

Sequencing Vocabulary Instruction: Artificial vs. Real Users

Samuel R.H. Joseph
University of Hawai'i,
USA
srjoseph@hawaii.edu

Stephen H. Joseph
University of Sheffield,
UK
s.joseph@sheffield.ac.uk

Michael H. Joseph
University of Leicester,
UK
mhj1@leicester.ac.uk

Abstract. There are various widely researched strategies that appear to be helpful in some, but not necessarily all vocabulary learning situations. However, an early report suggested that an extremely simple strategy, in which only the ordering of the material presented is varied, might have very substantial effects on learning and recall. These observations have been used as the basis of many subsequent developments, but rarely been subject to rigorous examination and replication. We have recently been examining both the theoretical foundation, and the practical implementation, of this latter approach. In this paper we present a comparison of data obtained using virtual users, operating in accordance with the underlying theory of memory, with the earlier experimental data obtained with real users.

1. Introduction

Sequencing of vocabulary instruction has tended over the years to be a somewhat imprecise art. With a few exceptions most examples in the psychology of learning literature have promoted a variety of heuristics to optimise the retention of vocabulary. One prominent exception is the work of Atkinson [1], which not only presents a detailed mathematical approach, but also a theoretical framework relating to the nature of memory. This work appeared to demonstrate that a teaching programme designed on the basis of that framework could dramatically improve retention rates in paired associate learning. Although the simple model of short-term memory employed in the work does not completely explain all subsequent experimental results, the original result remains a powerful demonstration of the possibilities of building a teaching system based on a well-defined model of memory. We have thus been led to examine the original model, to try to understand how it works and to see how it can be used to help support the design of learning programmes. After reviewing some related work this paper gives an overview of the Atkinson model, followed by our own analysis and the results of simulations using artificial users that precisely embody Atkinson's memory model.

2. Learning Vocabulary

One might argue today that the Atkinson Model is a limited model of vocabulary learning, because since it's creation in the 60's and 70's an extensive literature has developed on the many different factors that can affect vocabulary learning. For example there are results to indicate that associating sentences with vocabulary or requiring learners to perform generative tasks improves retention [5,10]. Other experiments have confirmed the widely known memory boosting effect of mnemonic strategies [11,15] as well as indicating that a scripted pair-learning/testing format can provide additional benefits [9,11]. Conversely, some studies have indicated that visual repetition of vocabulary items correlates negatively with performance [8], and emphasize the positive effects of meta-cognitive strategies such as "Self Initiation" and "Selective Attention". There is also a great deal of evidence to support the notion that

“distributed” practice is more effective than “massed” [4,6]. In addition de Groot [7] has shown that methods designed to encourage deep processing of vocabulary reduced retention loss two to three weeks after initial presentation.

As a result the benefit of replicating Atkinson’s approach may not be immediately apparent. One might argue that the Atkinson model would be a poor choice for instructional design, since it does not explicitly handle long term memory decay, interference between items, or phonological encoding that might allow the advance prediction of errors. However, Atkinson obtained very striking improvements in vocabulary recall by using relatively simple strategies based upon the sequencing and frequency of representation of individual items during learning. His best results were obtained with an algorithm that required information on the difficulty of individual items for the target population, and the approach was based upon an explicit theory of memory function. Firstly it is important to determine whether such results are reproducible. If they are, then this alone could have important implications for practical vocabulary teaching, although naturally further work would be required in order to combine Atkinson’s model with the other factors necessary to make a complete instructional approach. In addition we believe that Atkinson’s method of model formation and testing is likely to complement a purely empirical approach, which can show whether one procedure is superior to another, but not why.

3. Paired Associate Paradigm

The paired-associate learning task is a standard procedure for assessing human explicit memory. For example, randomly paired elements such as words and letter strings are presented to subjects, who are then asked to recall one half of the pair from the corresponding other half, after which different types of feedback may be made available [19]. Rizzuto & Kahana [16] provide a summary of some different approaches to modelling of the paired associate learning task, as well as their own auto-associative neural network model. Nesbit & Yamamoto [14] showed that grouping together similar paired associates in sub-lists, caused subjects to generate more practise errors, but overall retention was better (around 20% over 64 test subjects).

In this paper we focus on the approach presented by Atkinson and Crothers [2] that was based on a model incorporating concepts of both short and long term memory. In their original study the predictions of different models of the day were compared with the results of a variety of different paired associate experiments, using tri-grams, Greek letters, digits and normal letters. Having demonstrated the explanatory superiority of a three state model that distinguished between long-term and short-term memory, as well as including interference based forgetting, Atkinson [1] showed how the model could be applied to vocabulary learning.

4. Atkinson Model

The Atkinson Model takes a multiple state memory model that effectively distinguishes between long-term memory (LTM) and short-term memory (STM). It is comparable to the Knowledge Tracing model of Anderson & Corbett [3]; the difference being an additional short-term memory state. In Atkinson’s model paired associate items comprising of cue and response may be in LTM (state P), a permanent state, or in STM (state T) a temporary state where the association may be forgotten, becoming unknown (state U). The assumption is that a learner will give a correct response when presented with any cue from a paired associate item that they have in either state P or T. Conversely, if that item is in state U they will give an incorrect response.

	P	T	U
P	1	0	0
T	x	1-x	0
U	y	z	1-y-z

	P	T	U
P	1	0	0
T	0	1-f	f
U	0	0	1

Fig. 1: Presented Item (left), and Other Item (right) Probability Transition Matrices (P=Permanent, T=Temporary, U= Unknown), showing probability of transition from one memory state to another

The matrices in figure 1 show the probability of transition from one state to another with the left hand column being the state before presentation and the top row being the state after presentation. The presented item transition matrix in figure 1 is applied whenever an item is presented, e.g. if the presented item is currently in state U, then the likelihood of transferring to state P is y. The transition matrices are defined in terms of a number of parameters, x, y, z, and f which indicate how difficult it is to learn or how easy it is to forget each item. The second matrix is applied to those items that are not being presented, on each presentation that leads to an incorrect response. The implication is that interference from other items in the unknown state can cause an item to drop out of short-term memory. There is also a fifth parameter g which defines the probability that a subject already has the item in state P before the start of the experiment.

Atkinson [1] created a teaching system based on this model that would choose items for presentation that were most likely to be transferred into the P state¹. Given the sequences of correct and incorrect responses from the user so far, the model would estimate the likelihood of each item being in a particular state. Given knowledge of the transition parameters, the system could then infer which item, if presented next, would most likely be transferred to state P. The assumption was that items in state P would remain in permanent store and thus be available for subsequent recall a week later, while items in state T would not. In experiments using German-English word pairs Atkinson's optimal strategy condition significantly outperformed subjects selecting their own study order ("self selection"), and random presentation (fig 2). Atkinson's experimental procedure involved presenting seven sets of 12 German cue words in round robin fashion. Each list of German cues numbered 1...12 was projected onto the wall in turn, and the subjects were presented with the number of a cue on a teletype. The subjects would then type in what they believed was the English response, and the teletype would respond with the correct response. A delayed test session a week later was in a similar format, except no feedback was given. There were in fact two types of Atkinson algorithm, with one setting the x,y,z,f, and g parameters equal across all items; the other allowing them to vary. It was this latter algorithm, referred to as the "optimal (unequal)" approach that proved the most effective. The former or "optimal(equal)" approach performed similarly to the self-selection condition

¹ The precise equation is $P(U)*P(U \rightarrow P) + P(T)*P(T \rightarrow P)$

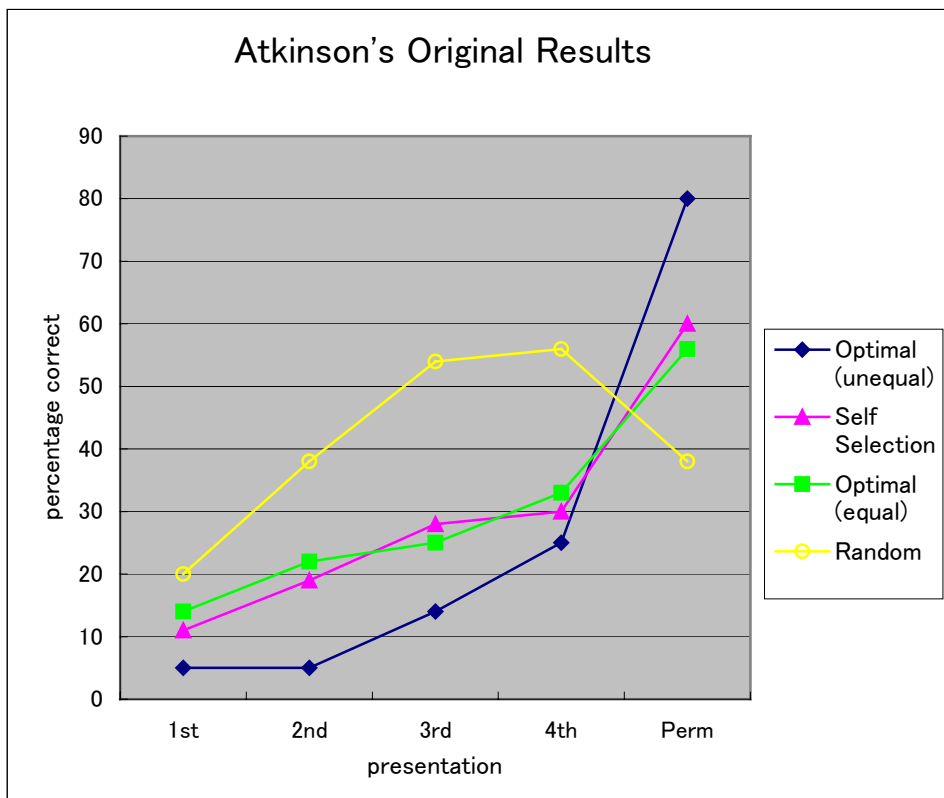


Fig. 2: Results reproduced from Atkinson's 1972 paper showing the percentage correct for the 1st, 2nd, 3rd, and 4th sets of 84 presentations (4 round robin repetitions of 7 lessons of 12 words) for each of the different experimental conditions (see text for more details)

The results also showed a clear inverse relation between the performance during training, and subsequent test, e.g. the random condition subjects performed best during instruction, but worst at test, while the optimal (unequal) condition subjects performed badly during training but were the best during subsequent recall. The contrast between training and recall is most remarkable under the optimal (unequal) condition (see fig 2). The performance of the subjects in the optimal (unequal) condition is in fact extraordinarily low during training. One possible explanation would be that the algorithm was re-presenting extremely difficult items again and again, such that most of the subjects responses were incorrect.

Setting the parameters for Atkinson's model requires pilot studies to be performed on the same word pairs. A minimization algorithm must then be employed to find the parameters that best fit the observed experimental data. In previous studies [12] we were unable to replicate Atkinson's results precisely. This may well be due to the lack of detailed information about the procedures used in the optimization process. There were other differences in our experimental setup such as the use of Japanese words as opposed to German, and alternative blocking of the lessons which must also be expected to have influenced our results. In Atkinson's experiments sets of 7 lessons were presented in round-robin order, so a subject would see one item from lesson 1, then one from lesson 2 etc. Our initial studies presented all items from lesson 1 before moving to lesson 2.

Subsequent communications with the original author have cleared up many of the ambiguities and we are confident that our current user studies come much closer to replicating the original experiments. Organising and managing human studies takes time and overhead, and in the meantime we have employed a mixture of analysis and simulated users to try and understand the Atkinson model in more detail.

5. Analysis of the Atkinson Model

The Atkinson model has influenced various authors, but with the exception of Katsikopoulos [13] few have replicated the algorithms in full. For example, while Seigel & Misselt [18] refer to Atkinson's work, they reject the use of a theoretical model in favour of a selection of heuristics based on instructional design strategies. Van Bussel [19] created a modified version of Atkinson's original procedure called the "a priori knowledge (APK) sequencing procedure", which more frequently presents items that are difficult to learn, as measured by the number of mistakes made by a particular user. This fits in with other work such as that of Schneider et al [17] suggesting that focusing on difficult tasks can lead to better retention. However Van Bussel's approach overlooks the fact that the most effective Atkinson procedure does not necessarily present the most difficult items more frequently, since it may in fact avoid presenting them at all if there are other items that have a higher likelihood of entering the permanent state. Van Bussel's results are extremely interesting however, showing that performance between the APK and fixed presentation strategies can only be distinguished if the users' self-regulated versus externally regulated learning styles are taken into account. As a result we plan to incorporate the same learning styles questionnaire used by Van Bussel into our current human studies.

The relationships between the original model, the heuristic alternatives, and the modified model remain unclear. The customary approach to the modelling problem is to formulate analytical or numerical solutions to the particular conditions of the experiment, and compare the predictions and results according to some chosen measure. We have found that it is possible to understand how the model works, and thus demonstrate some of its general properties, and so compare them with the operation of other systems, by considering the following points:

1. Paired associates can be thought of as being in one of three states:

- i. Un-tried: not yet presented to the subject
- ii. E-tried: presented and most recent response was erroneous
- iii. C-tried: presented and most recent response was correct

2. Let us first ignore the T state and any forgetting processes, then if g is the probability of a prior known and y is the probability of transition from U to P ($P(U \rightarrow P)$) then the merit (i.e. the Probability of a transition to P if the word is tried) of the different types of word in the Optimal (Equal) condition are:

- i. Un-tried: $(1-g).y = P(\text{Unknown}) * P(U \rightarrow P)$
- ii. E-tried: $(1-y).y = P(U \rightarrow U) * P(U \rightarrow P)$
- iii. C-tried: 0

3. If $g < y$, indicating the probability of learning an item is greater than the probability of already knowing it, Un-tried words will have the highest merit. Thus an Optimal (Equal) approach will present all the Un-tried items, followed by the E-tried items, continuing until everything has been responded to correctly - i.e. a correct response will lead to dropping that item from consideration for subsequent presentation. However one should note that there is no guarantee that the remaining E-tried items will be presented again with uniform frequency at any point, since they are all equally likely to be selected for presentation.

4. If $g > y$, indicating the probability of knowing an item is greater than the probability of learning it, E-tried words will have the highest merit. Thus an Optimal (Equal) approach will present Un-tried items until an erroneous response is received, after which it would focus on that item until it was responded to correctly. However it is important to note that the round-robin operation would prevent the subject from being presented the same item more than once every seven presentations so there would be a chance for items on other lists to enter the E-tried state.
5. Continuing with the same assumptions the merit of the different types of word in the Optimal (Unequal) condition now depend also on the individual variation of their parameters:
 - i. Un-tried - easy to learn words (high y) and non-obvious (low g) will be tried first
 - ii. E-tried - words of middling difficulty ($y=0.5$) will be favoured
 - iii. C-tried - as above
6. Thus in the Optimal (Unequal) condition easy to learn and non-obvious words will be presented first, while C-tried items will be dropped as before. Sufficiently obvious and difficult to learn words may conceivably be excluded altogether. Once the set of suitably easy to learn and non-obvious Un-tried items have been presented then E-tried items will start to be presented with a general emphasis on those items with middling difficulty.
7. If we now include the T state. The merit of the different types of word in the Optimal (Equal) condition are:
 - i. Un-tried - $(1-g).y = P(\text{Unknown}) * P(U \rightarrow P)$
 - ii. E-tried - $(1-y-z).y + z.x = P(U \rightarrow U) * P(U \rightarrow P) + P(U \rightarrow T) * P(T \rightarrow P)$
 - iii. C-tried - if first round 0, or possibly non-zero if failure frequency is low²
8. Since forgetting only reduces T and increases U without changing P, the merit of an E-tried item will go up or down depending on the relative value of $P(U \rightarrow P)$ and $P(U \rightarrow T)$
9. If $x > y$, i.e. $P(T \rightarrow P) > P(U \rightarrow P)$, more recent E-tried items will have a higher merit, and thus the Optimal (Equal) system will be likely to re-present recent E-tried items.
10. If $y > x$, i.e. $P(U \rightarrow P) > P(T \rightarrow P)$, older E-tried items will have a higher merit, and thus the Optimal (Equal) system will be likely to re-present older E-tried items over more recently presented ones.
11. It is interesting to note the difference between 9&10 and 4 above, in as much as while they both try to re-present E-tried items the effects of 9&10 wear off so that while the system described in 4 would keep repeating an item until it received a correct response,

² We remark that the effect of the T state will only be detectable if the frequency of presentation of a paired associate is of the same order as the forgetting rate; for the data this would only be so in the latest stages of the rehearsal, since as Atkinson's results show, the majority of presentations lead to incorrect responses and thus a high forgetting rate.

the parameters in 9&10 might be such that the system oscillated between presenting E-tried and Un-tried items.

12. Continuing with the same assumptions the merit of the different types of word in the Optimal (Unequal) condition now depend also on the individual variation of their parameters:

- i. Un-tried - easy to learn words (high y) and non-obvious (low g) will be tried first [as before]
- ii. E-tried - words of middling difficulty ($y=0.5$) will be favoured, as will words that are easy to learn using the T state (high x and z)
- iii. C-tried - are more likely to be repeated if easily forgotten (high f) and are likely to pass through the intermediate T state (high z)

13. Thus in the Optimal (Unequal) condition we are likely to see similar effects as 7 above, however certain categories of items will have a much higher likelihood of repeated presentation - those that are easily forgotten, and those that are likely to be learnt by passing through the intermediate T state. All this will be in combination with the effects described in 9&10.

The consequence of these considerations is that the Atkinson algorithm is not necessarily approximated by repeatedly presenting items in proportion to how many times the user has answered them incorrectly. The likelihood of an item being presented depends much more on the actual parameter settings associated with that item.

6. Using Artificial Users

In order to test our understanding of the Atkinson model we developed simulated users that embody precisely the theoretical basis of the Atkinson model. Specifically these users maintain a set of three states (P, T & U) and the transition of items between the three states takes place as described in the Atkinson model. Although the simulation did not model feedback to the user explicitly, the ability of the artificial user to learn an item implies that some sort of feedback must be present. Using these artificial users we were able to replicate a subset of Atkinson's original results (figure 3). We created seven lessons, each consisting of 12 pairs of nonsense words, and gave each pair x , y , z and f parameters selected randomly from a Uniform distribution between 0 and 1, with the additional constraint that $y+z < 1$. The g parameter was selected from a Uniform distribution between 0 and 0.4 to achieve similar first round success levels as seen in Atkinson's original results.

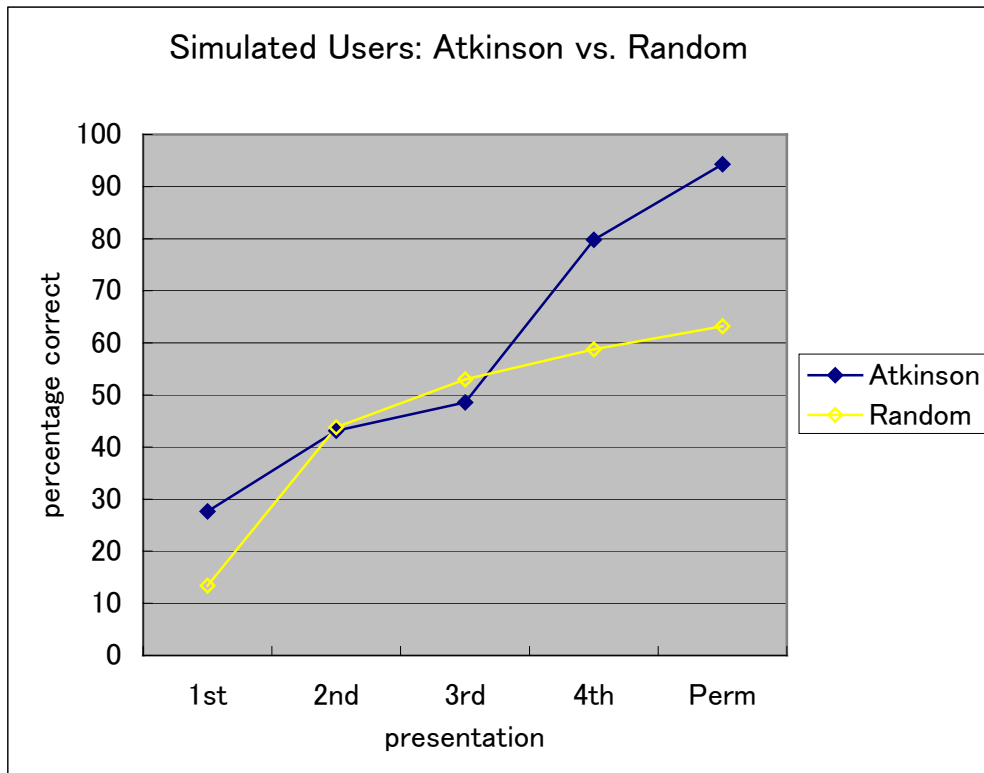


Fig. 3: Results generated by 5 artificial users that embody Atkinson’s theoretical , showing the percentage correct for the 1st, 2nd, 3rd, and 4th sets of 84 Training presentations (4 round robin repetitions of 7 lessons of 12 words) for both of the different experimental conditions. The Perm presentation simulates the results of Test a week later, showing the percentage of items in the permanent state at the end of Train.

An inspection of an individual run bears out the predictions of our analysis, with the first 14 items presented all having y values in excess of 0.6. Interestingly all but one of them was responded to incorrectly, and half of them were never presented again. It was also clear that some items were repeated as much as 10 or 11 times, and these items tended to be easy to forget (high f), or ones that were easier to learn via the T state ($z > y$ and $x > y$).

An ANOVA showed that the differences between the number of items in the permanent state after the first Training presentation and the final testing round are significant, $F(1,4) = 13.919$, $p < 0.01$, $F(1,4) = 41.424$, $p < 0.01$, and confirm the potential efficacy of employing the Atkinson algorithm as opposed to a purely random presentation order. However the results indicate that the % correct order of the results at Test roughly reflect the ordering during Train, the inverse of Atkinson’s original results. For these results to have mirrored Atkinson’s original we would have expected the Random condition to do better on 4th round of training, and then worse in terms of the number of items in the Permanent state.

The extreme behaviour of the Atkinson algorithm, whereby many items are only presented once, and others are repeated again and again, does go some way towards explaining the remarkably high error rates in Atkinson’s optimal condition, along with the subsequent high performance at test. The algorithm is presenting items that can be learnt on a single trial, where the user likely makes a mistake, but the algorithm anticipates that the item has been learnt and need not be presented again, thus the high error rate – i.e. the items that the user would answer correctly are not presented again, and the algorithm focuses on other items that are easily forgotten, or are learnt via the short term memory route. Items are thus transferred to the Permanent state without being presented any more times than necessary.

7. Discussion

The artificial user studies not only provide an essential check that the algorithms are correctly implemented, but also test the model under ideal conditions. The improved performance of real users under the Optimal (Unequal) condition may not be due to the validity of the Atkinson model for those users. For the artificial user, however, the improvement *due only to the model* can be measured accurately, and so separated out from that which is a by-product of the condition for the real user. It appears from our results that the effects of the Optimal (Unequal) condition on the real user are not entirely due to the model. Specifically the inverse relationship between performance at test and train present in Atkinson's experiments does not appear in our simulations or interpretation of the model.

This conclusion remains provisional, however, until the completion of our trials with real users, and of more extensive investigations of artificial users. It would seem reasonable to expect that real users cannot be completely modelled by a three state memory model with constant transition rates. However, the contribution of this model is hard to elucidate in the complex context of real teaching programmes on real subjects. The facility for the creation of larger numbers of artificial users, with appropriate distributions of parameters, and subsequent testing of their performance under various programmes, will be of great assistance in critiquing the model and furthering its development. In this way we hope to extend the range of real responses that can be modelled, and perhaps explore the limitations of what is possible with such models.

Acknowledgements

Many thanks to the anonymous reviewers for insightful feedback, and to Luke & Aya Joseph for support during the writing of this paper.

References

- [1] Atkinson, R. (1972) Optimizing the learning of a second language vocabulary. *Journal of Experimental Psychology*, 96, 124-129.
- [2] Atkinson, R. & Crothers, E. (1964) A comparison of paired-associate learning models having different acquisition and retention axioms. *Journal of Mathematical Psychology*, 1, 285-315.
- [3] Corbett, A. and J. Anderson, (1995) Knowledge tracing: Modeling the acquisition of procedural knowledge. *User modeling and user-adapted interaction*, 4: p. 253-278.
- [4] Cull, W.L. (2000) Untangling the benefits of multiple study opportunities and repeated testing for cued recall *Applied Cognitive Psychology* 14 (3): 215-235
- [5] Grace, C. (1998) Retention of word meanings inferred from context and sentence-level translations: Implications for the design of beginning-level CALL software. *Modern Language Journal* 82 (4): 533-544.
- [6] Greene, R (1989) Spacing effects in memory: evidence for a two-process account. *Journal of Exp. Psy.: Learning, Memory & Cognition*, 15 (3): 371-377.
- [7] de Groot AMB, Keijzer R (2000) What is hard to learn is easy to forget: The roles of word concreteness, cognate status, and word frequency in foreign-language vocabulary learning and forgetting *Language Learning* 50 (1): 1-56
- [8] Gu, Y & Johnson R (1996) Vocabulary learning strategies and language learning outcomes. *Language Learning*, 46 (4): 643-679.
- [9] Hansen, L., Umeda, Y. & McKinney, M. (2002) Savings in the relearning of second language vocabulary: The effects of time and proficiency. *Language Learning*, 52 (4): 653-678.
- [10] Joe, A (1998) What effects do text-based tasks promoting generation have on incidental vocabulary acquisition? *Applied Linguistics*, 19 (3): 357-377.
- [11] Jones, M., Levin M., Levin, J. & Beitzel, B. (2000) Can vocabulary-learning strategies and pair-learning formats be profitably combined? *Journal of Educational Psychology*. 92 (2): 256-262.

- [12] Joseph, S., Smith Lewis, A. & Joseph, M.H. (2004) Adaptive Vocabulary Instruction. IEEE International Conference on Advanced Learning Technologies, 141-145.
- [13] Katsikopoulos, K.V., Fisher, D.L. (2001) Formal requirements of Markov state models for paired associate learning. *Journal of Mathematical Psychology* 45 (2): 324-333
- [14] Nesbit, J.C. & Yamamoto N. (1991) "Sequencing Confusable Items in Paired-Associate Drill" *Journal of Computer-Based Instruction*, 18-1, 7-13.
- [15] Raugh, M.R., Schupbach, R.D. & Atkinson, R.C. (1977) Teaching a large Russian language vocabulary by the mnemonic keyword method. *Instructional Science* 6:100-221.
- [16] Rizzuto, D. & Kahana, M. (2001) An autoassociative neural network model of paired-associate learning. *Neural Computation*, 13 (9): 2075-2092.
- [17] Schneider, V., Healy, A. & Bourne, L. (2002) What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language*, 46 (2): 419-440.
- [18] Siegel, M.A. & Misselt A. L (1984) Adaptive feedback and review paradigm for computer-based drills *Journal of Educational Psychology* 76(2):310-317
- [19] Van Bussel, F.J.J. (1994) Design rules for computer aided learning of vocabulary items in a second language. *Computers in Human Behaviour* 10:63-76