Y. Ma, L. Agnihotri, R. Baker, and S. Mojarad, "Effect of student ability and question difficulty on duration," *Journal of Educational Data Mining*, pp. 135–142, 2016.

[14] J. Beck, *Engagement tracing: using response times to model student disengagement*, vol. 125. Ios Press, 2005.

[15] M. Cocea and S. Weibelzahl, "Cross-system validation of engagement prediction from log files," in *Proc. European Conference on Technology Enhanced Learning*, pp. 14–25, Springer, 2007.

[16] M. Cocea and S. Weibelzahl, "Log file analysis for disengagement detection in e-learning environments," *Journal of User Modeling and User-Adapted Interaction*, vol. 19, no. 4, pp. 341–385, 2009.

[17] S. Reddy, I. Labutov, S. Banerjee, and T. Joachims, "Unbounded human learning: Optimal scheduling for spaced repetition," in *Proc. ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 1815–1824, ACM, 2016.

[18] N. Kornell, "Optimising learning using flashcards: Spacing is more effective than cramming," *Applied Cognitive Psychology*, vol. 23, no. 9, pp. 1297–1317, 2009.

[19] D. Keder, *Computer-assisted language learning using spaced repetition*. PhD thesis, Masarykova univerzita, 2009.

[20] G. Vigliocco, D. P. Vinson, J. Druks, H. Barber, and S. F. Cappa, "Nouns and verbs in the brain: a review of behavioural, electrophysiological, neuropsychological and imaging studies," *Neuroscience & Biobehavioral Review*, vol. 35, no. 3, pp. 407–426, 2011.

[21] Z. Yücel, P. Supitayakul, A. Monden, and P. Leelaprute, "An algorithm for automatic collation of vocabulary decks based on word frequency," *IEICE Transactions on Information and Systems*, vol. E103-D, no. 8, pp. 1–10, 2020.

[22] D. S. Kerby, "The simple difference formula: An approach to teaching nonparametric correlation," *Comprehensive Psychology*, vol. 3, pp. 11–IT, 2014.

[23] D. Elmes, "Anki - friendly, intelligent flashcards." `https://ankiweb.net/about`, 2019. [Accessed 2019-03-15].

[24] C. Altiner, "Integrating a computer-based flashcard program into academic vocabulary learning," Master's thesis, Iowa State University, 2011.

[25] U. Olsson, F. Drasgow, and N. J. Dorans, "The polyserial correlation coefficient," *Psychometrika*, vol. 47, no. 3, pp. 337–347, 1982.

[26] V. Arkorful and N. Abaidoo, "The role of e-learning, advantages and disadvantages of its adoption in higher education," *International Journal of Distance Education Technologies*, vol. 12, no. 1, pp. 29–42, 2015.