Individual differences in lexical learning and use of a pronunciation feature in REAP, an intelligent tutoring system that teaches ESL vocabulary through reading new words in context

Carol Sisson
csisso@cs.cmu.edu
Language Technologies Institute
Carnegie Mellon University
Pittsburgh Science of Learning Center

INTRODUCTION

This study tests the incorporation of text-to-speech (TTS) generated audio in REAP, an intelligent tutoring system for reading and vocabulary acquisition, and investigates student and word characteristics which affect learning, in an attempt to enhance the student model and improve personalization of the system.

In REAP, students read authentic texts from the web and answer practice questions about targeted vocabulary words. Topics of interest and focus words determined by a pre-test initialize a student model, which is dynamically modified based on performance during practice. Based on this model, appropriate documents are automatically selected. As students read, they have access to built-in dictionary definitions and post-reading practice questions, for which they are given immediate feedback. All student actions are logged by the system and incorporated into the dynamic student knowledge model.

BACKGROUND

Native Language (L1) Writing System and Word Processing Strategies

The students in this study are all adults who read and write fluently in their respective L1s thus they have already developed important skills, that are effective and efficient in their L1s.

Writing systems differ is in terms of how systematically the phonological representation of a word can be accessed based on its orthographic representation. When the mapping is strong and reliable, a language has a high **grapheme-to-phoneme** correspondence. The orthographies of languages where this mapping is very regular are said to have high **phonological recoverability** (Koda, 1998).

The effect of L1 writing systems on second language (L2) word processing strategies is well documented. Until recently, reading researchers believed that phonological information was accessed prelexically by native readers of all languages. However, in the last few years, many studies on word recognition have shown that transfer from nonalphabetic L1s to English is different from transfer from an alphabetic L1 to English (Fender 2003, Wang, Koda and Perfetti 2003, Wade-Woolley 1999). These and other recent studies found that that the nonalphabetic L1 subjects relied less on phonology in pre-lexical processing than did those from alphabetic L1s.

Conversely they found that subjects with nonalphabetic L1s were faster and more accurate on word recognition tasks. These differences are explained through the transfer of the most effective L1 word processing strategies to L2 English reading: **visual processing** for nonalphabetic languages and **phonological processing** for alphabetic languages.

Clusters and Word Level Processing

The effect of consonant clusters on textual word processing is unclear. There is a wide body of research showing that learners from nonalphabetic L1 backgrounds have difficulty with perception, production and metalinguistic manipulation of L2 clusters that violate the phonotactic (syllable formation) constraints of their L1.

However, results from many reading studies suggest that readers from nonalphabetic L1 backgrounds are less sensitive to intraword information than are those from alphabetic backgrounds (Koda 1990, Koda 1999 and Muljani, Koda & Moates 1998). It appears that the effects of clusters and intraword information found in tasks requiring phonological processing, are minimized or non-existent in the more passive process of reading, where visual processing is effective.

Data from previous REAP studies motivated us to further investigate the effect of clusters on vocabulary acquisition through reading. In the spring 2007 semester, 32 level 4 students were pre-tested on a random 200-word subset of the academic wordlist. In the pre-test they were asked to indicate which words they already knew.

Word length appeared to have significant effect on the likelihood that students knew a word. The average length of the known words was significantly shorter than that of the unknown words. Controlling for word length in syllables, we compared the prior knowledge of words from different cluster categories, defined as follows:

Easy Cluster - One or less cluster

Hard Cluster - Three or more clusters or two clusters, where one is longer than two letters

As seen in the table below, easy cluster three- and four-syllable words were more likely to be known than hard cluster words of the same length.

Overall Pretest Knowledge

Cluster Category	2 syllables		3 syllables		4 syllables	
	% known	sample size	% known	sample size	% known	sample size
Easy	49.3	672	60.9	450	67.3	199
Hard	50.8	490	41.3	520	50.9	216
p-v alue	0.299		4.55E - 10		3.18E - 4	

These pre-test results suggested that in the ESL classroom and in everyday exposure to English, students are either less likely to be exposed to words with more clusters or that they are less likely to retain such words.

The post-test dataset was more limited in size. In a comparison of overall post-test performance on three-syllable words, which was the only substantial word length group represented, performance was highest on words with fewer clusters. Words which were classified as hard cluster words were learned significantly less than words classified as medium cluster words ($p = 0.023$) and words classified as easy cluster words ($p = 0.019$). The results are shown in the table below.

Post-test Performance

Cluster Category	3 syllable words	
	% correct	sample size
Easy	52.1	48
Medium	45.7	208
Hard	33.7	95

EXPERIMENT

L1	n
Arabic	10
Chinese	7
Japanese	8
Korean	20
Other	11
Total	56

Subjects

The subjects in this study were the 56 adult students enrolled in the intermediate and advanced level reading classes at the English Language Institute (ELI) of the University of Pittsburgh. The table below shows the native languages of the subjects.
The table below contrasts the phonotactic constraints of the languages and also shows the basic unit of representation in the writing systems of the languages.

Comparison of Primary L1 groups and English

Syllable Inventory	Language	Writing system	Representational unit
CV, CVV, CVC, CVVC, CVCC, VC1	Arabic (L1)	alphabetic	phoneme (consonants)
CV, CVN, CGV, CGVN2	Chinese (L1)	nonalphabetic (logographic)	syllable, morpheme (often word)
V, VV, CV, CVV, CVC3	Japanese (kanji) (L1)	nonalphabetic (logographic)	morpheme, syllable
V, CV, CVC4	Korean (L1)	alphabetic	phoneme, syllable
(C)(C)(C)V(C)(C)(C)(C)	English (L2)	alphabetic	phoneme

- Phonotactic constraints in the L1s more strictly limit complex syllables than in English
- Fewer consonant clusters are possible within and between syllables in the L1s
- Arabic and Korean have higher phonological recoverability than does English
- English grapheme-to-phoneme rules are less systematic than those of many alphabetic languages
- Chinese and Japanese, which are nonalphabetic, have very low phonological recoverability

Materials

The target word list consisted of 30 multisense words from the Academic Wordlist (Coxhead, 2000). The words were deliberately chosen to be hard or easy cluster words.

The students began the semester reading at the seventh or eighth grade level. Documents were selected from the pre-crawled corpus based on reading level, text quality, presence of focus words and topic. The target length of a document was about 1000 words.

The audio consisted of pre-recorded TTS mp3s, generated using an American male voice built by the Festival speech engine (http://festvox.org) using unit selection concatenative synthesis. The use of TTS made it easy to generate pronunciations of all 30 words and the 79 inflectional variants which were also highlighted in the documents. This allowed us to investigate the adequacy of current word-level TTS in foreign language CALL applications.

Hard Cluster Words	Easy Cluster Words
issue	aid
complex	code
shift	pose
brief	tape
suspend	bond
channel	major
trigger	panel
depress	volume
contract	factor
conceive	manual
function	monitor
principal	qualify
supplement	parallel
appreciate	procedure
foundation	
transmission	

Method

The within-subjects variable was the availability and use of the audio feature; for each subject, audio was available for half of the focus words. The between-subjects variable was L1.

On the last day of the semester, students were given a post-test. Our results are based on responses to 38 cloze questions which were written and reviewed by several ELI reading teachers and the researchers. To supplement the post-test of slower readers, who saw fewer words over the course of the semester, some post-test questions for unseen or unpracticed words or senses were added to total 38 questions per student. In the analysis of the data we divide between **practiced** and **unpracticed** results.

HYPOTHESES

Hypothesis 1 - Students with an alphabetic L1 background, who rely heavily on phonological processing in reading, will make use of the audio feature more often that those with a nonalphabetic L1 background, who rely more on visual processing.

Hypothesis 2 - The audio feature will be used more on hard cluster words than on easy cluster words. All of the principal L1s have more limited clusters than English. Koreans are most likely to show this trend because Korean both limits clusters and incorporates a high level of phonological processing in reading.

Hypothesis 3- Those who use the audio feature frequently will have greater post-test accuracy on the words where audio was available than on words where audio was unavailable. The use of the audio feature will help them retain new words.

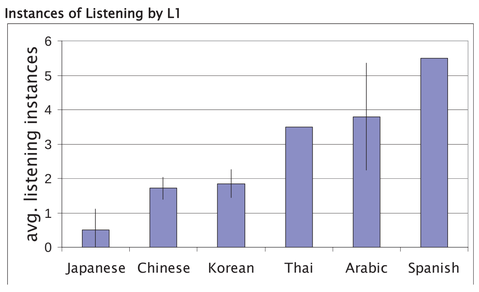
Hypothesis 4 - The Students with L1s that strictly limit clusters (Japanese and Chinese and Korean) will have a greater performance differential based on cluster category than will those from L1s that allow more clusters (Arabic).

RESULTS

The use of the audio feature varied greatly. Even with substantial encouragement from teachers to try out the new feature, 13 of the 56 students did not listen to a single word. At the other extreme one Arabic speaker listened on 16 different instances.

Hypothesis 1

Consistent with our hypothesis, students with alphabetic L1s used the audio feature more than those with nonalphabetic L1s. The average number of listening instances per alphabetic student was 2.74 (93 total instances) while that of nonalphabetic students was 1.15 (15 instances). This difference is significant ($p = 0.00619$).



The graph to the left shows the average number of listening instances for each L1. Thai (alphabetic) and Spanish (alphabetic) do not show standard error because they were each represented by only two students.

Hypothesis 2

There was no significant difference in the use of the audio feature on hard and easy cluster words. Data was sparse because there was very limited use of the audio feature.

Hypothesis 3

Only six students used the audio feature five or more times and took the post-test. In this very limited dataset, there was no consistent or significant effect of choosing to listen to a word.

Hypothesis 4

The table below compares the percentage accuracy of each L1 group on all easy cluster words and hard cluster words. It is based only on post-test questions for practiced words.

Post-test accuracy (%) and cluster category

Cluster Category	Arabic	Chinese	Japanese	Korean
easy	71.7%	67.4%	74.7%	73.6%
hard	67.0%	67.2%	67.5%	66.2%
differential	4.8%	0.3%	7.3%	7.4%

These results are largely consistent with our hypothesis.

- The Korean group has the largest differential, suggesting difficulties with hard cluster words.
- The Chinese group shows no difference in performance based on cluster category, although Chinese clusters are very restricted. This suggests visual processing.
- The Japanese group has a large performance differential. This result is surprising because although clusters are limited in the L1, we expected Japanese readers to rely most heavily on visual processing and to perform similarly to the Chinese group.
- Word length probably accounts for some of this differential.

The table at the top of the next column breaks down L1 groups' performance by whether or not students had practiced the post-tested word and also by whether or not they listened to the word or did not listen (either by choice or because audio was unavailable).

Listening, Practice and Post-test Accuracy

			Percentage Accuracy			
			Arabic	Korean	Chinese	Japanese
Practiced	Listened	easy cluster	70.6	72.2	N/A	
		hard cluster	69.2	60.0		
	Silent	easy cluster	72.0	73.7		
		hard cluster	66.7	66.8		
Unpracticed		easy cluster	35.5	49.4	63.6	59.0
		hard cluster	48.6	55.1	61.8	63.8

- The Arabic, Korean and Japanese groups have higher accuracy on unpracticed words, when the word is a hard cluster word. This suggests that there is a bias towards choosing hard cluster words when the answer is unknown.
- The Korean students were 12.2% more accurate on easy cluster words they listened to than hard cluster words.
- For both th Korean and Arabic groups, words without audio had higher accuracy if they were easy cluster words.

Part of Speech

Overall post-test performance was significantly (Chi-square is 62.831 with 2 degrees of freedom and $p < 0.0005$) affected by the part of speech to be supplied in the cloze questions. Distracters for the cloze questions had the same part of speech as the correct answer. The table below shows this result.

Part of Speech and Overall Post-test Accuracy

	correct	incorrect	n	% correct
Adjective	211	65	276	76.4
Noun	669	249	918	72.9
Verb	327	265	592	55.2

DISCUSSION

Many of the results of this study are suggestive, but not statistically significant. Part of this may be due to data sparseness in some cases or to the fact that the data was collected in a real classroom, unlike the rigorously controlled the lab studies described in the background section.

The results support theories that relate L2 reading processes to L1 scripts. The alphabetic L1 groups with high phonological recoverability do appear to rely more heavily on phonological processing when reading in English than the nonalphabetic groups. This difference is supported by the more frequent usage of the audio feature by alphabetic students. Also, the lack of effect of cluster difficulty on the Chinese group is evidence that they do rely principally on visual processing and are unhindered by phonology in reading. The Japanese group showed more evidence of phonological processing than expected in the comparison of overall performance on hard and easy cluster words. The use of different types of scripts in Japanese makes it more difficult to classify the processes used in reading.

Based on these results, the REAP student model could be improved by giving more practice to words based on length, part of speech and cluster density (for alphabetic students). Additionally, this project suggests several areas of interesting future research. The lack of evidence for phonological processing among Chinese speakers could be tested with more interactive, listening-based practice questions. A study of vocabulary transfer from reading to speaking would also be interesting. Additionally, within the framework of REAP, long-term retention as related to listening and cluster difficulty could easily be tested.

ACKNOWLEDGEMENTS

Funded by the National Science Foundation through the Pittsburgh Science of Learning Center and the Barbara Lazarus Women@IT Fellowship.