Computer Assisted Language Instruction Consortium

CALICO JOURNAL (ONLINE) ISSN 2056–9017

Article

# The effectiveness of computer-based spaced repetition in foreign language vocabulary instruction: a double-blind study

*Evgeny Chukharev-Hudilainen*[1] *and Tatiana A. Klepikova*[2]

## Abstract

*The purpose of the present paper is twofold; first, we present an empirical study evaluating the effectiveness of a novel CALL tool for foreign language vocabulary instruction based on spaced repetition of target vocabulary items. The study demonstrates that by spending an average of three minutes each day on automatically generated vocabulary activities, EFL students increased their long-term vocabulary retention rate three fold. Second, we demonstrate that the double-blind experiment design, which has become standard research practice in such extremely high-stakes fields as pharmacology and healthcare, has the potential of being successfully implemented in CALL research.*

KEYWORDS: ACTIVITY GENERATION; DOUBLE-BLIND DESIGN; VOCABULARY ACQUISITION

## Introduction

The importance of building a working knowledge of vocabulary in learners of English as a foreign language (EFL) cannot be overestimated. Because students are not immersed in the target language environment, diminishing the possibilities for incidental vocabulary acquisition, vocabulary learning becomes

**Affiliation**

[1]Program in Applied Linguistics and Technology, Iowa State University, Ames, IA, USA.

[2]Department of English and Translation Studies, St. Petersburg State University of Economics, Russia.
email: evgeny@iastate.edu (corresponding author)

CALICO JOURNAL VOL 33.3 2016 334–354

doi : 10.1558/cj.v33i3.26055

a time- and effort-consuming task. Limited contact time in many EFL classes makes this challenge even more salient.

In addressing the challenge of foreign language (FL) vocabulary learning, the potential of various CALL approaches has been investigated: training students to use online corpora (Gordania, 2012) or web-based dictionaries (Al-Jarf, 2007; Ranalli, 2013), employing CALL tools to provide students with multimodal annotations (glosses) for vocabulary items (Gorjian, Moosavinia, Kavari, Asgari, & Hydarei, 2011; Poole, 2012), designing mobile applications with pre-created sets of activities (Stockwell, 2007), or using flashcard-based spaced repetition programs (Nakata, 2011; Zhu, Fung, & Wang, 2012), which rely on pair-associate vocabulary learning. The present paper will focus on the latter approach, dealing with pair-associate learning, because it can be efficiently applied to any set of vocabulary items without the need for the instructor to develop specific activities manually.

## Literature review

The history of spaced repetition (the concept underlying flashcard-based software) traces back to the nineteenth century, when Ebbinghaus (1885) hypothesized that the rate at which people forget information increased exponentially with time, but if an item was repeatedly revised, it tended to be forgotten at a slower rate, the latter gradually decreasing with each repetition. Further research in cognitive psychology of memory and learning articulated the following two principles (Nakata, 2008, pp. 5–6):

1. a successful recall from memory yields superior retention to mere presentation of the target item; and
2. successfully recalling an item from memory after a delay is more effective than recalling it immediately after we learn it.

The two principles are contradictory, because the former calls for intervals between repetitions *small enough* so that the item may be successfully recalled, while the latter, on the other hand, recommends using *longer* intervals. It can be inferred, therefore, that in an efficient learning strategy, items should be reviewed *just when they are about to be forgotten*. It is, of course, a challenge to accurately identify the moment in time when the item is on the verge of oblivion, but remains in the learner's memory.

A number of algorithms have been devised to estimate such moments and thus create optimal revision schedules based on the learner's prior experience with the items to be learned. Two prominent examples include "Leitner's learning box" (Leitner, 2011) and SuperMemo (Wozniak, 1990), with the latest (fifteenth) revision of the latter implemented in 2011 (www.supermemo.com). Although authors of particular spaced-repetition algorithms may argue for

their uniqueness and high relative efficiency, research suggests that differences between spacing schedules may not be that important. Specifically, when Karpicke and Bauernschmidt (2011) explored various relative spacing schedules (expanding, equal, and contracting), they found no evidence for a particular schedule being inherently superior to another. Interestingly, although expanding schedules afforded a pattern of increasing retrieval difficulty across repeated tests, this did not translate into gains in long-term retention.

In a comprehensive review of nine flashcard programs for learning vocabulary based on spaced repetition, Nakata (2011) notes that while, in general, such programs are developed in a way that maximizes vocabulary learning, they all have some room for improvement. According to the researcher, the most notable shortcoming of the existing vocabulary learning software is that "none of the programs is designed to encourage generative use of target words" (Nakata, 2011, p. 33). Additionally, the programs were found to be limited in their ability to increase retrieval effort along the course of instruction: most of the programs do not provide sufficient support for data entry, do not take advantage of the word frequency information, and have limited support for multiple-choice exercises (Nakata, 2011). The fact that the existing programs mostly aim at autonomous learners may be viewed as an advantage, but at the same time, teachers willing to use flashcard software in a classroom setting are given little (if any) control over the material presented to the students, therefore the extent to which the existing tools can be integrated into the curriculum may be limited. In addition to the limitations of the existing flashcard-based space repetition software, researchers have drawn attention to the more fundamental controversy around the essentially behavioral nature of the pair-associate paradigm (Hulstijn, 2001). While research suggests that students find flashcards to be a useful learning tool (Wissman, Rawson, & Pyc, 2012), there is no denying that vocabulary acquisition is a complex process encompassing many aspects of the word knowledge beyond the simple "form-meaning" mapping (Nation, 2001).

## Our approach

The limitations outlined above, along with the lack of research into the psycholinguistic adequateness and comparative practical benefits of different spaced-repetition algorithms, prompted the authors of this paper to develop a new computer-based tool for vocabulary learning. The tool (named Linguatorium) takes the form of a web-based system that runs in any modern Web browser on both personal computers and touch-screen tablets. The system addresses the above-mentioned limitations of the existing vocabulary learning tools in the following ways.

1. The system was designed to generate custom exercises (activities) for the students, including multiple-choice exercises with automatic selection of distractors along with spelling and "fill-in-the-blank" exercises that promote some degree of generative use of vocabulary. For a discussion of activities that the system can generate see "Generating Activities" below.

It should be noted that the absence of vocabulary production exercises in many programs and the limited extent to which other tools (including ours) support the productive use of vocabulary is to a large extent explained by the inherent limitations of automated processing of natural language semantics in computational systems (Piotrowski, 1999, p. 229). Being aware of this limitation, as well as the controversy around the behavioral "pair-associate paradigm" cited above, we intended our system as a supplemental tool, rather than a comprehensive vocabulary learning solution. Our aim, therefore, was to minimize the time the student would be required to spend on working with the system, so that adequate room for other activities could be made in the curriculum.

2. The system allows the instructor to enter custom wordlists (called "lexical themes") based on his or her curriculum, while supporting the data entry process by automatically retrieving definitions and semantic information from online dictionaries and WordNet (Miller, 1995), performing automated Google searches to retrieve images that the instructor might want to consider including in the lexical unit cards for "imageable" vocabulary, and employing frequency data from corpora. The system allows the instructor to assign multiple wordlists to the students and specify "due dates," that is, dates by which each of the lists should be acquired by the students. The system will prioritize the order in which lexical items are introduced to the students based on the due dates. This capability, along with individual and aggregated progress reports, provides the flexibility needed for the integration of a CALL tool into the classroom.

3. The system employs an adaptive tutoring algorithm developed by one of the authors. The algorithm is driven by computational models of student lexical memory, and is presented in detail in "The adaptive tutoring algorithm" below. While several studies (Labrie, 2000; Nakata, 2008; Oberg, 2011) have found no statistically significant differences in learning outcomes between CALL-based vocabulary learning and paper-based approaches, it would be reasonable to expect that implementing an adaptive tutoring algorithm (impossible or at least impractical in the paper-based paradigm due to the associated computational complexity) may add value to a CALL tool. Student modeling has been long recognized as an essential component of an effective intelligent tutoring system. Since information about the learner's knowledge of L2 is not directly accessible by the system, it needs to be inferred from the learner's responses to practice tasks (Brown, 2002, pp. 343–344).

Theoretically, the development of Linguatorium was rooted in the concept of Linguistic Automaton and the cybernetic approach to instruction as regulation and control. Linguistic Automaton (Piotrowski, 1999; Piotrowski & Beliaeva, 2005) is a theoretical framework for creating computational models of the human verbal-mental activity. The cornerstone of the model is modular architecture, allowing for "graceful fallbacks" in case the linguistic information fails to be processed at a certain stage of analysis. Initially devised primarily for the purposes of machine translation, the concept has been successfully applied to linguistically aware intelligent tutors (Beliaeva, 2007). An example of graceful fallback in the context of a vocabulary tutoring system would be a case when the system might not be able to use a lexical unit in certain types of activities due to the lack of semantic information available about the unit, but would still incorporate it into simpler activity types that do not rely on computational semantics. The cybernetic approach (Rastrigin & Erenštejn, 1988) treats the process of instruction as a complex regulatory system, wherein every instructional session is formalized as a regulatory action, through which the tutor affects certain internal parameters of the student in a way desirable for achieving the learning goals. This approach, originally developed specifically for FL vocabulary learning, provides a formal framework for conceptualizing and implementing various tutoring systems.

Our system was designed with an application programming interface (API) for automated manipulation of the tutoring algorithm parameters. Specifically, the API allows for arbitrary internal labels to be randomly assigned to both participants and lexical units, and for the behavior of the system to be altered based on combinations of such labels. This functionality provided the infrastructure required for double-blind empirical studies of vocabulary acquisition.

The need for double-blind studies in CALL research required some justification. It is well known that in such high-stakes fields as evidence-based medicine or pharmacology, double-blind and triple-blind clinical trials have become the standard of experimental design, protecting researchers against placebo effects, observer bias, and conscious deception (Davidoff, Haynes, Sackett, & Smith, 1995). In pedagogical research, non-blind randomized controlled trials to date have been the highest standard of evidence-based research design (What Works Clearinghouse, 2011, pp. 11–16). Undoubtedly, blinding is difficult to achieve in pedagogical research because the participating instructors are aware of the teaching methods used in the intervention and the comparison groups (Jones, Gebski, Onslow, & Packman, 2002). Arguably, the introduction of blinding procedures into experimental design may prevent the instructors' personal enthusiasm about and attitude towards particular teaching methods from adversely affecting the study outcomes. However, to date, the feasibility of blinding in applied linguistic research has not been investigated.

Based on the above, we aimed the present study at answering two research questions.

1. How do automatically generated supplemental activities based on spaced repetition improve vocabulary learning gains in EFL students?
2. Can double-blind experimental design be successfully implemented in a CALL-based pedagogical intervention study?

In the remaining parts of the present paper, we will discuss the details of the adaptive tutoring algorithm used in the system, paying special attention to the design of the student memory model; describe the methodology of the present study; and, finally, present and discuss the empirical findings.

## The adaptive tutoring algorithm

As mentioned above, Linguatorium is based on a novel adaptive algorithm, which controls spaced repetition of target vocabulary items. The algorithm operates on vocabulary lists (lexical themes) provided by the instructor. To ensure proper synchronization with classroom activities, it is recommended that the content of the lexical themes be selected to correspond to units of instruction in the curriculum. The instructor enters lexical units into the system along with glosses (translations or definitions, images, associations, and additional comments) and context usage examples.

Each student's lexical memory is formalized as four non-intersecting sets of lexical units:

*N* – units assigned to the student and pending introduction;
*P* – units in the process of active acquisition;
*S* – units in the student's short-term memory;
*L* – learned units (in the long-term memory).

The term *short-term memory* has been used in cognitive psychology to describe the capacity for holding a small amount of information in mind for a short period of time, in the order of seconds, which can be prolonged if the information is rehearsed (cf. Davelaar, Goshen-Gottstein, Haarmann, & Usher, 2005; Murdock, 1972). For the purposes of the present algorithm, a lexical unit is deemed to be in the student's short-term memory after it has been successfully recalled at least four times in various types of activities (of increasing difficulty), without significant intervals between successive repetitions. In contrast, a unit is considered to be in the long-term memory (in other words, to be "fully learned") after it has stayed in the short-term memory for at least seven days, and is still successfully recalled by the student. The authors of this paper are unaware of any prior work on distinct stages of word acquisition

in vocabulary tutoring algorithms, therefore, the above-mentioned thresholds were selected arbitrarily with the intention to refine them through further experimentation. The terms *short-term memory*, *long-term memory*, and *learned* are used operationally under the definitions described above. The relationship between these operational terms and psycholinguistic reality, including the distinction between tacit and explicit knowledge (Collins, 2010), is not investigated in the present study.

Initially, set *N* is populated with all lexical units that are assigned by the instructor, and sets *P*, *S*, and *L* are empty. When the instructor assigns a new lexical theme to the student, lexical units constituting the theme are added to set *N*. Regardless of the set to which a lexical unit belongs, it has a numerical vector (*p*) associated with it, describing the unit's anticipated complexity, the degree to which the unit has been learned by the student at every moment in time, and other parameters. As the student works on the lexical units, they are first moved from *N* to *P* (introduction of new material), then from *P* to *S* (activation in short-term memory), and, finally, from *S* to *L* (long-term retention). The only backward movement of lexical units permitted by the model is that from *S* to *P*, which may happen if the student fails to recall the unit.

While the lexical unit is in set *P*, it is presented frequently to the student in activities of increasing difficulty levels until the student starts to recall it reliably (specific vocabulary learning activities implemented in the system are described below). While in set *S*, intervals between the consecutive presentations of the lexical unit are increased to at least an hour; if the student practices once a day as recommended, it effectively means that the unit will only be shown once daily. The units in set *L* are presented to the student only occasionally.

### Step 1: Introducing new words

Let us suppose that a certain lexical unit has been assigned to the student, introduced, but then for some reason is dropped from the active acquisition process and re-introduced at a later time. In such a case, the time and effort spent by the student between the initial introduction and the temporary removal from the acquisition process would be, effectively, wasted. After the re-introduction, the student would need to start the learning process all over.

In order to avoid this sort of inefficiency, the adaptive tutoring algorithm will always have the student fully acquire all lexical units that have been introduced to them. The only exception to this general rule is made when the instructor cancels the assignment and the lexical unit in question becomes no longer relevant for the student. An important implication is that the tutoring

algorithm must make all prioritization decisions at the time of selecting material from the pool of the assigned units (set *N*) to be introduced to the student. Once a unit has entered the acquisition process (set *P*), it is treated equally with other units that are already in the process, regardless of the priorities set by the instructor for the corresponding lexical themes.

When introducing new words, it is important not to overflow the student's working memory. Miller (1956) established 7 ± 2 items as an estimate of working memory capacity, but this value was later shown to be age dependent, increasing during childhood development and decreasing with ageing (cf. Morraa, Viglioccob, & Penelloc, 2001). In our system, Miller's estimate of 7 ± 2 items is adopted: under normal circumstances, the tutoring algorithm will not introduce more than seven new lexical units per training session, but if the deadlines specified by the instructor are very tight, or the student is falling behind, this limit may be automatically increased to nine.

New vocabulary items are initially presented in a multiple-choice matching activity: the target unit is shown along with several different L1 translations, and the student is prompted to select the translation that matches the target word. The adaptive algorithm recognizes that some of the words may be known to the student by the time they are first introduced within the system. If in this initial activity the student performs as if he or she already knew the word, the system will ask them to confirm if it is indeed the case. If an affirmative answer is given, the system will test the student on this lexical unit one more time (after a certain time interval has elapsed) and then move the unit directly to set *L*. However, if the student makes a mistake, the lexical unit will be dealt with as if it were initially unknown to the student.

After vocabulary items have been introduced to the student, they all enter the same process of active acquisition, supported by various activities generated by the system. These activities will be discussed in more detail in the following section.

## Step 2: Generating activities

During each training session, the tutoring algorithm selects a subset of lexical units from sets *P*, *S*, and *L* and automatically generates activities based on these words, trying to optimize for the predicted learning gains. Each of the three sets is allocated its own quota within the training session time. In the present study, these quotas were set arbitrarily, again with the intention of refining them in our continued research.

Once the target words are selected, they are used to generate activities. At the time of the experiment, the system was capable of generating seven types of activities:

- multiple-choice matching of the target word to its L1 translation;
- multiple-choice matching of the L1 translation to the target word;
- spelling of the target word based on an L1 prompt;
- listening comprehension-1: the target word is presented audibly, and the student is prompted to select an appropriate L1 translation among distractors (see Figure 1);
- listening comprehension-2: the target word is presented audibly, and the student is prompted to spell the target word;
- semantic classification (sorting, see Figure 1);
- fill-in-the-blank sentence completion.

An important part of the generation of multiple-choice (matching) activities is *collision prevention*. By collisions we mean instances when not only the target item, but also one or more of the distractors may be reasonably perceived as valid responses to the prompt. For example, English words *landing* and *boarding* are both translated by the Russian word *посадка* [pɐˈsatkə]. When generating a translation-to-word matching activity with *посадка* as the prompt and *landing* as the target word, the selection of *boarding* as one of the distractors would lead to a collision. To avoid collisions, WordNet data (Miller, 1995) is used to exclude synonyms and direct hypernyms/hyponyms of the target word from the inventory of possible distractors, along with heuristics based on Levenshtein distances between pairs of glosses. Paronyms, on the other hand, are given preference in distractor lists, to help the students learn to distinguish between them.

Examples of activities are presented in Figure 1.

### Step 3: Updating student models

After the student completes an exercise, instant feedback is displayed together with additional information about the target lexical unit, such as usage examples or comments. Upon reviewing the feedback at his or her convenience, the student can proceed to the next exercise. The information about all completed activities is stored in a detailed log file, including the activity type, the target lexical unit, the distractors (if applicable), and the student's answer to the prompt. Additionally, the following timings are stored in every log entry: the timestamp of the moment the activity was completed, the time interval in seconds between the moment when the prompt was displayed to the student and the moment when the student responded to it, and the time interval between the moment feedback was shown to the student and the moment when the student finished reviewing the feedback and proceeded to the next activity. The log files are used to generate student models, which, in turn, are used to generate further exercises. For research purposes, Linguatorium also has the capability of restoring a student model as of any specified moment in time.

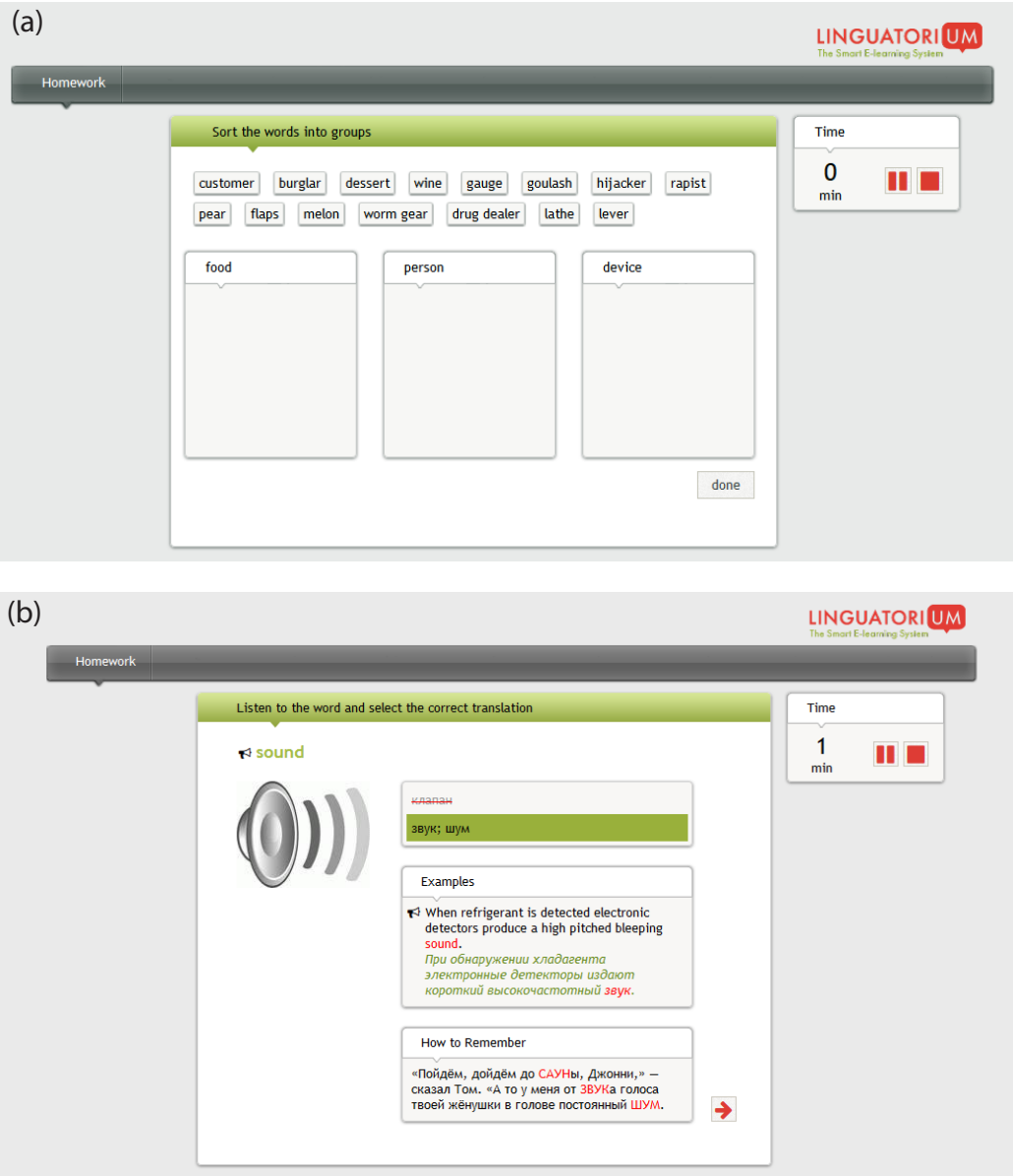


**Figure 1:** Sample activities in Linguatorium: (a) semantic sorting; (b) listening comprehension-1.

## Methodology

### Classroom environment

The classroom where Linguatorium was used for vocabulary teaching was at a state maritime academy in Russia. The Marine Engineering Department of the institution is one of the oldest and most renowned among similar educational establishments in Russia. According to its graduation requirements, proficiency in English is one of the conditions of being granted a professional degree. Educational standards in this respect are based on internationally

accepted requirements for competence with regard to communication on board ocean-going ships. As stated by the International Maritime Organization Convention on Standards of Training, Certification, and Watchkeeping for Seafarers (STCW), the ability to communicate in a clear and concise way on matters related to successful performance of watch is considered to be one of the most important among various maritime skills required to qualify as a watchkeeping rating or an Able Seafarer Engine.

The aim of the EFL class, therefore, is to develop communicative competence relevant to the future professional demands of academy graduates. The areas of competence fall under several categories: graduates should be proficient in speaking, listening comprehension, reading, and writing (within the scope of their professional duties). As specified by the STCW Convention, cadets should be trained to use correct terminology in machinery spaces and correct names of machinery and equipment, understand orders and be understood in matters relevant to watchkeeping duties, and record information, as instructed, in the engine-room log book, engine movements log, and other record books. Communication should be clear and concise, and advice or clarification should be sought from the officer of the watch where watch information or instructions are not clearly understood. At the advanced level, communication within the operator's area of responsibility should be consistently successful.

Taking into consideration the certification requirements, the task of developing lexical competence should not be underestimated. However, the lack of FL exposure, limited time allocated for EFL classes, low initial level of language competence displayed by most marine engineering cadets, and other disadvantages make the task of practical acquisition of marine engineering English a demanding one. The current EFL pedagogy at the academy relies on reading, translation, and discussion of specialized texts, followed by vocabulary, grammatical, and communicative activities and homework assignments (Nicholls & Potapova, 2010). In the present study, Linguatorium was used as a supplementary vocabulary learning tool.

## Participants

Our participants were 22 cadets enrolled in an EFL class during the spring semester of 2012. All participants were third year cadets, studying marine engineering as their specialty. Their ages ranged between 20 and 22 years, averaging at 20.6 years. There was only one female student among the participants, which reflects the general gender distribution among the cadets. All participants were native speakers of Russian. The second author of the present paper taught the class as an adjunct professor. Each cadet signed up for a student account with Linguatorium and accepted a service contract offer with a provision of consent to participate in the present study.

According to the current curriculum, the following lexical themes were covered in class: "Refrigerating and Air Conditioning Plant," "Pumps," "Leak Detection," and "Compressors." A total number of 155 words were selected by the instructor to be supplemented with computer-based activities, from which a subset of 112 items was chosen for the present experiment. All 155 lexical units were entered into the system, along with their Russian translations, images, and usage examples extracted from the materials used in the traditional classroom teaching. Conventional in-class and homework activities provided the main exposure to the target vocabulary, while Linguatorium was used as a supplementary activity. All participants were asked to spend at least 10 minutes each day on working with the system. This time was chosen arbitrarily, under the common sense assumption that 10 minutes a day would not be regarded as exceedingly burdensome for a supplementary activity.

## Randomization and blinding

The present study followed the Independent samples *t*-statistic design with double blinding (Jones et al., 2002; Kirk, 2009), strengthened by creating multiple control (comparison) and experimental (intervention) groups. For each target lexical unit $u_1$, $u_2$, $u_3$, ..., $u_{112}$, students were split into two groups: control and experimental, resulting in a total of $M = 224$ groups ($C_1$, $C_2$, $C_3$, ..., $C_{112}$; $E_1$, $E_2$, $E_3$, ..., $E_{112}$). For each lexical unit $u_i$, every participant was randomly assigned to either the corresponding control group $C_i$, or the experimental group $E_i$. Each participant, therefore, was assigned to 112 different groups, some of them being control groups ($C_i$), others experimental ($E_i$).

In each of the control groups $C_i$, the respective lexical unit $u_i$ was presented in class and practiced through conventional homework assignments. In the corresponding experimental group $E_i$, our system was used as an additional tool for the acquisition of that particular lexical unit $u_i$. With $M = 224$ distinct control and experimental groups and $N = 22$ participants, $M \cdot N = 2{,}464$ participant-to-group mappings were obtained. The randomization process was performed automatically, by using a program developed by one of the authors. The program accessed the Linguatorium server via the API to create a key that mapped the participants to their corresponding groups. A quasi-random number generator was used to determine the assignments. The key was stored in the database and was not disclosed to the researchers until the study was completed. This process allowed us to keep all 112 conditions independent for each of the participants throughout the experiment. As a result, every participant-to-group mapping was treated as an independent observation. Another way of looking at it is as if we had $M = 112$ concurrent, yet distinct experiments (one per each target lexical unit), involving the same set of $N = 22$ participants.

This approach to randomization had two limitations. First, it did not allow us to quantify the within-participant and across-participant variance of performance measures. In this study, however, we did not aim at identifying multiple factors that could contribute to better learning gains; instead, we intentionally limited the study to a single factor, that is, the use of the vocabulary tutoring system. Second, the large number of observations could inflate the statistical significance of the results: the differences between the control and experimental groups may be found statistically significant, but not necessarily practically important. The discrepancy between statistical and practical significance has been long recognized in medical research (e.g., Hays & Woolley, 2000). We will address this potential discrepancy in the Discussion section, below.

A schematic illustration of randomization and blinding is presented in Figure 2.

| Lexical Units | Participants | $p_1$ | $p_2$ | $p_3$ | $p_4$ | $p_5$ | $p_6$ | $p_7$ | $p_8$ | $p_9$ | $p_{10}$ | $p_{11}$ | ... | $p_{22}$ |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| $u_1$ | reciprocating compressor | *C* | *C* | *E* | *C* | *C* | *E* | *E* | *C* | *C* | *C* | *E* | | *E* |
| $u_2$ | auxiliary | *E* | *E* | *C* | *C* | *E* | *C* | *C* | *E* | *E* | *E* | *C* | | *E* |
| $u_3$ | high-pitched | *E* | *C* | *E* | *E* | *C* | *E* | *E* | *E* | *E* | *E* | *C* | | *E* |
| $u_4$ | operating procedure | *E* | *C* | *E* | *E* | *C* | *E* | *E* | *E* | *E* | *E* | *C* | | *E* |
| $u_5$ | supply | *C* | *E* | *E* | *E* | *C* | *E* | *E* | *C* | *E* | *E* | *C* | . | *C* |
| $u_6$ | positive displacement | *C* | *E* | *C* | *E* | *E* | *E* | *E* | *E* | *E* | *C* | *E* | . | *C* |
| $u_7$ | air start valve | *E* | *E* | *C* | *C* | *E* | *C* | *E* | *E* | *E* | *C* | *E* | . | *C* |
| $u_8$ | torch | *E* | *E* | *C* | *C* | *E* | *C* | *C* | *E* | *E* | *C* | *E* | | *C* |
| $u_9$ | barrier cream | *E* | *E* | *E* | *C* | *C* | *C* | *C* | *E* | *E* | *C* | *E* | | *E* |
| $u_{10}$ | gauge | *C* | *C* | *E* | *C* | *E* | *E* | *C* | *C* | *E* | *C* | *C* | | *C* |
| $u_{11}$ | dust mask | *E* | *C* | *E* | *C* | *C* | *E* | *C* | *C* | *E* | *C* | *C* | | *C* |
| $u_{12}$ | hazardous | *E* | *C* | *E* | *E* | *C* | *E* | *E* | *C* | *C* | *E* | *C* | | *E* |
| ... | | | | | | . | . | . | | | | | | |
| $u_{122}$ | refrigerating capacity | *C* | *E* | *E* | *E* | *E* | *E* | *E* | *C* | *C* | *E* | *E* | | *E* |

**Figure 2:** Randomization and blinding:
*E* – experimental groups;
*C* – control groups.

## Outcome assessment

The present study followed the randomized controlled double-blind design; therefore the differences in the outcomes could be attributed to the treatment variations in the absence of a pretest (Jones et al., 2002). A paper-based vocabulary test was designed following the evaluation practices currently in use at the academy (Nicholls & Potapova, 2010). A graduate student of

linguistics from a different school was employed to deliver the test at the end of the semester. Participants were informed that the test was given to them for research purposes only, and their performance at the test would not influence their grade in class. During the test, participants were presented with Russian terminology and general vocabulary units and prompted to provide their corresponding English equivalents in writing. Participants were allowed one hour to complete the test. Responses to each of the test items were graded as follows: *no credit*, if the item was translated incorrectly; *partial credit*, if there was a spelling error in the translation, or the part of speech was chosen incorrectly; *full credit*, if the translation was correct. The above-mentioned graduate student, who did not know the cadets personally and was blind to the distribution of the participants across experimental and control groups, performed all grading.

### Data analysis

After the administration of the posttest, a detailed log file documenting the participants' interaction with the system throughout the semester was exported from the server database. The key mapping the lexical units to their corresponding groups was also retrieved, and the data were supplemented with the cadets' test results. At this time, all personally identifiable information was irreversibly removed from the dataset, and the key was destroyed.

Based on the log data, every student model was traced back to the first day of learning, and then day-to-day progress was incrementally reconstructed. This allowed us to estimate the amount of time the students spent on computer-based activities to acquire each of the target lexical units, as well as identify the status of each unit in the students' lexical memory as per the student memory model at the time of posttest administration. The one-tailed *t*-test for proportions was employed to assess the statistical significance of the differences in the posttest scores (observable dependent variable) across the controlled conditions (independent variables).

## Results

Although the participants were asked to spend at least 10 minutes a day working with the system, activity log analysis revealed that, on average, they actually spent 174 seconds (just under 3 minutes) per day. For each of the lexical units, the combined activity time was calculated from the moment the unit was first introduced (i.e. moved from set *N* to set *P*) up to the moment when it was deemed fully learned (i.e. moved from set *S* to set *L*), yielding a mean of 107 seconds ($SD$ = 159) and a median of 71 seconds. This is an estimate of the average time it takes a lexical unit to make

its way to the student's long-term memory. A typical pattern of a student model changing in time is presented in Figure 3. It is evident that the rate at which the new lexical units are introduced to the student varies slightly; this variation is the result of the algorithm adapting to the 'due dates' specified by the instructor. Whenever a deadline for a certain lexical theme was approaching, the rate of introducing new vocabulary slightly increased in an effort to meet the deadline.

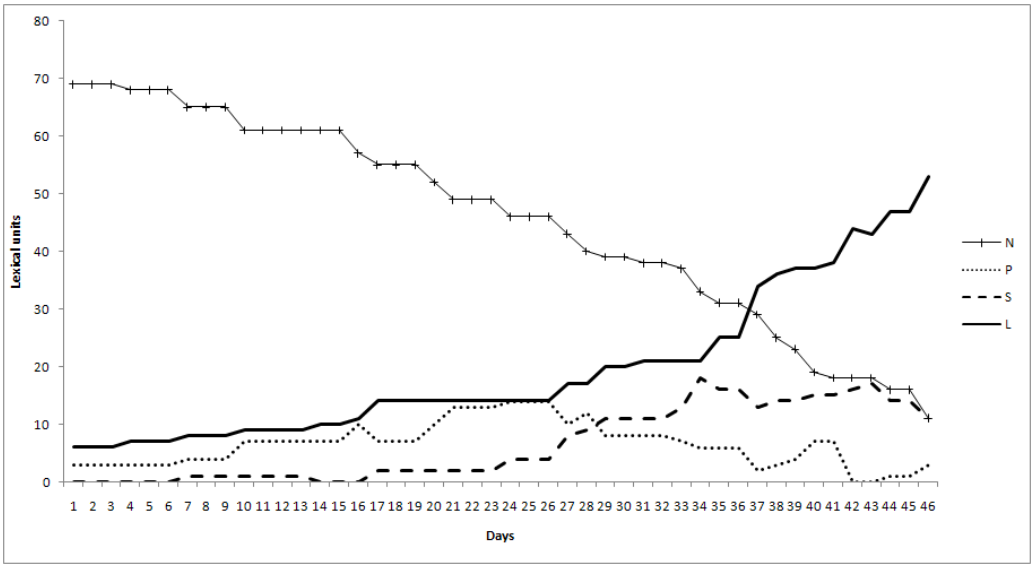


**Figure 3:** A student model changing in time:

*N* – number of units pending introduction;
*P* – number of units in active acquisition;
*S* – number of units in short-term memory;
*L* – number of units in long-term memory.

Because our participants were not perfectly diligent, a portion of lexical units in the corresponding experimental groups only managed to progress to sets *P* or *S* by the time of the posttest (see "The adaptive tutoring algorithm" above for details), while the majority reached set *L* and therefore were considered 'learned'. Table 1 shows raw counts of lexical units in the control groups, sets *P*, *S*, and *L* at the time of the test administration (note that lexical units in set *L*, which were identified as previously known to the students at the time they were first introduced within the system, are counted separately). Lexical units in each category are split by the score given in the posttest: *no credit*, *partial credit*, and *full credit*. Proportions of lexical units that received some credit (full or partial) within the corresponding groups are also presented.

**Table 1:** Lexical Unit Statuses in Student Models and Posttest Credit

| | Raw counts of lexical units | | | | Some credit, %[a] |
|---|---|---|---|---|---|
| | *M*·*N*[b] | No credit | Part credit | Full credit | |
| Control groups | 596 | 495 | 52 | 49 | 16.9 |
| Experimental groups | 1,868 | 936 | 306 | 626 | 49.9 |
| *Breakdown by final status:* | | | | | |
| Active acquisition (*P*) | 135 | 113 | 11 | 11 | 16.3 |
| Short-term memory (*S*) | 150 | 107 | 19 | 24 | 28.7 |
| Long-term memory (*L*) | 884 | 441 | 157 | 286 | 50.1 |
| Previously known units | 699 | 275 | 119 | 305 | 60.7 |

[a]Some credit – proportion of lexical units that received 'partial credit' or 'full credit' among all lexical units the corresponding group (*M*·*N*).
[b]*M*·*N* is the number of participant-to-group mappings treated as independent observations. See "Randomization and blinding" for details.

The difference in test scores between fully acquired lexical units in the experimental group (set *L*) and those known to the students prior to introduction in Linguatorium was not very large (50.1% vs. 60.7%), but significant ($p < 0.001$, one-tailed), which indicates that the students may have failed to attain the level of vocabulary knowledge identical to the previously known lexical units.

In terms of the test scores, the difference between the lexical units in the control group and set *P* was not significant, while the difference between sets *P* and *S* was significant ($p < 0.01$, one-tailed), and so was the difference between sets *S* and *L* ($p < 0.001$, one-tailed). This finding provides support for the criteria that were chosen for moving lexical units from *P* to *S* (initial activation in the short-term memory) and from *S* to *L* (long-term retention). Based on the test results, the difference between vocabulary items considered fully learned by the adaptive tutoring model (the experimental group) and the control group was almost threefold (50.1% vs. 16.9%) and statistically significant ($p < 0.001$, one-tailed).

## Discussion

The focus of this paper was not on critiquing pedagogical approaches that are typically used for teaching vocabulary in an EFL classroom, such as reading and discussing specialized texts. Rather, we explored the effects of a supplementary activity that took a few minutes a day and did not require any modifications to the usual pedagogy, which remained unchanged for the duration of the study. As seen in the results, although students only spent an average of 3 minutes per day on working with the system, the recommendation being

10 minutes per day, the rate of long-term vocabulary recall in the experimental groups was found to increase almost three times compared to that in the corresponding control groups. The acquisition of each lexical unit took the students a median of 71 seconds overall in Linguatorium-based activities. Intuitively, this is a reasonably short time, which suggests a high practical value of the developed system. Yet, supplementary in nature, our system did not aim at substituting the existing methods of vocabulary learning, thus eliminating the ground for controversy related to the behavioral nature of the pair-associate instructional paradigm (Hultsijn, 2001) and the confinement of flashcard-based vocabulary software to the form-meaning matching of lexical units (Nakata, 2011).

In this paper, we also demonstrated the feasibility of randomized controlled double-blind trials in CALL research, wherein both the instructors and the students are unaware of the participant assignment to groups—a rare case in pedagogical intervention research (Jones et al., 2002). This approach not only contributed to the credibility of our conclusions, but also allowed us to study the effect of computer-based spaced repetition on vocabulary learning gains *independent* of teachers' and students' beliefs about and attitudes towards different learning methods. Such beliefs and attitudes are known to be difficult to control for, and as a result, many educational researchers prefer to avoid predictive experiments altogether (Hoadley, 2004). Furthermore, unlike quasi-experimental designs, randomized controlled trials do not require that a pretest be administered (cf. Jones et al., 2002), which is beneficial due to the questionable pedagogical value of the pretest.

## Conclusion

In this paper, we have demonstrated that (1) automatically generated supplemental activities based on spaced repetition can lead to a nearly threefold improvement of vocabulary learning gains in EFL students without any changes to the rest of pedagogy in the classroom, and (2) a double-blind experimental design can be successfully implemented in a CALL-based pedagogical intervention study.

Our work, however, had several limitations. First, due to the implemented randomization and blinding procedures, we could not quantify the within-participant and across-participant variance of performance measures. In future work, these procedures should be improved to allow for such quantification. Second, the recommended duration of activities (10 minutes per day) was not sufficiently substantiated. Further research should be done to refine the recommended duration of activities in terms of the cost/benefit ratio. Third, our study was carried out in a specific purpose context (marine engineering English), and the lexical material represented a mix of general

vocabulary and special terminology. In our future work, we plan to study the differences between general and special vocabulary in terms of the effectiveness of computer-based spaced repetition, which might clarify the prospects of using this method in general EFL contexts.

Finally, when modeling the three stages of vocabulary learning (active acquisition process, activation in short-term memory, and long-term retention), we were unable to find any literature on formal models of word acquisition stages in vocabulary tutoring algorithms. Therefore, we had to establish our own formal procedures identifying such stages, with a hope to refine them experimentally. We were fortunate to discover that the learning stages identified by our algorithm corresponded to statistically significant changes in the long-term vocabulary retention as shown by the posttest scores, which supports the validity of our model. In a follow-up study, we plan to manipulate the thresholds used to identify the stages, to see if any adjustments might lead to improved efficiency of the system. Furthermore, additional research will explore the relationships between the algorithmic stages and the psycholinguistic reality of both the process of vocabulary acquisition and the state of tacit and explicit knowledge of vocabulary (Collins, 2010).

## Acknowledgement

The authors are grateful to Serguei Bakhmoutov for his valuable assistance with preparing the vocabulary lists used in this study; Julia Kharinova, who administered and graded the test; Prof. Carol A. Chapelle, Prof. Volker Hegelheimer, and the four anonymous reviewers for their helpful comments on earlier drafts of this article.

## About the authors

Evgeny Chukharev-Hudilainen is an assistant professor in the Applied Linguistics and Technology program at Iowa State University. He holds BSc and MSc degrees in computer engineering from Northern Federal University of Russia and a PhD in applied and computational linguistics from Herzen State Pedagogical University. Prior to joining Iowa State in 2012, he spent more than six years working as a senior software engineer at the Central Bank of Russia. His current research interests lie at the intersection of applied linguistics, psycholinguistics, and computational linguistics. He is leading research and development of a web-based automated writing evaluation system and teaching graduate and undergraduate courses in linguistics at ISU.

Tatiana A. Klepikova is a professor in the Department of English and Translation Studies at St Petersburg State University of Economics. She holds PhD and DrSc degrees in Germanic linguistics. Her areas of interest include cognitive linguis-

tics, corpus linguistics, and computer-assisted language learning, with special emphases on the theory of linguistic metarepresentations, construction grammar, vocabulary acquisition in foreign language pedagogy, and discipline-specific language. She has published more than 100 papers in journals and edited collections in Russia and internationally, and is an editorial board member for the journal *Issues in Cognitive Linguistics*. She teaches both theoretical courses in linguistics and practical courses of English for specific purposes.

## References

Al-Jarf, R. (2007). Teaching vocabulary to EFL college students online. *CALL-EJ Online*, *8*(2). Retrieved from www.callej.org/journal/8-2/al-jarfl.html

Beliaeva, L. N. (2007). *Lingvističeskie avtomaty v sovremennyx gumanitarnyx texnologijax* [Linguistic automata in modern humanitarian technologies]. St Petersbug: Knižnyj dom.

Brown, C. G. (2002). Inferring and maintaining the learner model. *Computer Assisted Language Learning*, *15*(4), 343–355. Retrieved from http://dx.doi.org/10.1076/call.15.4.343.8269

Collins, H. (2010). *Tacit and explicit knowledge*. Chicago, IL: University of Chicago Press. Retrieved from http://dx.doi.org/10.7208/chicago/9780226113821.001.0001

Davelaar, E. J., Goshen-Gottstein, Y., Haarmann, H. J., & Usher, M. (2005). The demise of short-term memory revisited: Empirical and computational investigation of recency effects. *Psychological Review*, *112*(1), 3–42. Retrieved from http://dx.doi.org/10.1037/0033-295X.112.1.3

Davidoff F., Haynes B., Sackett D., & Smith R. (1995). Evidence based medicine. *BMJ*, *310*(6987), 1085–1086. Retrieved from http://dx.doi.org/10.1136/bmj.310.6987.1085

Ebbinghaus, H. (1885/1913). *Memory: A contribution to experimental psychology.* New York: Columbia University.

Gordania, Y. (2012). The effect of the integration of corpora in reading comprehension classrooms on English as a Foreign Language learners' vocabulary development. *Computer Assisted Language Learning*, *26*(5), 1–16.

Gorjian, B., Moosavinia, S. R., Kavari, K. E., Asgari, P., & Hydarei, A. (2011). The impact of asynchronous computer-assisted language learning approaches on English as a foreign language high and low achievers' vocabulary retention and recall. *Computer Assisted Language Learning*, *24*(5), 383–391. Retrieved from http://dx.doi.org/10.1080/09588221.2011.552186

Hays, R. D., & Woolley, J. M. (2000). The concept of clinically meaningful difference in health-related quality-of-life research. *Pharmacoeconomics*, *18*(5), 419–423. Retrieved from http://dx.doi.org/10.2165/00019053-200018050-00001

Hoadley, C. M. (2004). Methodological alignment in design-based research. *Educational Psychologist*, *39*(4), 203–212. Retrieved from http://dx.doi.org/10.1207/s15326985ep3904_2

Hulstijn, J. H. (2001). Intentional and incidental second-language vocabulary learning: A reappraisal of elaboration, rehearsal and automaticity. In P. Robinson (Ed.), *Cogni-*

*tion and second language instruction* (pp. 258–286). Cambridge: Cambridge University Press. Retrieved from http://dx.doi.org/10.1017/CBO9781139524780.011

Jones, M., Gebski, V., Onslow, M., & Packman, A. (2002). Design of randomized controlled trials: Principles and methods applied to a treatment for early stuttering. *Journal of Fluency Disorders*, *26*(4), 247–267. Retrieved from http://dx.doi.org/10.1016/S0094-730X(01)00108-5

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(5), 1250–1257. Retrieved from http://dx.doi.org/10.1037/a0023436

Kirk, R. (2009). Experimental design. In R. E Millsap & A. Maydeu-Olivares (Eds.), *The SAGE handbook of quantitative methods in psychology* (pp. 24–47). Thousand Oaks, CA: Sage Publications. Retrieved from http://dx.doi.org/10.4135/9780857020994.n2

Labrie, G. (2000). A French vocabulary tutor for the Web. *CALICO Journal*, *17*(3), 475–499.

Leitner, S. (2011). *So lernt man lernen: Der Weg zum* (18th ed.). Auflage. Freiburg: Verlag Herder.

Miller, G. A. (1956). The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, *63*(2), 81–97. Retrieved from http://dx.doi.org/10.1037/h0043158

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, *38*(11), 39–41. Retrieved from http://dx.doi.org/10.1145/219717.219748

Morraa, S., Viglioccob, G., & Penelloc, B. (2001). M capacity as a lifespan construct: A study of its decrease in ageing subjects. *International Journal of Behavioral Development, 25*(1), 78–87. Retrieved from http://dx.doi.org/10.1080/01650250042000050

Murdock, B. B., Jr. (1972). Short-term memory. *Psychology of Learning and Motivation*, *5*, 67–127. Retrieved from http://dx.doi.org/10.1016/S0079-7421(08)60440-5

Nakata, T. (2008). English vocabulary learning with word lists, word cards, and computers: Implications from cognitive psychology research for optimal spaced learning. *ReCALL Journal*, *20*(1), 3–20. Retrieved from http://dx.doi.org/10.1017/S0958344008000219

Nakata, T. (2011). Computer-assisted second language vocabulary learning in a paired-associate paradigm: A critical investigation of flashcard software. *Computer Assisted Language Learning*, *24*(1), 17–38. Retrieved from http://dx.doi.org/10.1080/09588221.2010.520675

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

Nicholls, A. P., & Potapova, Y. B. (2010). *Insight into marine engineering English*. St. Petersburg: GMA im. adm. S. O. Makarova.

Oberg, A. (2011). Comparison of the effectiveness of a CALL-based approach and a card-based approach to vocabulary acquisition and retention. *CALICO Journal*, *29*(1), 118–144. Retrieved from http://dx.doi.org/10.11139/cj.29.1.118-144

Piotrowski, R. G. (1999). *Lingvističeskij avtomat (v issledovanii i nepreryvnom obučenii)* [The Linguistic Automaton (in research and continuous instruction)]. St Petersburg: RGPU.

Piotrowski, R. G., & Beliaeva, L. N. (2005). Linguistic automaton. In R. Köhler, G. Altmann, & R. G. Piotrowski (Eds.), *Quantitative Linguistik: Ein internationales Handbuch* (pp. 921–931). Berlin: de Gruyter.

Poole, R. (2012). Concordance-based glosses for academic vocabulary acquisition. *CALICO Journal*, *29*(4), 679–693. Retrieved from http://dx.doi.org/10.11139/cj.29.4.679-693

Ranalli, J. (2013). Designing online strategy instruction for integrated vocabulary depth of knowledge and web-based dictionary skills. *CALICO Journal*, *30*(1), 16–43. Retrieved from http://dx.doi.org/10.11139/cj.30.1.16-43

Rastrigin, L. A., & Erenštejn, M. X. (1988). *Adaptivoe obučenie s model'u obučaemogo* [Adaptive learning with a learner model]. Riga: Zinatne.

Stockwell, G. (2007). Vocabulary on the move: Investigating an intelligent mobile phone-based vocabulary tutor. *Computer Assisted Language Learning*, *20*(4), 365–383. Retrieved from http://dx.doi.org/10.1080/09588220701745817

What Works Clearinghouse (2011). *Procedures and standards handbook* (v. 2.1). Retrieved from http://ies.ed.gov/ncee/wwc/pdf/reference_resources/wwc_procedures_v2_1_standards_handbook.pdf

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2012). How and when do students use flashcards? *Memory*, *20*(6), 568–579. Retrieved from http://dx.doi.org/10.1080/09658211.2012.687052

Wozniak, P. A. (1990). Optimization of learning: Simulation of the learning process conducted along the SuperMemo schedule. Retrieved from http://www.supermemo.com/english/ol.htm

Zhu, Y., Fung, A. S., & Wang, H. (2012). Memorization effects of pronunciation and stroke order animation in digital flashcards. *CALICO Journal*, *29*(3), 563–577. Retrieved from http://dx.doi.org/10.11139/cj.29.3.563-577