PAPER

# An Algorithm for Automatic Collation of Vocabulary Decks Based on Word Frequency

Zeynep YÜCEL[†a)], Parisa SUPITAYAKUL[†], *Nonmembers*, Akito MONDEN[†],
*and* Pattara LEELAPRUTE[††], *Members*

**SUMMARY** This study focuses on computer based foreign language vocabulary learning systems. Our objective is to automatically build vocabulary decks with desired levels of relative difficulty relations. To realize this goal, we exploit the fact that word frequency is a good indicator of vocabulary difficulty. Subsequently, for composing the decks, we pose two requirements as *uniformity* and *diversity*. Namely, the difficulty level of the cards in the same deck needs to be uniform enough so that they can be grouped together and difficulty levels of the cards in different decks need to be diverse enough so that they can be grouped in different decks. To assess uniformity and diversity, we use rank-biserial correlation and propose an iterative algorithm, which helps in attaining desired levels of uniformity and diversity based on word frequency in daily use of language. In experiments, we employed a spaced repetition flashcard software and presented users various decks built with the proposed algorithm, which contain cards from different content types. From users' activity logs, we derived several behavioral variables and examined the polyserial correlation between these variables and difficulty levels across different word classes. This analysis confirmed that the decks compiled with the proposed algorithm induce an effect on behavioral variables in line with the expectations. In addition, a series of experiments with decks involving varying content types confirmed that this relation is independent of word class.

***key words:*** *e-learning, vocabulary learning, log file analysis*

## 1. Introduction and Motivation

This study focuses on e-learning systems, where a user utilizes a computer-based platform to study, review or practice a certain subject. Recently, such systems got quite popular at various levels of education, i.e. from elementary to high schools, as well as universities. In addition to their use in (organized) educational institutions, they are also a popular choice as learning medium for voluntary self-motivated pursuit of knowledge (i.e. lifelong learning).

The rapid diffusion of such systems is suggested to be due to a variety of reasons. First of all, students do not need to meet at the same time and place, which makes such systems flexible and beneficial particularly for off-curriculum studying (e.g. as a hobby) [1]. Moreover, they are economical, since they cut costs due to hiring of professionals or rooms [2]. In addition, the ease of access to a diverse range of customizable materials makes them suitable for people of all ages, experiences, and interests [3]. Furthermore, this expansion of target user group has profited substantially from the recent rapid proliferation of mobile information devices (e.g. smart phones, tablets) into daily life [4].

However, the compilation of adequate study material is quite a challenging task in e-learning systems. Fortunately, the learning platform (i.e. computer) is quite feasible for observing various reactions of the users, which provide worthful information for measuring adequacy of the study material in meeting their needs. In that respect, a substantial resource is provided by activity logs due to the fact that they can be easily recorded from a large number of users. Namely, they enable detection of regularities or deviations within users [5] allowing a comprehensive evaluation of learning systems.

These logs are particularly beneficial in obtaining certain quantitative measures of difficulty that the learners experience in remembering the information. Such measures can potentially be used in building user-specific learning material (at the appropriate difficulty level), as well as in subjective evaluation of difficulty and in pro-active reprogramming of study schedule.

In this respect, we focus on e-learning systems targeting foreign language learning [6], [7]. We specifically consider the task of memorization of vocabulary, which is a substantial part of language education. Our objective is to build vocabulary decks with desired levels of difficulty. To that end, we utilize the fact that word frequency is a good indicator of difficulty and propose an iterative algorithm to compose decks based on frequency values. In doing that, we rely on rank-biserial correlation [8] and compose decks by choosing words from word frequency lists constructed according to their numbers of occurrence in daily settings.

In experiments, we evaluated the efficacy of this approach by presenting several decks composed of non-linguistic content and linguistic content (i.e. vocabulary) with different difficulty levels. We used a spaced repetition flashcard software to display those to several participants [9]. Based on the activity logs recorded from these participants, we derived a set of behavioral variables and investigated the effect of difficulty on those variables [10]. In this manner, we confirmed that the decks with the targeted levels of difficulty induce an effect in line with the expectations. In addition, we evaluated the effect of various *word classes* (also referred as part-of-speech [11]) on behavioral variables and ascertained that the proposed method does not

Manuscript received October 14, 2019.
Manuscript revised February 6, 2020.
Manuscript publicized May 7, 2020.

[†] The authors are with Okayama University, Okayama-shi, 700–8530 Japan.

[††] The author is with Department of Computer Engineering, Faculty of Engineering, Kasetsart University, Bangkok, Thailand.

a) E-mail: zeynep@okayama-u.ac.jp

DOI: 10.1587/transinf.2019EDP7279

present significant distinction in relation to word class and can confidently be used independent of the lexical content characteristics of source corpus.

## 2. Background and Related Work

Computer based tutoring software and multimedia learning systems have a large diversity. For instance, they can have various interaction frameworks such as individual access or collaborative activities. Moreover, they may be hosted online or offline. In addition, they may use different learning objects to teach or practice such as video, test etc. [12], [13].

However, regardless of such specifics, these systems have certain common points. Here, we would like to address two particular features as (i) diversity of learning material and (ii) continuous tracing of user behavior. The diversity feature is in close relation to the flexibility of computer based learning platforms in adjusting learning material. Namely, a diverse range of materials can be upheld such that people of various ages, experiences, and interest can profit from them [3]. The second feature relates the feedback collected from users, i.e. information on users' actions and course of progress (e.g. frequency of logins, number and frequency of responses/views, time spent online) [13]–[15].

Nevertheless, this abundance of information is hard to handle, especially at the design stage. For instance, although a diverse range of material can be made available to users, it is quite difficult to decide what material is most adequate for each person. In specific, regarding our particular focus of computer based vocabulary learning, compilation of vocabulary lists out of an abundant number of words, which provide a decent correspondence to learners' current level, target level, and desired pace, stands out as a significant challenge. Note that, as *level*, we consider vocabulary difficulty rather than text coverage or reading comprehension [16].

Ideally, the learning material should be composed such that it does not frustrate the user by being too difficult, too simple, dull or monotonous. Fortunately, once learning material is compiled, the latter of the above-mentioned features (i.e. traces of actions) can be used to judge its adequacy.

In order to address the issue of building vocabulary lists with desired level of difficulty, various markers have been suggested in literature (e.g. word length and frequency) [17]. Among these markers, especially frequency has been treated in detail and ascertained multiple times to have a strong relation to word difficulty based on a variety of evaluation methods ranging from decision trees to deep recurrent neural networks [18], [19] and not only in English but other languages as well [20], [21]. This relation is possibly due to the fact that frequent use of a word or word-family can enhance peoples' familiarity to it (e.g. by associations), which in turn affects the perceived or experienced level of difficulty [22].

From a particularly didactic perspective, the adequacy of word frequency in measuring word difficulty is investigated by [23] and [24] within the context of vocabulary test construction and textbook preparation. The results of these studies confirm that word frequency is a sufficiently good and practical measure of word difficulty. Nevertheless, these studies apply to existing tests and textbooks; and do not aim to compile new material. In that respect, this study distinguishes itself from previous works by aiming compilation of new -vocabulary- learning tasks with desired difficulty relations, which is also shown to have the potential of being transferred to other study subjects than vocabulary learning.

Obviously, the main challenge here is to assess word difficulty in a fair manner and in right correspondence to users' skills. Here, we consider difficulty to be based on *de facto* properties admissible to generic users (of similar levels), unlike systems with user-specific adaptations [25].

In assessment of word difficulty, numerous previous studies rely on human coding. For instance, Sohsah et al. ask several language teachers to define word difficulty for a set of 7000 words [26], whereas Rudell relies on subjective ratings of various volunteer judges on a set of 840 words [27]. Although these studies confirm the assessed (i.e. coded) difficulty, they do not propose statistical methods for automatically *assessing difficulty*. In addition, their assessments are in strict relation to predefined grade levels. In this respect, we aim at contributing to literature by (i) *automatically assessing difficulty relation* rather than confirming consistency of codings and (ii) quantifying difficulty relation in a continuous range rather than evaluating it in relation to discrete grade levels.

On the other hand, broadly speaking, employment of computer recorded actions of the user has a long history in technology mediated learning. However, most studies consider markers of a single particular task [28] and do not consider variations of content type in the conceptual or semantic sense. Therefore, an additional and more profound evaluation of learning systems in direct relation to variations of task content, has the potential to evaluate learning systems in a more efficient manner, improving contributions of the conventional methods [29]–[31].

## 3. Experiments

We performed a set of experiments with a total of 6 subjects, 3 males and 3 females, with an age of 45 ± 5.6[†]. The subjects come from a diverse background, speak English as a foreign language[††], and reported to have similar skill degree in English corresponding to high school graduate level.

### 3.1 E-learning Software

In the experiments, we deployed free and open-source flashcard software called "Anki" [9]. Anki relies on spaced repetition and presents the memorization task to the users by mimicking conventional (i.e. physical paper-based) flashcards. Namely, each virtual flashcard emulates two "sides"

[†] The subjects are informed in a clear manner about the nature and method of the research, volunteered to participate in the experiments, gave their permission for data recording and are reimbursed with an honorarium according to the rules of the university.

[††] The subjects are all mother tongue Japanese speakers.

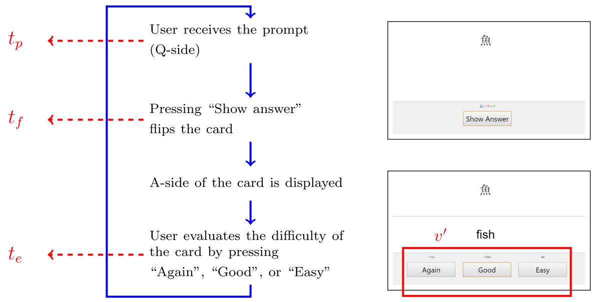

$t_p$
User receives the prompt (Q-side)
$t_f$
Pressing "Show answer" flips the card
A-side of the card is displayed
$t_e$
User evaluates the difficulty of the card by pressing "Again", "Good", or "Easy"
魚
21 + 18 + 0
Show Answer
魚
$v'$
fish
<1m
<10m
4d
Again
Good
Easy

**Fig. 1** (a) Front (Q-side) and (b) back (A-side) of a sample card from the deck of concrete nouns. For readability, the interface is displayed in English but in the actual experiments, we used a Japanese interface.

**Table 1** Probes used in monitoring user activity.

| Variable | Explanation |
|---|---|
| $t_p$ | Time of prompt |
| $t_f$ | Time of flip |
| $t_e$ | Time of evaluation |
| $d$ | Deck ID |
| $k_d$ | Card ID from deck $d$ |
| $v'$ | Evaluated level of difficulty |

as "front" and "back", which we call Q-side (i.e. question side) and A-side (i.e. answer side), respectively.

Analogous to physical flashcards, Q-side involves the "prompt", which bestirs the users towards the information on the A-side (see Fig. 1). Specifically, after the users are exposed to the prompt, they think and try to remember the corresponding answer. At this step, they are allowed to take as long time as they need. When they decide to "flip" the card, they can press "Show answer" button at the bottom of the card (see Fig. 1) and view the correct answer. After viewing the correct answer on the A-side, they can spend as long time as they wish to either memorize the answer (if they do not know it) or to confirm it (if they could correctly remember it). Before proceeding to the next card, users evaluate the ease of the current card by choosing either "Again", "Good" or "Easy", depending on how much confident they feel to "have this card committed to their memory" (see Fig. 1). This can be considered as a *subjective evaluation of difficulty* at the card level. In addition, we do not enforce a time limit on a card basis, but we let the users study a group of 30 cards, called a *deck*, for no more than 15 minutes.

## 3.2 Data Logs

The software registers several actions of the user into a log file. Specifically, the log file involves the information presented in Table 1. Here, $t_p$ denotes the instant that prompt is presented, i.e. display time of the Q-side. In addition, $t_f$ stands for the time of flip, i.e. the instant when the user presses "Show answer". Moreover, $t_e$ expresses time of subjective evaluation. Namely, it represents the instant when the user assesses the difficulty of the card by pressing "Again", "Good" or "Easy" (see Fig. 1). All variables relating time are registered as standard Unix time in milliseconds.

In addition to these values relating time course of users' actions, several identification information relating the decks and cards are registered as well. This identifying information involves $d$ and $k_d$, which are both 13-digit integer codes representing deck ID and card ID (from deck $d$), respectively. In addition, $v' \in [1, 3]$ is an integer representing the subjectively evaluated level of difficulty. In particular, $v' = 1$ denotes pressing "Easy", $v' = 2$ denotes pressing "Good", $v' = 3$ denotes pressing "Again" (see Fig. 1).

In our experiments, by design, each deck consists of 30 cards from the same *content type* and with same level of *difficulty*. Based on this design principle, we prepared 12 decks with different combinations of difficulty and content type, whose details are explained in Sect. 3.3.

## 3.3 Content Types

As for content types, we consider 4 categories. One particular category relates to non-linguistic knowledge, namely common human geography facts on country-capital associations. Three other categories relate linguistic knowledge with different word classes of English vocabulary as (i) abstract nouns, (ii) concrete nouns and (iii) verbs. Note that since not all the decks involve "vocabulary", we use the term *content type* rather than *word class*, when referring to the variations in the contents of the decks.

The reason for considering a non-linguistic content type in addition to three linguistic content types is two fold. First of all, we would like to try an alternative subject (in addition to foreign language learning). In that respect, country-capital association is a good choice, since it does not require any specific expertise or education. In addition, we use this as a step to familiarize the users with the learning tool prior to the task of vocabulary learning. Namely, we assume the users do not necessarily have any previous encounter with the particular e-learning software that we deploy. Thus, before observing their vocabulary learning behavior, we ask them to carry out a non-linguistic task to become familiar with the learning platform. This is supposed to eliminate behavioral variations due to users' computer skills.

## 3.4 Assessment of Difficulty Levels

Concerning the difficulty of a deck, we assign one of the three levels, i.e. easy, medium or hard, denoted by E, M and H, respectively. In that regard, we adjust the composition of the decks such that E is performed in a facile manner, while M is in reasonable correspondence with the skills of the users, and H requires substantial effort. Specifically, we benchmark M at the skill level of the participants, such that it is in fair correspondence with their proficiency. In addition, we build two other decks, where one is relatively easier (E) and the other is relatively harder (H). In particular, for building E, M, and H, we utilize three lists containing 200+

words each, which we compiled from reference materials (such as practice test books and online study materials). A collection of these lists constitute the *source list* $S$. For ranking the cards, we use word frequency lists by Wiktionary.

To confirm the above-mentioned correspondence between difficulty of the decks and required effort by the subjects, we pose two requirements on the composition of decks, namely *uniformity* and *diversity*. Uniformity suggests that the difficulty of the cards in the same deck needs to be uniform, such that they can be grouped together, whereas diversity requires the difficulty of the cards in different decks to be diverse, such that they can be grouped in different decks.

However, coordination of difficulty is challenging due to an additional condition relating cross-content relations, i.e. *comparability*, which implies that the difference in difficulty across different content types needs to be similar. In other words, unless the difficulty differences are well-defined across different content types, the effects due to content type cannot be distinguished.

### 3.5 Task Schedules

The experiments are performed over a 2 weeks time period and each participant studied exactly one content type on a single day. Since we limit the time devoted on each deck to 15 minutes, studying a single content type (with three decks) yields a maximum of 45 min for each user, which is regarded as a reasonable duration for daily off-curriculum activity.

In addition, as explained in Sect. 3.3, the users studied non-linguistics decks prior to vocabulary learning. This task is performed one calendar week before the other three, which are performed all in the same week (but not necessarily on consecutive days). This kind of temporal spacing is considered to guarantee that participants achieve familiarity beforehand and retain it until the experiment is over.

Moreover, the three difficulty levels regarding each content type are presented in a random order. In other words, we did not present the tasks with any gradual effort requirement (such as first E, then M, and finally H; or any other way around). On the contrary, we adjusted the schedule of the experiments such that any permutation sequence of difficulty levels is presented equal number of times over the entire experimentation period. This sort of scheduling is assumed to eliminate any bias due to the sequence of difficulty.

## 4. Proposed Method

This section presents details of the proposed method and in particular describes how we satisfy the three requirements introduced in Sect. 3.4 at once. Specifically, we follow a two stage approach. At the first stage, we build some "potential" decks based on a set of suppositions. Then, at the second stage, we revise (i.e. verify and modify) the cards in each deck in a repetitive way such that we achieve satisfactory levels of uniformity, diversity and comparability proven through quantitative evidence.

**Algorithm 1:** Building potential decks.

```
Input: Source list $S$, ranks $R$ of cards in $S$, desired deck size $D$
Output: Decks $E, M, H$,
        ranks of cards in these decks $R_E, R_M, R_H$
$N_S = |S|$                    // Size of source set $S$
$I_S = \{1, \ldots, N_S\}$          // Integers from 1 to $N_S$
$I' = I_S$              // Set of available cards $I'$
/* Arrays of decks $\Sigma$, arrays of ranks $R$, and
   median ranks $\mu$ are set to empty set.     */
$\Sigma = \emptyset, R = \emptyset, \mu = \emptyset$
for $i \leftarrow 1 : 3$ do                        // Build 3 decks
    $J = \emptyset$              // No chosen indices
    for $d \leftarrow 1 : D$ do        // Sample $D$ indices in $I'$
        $J+ = \textbf{sample}(q \in I')$
        $I' - = q$
    $X = S[J]$                // Random deck, $X$
    $r_X = R[J]$        // Ranks of cards in $X$, $r_X$
    $m_X = \textbf{median}(r_X)$        // Median of $r_X$, $m_X$
    $\Sigma+ = X, R+ = r_X, \mu+ = m_X$      // Extend arrays
$[u, v, w] = \textbf{argsort}(\mu)$    // Sort in descending order.
$E = \Sigma[u], M = \Sigma[v], H = \Sigma[w]$
$R_E = R[u], R_M = R[v], R_H = R[w]$
$S' - = \{E, M, H\}$   // Available cards in source set.
```

Obviously, non-linguistic and linguistic decks require different strategies at the first stage, since their nature bears different sets of suppositions. However, once potential decks are formed, second stage is carried out in the same manner.

### 4.1 Building Potential Decks

The assignment of difficulty levels mentioned in Sect. 3.4 is quite subjective without reference to a skill degree. This section elaborates on the details of our approach in setting reference degrees and initialization of decks.

For building the potential decks of non-linguistic content type, randomly sample three mutually exclusive sets of pairs out of all country-capital pairs[†]. For building the potential decks of linguistic content types, we predicated on the Japanese standardized English language test, STEP Eiken [33], which is benchmarked to the standard curricula of organized education institutions. In particular, we examined reference materials and compiled a source list with ~ 900 words spanning the three levels.

Specifically, for building potential decks (see Algorithm 1), we require as input, a source list $S$, ranks of cards in source list $R$, and desired deck size $D$. By choosing 3 sets of arbitrary indices and sorting them with respect to median ranks of concerning words, we initialize $E$, $M$, and $H$[††].

[†] There are 193 member states of the United Nations [32].

[††] Here, **sample**($a \in A$) refers to choosing an arbitrary element $a$ from set $A$; and $B+ = b$ refers to adding an element $b$ to set $B$.

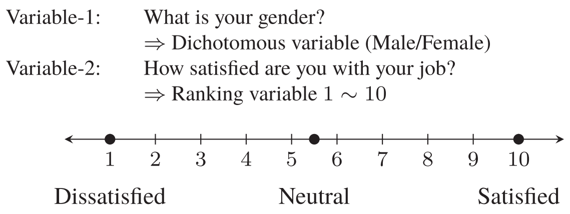

Variable-1: What is your gender?
⇒ Dichotomous variable (Male/Female)
Variable-2: How satisfied are you with your job?
⇒ Ranking variable 1 ∼ 10

**Fig. 2** An example case for evaluating RBSC.

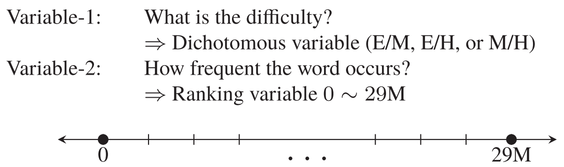

Variable-1: What is the difficulty?
⇒ Dichotomous variable (E/M, E/H, or M/H)
Variable-2: How frequent the word occurs?
⇒ Ranking variable 0 ∼ 29M

**Fig. 3** Application of RBSC on two decks with varying difficulty. In this example, Variable-2 relates linguistic content types. For non-linguistic content type, Variable-2 would be number of Google news search results.

## 4.2 Revision of Potential Decks

For revising the potential decks, we followed a strategy based on “rank bi-serial correlation” (RBSC). In what follows, we first introduce briefly the fundamental concepts of RBSC and then explain the details of our specific practice.

### 4.2.1 Rank Bi-Serial Correlation

RBSC defines a correlation coefficient relating a dichotomous variable and a ranking variable. As an example consider the case presented in Fig. 2, where the relation between gender and job satisfaction is investigated. Here, the dichotomous variable is “gender” (male or female), and the ranking variable is “satisfaction” (rated between 1 and 10).

Consider that RBSC is to be evaluated for checking the relevance of a hypothesis that “Men are more satisfied with their jobs”. Suppose that $n_m$ male subjects and $n_f$ female subjects rate their job satisfaction on a Likert scale of $[1, 10]$. To evaluate the validity of the hypothesis based on this sample, initially the job satisfaction ratings of each pair (i.e. one male subject and one female subject) are compared. Consider that the rating of a male subject $i$ is represented with $m_i$, where $m_i$ is an integer in the range $[1, 10]$ and $1 \leq i \leq n_m$. Similarly, let the rating of the female subject $j$ be represented with $f_j$, $f_j \in [1, 10]$ and $1 \leq j \leq n_f$. Let $S$ stand for the number of evidence supporting the hypothesis, (i.e. male subject is more satisfied than female subject),

$$S = \sum_{\forall i,j} \frac{\text{sign}\,(m_i - f_j) + 1}{2},$$

where sign(·) is the signum function. Assume that $C$ represents the number of evidence contradicting the hypothesis (i.e. female subject is more satisfied than male subject),

$$C = \sum_{\forall i,j} \frac{\text{sign}\,(f_i - m_j) + 1}{2}.$$

After obtaining the number of evidence for and against the hypothesis in this manner, this study employs *Simple Difference Formula* proposed by Kerby [8] to compute RBSC coefficients. According to this approach, the nonparametric correlation equals the simple difference between the proportion of “favorable” and “unfavorable” evidence, where favorable stands for the pairs supporting the hypothesis and unfavorable for the ones disagreeing with the hypothesis. In explicit terms, RBSC coefficient $\rho$ is computed simply as,

$$\rho = \frac{S - C}{S + C}.$$

and is bounded in the range from −1 to 1. If the data are all favorable, then $\rho$ is exactly 1. On the contrary, if the data are all unfavorable, $\rho$ is −1, whereas a $\rho$ of 0 indicates equal amount of favorable and unfavorable evidence.

### 4.2.2 RBSC in Practice

Following the logic of the simple example given in Sect. 4.2.1, dichotomous variables in our case are any two decks, which are contrasted between them, and ranking variables are positions of the cards, when sorted with respect to certain quantitative properties, which we will explain below.

Specifically, for sorting the cards in the non-linguistic decks, we used the number of Japanese Google news search results regarding capitals. This criterion is assumed to be a clear indicator of how much an average person can be exposed in his/her daily life to the information on the A-side of the cards from the non-linguistic decks.

Regarding linguistic content types, recently several large scale online resources have been released [34], [35]. We examined these data sets and decided to use word frequency lists provided by Wiktionary [35]. Namely, the number of occurrences of the A-sides of the cards in [35] is chosen as the ranking criterion. These frequency lists are built based on TV and movie scripts/transcripts and are quite comprehensive involving over 29 million words. Note that the same word may appear in different forms (e.g. conjugated or plural). In this respect, we account for inexact matches by considering the total number of occurrences of a -base- word irrespective of conjugations, plurals etc. In the literature, this sort of treatment is referred as “lemmatization” [23].

In our specific application, one of the hypotheses is that the cards in E deck emerge more frequently than those in M deck (either in Google news search or in daily use of English, depending on content type). In addition, we test two other similar hypotheses relating E-H and M-H pairs of decks from same content type (see Fig. 3).

Here, we consider RBSC to be an indicator of how diverse the cards in different decks are in terms of difficulty. This can also be seen as a measure of similarity of a card to other cards in the same deck as contrasted to the ones in other decks. The closer $\rho$ is to 1, the more diverse the cards

**Algorithm 2:** Getting required updates on decks.

```
Input: Decks $E$, $M$, $H$, ranks of cards $R_E$, $R_M$, $R_H$, desired
       RBSC, $\rho^*$, tolerated error $\epsilon$
Output: Required updates on decks $U_E$, $U_M$, $U_H$
$U_E$=null, $U_M$=null, $U_H$=null
$\rho_{EM} = RBSC(R_E, R_M), \rho_{MH} = RBSC(R_M, R_H)$
if $\rho_{EM} < \rho^* - \epsilon$ then $U_E$ = harder
else if $\rho_{EM} > \rho^* + \epsilon$ then $U_E$ = easier
if $(\rho_{EM} < \rho^* - \epsilon) \wedge (\rho_{MH} > \rho^* + \epsilon)$ then $U_M$ = harder
else if $(\rho_{MH} < \rho^* - \epsilon) \wedge (\rho_{EM} > \rho^* + \epsilon)$ then
    $U_M$ = easier
if $\rho_{MH} < \rho^* - \epsilon$ then $U_H$ = harder
else if $\rho_{MH} > \rho^* + \epsilon$ then $U_H$ = easier
```

**Algorithm 3:** Updating a deck.

```
Input: Deck $X$, ranks of cards in $X$ $R_X$, available cards $S'$
Output: Updated deck $X$
Get required update $U_X$ on $X$   // Easier, harder, null
if $U_X$ is easier then                  // Make $X$ easier
    /* Pick a hard card $c_h$ in $X$ to remove */
    sample($c_h \in X \mid r_h < m_X$)
    /* Pick an easy card $c_e$ in $S'$ to add to $X$ */
    sample($c_e \in S' \mid r_e > m_X$)
    $X += c_e - c_h$           // Add $c_e$, remove $c_h$
    $S' += -c_e + c_h$
else if $U_X$ is harder then             // Make $X$ harder
    /* Pick an easy card $c_e$ in $X$ to remove */
    sample($c_e \in X \mid r_e > m_X$)
    /* Pick a hard card $c_h$ in $S'$ to add to $X$ */
    sample($c_h \in S' \mid r_h < m_X$)
    $X += -c_e + c_h, S' += c_e - c_h$     // Update $X$, $S'$
```

in different decks are. Equivalently, this can be considered as a reduction in uniformity, i.e. interference of difficulty levels.

In order to address these issues, we propose to revise (i.e. update) the potential decks so as to adjust RBSC coefficients $\rho$, relating the three permutations of difficulty (E-M, M-H and E-H), $\rho_{EM}$, $\rho_{MH}$, $\rho_{EH}$. In particular, we set a desired RBSC value $\rho^*$ and determine the sorts of updates on each deck, which make $\rho$ converge to $\rho^*$ (see Algorithm 2). The required updates can be either *easier*, *harder* or *null* (i.e. no-update). Subsequently, for applying a particular update on a deck $X$, e.g. for making $X$ *easier*, we remove a hard card from $X$ (i.e. a card with rank below the median rank of $X$, $m_X$) and return it to the set of available cards $S'$. We then choose an *easy* card from $S'$ (i.e. with a rank larger than $m_X$) and append it to $X$ (see Algorithm 3).

We alternate between Algorithms 2 and 3, until the termination criterion is achieved, namely, $\rho_{EM}$ and $\rho_{MH}$ are in $\pm\epsilon$ interval around $\rho^*$. In our experiments, we observed that the termination criterion is satisfied in several runs, since our source set is sufficiently large as compared to the desired deck size. Nevertheless, we assume that the size of source set is unlikely to be a problem, particularly for vocabulary learning, since there is always an abundant number of words to be learned at any level. In addition to source set size, ill-conditioned ranks can be a bottleneck for the proposed method. For instance, if we have same number of occurrences in Wiktionary for most words or same number of search results in Google news, it may not be possible to choose cards (or words) to add or replace. Nevertheless, we consider such pathological cases to be virtually non-existent.

**Table 2** RBSC coefficients of potential and revised decks.

| | Potential decks | | | Revised decks | | |
|---|---|---|---|---|---|---|
| | **E-M** | **M-H** | **E-H** | **E-M** | **M-H** | **E-H** |
| Non-linguistic | 0.19 | 0.15 | 0.28 | 0.82 | 0.79 | 0.96 |
| Abstract nouns | 0.22 | 0.29 | 0.99 | 0.76 | 1.00 | 0.98 |
| Concrete nouns | 0.19 | 0.41 | 0.91 | 0.75 | 0.98 | 0.87 |
| Verbs | 0.41 | 0.50 | 0.85 | 0.70 | 1.00 | 1.00 |

Table 2 presents $\rho$ values regarding the potential decks as well as the revised decks. Here, the values relating the potential decks can be considered as performance of a *baseline* approach based on random sampling, whereas the contribution of the proposed method can be judged through the values relating revised decks. It can clearly be seen that for non-linguistic decks, where there is a single source set for all three decks, the initial $\rho$ values can be quite lower (0.19~0.28) than what is possible to achieve (0.79~0.96). On the other hand, when there is limitation on the source set (regarding what can be E, M or H), the initial values are slightly higher for nouns regarding E-M and M-H pairs (0.19~0.50), and much higher E-H pairs (0.85~0.99). This is due to the fact that the words at E and H source lists are, as expected, already quite different in difficulty due the predication on Eiken grade levels, limiting the contribution of the algorithm. Nevertheless, the benefits of the algorithm are quite clear, when one contrasts $\rho$ relating E-M and M-H pairs of potential decks to those of revised decks.

In addition, comparing $\rho$ relating deck pairs from different content types, we can see that, for instance, non-linguistic E-M pair has lower uniformity and diversity than abstract noun E-M pair. At this point, we would like to point out the difficulty of dealing with small corpora. Specifically, there is a slight discrepancy between the $\rho$ values of M-H pairs from different content types. Namely, $\rho_{MH} = 0.79$ relating non-linguistic content type is further away from the values concerning linguistic content types. The reason is suggested to be the fact that linguistic decks are built out of a larger pool of options, while the non-linguistic decks inherently appertain to a much more limited repertoire. In this small repertoire, we have very little space for freedom, which clearly reflects on comparability of deck pairs.

To the best of our knowledge, there is no guideline for distinctness (or interference) in terms of $\rho$. Nevertheless, we consider the values of revised decks in Table 2 to be adequate for covering a certain range of diversity (or uniformity). Moreover, the similarity of the values on a column basis is considered to indicate that the comparability requirement is satisfied considerably well. Thus, the constructed learning material is assumed to stimulate diverse behavioral patterns.

## 5. Analysis of Subjects' Behavioral Patterns

While the participants study the decks constructed as explained in Sect. 4, we collect the data logs introduced in Sect. 3.2. From these logs, we derive several behavioral variables summarizing user reactions and investigate the correlation of these variables with varying difficulty across content types as explained below.

### 5.1 Behavioral Variables

The log files are analyzed and certain markers describing the level of required effort by the subjects to practice each deck are derived. Specifically, a total of 4 behavioral variables are considered, where 2 of them are collected at *deck level* and 2 of them are collected at *card level* and then averaged.

In particular, at the deck level, we collect the following:

- $n_{\mathrm{tot}}$ denotes the total number of card displays, i.e. how many cards are viewed by the user from a single deck, allowing for multiple counts of the same card.
- $n_{\mathrm{avg}}$ denotes the average number of displays over all cards. Namely, the participant may choose to view a card that he/she does not feel confident in remembering. As a results, such cards are displayed multiple times. In that regard, we count how many times each card is displayed and compute their average denoted with $n_{\mathrm{avg}}$.

Obviously, if the study material (i.e. deck content) is facile with respect to the skill level of the subject, he/she is likely to finish studying it within the allocated time. However, if the deck is difficult, the subject may consume the allocated time without accomplishing the goal (of exhausting the cards).

In that case, a high value of $n_{\mathrm{tot}}$ is likely to be attained. For this reason, $n_{\mathrm{tot}}$ can be considered to be in indirect relation to the level of difficulty of the deck. On the other hand, $n_{\mathrm{avg}}$ is supposed to increase steadily for each card as difficulty level increases. In that regard, it enables testing the uniformity requirement. In other words, both of the two behavioral variables at the deck level, are expected to increase with deck difficulty, but only $n_{\mathrm{tot}}$ is affected by the ability of the user to finish the deck.

In addition, we collect three behavioral variables collected at the card level as follows. Let $D$ represent the set of all deck IDs and consider $d$ to be an arbitrary deck ID $d \in D$. Furthermore, suppose $K_d$ represents the set of card IDs belonging to deck $d$. In addition, suppose that $k_d$ represents an arbitrary card ID from deck $d$, $k_d \in K_d$.

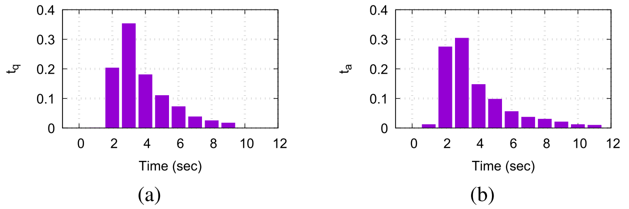
Bearing these definitions in mind, the two behavioral variables at card level can be expressed explicitly as follows:

- $t_q[d]$ is the mean value of time periods, which the user spends on the Q-sides of the cards belonging to deck $d$,

  $$t_q[d] = \frac{\sum_{k \in K_d} t_f[d,k] - t_p[d,k]}{|K_d|},$$

  where $|\cdot|$ stands for cardinality
- $t_a[d]$ is the mean value of time periods, which the user spends on the A-sides of the cards belonging to deck $d$,

  $$t_a[d] = \frac{\sum_{k \in K_d} t_e[d,k] - t_f[d,k]}{|K_d|}.$$



(a) (b)

**Fig. 4** The distribution of time spent on (a) Q-side and (b) A-side of the cards relating all the content types and difficulty levels after pre-processing.

### 5.2 Pre-Processing

As we examined the distribution of behavioral variables collected at card level, we noticed several outliers. In particular, some cards have unusually long $t_q$ or $t_a$ values. This is possibly due to the users, who take a break from studying and rest for a few minutes. Since these cases are quite few in quantity and also not represent an act of studying, we regard them not to represent the "normal" learning behavior. In order to filter them out, rather than setting an explicit hard threshold value on $t_q$ or $t_a$, we simply preserve the data belonging to the lower 95% of all the values.

Suppose that for a particular behavioral variable at card level, i.e. $t_q$ or $t_a$, we have the observation set $\boldsymbol{\alpha} = \{\alpha\}$ relating all $d$ and $K_d$. We compute the relating normalized histogram $h_\alpha$ with a certain bin size $\delta$. Obviously,

$$h[i] = \frac{|\{\alpha | \alpha \in \boldsymbol{\alpha}, i\delta < \alpha \leq (i+1)\delta\}|}{|\boldsymbol{\alpha}|\delta}.$$

The observations belonging to the lower 95% of all the values should be below a value $T$ such that

$$T = \arg\max_{\tau} \left( \sum_{i=1}^{\tau} h[i]\delta < 0.95 \right).$$

After preserving the observations falling in the desired range (i.e. less than $T$), probability density functions of $t_q$ and $t_a$ regarding all decks and participants are found as in Fig. 4.

### 5.3 Correlation of Behavioral Variables and Difficulty

Polychoric and polyserial correlations are measures of bivariate association, where at least one of the variables is an ordinal random variable. Namely, polychoric correlation deals with two ordinal variables, whereas polyserial correlation defines the correlation between a quantitative variable and an ordinal variable [36]. In our case, since word difficulty is ordinal and behavioral variables are quantitative, we employ polyserial correlation to determine the nature and degree of the relation between them. Specifically, in computing polyserial correlation, one assumes the joint distribution of the quantitative variable and a latent continuous variable underlying the ordinal variable is bivariate normal.

## 6. Results and Discussion

Polyserial correlation values illustrated in Table 3 represent the relation between the proposed set of behavioral variables and three difficulty levels concerning each content type.

In Table 3, behavioral variables computed at the deck level (i.e. $n_{\text{tot}}$ and $n_{\text{avg}}$) have in general a higher correlation with difficulty, than those computed at the card level (i.e. $t_a$ and $t_q$). Moreover, comparing columns 1 and 2 of Table 3, it is seen that the correlation relating $n_{\text{avg}}$ is always larger than the correlation relating $n_{\text{tot}}$, although the difference is quite limited. From these observations, we can deduce that the more difficult the deck is, the more often its cards appear, confirming our expectations that with increasing level of difficulty, the user needs to study more (i.e. viewing cards a higher number of times). In addition, comparing columns 3 and 4 of Table 3, it is observed that the correlation relating $t_a$ is quite higher than that of $t_q$ for all content types. Although $t_a$ covers both review and evaluation of the A-side, we expect that reviews would be a more dominant contributor in $t_a$ than evaluations, due to their variability (as compared to the monotony of evaluation). In addition, such variability is expected to have a stronger effect, in particular, as users improve their familiarity with the e-learning tool. Thus, we assume that the subjects spent a longer duration of time on the A-side of card (possibly for memorizing or confirming) proportional to its level of difficulty. However, for the Q-side of the cards (i.e. receiving the prompt), they need an approximately equal duration of time regardless of level of difficulty. In other words, $t_q$ does not depend strongly on the level of difficulty. Nevertheless, we observe that all values in column 3 are larger than 0.

In Tables 4-(a), (b) and (c) the polyserial correlation values are illustrated, representing the relation between various behavioral variables and a pair of difficulty levels, which are E-M, M-H and E-H, respectively.

Judging from the values shown in Table 4-(a), subjects' behaviors both at deck level (i.e. $n_{\text{tot}}$ and $n_{\text{avg}}$) and at card level (i.e. $t_a$) are quite different, which suggests that E and H decks have a quite distinct composition. In other words, the high values at columns 1, 2, and 4 of Table 4-(a) indicate that as the difficulty increases a significant behavioral variation is observed such that a higher number of cards are studied and for a longer duration of time.

Although the M-H pair has in general the lowest correlation values, particularly the non-linguistic content type in the M-H pair sticks out as the least correlated content type of this difficulty pair (see row 1 of Table 4-(c)). This is considered to be due to the fact that the repertoire of the non-linguistic content type is quite narrow, especially in comparison to the word frequency lists used to compile the decks of other three content types. This effect is related to the limited freedom in choosing Q-A pairs (i.e. presenting fewer options).

**Table 3** Polyserial correlation coefficients between behavioral variables and difficulty levels for varying content types.

| Content type | $n_{\text{avg}}$ | $n_{\text{tot}}$ | $t_q$ | $t_a$ |
|---|---|---|---|---|
| Non-linguistic | 0.69 | 0.67 | 0.23 | 0.69 |
| Abstract noun | 0.89 | 0.84 | 0.55 | 0.92 |
| Concrete noun | 0.95 | 0.90 | 0.60 | 0.87 |
| Verb | 0.89 | 0.82 | 0.69 | 0.79 |

**Table 4** Polyserial correlation coefficients between behavioral variables and difficulty levels of (a) E and H, (b) E and M and (c) M and H pairs.

(a)

| Content type | $n_{\text{avg}}$ | $n_{\text{tot}}$ | $t_q$ | $t_a$ |
|---|---|---|---|---|
| Non-linguistic | 0.91 | 0.84 | 0.34 | 0.86 |
| Abstract noun | 0.99 | 0.97 | 0.68 | 0.99 |
| Concrete noun | 0.99 | 0.99 | 0.76 | 0.99 |
| Verb | 0.99 | 0.96 | 0.82 | 0.99 |

(b)

| Content type | $n_{\text{avg}}$ | $n_{\text{tot}}$ | $t_q$ | $t_a$ |
|---|---|---|---|---|
| Non-linguistic | 0.77 | 0.79 | 0.41 | 0.70 |
| Abstract noun | 0.99 | 0.99 | 0.26 | 0.85 |
| Concrete noun | 0.99 | 0.99 | 0.62 | 0.99 |
| Verb | 0.77 | 0.85 | 0.40 | 0.55 |

(c)

| Content type | $n_{\text{avg}}$ | $n_{\text{tot}}$ | $t_q$ | $t_a$ |
|---|---|---|---|---|
| Non-linguistic | 0.29 | 0.29 | -0.08 | 0.46 |
| Abstract noun | 0.85 | 0.66 | 0.55 | 0.87 |
| Concrete noun | 0.95 | 0.85 | 0.35 | 0.65 |
| Verb | 0.89 | 0.76 | 0.69 | 0.64 |

In addition, comparing rows 2, 3, 4 of each of Tables 4-(a), (b) and (c), it is seen that there is no obvious behavioral variation pattern depending on the linguistic content type. In other words, the proposed method yields equally distinct decks, regardless of the word class of the cards.

Furthermore, comparing corresponding entries of all three tables, we can see that in general the values are higher for the E-H pair, which is followed by the E-M pair and then by M-H pair. That is to say, E-H pair presents a much clear contrast than E-M, which in turn has a higher distinction than M-H. This could be due to the fact that certain rare words can be used and heard, but without full comprehension or with a conceptual interpretation (i.e. guessing from context).

Even though the number of participants (i.e. 6) is not large, previous studies investigating sufficient sample size for efficient identification of usability problems [37] indicate that standard deviation of estimations are already at reasonable levels for set sizes as in this study. In addition, we believe that careful selection and recruitment of participants with uniform specifications support ensuring reliability of the statistics presented in Tables 3 and 4.

## 7. Conclusion and Future Work

This study offers using word frequency as a marker for vocabulary difficulty and devises an iterative method for building decks with desired relative difficulty relations. To confirm the efficacy of the proposed method, we carried out extensive experiments with a diverse range of subjects, content types and difficulty levels. Based on our experimental results, we confirmed that there exists a significant parallel between the RBSC coefficients, which quantify the relation between difficulty and frequency, and polyserial correlation coefficients, which express the variations in subjects' behavior due to different levels of difficulty. Moreover, this relationship is more pronounced, when we contrast the pair of E and H, in comparison to the pair of E and M or to the pair of M and H. In addition, between the pairs of E-M and M-H, we can see that there is a stronger relation between frequency and difficulty for the E-M case, whereas the distinction between M-H pair is relatively lower. This could be due to over-estimation of the subjects regarding their self-assessment. Nevertheless, the polyserial correlation coefficients, being always on the positive side, suggest that word frequency is a good indicator of vocabulary difficulty. In addition, coefficient values relating different word classes show that this relation is independent of word class, and the proposed method can be used to compile vocabulary lists composed of any word class.

In addition, one potential capability of the proposed method, is expansion of word pools used in building decks to a larger set, which is not necessarily entirely coded. In other words, while benchmarking a deck to user's skills based on a coded list of words is essential, relatively easier or harder decks (in respect to the benchmark) can potentially be built using uncoded corpora. Moreover, for a study subject, that requires no specific training, the proposed method can actually entirely eliminate coding effort, provided that a reasonable ranking is available (e.g. number of search results).

E-learning systems that do not provide objective assessment of performance have to rely on subjective evaluation of difficulty to reprogram learning schedule. The objective observables proposed in this work, could be incorporated with such systems. For instance, future learning schedule can be based on a learning algorithm using weighted information from the proposed observables and users' self-evaluation.

## Acknowledgments

This work was supported in part by JSPS KAKENHI Grant number J18K18168, Japan.

### References

[1] A. Beinicke and T. Bipp, "Evaluating training outcomes in corporate e-learning and classroom training," Vocations and Learning, pp.1–28, 2018.

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

[3] 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.

[4] Y.-T. Sung, K.-E. Chang, and T.-C. Liu, "The effects of integrating mobile devices with teaching and learning on students' learning performance: A meta-analysis and research synthesis," Computers & Education, vol.94, pp.252–275, 2016.

[5] M. Cocea and S. Weibelzahl, "Can log files analysis estimate learners' level of motivation?," Workshop on Lernen, Wissensentdeckung, Adaptivität, pp.1–4, 2006.

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

[7] P.J. Groot, "Computer assisted second language vocabulary acquisition," Language learning & technology, vol.4, no.1, pp.56–76, 2000.

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

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

[10] D. Keder, Computer-assisted language learning using spaced repetition, Ph.D. thesis, Masarykova univerzita, 2009.

[11] C. Quiles, K. Kūriákī, and F. López-Menchero, A grammar of modern Indo-European, Indo-European Association, 2012.

[12] 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, 2017.

[13] 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.

[14] 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.

[15] 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.

[16] B. Laufer and G.C. Kalovski, "Lexical threshold revisited: Lexical text coverage, learners' vocabulary size and reading comprehension," Reading in a foreign language, vol.22, no.1, pp.15–30, 2010.

[17] J. Medero and M. Ostendorf, "Analysis of vocabulary difficulty using wiktionary," International Workshop on Speech and Language Technology in Education, 2009.

[18] H.M. Breland, "Word frequency and word difficulty: A comparison of counts in four corpora," Psychological Science, vol.7, no.2, pp.96–99, 1996.

[19] G. Paetzold and L. Specia, "Semeval 2016 task 11: Complex word identification," International Workshop on Semantic Evaluation, pp.560–569, 2016.

[20] S.M. Yimam, C. Biemann, S. Malmasi, G.H. Paetzold, L. Specia, S. Štajner, A. Tack, and M. Zampieri, "A report on the complex word identification shared task 2018," arXiv:1804.09132, 2018.

[21] J. Lee and C.Y. Yeung, "Automatic prediction of vocabulary knowledge for learners of chinese as a foreign language," International Conference on Natural Language and Speech Processing, pp.1–4, IEEE, 2018.

[22] I. Nation, "How large a vocabulary is needed for reading and listening?," Canadian Modern Language Review, vol.63, no.1, pp.59–82, 2006.

[23] K. Chujo, "Measuring vocabulary levels of english textbooks and tests using a bnc lemmatised high frequency word list," in English corpora under Japanese eyes, pp.231–249, Brill Rodopi, 2004.

[24] J.M. Tamayo, "Frequency of use as a measure of word difficulty in bilingual vocabulary test construction and translation," Educational and Psychological Measurement, vol.47, no.4, pp.893–902, 1987.

[25] Y. Ehara, N. Shimizu, T. Ninomiya, and H. Nakagawa, "Personalized reading support for second-language web documents," ACM T.

Intelligent Systems and Technology, vol.4, no.2, p.31, 2013.

[26] G.N. Sohsah, M.E. Ünal, and O. Güzey, "Classification of word levels with usage frequency, expert opinions and machine learning," British Journal of Educational Technology, vol.46, no.5, pp.1097–1101, 2015.

[27] A.P. Rudell, "Frequency of word usage and perceived word difficulty: Ratings of kučera and francis words," Behavior Research Methods, Instruments, & Computers, vol.25, no.4, pp.455–463, 1993.

[28] 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.

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

[30] J. Beck, "Engagement tracing: using response times to model student disengagement," Artificial intelligence in education: Supporting learning through intelligent and socially informed technology, vol.125, p.88, 2005.

[31] E. Yamamoto and H. Isahara, "Related word lists effective in creativity support," IEICE Trans. Inf. & Syst., vol.E90-D, no.10, pp.1509–1515, 2007.

[32] United Nations, "Member states." http://www.un.org/en/member-states/, 2019. [Accessed 2019-03-19].

[33] Eiken Foundation of Japan, "STEP Eiken." http://www.eiken.or.jp/eiken/en/, 2019. [Accessed 2019-03-19].

[34] B. Settles, "Data for the 2018 duolingo shared task on second language acquisition modeling." https://doi.org/10.7910/DVN/8SWHNO, 2018.

[35] Wiktionary, "Frequency lists." https://en.wiktionary.org/wiki/Wiktionary:Frequency_lists/TV/2006/explanation, 2019. [Accessed 2019-03-19].

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

[37] J. Nielsen and T.K. Landauer, "A mathematical model of the finding of usability problems," INTERCHI, ed. B. Arnold, G.C. van der Veer, and T.N. White, pp.206–213, ACM, 1993.

**Zeynep Yücel** is an assistant professor at Okayama University, Japan. She obtained her BS degree from Bogazici University, Istanbul, Turkey, and her MS and PhD degrees from Bilkent University, Ankara, Turkey in 2005 and 2010, in electrical engineering. She was a postdoctoral researcher at ATR labs in Kyoto, Japan, before being awarded a JSPS fellowship. Her research interests include robotics, computer vision, and pattern recognition.

**Parisa Supitayakul** is a master course student in the Division of Electronic and Information Systems Engineering, Graduate School of Natural Science and Technology, Okayama University. She received the BE degree in software engineering from Kasetsart University in 2018. Her research interests include human behavior understanding.

**Akito Monden** is a professor in the Graduate School of Natural Science and Technology at Okayama University, Japan. He received the BE degree (1994) in electrical engineering from Nagoya University, and the ME and DE degrees in information science from Nara Institute of Science and Technology (NAIST) in 1996 and 1998, respectively. His research interests include software measurement and analytics, and software security and protection. He is a member of the IEEE, ACM, IEICE, IPSJ and JSSST.

**Pattara Leelaprute** is currently an Assistant Professor at Department of Computer Engineering, Faculty of Engineering, Kasetsart University, Thailand. He received his B.E. (2001) in Information and Computer Science, M.E. (2003) in Computer Science, and Ph.D. (2006) in Information and Systems Engineering from Osaka University, Japan. His research interests include Feature Interactions, Home Network Systems, Big Data and Software Engineering. He is a member of IEICE.