

Text-to-speech in Vocabulary Acquisition and Student
Knowledge Models: a Classroom Study Using the
REAP Intelligent Tutoring System

Carol Sisson
Language Technologies Institute
Carnegie Mellon University
Technical report number :CMU-LTI-07-009

20 August 2007

Table of Contents

- 1 Language Technology and REAP
 - 1.1 REAP System Description
 - 1.2 Finding Documents
 - 1.3 Student Modeling
 - 1.4 Text-to-speech (TTS)
 - 1.4.1 Types of TTS
 - 1.4.2 TTS and Computer Aided Language Learning (CALL)
- 2 Linguistics Literature Review
 - 2.1 Phonology and Phonotactics
 - 2.2 Classroom Habits and Exposure
 - 2.3 Reading Process
 - 2.3.1 Arabic
 - 2.3.2 Chinese
 - 2.3.3 Japanese
 - 2.3.4 Korean
 - 2.3.5 Comparison of Reading Processes
 - 2.4 Clusters and Word Level Processing
- 3 Classroom Study
 - 3.1 Subjects
 - 3.2 Design
 - 3.3 Stimuli
 - 3.3.1 Words
 - 3.3.2 Documents
 - 3.3.3 Audio
 - 3.3.4 Post-test Questions
 - 3.4 Method
 - 3.5 Hypotheses
 - 3.6 Results
 - 3.6.1 Use of Audio Feature
 - 3.6.2 Clusters and Use of Audio
 - 3.6.3 Effect of Audio Quality
 - 3.6.4 Performance of Frequent Listeners
 - 3.6.5 L1-based Performance Differences
 - 3.6.6 Logistic Regression of Word and Student Characteristics
- 4 Discussion

1 Language technology and REAP

This paper describes REAP, an intelligent tutoring system for reading and vocabulary acquisition. The final sections describe a study of students using REAP at the English Language Institute (ELI) at the University of Pittsburgh in the summer of 2007. The study tests the incorporation of text-to-speech (TTS) generated audio in the system and investigates possible native language-related improvements to the student knowledge models.

1.1 REAP System Description

A key advantage of intelligent tutoring systems is the ability to automatically cater to the learning needs of the individual student. REAP combines several language technologies to create a personalized approach to vocabulary acquisition through reading.

In REAP, students read authentic texts from the web and answer practice questions about targeted vocabulary words. Students are pre-tested to determine what topics¹ interest them and which words they need to learn, creating a personalized focus word list. Based on a model of a given student's knowledge, appropriate documents are automatically selected for the individual student. 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.

REAP and other intelligent tutoring systems represent a rich opportunity for the application of several language technologies. The success of the system relies on solid information retrieval and language modeling, in order to find appropriate high-quality documents for students to read. Also at the core of the system, personalization requires statistical models of students' knowledge and interests. Other technologies could improve the system as well. Speech synthesis and recognition have the potential to make tutoring a multimodal experience. Natural language understanding could make it possible to automatically generate and grade more interactive question types, which require active vocabulary knowledge on the part of the student.

1.2 Finding Documents

To find the high quality, authentic and documents that are the foundation of REAP, researchers crawl the web for texts containing target vocabulary words. Inflectional variants of words are included in the search queries as well, such as *triggered*, *triggering* and *triggers* for the focus word *trigger*. The documents are annotated and indexed after they have been downloaded. Filters are applied for text quality, length and grade-level using language modeling techniques developed by Collins-Thompson and Callan (2004). Selecting documents that contain high level focus words in lower grade level contexts is a challenging task. High level focus words are most commonly found among other high level words. Documents are classified by topic

¹ Topic choice has been shown to improve students' interest and performance (Heilman et. al to appear)

using the SVM-Light toolkit (Heilman, Juffs & Eskenazi, to appear). About four million documents were downloaded and only 0.5% of them were found to be suitable (Le Zhao, personal communication August 19, 2007).

1.3 Student Modeling

Intelligent tutoring systems allow students to work at their own pace on the areas where they need the most practice. This personalization requires a detailed model of each student's knowledge. The foundation of the individual knowledge models is the initial focus word list and topics of interest, indicated by the student in a pre-test. The self-assessment pre-test used in REAP, where students read a list of words and indicate the ones that they already know, is a fast way to create a list of focus words for the student model. Self-assessment is accurate when students say that they do *not* know a word, but unfortunately it is not as reliable when they say that they do know a word (Heilman and Eskenazi, 2006). For words that have multiple distinct senses it is possible that students might know one sense without being aware that a word has other meanings. For this reason, REAP provides practice questions for all target words that appear in a document, even those that the student claims to know. The focus words determined by the pre-test are given priority in the student model which determines the next documents to be presented.

In the current version of REAP, student models are based entirely on a student's behavior in the system. Topics of interest, focus words, a history of which words have been viewed and performance on practice questions are combined to form a dynamic student model. An improved student model could take into account students' backgrounds, in addition to their behavior. Incorporating students' linguistic backgrounds and characteristics of the target words could complement the information from the pre-test, providing a greater degree of individualization. A more detailed student model could affect the order and priority of document and word presentation, the type of practice questions given and even the type of feedback provided.

For example, a Spanish speaker should need less practice than a Korean speaker in order to learn English words with Latin origins. In this case, the student model of the Spanish speaker could give lower priority to displaying documents containing such words. Chinese speakers might have difficulty with long words with complex onsets which are not permitted in Chinese, meaning that REAP should provide extra practice on words of this type, by displaying them more frequently in documents and providing more interactive practice questions.

The study described in this paper, along with data studies from previous semesters of REAP usage, investigates student and word characteristics that affect learning, in an attempt to enhance the student model.

1.4 Text-to-speech (TTS)

1.4.1 Types of TTS

Text to speech systems generate speech signals from text. Before synthesis takes place, the text is processed using a lexicon, letter-to-sound rules and models that disambiguate part of speech and define supra-segmental features. The output is usually synthesized using one of three principal methods: formant, articulatory and concatenative synthesis. Formant synthesis generates the speech signal based on an acoustic model of a human voice, while articulatory synthesis is based on a model of the vocal tract. These types of synthesis tend to be more flexible than concatenative synthesis, which combines units of real human speech from an indexed and labeled speech database. Concatenative synthesis generally has the advantage of higher quality, more natural sounding speech; however, it is depended on the quality and coverage of the speech database from which it is built.

The synthesis used in this study is built using unit selection concatenative synthesis with the festival speech synthesis system. Festival is a free engine that has hundreds of thousands of users throughout the world. It is documented at <http://festvox.org>.

As mentioned, concatenative synthesis depends on a database of human speech. The database must be designed to have sufficient coverage of all phonetic combinations that will occur in that language that is to be synthesized. Unit selection synthesis automatically chooses the most contextually correct sample of an appropriate unit, which is usually a single phone or sound. In order to create natural transitions between units, the synthesis engine selects a unit along with part of its context, modifies it to impose prosody according to a model and concatenates it smoothly with the adjacent units (Black & Lenzo). Concatenative synthesis tends to maintain the character of speech in the database. It is important that the speaker's volume and register as well as the recording quality and conditions remain consistent throughout the hours required to create the database.

As with many language technologies, limiting the domain results in higher quality. Synthesis is more challenging domains with many words that are not in the lexicon, such as proper names in news text or foreign words in literature. Applications such as dialogue systems and storytelling, which require highly accurate intonation, test the capabilities of prosodic modeling in current synthesis technology.

1.4.2 TTS and Computer Aided Language Learning (CALL)

There are several clear benefits of TTS over pre-recorded human speech. TTS can create large numbers of unique utterances that need not be specifically planned for in advance. TTS requires less storage space than large recorded speech systems, which must plan and store every desired utterance, despite the redundancy and inefficiency of doing so. In addition to the flexibility of content afforded by TTS, the prosodic features of utterances can be modified.

However, use of TTS in CALL remains limited (Egan & LaRocca, 2000; Sobkowiack, 1998). Synthesis quality remains a concern and evaluation of the adequacy

of the technology receives little attention (Handley & Hamel 2005). It is difficult to satisfactorily evaluate the somewhat subjective quality of a continually advancing technology. Acceptance of TTS in CALL also requires further study about how it affects the attitude and motivation of those who listen to it.

A common concern regarding TTS in CALL is that students will learn to imitate the mistakes and occasional unnatural sounds made by TTS (Alan Black, personal communication March 15, 2007). While this may be a concern for pronunciation training of beginners, who are still learning the distinctive phonemes of a language, or those who are learning fine-grained supra-segmental aspects of pronunciation, it does not rule out pronunciation training at the segmental level. The students who took part in the study described in section 3.1 are fairly advanced, but could still benefit from a tool that could tell them, for example, whether the second 'c' in *conceive* is pronounced as /k/ or /s/.

This application of TTS is demonstrated by talking dictionaries, which augment written dictionary entries with sound. In this case, the storage benefits of TTS are made clear; systems with recorded speech tend to only have space to include the pronunciation of the head word, while TTS makes it easy to include additional pronunciations of morphological variants of the head word, as well as examples of usage. Talking dictionaries using TTS are commercially available, such as the *Oxford-Hachette French Dictionary on CD-ROM* (2003), which uses the *RealSpeak* TTS synthesizer. In REAP, it is important for dictionaries to be built in to the system so that students' usage can be tracked and incorporated into the student models. The study described in section three tests whether the inclusion of a limited TTS-generated talking dictionary improves students' vocabulary acquisition.

If the feature is shown to be beneficial it could be extended to all words in a document and TTS could also be used to develop new types of questions. To increase students' motivation to learn the phonetic form of words they read, we could use cloze questions where the student must listen to several options before typing the missing word into the blank. Similarly we could develop exercises that require students to take dictation or to match written and auditory forms of words.

2 Linguistics and Literature Review

In the adult second language classroom two pieces of a student's background are of critical importance to learning: the native language of the student and the learning strategies he or she already relies upon.

The process of learning a second language (L2) as an adult is highly affected by transfer from the learner's first language (L1). Features of the L2 which are also found in the L1 undergo positive transfer and are learned more easily than features which are newly encountered in the L2. New features of the L2 are often affected by negative transfer, requiring a longer acquisition period and more frequent errors. Negative and positive transfer effects are well documented in the syntax and phonology of L2 learners (Hawkins, 2001). This study investigates L2 vocabulary acquisition, making word-level phonological transfer and phonotactic constraints relevant.

Also important in the case of adult vocabulary acquisition through reading is the reading process and word processing skills inherently learned through native fluency in

the writing system of the L1. Researchers have found that readers from typographically different L1s rely on different skills and even different areas of the brain (Perfetti and Liu 2005) as they read.

Finally, the classroom environment of this study makes the study habits and classroom behavior developed in the student’s country of origin pertinent.

2.1 Phonology and Phonotactics

Phonological and phonotactic constraints define the viable syllable shapes and phoneme combinations in a given language. English has a larger syllable inventory than any of the L1s represented by the students investigated in this study (hereafter referred to as the L1s), resulting in many more possible consonant clusters within and across syllables. Table 2.1.1 shows the syllable inventory of the L1s in contrast with the large syllable inventory of English shown in table 2.1.2.

Table 2.1.1

Language	Syllable Inventory
Arabic	CV, CVV, CVC, CVVC, CVCC, VC ²
Chinese	CV, CVN, CGV, CGVN ³
Japanese	V, VV, CV, CVV, CVC ⁵
Korean	V, CV, CVC ⁶

Table 2.1.2

English Syllable Inventory
(C)(C)(C)V(C)(C)(C)(C) ⁴

Chinese, Japanese and Korean permit a single coda consonant, restricted to a subset of the consonants in the language. Arabic can have up to two coda consonants. English, on the other hand, can have as many as four coda consonants. English also permits up to three onset consonants in a single cluster, while none of the L1s allow complex onsets.

This study investigates clusters at the word level, with a limited focus on syllable structure. The syllable inventories of the L1s are discussed in order to characterize the possible word-level clusters. When these syllables combine to form words, the L1s cannot have medial consonant clusters longer than two, or three, in the case of Arabic. English has medial clusters of up to four consonants, often in compounds like *catchphrase*. Word initial clusters have the same limitations as onset clusters and are not

² VC in words like “al” pronounced with no prior context. When in fluent speech, it would be elide with the previous word, either by dropping the V before an open syllable or borrowing the previous consonant of a closed syllable.

³ G is a glide ([j], [ɥ], or [w]) and N is a nasal ([n] or [ŋ]). Cantonese has one more nasal sound that can occur at the end of syllables, namely, [m]. Differing again from Mandarin, in addition to the nasal endings, Cantonese has three stop endings, i.e., [p], [t], and [k]. (unreleased)

⁴ There are restrictions on large clusters in English. The outer-most consonants in large clusters are a appendices, which are usually an /s/ as in *strengths*.

⁵ The coda can only consist of a nasal or the first part of a geminate consonant.

⁶ The coda can only consist of nasals [m], [n], [ŋ], the liquid [l] and unreleased stops [p], [t] and [k].

found in any of the L1s. Word final clusters, like coda consonant clusters, are not found in Chinese, Japanese or Korean, and they are limited to two in Arabic.

Given the rarity or impossibility of consonant clusters in the L1s, we hypothesize that English words with multiple clusters or long clusters of consonants will be more difficult to learn. This hypothesis is supported by a wealth of related literature on L2 perception and production of consonant clusters. Many of these studies focus on the strategies L2 learners use to produce clusters which are illegal in their L1, such as breaking up or resyllabifying clusters through epenthesis or deletion (Kabak 2003). There is also evidence that L2 learners perceive additional epenthetic vowels when listening to words which contain clusters that violate the phonotactics of their L1 (Dupoux et. al. 1999).

An important qualification here is that the analysis of syllable structure, as well as the studies on L2 perception and production of clusters, is limited to analysis of *phonetic* consonant clusters. Students in this study are learning vocabulary through reading, making orthography very important. The mapping of English letters to phonemes is often not one-to-one, as in the case of words like “eighths,” /ejθs/. Because of the focus on reading, consonant clusters are counted in terms of orthography. However, the previous description of permissible syllable structure and phonetic consonant clusters is still relevant is still a useful guideline.

This study emphasizes consonant clusters, however vowel clusters are also considered. English has many diphthongs, often written with adjacent vowel characters. It seems reasonable to hypothesize that such vowel clusters might be difficult for Arabic speakers whose orthography does not encode vowel information.

Further evidence for the difficulty presented by consonant clusters comes from the results of previous REAP studies, which suggest that words with more consonant clusters are less likely to be familiar to students initially and also less likely to be learned over the course of the semester.

In the spring 2007 semester, 32⁷ level 4 students were pre-tested on a random 200-word subset of the academic wordlist. Each student’s subset was unique. They were asked to indicate which words they already knew. Overall, the words that students knew were significantly shorter, in terms of letters, and had significantly fewer clusters than the words that they did not know. They were also significantly less likely to know words containing a long cluster of three or four characters. The p-values and means are in table 2.1.3

⁷ The L1s were as follows: 10 Korean, 9 Arabic, 6 Chinese, 2 Japanese, 2 Spanish, 2 Turkish, 1 Taiwanese

Table 2.1.3 Spring 2007 Pre-test knowledge

	length in letters	number of clusters	number of 3 or 4 character clusters	clusters-per-letter
unknown	7.78	1.38	0.194	0.175
known	7.39	1.28	0.170	0.172
p-value (one-tailed unequal variance t-test)	3.41E -17	7.34E -8	0.00292	0.0731

The number of clusters and the length of a word are related, and so we looked at clusters-per-letter (the number of clusters divided by the number of letters). The difference between the clusters-per-letter of the known words and unknown words was marginally significant, as shown in table 2.1.3.

Looking at length in terms of syllables in an analysis of three- and four-syllable words showed that words with fewer clusters were more likely to be known. For this analysis we defined three levels of predicted difficulty, based on number and type of cluster: easy, medium and hard. Both consonant and vowel clusters were counted. Table 2.1.4 describes the categories.

Table 2.1.4 Pre-test knowledge by cluster category

Cluster Category	Description
Easy	one or less cluster
Medium	two clusters of two characters in length or a single cluster three or more characters long
Hard	three or more clusters or two clusters, where one is longer than two characters

We compared the prior knowledge of words from different cluster categories, while controlling for word length in syllables. Table 2.1.5 shows the means, sample size and significance of this comparison for the universal group. Three-syllable easy cluster words were also more likely to be known in well-represented L1 groups: Arabic ($p=0.000925$), Korean ($p=0.00211$) and Chinese ($p=0.00872$).

Table 2.1.5 Pretest knowledge by cluster category, word length controlled

	2 syllables		3 syllables		4 syllables	
	% known	sample size	% known	sample size	% known	sample size
Easy Cluster	49.3	672	60.9	450	67.3	199
Hard Cluster	50.8	490	41.3	520	50.9	216
p-value (one-tailed unequal variance t-test)	0.299		4.55E -10		3.18E -4	

Taken together, these results indicate that there is some effect of clusters that is independent of word length. The pre-test results suggest 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.

Word length also appears to have significant effect on the likelihood that students will know a word. All three major L1s are affected in the same way, as shown in table 2.1.6.

Table 2.1.6 Pre-test knowledge, average number of letters

	Arabic	Chinese	Korean
Unknown	7.63	7.86	7.94
Known	7.29	7.42	7.35
p-value (one-tailed unequal variance t-test)	1.47E -4	6.92E -5	1.93E - 10

The post-test for this semester showed some trends about the difficulty of learning words with clusters and longer words, but few significant results. For all three main L1s, Arabic, Chinese and Korean, the average number of clusters in the words they got wrong on the post-test was higher (though not statistically significant) than that of the words they got right. The average length of the words which the Koreans got right was shorter than that of the words they got wrong (marginally significant, $p=0.067$). For the universal group, ignoring L1 differences, words with clusters of three or four consonants were less likely to be learned (marginally significant, $p= 0.069$).

Looking only at the three-syllable words that occurred on the post-test for the universal group, words which were classified as hard cluster words were learned significantly less than words classified as medium cluster words (one-tailed t-test, unequal variance assumed, $p = 0.023$) and words classified as easy cluster words ($p=0.019$). The success rates and sample sizes are shown in table 2.1.7.

Table 2.1.7 Post-test success

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

2.2 Classroom Habits and Exposure

The students in this study are all adults whose learning strategies have been shaped by previous classroom experience. Pelletreau (2006) did a qualitative analysis of 13 students' use of REAP. He found that students fell into two distinct groups while using REAP: those who took notes on vocabulary words and focused on target words and those who asked the teacher vocabulary questions and concentrated primarily on non-target words. Only one student used both practices. The four Arabic speakers

represented the largest L1 group among the 13 students and they all fell into the group which preferred asking the teacher vocabulary questions.

In addition to the linguistic transfer discussed previously, students can be affected by conditions of their education and cultural exposure in their native countries. South Koreans are familiar with the Latin alphabet from use of the McCune Reischauer transliteration system on signs and in advertisements (Lee, 2001). English classes in Asia tend to emphasize reading and writing more than speaking. Students from Asia tend to lag behind in speaking while Arabic students tend to have better speaking skills. Borrowing of words can also give some students an occasional advantage, although often borrowings are adapted to fit the phonotactic constraints of the borrowing language so that the connection to the English word is no longer transparent. Korean and Japanese are also more open to borrowing English words than are Chinese and Arabic.

2.3 Reading Process

The students in this study are all adults who read and write fluently in their respective L1s thus they have already developed reading skills and strategies that are effective and efficient in their L1s. These skills are language-dependent and so not all students come to the ESL classroom relying on the same skill set.

L2 textual word processing strategies have been shown to be highly affected by the writing system of the L1. An important way that writing systems can differ is terms of how systematically the corresponding phonological representation of a word can be accessed, given its orthography representation. The orthographies of languages where this mapping is very regular are said to have high phonological recoverability (Koda, 1998). Following is a description of the writing system of each of the L1s, which is summarized in table 2.3.5.1.

2.3.1 Arabic

Arabic is written alphabetically in a linear, right-to-left form. Only the consonants are represented by unique graphemes. Young native speakers of Arabic learn to read and write using a phonologically straightforward orthography where vowel information is represented by diacritic marks on the consonants. More mature readers and writers of Arabic drop these diacritics, using a system which underspecifies the form of a word and create ambiguities (Abu-Rabia 1997a, 1997b, 1999). In terms of consonants, Arabic has much more reliable grapheme-to-phoneme correspondence than does English.

2.3.2 Chinese

Chinese has a logographic orthography which maps a character to a monosyllabic morpheme. It has a nonlinear layout and does not encode individual phonemes. Two characters can have the same pronunciation without having any resemblance to each other (Perfetti 2005).

Chinese children are taught Standard Mandarin writing with the aid of Pinyin (or similar phonetic systems), an alphabetic Romanization of Chinese that helps them

associate phonology with orthography. Pinyin is also useful for typing Chinese text on computers. Despite the early use of this phonetic system, phonology plays only a minor role in Chinese reading. Tan, et al (2005) found that native Chinese reading skills are strongly associated with writing skills and not with phonological sensitivity, contrary to the common belief that phonological sensitivity is a universal predictor of reading ability (Tan et. al., 2005). The connection between reading and writing in Chinese is manifest in the common practice of asking school children to repeatedly copy characters with attention to their internal structure (Tan et. al., 2005).

2.3.3 Japanese

Japanese is written with a combination of kanji symbols, borrowed from Chinese characters, and two syllabic scripts called hiragana and katakana. With the exception of some media like advertisements, kanji is the dominant orthography so Japanese writing is considered to be typographically similar to Chinese.

Japanese has a large lexicon, with many homonyms meaning that Japanese speakers are accustomed to noting the orthographic form of new vocabulary (Thompson, 2001). As in Chinese, there is an emphasis on memorizing new kanji words as whole entities.

2.3.4 Korean

Hangul, the Korean writing system, combines letters in a square formation called a kulja, which corresponds to a syllable. All syllables are written in one of six possible kulja orientations, each containing two to four letters (Yoon et. al., 2002). Chinese characters are also used in Korea, although they are not taught in school. According to Taylor (1995), Chinese characters make up about 10 percent of the words in the average daily South Korean newspaper. Korean has a strong grapheme-to-phoneme relationship.

2.3.5 Comparison of Reading Processes

Table 2.3.5.1 summarizes the writing systems of the four major L1s represented by the students. Spanish and Thai are only represented by two student and are included in table 2.3.5.1 for completeness.

Table 2.3.5.1

Language	Writing system	Representational unit
Arabic	alphabetic	phoneme (consonants)
Chinese	nonalphabetic (logographic)	syllable, morpheme (often word)
Japanese (kanji)	nonalphabetic (logographic)	morpheme, syllable
Korean	alphabetic	phoneme, syllable
Spanish	alphabetic	phoneme
Thai	alphabetic	phoneme
English	alphabetic	phoneme

The orthographies of the two alphabetic L1s, Arabic and Korean, have higher phonological recoverability than English. English grapheme-to-phoneme rules are less systematic than those of many alphabetic languages, sometimes making it necessary for readers to look at longer combinations of letters to find consistent grapheme-to-phoneme mappings (Treiman et. al. 1995). The nonalphabetic L1s, Chinese and Japanese, have very low phonological recoverability in their orthography.

The effect of L1 writing systems on L2 word processing strategies is well documented. Until recently reading researchers believed that phonological information was accessed pre-lexically 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) investigated English word recognition skills of Arabic and Japanese speakers, Wang, Koda and Perfetti (2003) compared word recognition skills of Korean and Chinese ESL learners and Wade-Woolley (1999) compared Japanese and Russian speakers on several word-level reading tasks.

These aforementioned studies consistently found 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 to be 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. Similar to these word-level studies, Akamatsu (2003) found that the reading speed of Persian (alphabetic) speakers were less affected than that of Chinese and Japanese speakers when reading a case-alternated English text. The visual interruption of case alternation threw the nonalphabetic L1 readers because of their reliance on visual orthographic patterns. This suggests that in REAP, where phonological processing of new vocabulary is not tested, students from nonalphabetic L1s who rely on strong visual processing skills are at no disadvantage.

2.4 Clusters and Word Level Processing

The effect of consonant clusters on textual word processing is unclear. As mentioned, there is much evidence of the difficulties encountered in L2 perception and production of clusters not found in the L1. Learners with nonalphabetic L1s have trouble on metalinguistic tasks like phoneme deletion which requires them to auditorily isolate a target phoneme in a word and then produce the word that results from the deletion of that phoneme. Wade-Woolley (2003) found that Japanese native speakers had more difficulty at this task when they were asked to delete a phoneme from a complex onset. These results indicate that learners from nonalphabetic L1s have difficulty manipulating phonological intraword information in auditory and oral tasks.

However, results of many reading studies suggest that readers from nonalphabetic L1 backgrounds are less sensitive to intraword information than are those from alphabetic backgrounds. Koda (1990) found that the L2 English reading speed of Arabic and Spanish speakers was slowed down by the distraction of unpronounceable key words (Sanskrit symbols) while the speed of Japanese readers was unaffected by the pronounceability of the key words in the text. Koda (1999) and Muljani, Koda & Moates

(1998) found similar results. It appears that the effects of clusters and intraword information found in perception and production studies are minimized or non-existent in the more passive process of reading.

3 Classroom Study

3.1 Subjects

Subjects in this study were the students enrolled in the intermediate and advanced level reading classes at the English Language Institute (ELI)⁸ of the University of Pittsburgh. There were a total of 56 students, 39 in the level 4 classes and 17 in the level 5 classes. Level 5 is the highest level offered at the ELI and level 2 is the lowest. The majority of students at the ELI study English full-time. When they arrive, they take the Michigan Test of English Language Proficiency (MTELP) and are placed in reading, writing, grammar, listening and speaking classes based on their MTELP score, teacher recommendations and their scores on a writing test and a listening test. Table 3.1.1 shows demographic information for the students who took the post-test.

Table 3.1.1 Subjects

L1	Level 4		Level 5	
	n	Avg. MTELP	n	Avg. MTELP
Arabic	5	52.0	5	53.2
Chinese	5	56.8	2	45.0
Japanese	7	70.5 ⁹	1	57.0
Korean	14	59.6	6	62.8
Spanish	2	68.0	0	n/a
Thai	2	59.5	0	n/a
Unknown	4	n/a	3	n/a

3.2 Design

The within-subjects variable is the availability and use of the audio feature. The between-subjects variable is L1. An additional between-subjects variable is the Word Sense Disambiguation (WSD) group¹⁰. Half of the students have been assigned to the control group and half to the WSD group, where the definitions are ordered so that the most likely definition of a word appears first, based on the context in which of the word appeared when the student clicked on it to see the definition.

⁸ <http://www.eli.pitt.edu/>

⁹ Only 4 of the 7 level four Japanese students had available MTELP scores.

¹⁰ The WSD intervention was run by another researcher. We ran simultaneous experiments in order to have enough data.

3.3 Stimuli

3.3.1 Words

The target word list consisted of 30 multi-sense words from the Academic Wordlist (Coxhead, 2000). The words were deliberately chosen to be hard or easy cluster words. In order to find words which also met the multi-sense requirement of the simultaneous Word Sense Disambiguation study, the definitions of hard and easy cluster were slightly different than the three-way classification described in section 2.1. Hard cluster words are words with two or more clusters or a cluster of at least three characters while easy clusters words are still limited to no more than one two-character cluster. Repeated letters, as in “parallel” are counted as clusters. Table 3.3.1.1 shows the base form of the words in the study.

Table 3.3.1.1 Word List

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	

It was not possible to meet the criteria of clusters, multi-senses and grade-level while also controlling the length of the words. The average length of the easy cluster words is 5.57 letters while that of the hard cluster words is 7.81 letters.

The teachers at the ELI avoided these 30 words in all other ELI classes so that students’ exposure to the words was uniformly captured by their use of REAP.

3.3.2 Documents

The level 4 students begin the semester reading at the seventh grade level and the level 5 students begin at the eighth grade level. Depending on student feedback about difficulty, reading level can be modified during the semester. In REAP, documents are selected from the pre-crawled corpus based on reading level, text quality, presence of focus words and topic¹¹. The target length of a document was about 1000 words. In cases of longer documents, a message saying “Stop reading here” was inserted into the document after 1000 words.

3.3.3 Audio

The audio consisted of pre-recorded TTS mp3s, generated using an American male voice built from an example database using the Festival speech engine (<http://festvox.org>). The use of TTS made it easy to generate pronunciation of all 30 words and the 79 inflectional variants which were also highlighted in the documents (See appendix A). If a student clicked on an uninflected focus word, the audio consisted of only that word. However, if they clicked on an inflected form, the audio consisted of the uninflected form followed by the inflected form (e.g. “*trigger...triggered*”).

The Festival voice’s pronunciation of individual words was not tuned because hand-tuning would not be feasible in a large-scale application. This means that the pronunciation quality is not completely consistent from one word to another. However, we believe that it represents a good sample of the overall quality of TTS at the word level. This allows us to investigate the adequacy TTS, in its current state, in foreign language CALL applications.

In the case of the word *contract*, there are multiple pronunciations. The simple verb form (a and b) has stress on the second syllable, while the particle verb (c) is stressed on the first syllable, as is the noun (d) form.

- a. They will conTRACT malaria.
- b. The lungs expand and conTRACT.
- c. Should we CONtract out the job?
- d. He will sign the CONtract.

Because this was the only word among the 30 with such a distinction in pronunciation, we used the noun stress pronunciation in all cases.

3.3.4 Post-test questions

The cloze questions were written and reviewed by several ELI reading teachers and the researchers. Each question targeted a specific sense of a word, distinguished by the WSD model. Every question had two distracters from among the other focus words

¹¹ Topic choice, which students indicated in a pretest survey, has been shown to improve post-test scores. See (Heilman et. al., to appear).

and two distracters from a list of non-focus words covered during the classroom time of the reading class.

3.4 Method

On the first day of the semester the students took a brief, REAP-based pre-test where they were shown a list of the thirty words and asked to indicate which words they already knew well enough to use in a sentence. The pretest also prompted students to indicate their level of interest in several topic categories.

Over the course of the eight-week summer semester, the ELI reading classes used REAP for 40 minutes per week, under the supervision of their reading teachers. Occurrences of the focus words were highlighted in the text to attract students' attention. REAP incorporates easy dictionary access; students can click on a word or type in a word to see a pop-up definition from *Cambridge Advanced Learner's Dictionary*, which also includes example sentences. After each reading, students answer multiple choice questions where they choose the appropriate definition¹² for each focus word that appeared in the reading. Students were also asked to indicate how interesting and difficult a reading was, in order to help the system select the most appropriate documents.

Audio pronunciation of some of the focus words is available in the pop-up definition window that appears upon clicking on a word. For every student, a randomized 15-word subset has audio available. In these cases, students can click on a button to hear synthetic voice pronounce the word. Image 3.4.1 shows an example of the definition window for a word with audio.

¹² The definition questions were automatically generated from Longman's Dictionary.

Image 3.4.1

Listen to pronunciation: channel

channel (n)

- a passage for water or other liquids to flow along, or a part of a river or other area of water which is deep and wide enough to provide a route for ships to travel along
- a way of communicating with people or getting something done
- a television station
- a route or way out of an airport or port where travellers' bags are examined

channel (v)

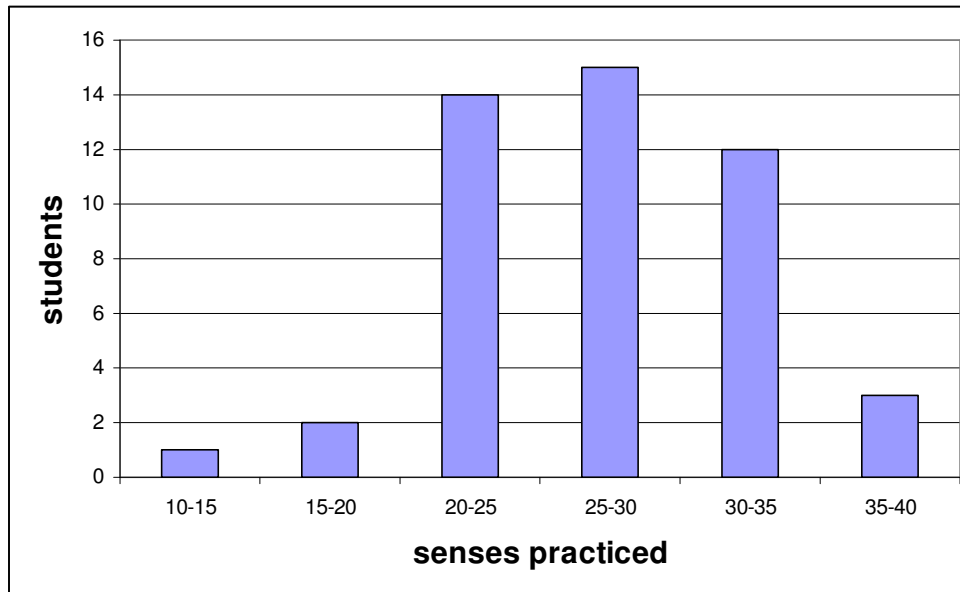
- to direct something into a particular place or situation

Examples:

- There are [drainage/irrigation] channels all over this flat agricultural land.
- The boats all have to pass through this narrow channel.
- We must open the channels [of communication] between the

On the last day of the semester, students were given a post-test. First they wrote 10 sentences using 10 pre-selected words from REAP. Then they answered 38 cloze questions. The students were asked a cloze question for every unique focus word sense they were exposed to over the course of the semester. All but two students had practiced fewer than 38 senses during the semester, as shown by graph 3.4.2.

Graph 3.4.2 Histogram of word senses practiced during the semester



In these cases the post-test was augmented with more cloze questions that covered senses or words to which the student had not been exposed.

3.5 Hypotheses

The current study investigates the utility of an audio pronunciation feature and looks for differences in performance based on the L1 of the student and certain features of the stimuli. Based on the earlier discussion, we make the following hypotheses:

- 3.5.1 Students with an alphabetic L1 background, who rely heavily on phonological processing in reading, will make use of the audio feature more often than those with a nonalphabetic L1 background, who rely more on visual processing.
- 3.5.2 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.
- 3.5.3 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.

3.5.4 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).

3.6 Results

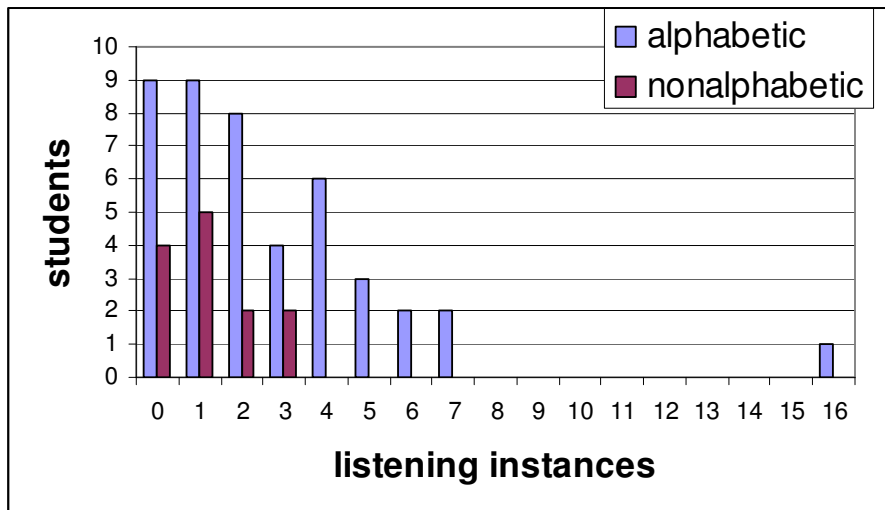
3.6.1 Use of Audio Feature

Students did not use the audio feature as much as we anticipated. This is due in part to an apparent lack of interest in the feature on the part of some students. Also, the average student read only 12.7 documents during the semester, which provided relatively few opportunities to listen.

Overall, 20% of the 6385 words looked up during the semester were focus words. This is an average of 23 focus word per student. The availability of audio was randomly assigned to half the words for each student. This means that there were approximately 650 lookups where the student had to option to use the audio feature. Overall there were 133 instances of listening, which is approximately one listen for ever six opportunities. On average, an individual student looked up 12 words where audio was available.

The use of the audio feature was extremely variable, as seen in graph 3.6.1.1. 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, including a few instances of the same word on different occasions.

Graph 3.6.1.1 Distribution of Listening Instances



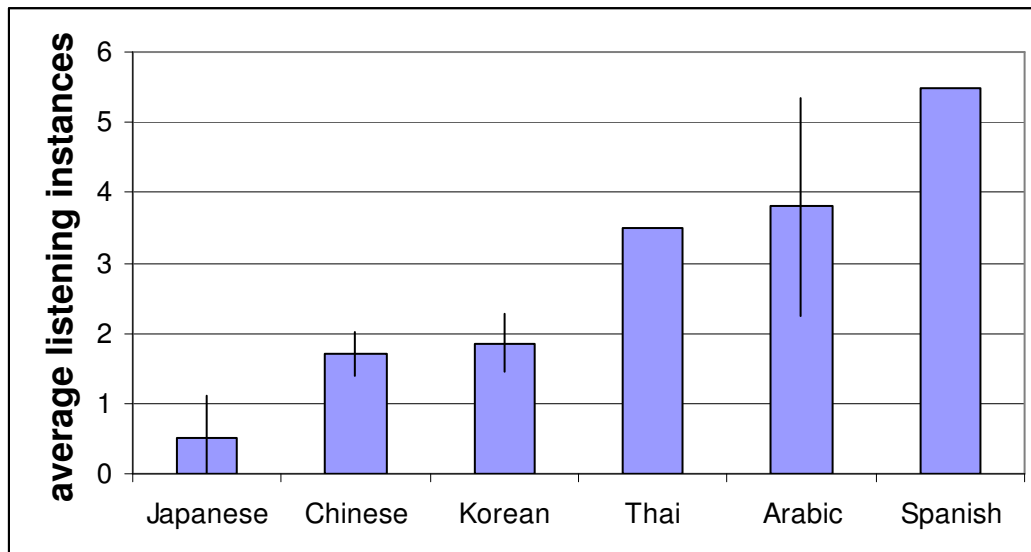
Consistent with hypothesis 3.5.1, students from alphabetic L1s did use the feature more than those from nonalphabetic backgrounds. The average number of listening instances per alphabetic student was 2.74 while that of nonalphabetic students was 1.15. This difference is significant ($p = 0.00619$) and is summarized in table 3.6.1.2.

Table 3.6.1.2 Instances of Listening

	number of listening instances	number of students	listening instances per student
alphabetic	93	34	2.74
nonalphabetic	15	13	1.15

Graph 3.6.1.3 shows the average number of listening instances for each L1. When the one Arabic student who listened 16 times is excluded, the Arabic average falls from 3.8 words to 2.44, which is still higher than that of the Korean students.

Graph 3.6.1.3 Instances of Listening by L1



Thai and Spanish do not show standard error because they were only represented by two students. These results are in accordance with the prediction that those with alphabetic L1s will use the feature more than those from non-alphabetic script backgrounds.

3.6.2 Clusters and Use of Audio

We predicted that the audio feature would be used more frequently on hard cluster words which break the phonotactic constraints of the principal L1s. However, there was no significant difference in the use of the audio feature on hard and easy cluster words. Overall, 64 easy cluster words and 69 hard cluster words were listened to. The Koreans listened to 16 easy cluster words and 19 hard cluster words. These results are based on a very small sample size.

3.6.3 Effect of Audio Quality

The highly variable usage of the feature made us look for an effect of audio quality on subsequent usage. In order to measure the quality of the synthesis, we had seven native speakers listen to all 30 words. The native speakers tried to identify the words and ranked them on intelligibility and naturalness, on a scale of one to five, where five was high. The full results can be found in appendix B. All seven native speakers listened to the 30 words in the same order, so it is possible that their ability to judge naturalness changed as they listened. For this reason, the rank of each word in the evaluation task is included.

We hypothesized that students who heard a poorly synthesized word the first time they used the feature, might be less likely to use it again. Surprisingly, the opposite appears to be true, although we have a very small sample size. Table 3.6.3.1 includes 10 words, each of which was the first listening instance of more than one student. It shows the number of students who accepted the feature and went on to use the audio feature again, after having heard the particular word as their first exposure to the feature.

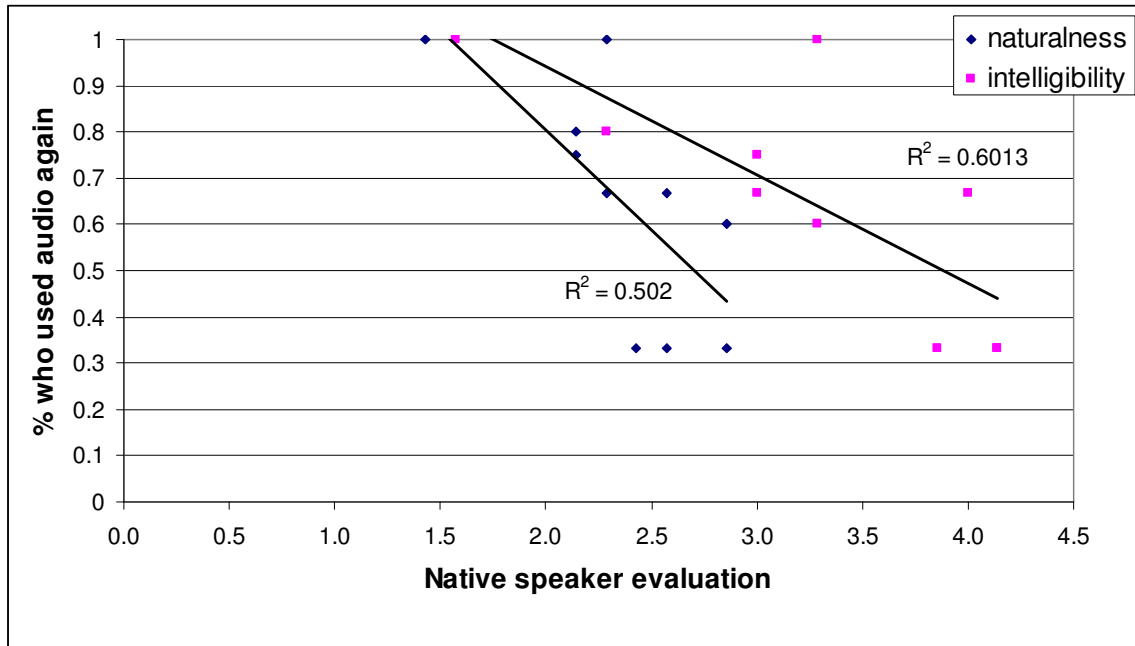
Table 3.6.3.1

First word heard	Rank in evaluation task	Native speakers who correctly identified	Average intelligibility ranking	Average naturalness ranking	Students who used audio feature again	Students who did not use audio feature again	% of students who used feature again
conceive	2	6	3.857	2.571	1	2	33.3
principle	26	7	4.143	2.429	1	2	33.3
qualify	5	7	4.143	2.857	1	2	33.3
aid	18	6	3.286	2.857	3	2	60
major	11	7	3.000	2.286	2	1	66.7
monitor	8	7	4.000	2.571	2	1	66.7
procedure	7	7	3.000	2.143	3	1	75
trigger	22	3	2.286	2.143	4	1	80
complex	19	4	3.286	2.286	2	0	100
panel	20	4	1.571	1.429	6	0	100

Panel had the lowest intelligibility score of all 30 words and only four of seven native speakers were able to identify the word, yet after hearing it, six of six students went on to use the audio feature again.

Graph 3.6.3.2 shows the surprising correlation between low intelligibility and naturalness and acceptance of the feature.

Graph 3.6.3.2 Acceptance of Audio Quality



3.6.4 Performance of Frequent Listeners

We predicted that those who used the feature frequently would perform better on words they listened to. Table 3.6.4.1 breaks down the post-test results of the six students who used the feature five or more times and took the post-test.

Table 3.6.4.1

user	L1	unpracticed		practiced					
				no audio		audio			
		n	%	n	%	listened		silent	
				n	%	n	%	n	%
kyc4	Korean	11	54.5	14	64.3	9	44.4	4	50
fma5	Arabic	8	100	15	73.3	10	90	5	80
ekl9	Thai	15	66.7	14	71.4	5	60	4	75
sad41	Arabic	n/a	n/a	20	50	4	50	14	71.4
jap100	Spanish	9	66.7	12	83.3	9	100	8	87.5
kak119	n/a	11	45.5	13	84.6	11	63.6	3	33.3

Students fma5 (Arabic), jap100 (Spanish) and kak119 (unknown) were more accurate on words where listening was available when they chose to use the feature. Students kyc4, ekl9 and jap100 actually did worse on words they listened to. More data is needed to evaluate the individual benefits of listening. It is very unfortunate that the Arabic speaker who listened 16 times did not take the post-test as he afforded the best opportunity for comparison of performance based on the availability of audio.

3.6.5 L1-based Performance Differences

Hypothesis 3.5.4 predicted that students with L1s that strictly limit clusters (Japanese and Chinese and Korean) will be more negatively affected by hard clusters than those from L1s that allow more clusters (Arabic). Table 3.6.5.1 compares the percentage accuracy of a complete L1 group on all easy cluster words and hard cluster words. It is based only on the post-test questions of words the students had been exposed to during the semester.

Table 3.6.5.1 Post-test accuracy and cluster category

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

As predicted, the Korean group has the largest differential, suggesting difficulties with hard cluster words. Word length probably accounts for some of this differential as well because hard cluster words tend to be slightly longer than easy cluster words. Japanese, like Korean, does not allow consonant clusters and the Japanese group also has a large differential. On the other hand, the Chinese group shows no difference in performance based on cluster category, although Chinese restricts clusters as much as Korean and Japanese do.

The reading process of native Chinese relies almost entirely on visual processing, rather than phonological processing. Korean, on the other hand, involves a great deal of phonological processing which explains why Koreans are affected by consonant clusters which are impossible in their L1. Japanese is classified as a nonalphabetic language, like Chinese, but these results suggest that Japanese learners are affected by clusters and phonological processing, meaning that their reading strategy is not as purely visual as that of the Chinese.

The Arabic group is less affected by consonant clusters. This is in accordance with our prediction, based on the fact that Arabic readers rely significantly on phonological processing and that Arabic allows more clusters (complex onsets) than the other L1s, but still less than English.

The results shown in table 3.6.5.2 below similarly group the Japanese results with Arabic and Korean, rather than with Chinese. 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. The Arabic, Korean and Japanese groups have higher accuracy on these 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.

Only the Arabic and Korean groups had a substantial number of students and significant use of the audio feature. The accuracy of these groups on practiced words (words they read and answered practice questions about during the semester) is further

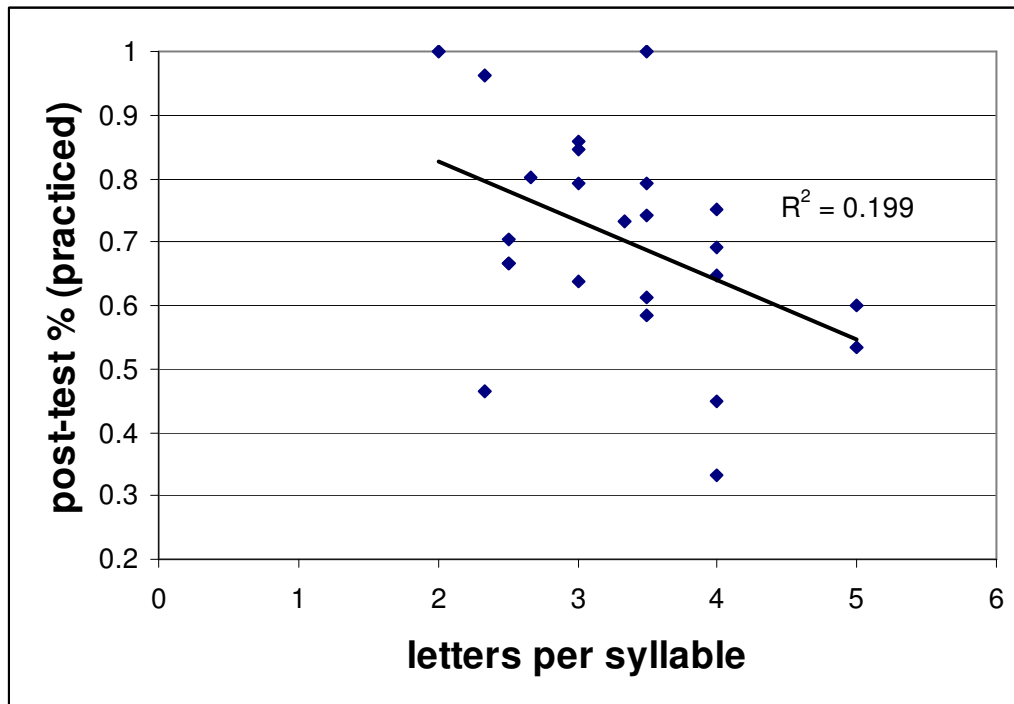
broken down in table 3.6.5.2 by whether or not the student answering the question had listened to the word. The percentages in the “listened” row represent two things: the choice to use the feature and the effect of hearing the word. The cluster difficulty of the words listened to by the Arabic students had no effect. However, the Korean students are 12.2% more accurate on easy cluster words they listened to than hard cluster words. For both languages, practiced words for which listening was either unavailable or unused had higher accuracy if they were easy cluster words.

Table 3.6.5.2 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

Because of the differences in the Korean group’s accuracy on the hard and easy cluster words and the large number of Korean students, we looked at the effect of clusters using another measure. Letters per syllable also represents the density of clusters in a word. Words with many letters per syllable must have some complex syllable structure and clusters. Graph 3.6.5.3 shows the post-test accuracy of the Korean group on words that were post-test for three or more students, who had practiced the word during the semester. It shows a negative correlation between accuracy and letters per syllable, supporting the hypothesis that Korean students are have greater difficulty learning words with more clusters.

Graph 3.6.5.3 Korean letters per syllable and post-test accuracy



Grouping together all Korean responses, including those that were practiced and unpracticed, heard and unheard, the success rate on easy cluster words (n=216) was 73.5% while the success rate on hard cluster words (n=284) was 66.2%.

3.6.6 Logistic Regression of Word and Student Characteristics

To comprehensively investigate the characteristics of words and students which affected the probability of a post-test question being answered correctly, we did a logistic regression¹³ at the post-test question level. We recognize that post-test results from a given student are correlated and we account for this by including the MTELP score of the relevant student in every line of data. MTELP scores have previously been shown to have a positive correlation with post-test success. We investigated the variables described in table 3.6.6.1.

¹³ We performed the regression using R.

Table 3.6.6.1: Definition of Variables

Word Features	
clust3or4	Number of clusters three of four characters long
letterCount	Number of letters in the base form of the word
syllables	Number of syllables in the base form of the word
vClust	Number of vowel clusters
POsv	The answer to the cloze question was a verb
POSn	The answer to the cloze question was a noun

Student Features	
alphabetic1	The L1 of the student is alphabetic
WSD	The student was in the WSD condition
mtelpB4	The student had an MTELP score in the that rounded to 40
mtelpB5	The student had an MTELP score in the that rounded to 50
mtelpB6	The student had an MTELP score in the that rounded to 60
mtelpB7	The student had an MTELP score in the that rounded to 70
mtelpB8	The student had an MTELP score in the that rounded to 80
mtelpB9	The student had an MTELP score in the that rounded to 90

Usage Features	
viewings	Number of documents read during the semester that contained the sense of the word being tested.
listened	The student listed to the word

For all 1786 post-test responses each additional practice instance had a positive effect on the log odds that a student would correctly answer the post-test question on that word. Clusters of three or four consonants and cloze questions eliciting a verb had a negative effect on the log odds of success. High proficiency students with high MTELP scores had had higher log odds of success while low proficiency students had lower log odds. The coefficients (estimates) and significance values can be seen in table 3.6.6.2.

Table 3.6.6.2: Universal group data

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	0.82418	0.46816	1.76	0.07833	+
viewings	0.14041	0.04974	2.823	0.00476	**
vClust	-0.17281	0.10566	-1.635	0.10194	
clust3or4	-0.53887	0.16845	-3.199	0.00138	**
letterCount	0.02307	0.05164	0.447	0.655	
syllables	0.08587	0.12086	0.711	0.47736	
WSD	-0.14236	0.1136	-1.253	0.21012	
listened	-0.06726	0.20633	-0.326	0.74443	
mtelpB4	-1.06823	0.40419	-2.643	0.00822	**
mtelpB5	-0.24671	0.43687	-0.565	0.57226	
mtelpB6	-0.01601	0.39978	-0.04	0.96806	
mtelpB7	0.37187	0.39702	0.937	0.34894	
mtelpB8	0.4333	0.4483	0.967	0.33378	
mtelpB9	2.83706	1.09082	2.601	0.0093	**
mtelpBnull	0.23224	0.4175	0.556	0.57803	
POSn	-0.08073	0.16954	-0.476	0.63398	
POSV	-0.81471	0.17591	-4.631	3.63E-06	***

For the 456 responses from nonalphabetic L1 (Chinese, Japanese) students, long consonant clusters again have a negative effect. Prior proficiency, represented by MTELP scores has the predicted effect, shown by the higher estimates for higher MTELP scores.

Table 3.6.6.3: Nonalphabetic data

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	0.7519	0.5639	1.333	0.182439	
viewings	-0.175	0.101	-1.733	0.083104	+
vClust	-0.3995	0.2071	-1.929	0.053718	+
clust3or4	-0.9301	0.2663	-3.493	0.000478	***
WSD	0.4796	0.2729	1.757	0.078899	+
listened	-0.1422	0.4305	-0.33	0.74122	
mtelpB6	1.3886	0.4079	3.404	0.000665	***
mtelpB7	1.3386	0.4058	3.299	0.000971	***
mtelpB8	1.9249	0.608	3.166	0.001545	**
POSn	-0.5472	0.3962	-1.381	0.167278	
POSV	-1.6746	0.4072	-4.112	3.92E-05	***

Data from the Arabic students (266 lines) does not fit the model of the universal group. The number of viewings and POSV do not have a significant effect on the log odds of success. Also different from the universal group's model is the fact that having word sense disambiguation (WSD) had a significant positive effect on post-test success.

Table 3.6.6.4: Arabic data

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	-0.69377	0.67161	-1.033	0.301609	
viewings	0.12473	0.10646	1.172	0.24133	
vClust	-0.34381	0.25705	-1.338	0.181043	
clust3or4	-0.75486	0.43947	-1.718	0.085853	+
letterCount	0.09527	0.1277	0.746	0.455642	
syllables	0.10929	0.29035	0.376	0.706606	
WSD	0.93928	0.27488	3.417	0.000633	***
listened	0.0483	0.45155	0.107	0.914817	
POSn	0.07154	0.4101	0.174	0.861511	
POSV	-0.24358	0.42172	-0.578	0.563537	

Koreans formed the largest group and so it is not surprising that the model based on the 684 lines of Korean data is similar to that of the universal group. Again the number of viewings had a positive effect and POSv had a negative effect.

Table 3.6.6.5: Korean data

	Estimate	Std. Error	z value	Pr(> z)	
(Intercept)	0.72061	0.39018	1.847	0.06477	+
viewings	0.19443	0.08372	2.322	0.02021	*
WSD	-0.36575	0.19721	-1.855	0.06364	+
mtelpB4	-1.10193	0.43142	-2.554	0.01064	*
mtelpB5	-0.25425	0.44103	-0.576	0.56429	
mtelpB6	-0.29906	0.43019	-0.695	0.48694	
mtelpB7	0.47943	0.39615	1.21	0.2262	
mtelpB8	0.57578	0.46182	1.247	0.21249	
mtelpB9	3.19318	1.09419	2.918	0.00352	**
POSV	-0.72761	0.18526	-3.928	8.58E-05	***

No model showed a significant effect of listening or the availability of audio on the log odds of success. As noted, few students used the audio feature frequently.

The effect of POSv was quite strong so we report the post-test accuracy based on the part of speech elicited. Post-test performance of the universal¹⁴ group was significantly¹⁵ affected by the part of speech to be supplied in the cloze questions. Accuracy was highest when the correct answer was an adjective and lowest when it was a verb.

¹⁴ This group includes all students, even those whose L1 is unknown.

¹⁵ Chi-square is 62.831 with 2 degrees of freedom and $p < 0.0005$.

Table 3.6.6.6 Part of Speech

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

4 Discussion

Many of the results of this study are suggestive, but not statistically significant. This may be due in part to data sparseness in some cases, such as the shortage of nonalphabetic students. Also, because this was a classroom study over the course of several weeks, it is less rigorously controlled than the lab studies reported on in section 2. For example students missed class occasionally and were able to ask the teacher questions that were not logged by the system. Talking to a teacher could easily yield the pronunciation of a word as an alternative to using the audio feature.

The results support theories that relate L2 reading processes to L1 scripts. The alphabetic L1 groups, whose L1s have high phonological recoverability, do appear to rely more heavily on phonological processing when reading in English than the nonalphabetic groups. This difference is evidenced 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.

It would be interesting to test the transfer of words learned through reading to words used when speaking. Because REAP tests semantic knowledge of new vocabulary, not phonological knowledge, the current study does not directly investigate differences in the reliance on orthographic processing as compared to phonological processing. Would the visual processing Chinese be negatively affected by consonant clusters, which violate the phonotactic constraints of their L1, when they had to speak them rather than write them? If that were the case, it would motivate practice exercises in REAP that engage phonological processing, even for students who are able to learn the written forms of word primarily through visual processing.

Tests of long-term retention as related to listening and cluster difficulty would also be interesting. The analysis of pre-tested word in section 2.1 found that students were less likely to know longer words and words with more clusters, which could be due to less long-term retention of these words. This is an interesting hypothesis that REAP could easily test.

We have not done a detailed analysis of guessing in REAP, but these results suggest that such an analysis would be worthwhile. The fact that the Arabic, Korean and Japanese groups actually performed better on hard cluster words than easy cluster words

when they were given post-test questions for words they had not practiced suggests a bias towards guessing hard cluster words. If this is the case, it might mean that the accuracy on practiced hard cluster words is also inflated through guessing and that the differential between knowledge of hard and easy cluster words is actually larger than the post-test performance suggests. This could be tested through a study of existing REAP data. If there is a relationship between guessing and choosing hard cluster words it could affect results of transfer studies as well because we would have less confidence that a correct post-test response for a hard cluster word indicated knowledge of that word. It is a challenge to design new question types that can be automatically generated while accurately measuring knowledge. More active question types, which require written production on the part of the student, are a better gauge of knowledge, but they are also more difficult to automatically evaluate.

An important result of this study was the effect of part-of-speech on learning success. Cloze questions eliciting verb had an overall success rate of only 55% while questions on nouns and adjective had over 72% accuracy. This could easily be factored in to the student knowledge model so that students are given more practice on verbs, either through more viewings or more exercises. Particularly long words or words with many or long clusters should also be given higher priority and more practice in the models of the alphabetic students.

In this study the variability in the usage of the audio feature brought to light the difference between treatment and intent to treat. As a result, it was impractical to evaluate the audio feature by simply comparing words where audio was available to words where it was unavailable. Automatically playing audio when students access dictionary definitions is one possible solution. However, given that students were sometimes reluctant to wear the headphones, as was necessary in the language lab, foisting the new feature on students does not seem likely to improve learning. A better solution might be to design questions where phonological knowledge and word recognition are practiced and tested, giving students increased motivation to use the audio feature. Another sound related-improvement might be to have a multimodal pre-test where students self-assess their knowledge of words as they both read and listen to them. It is possible that past students have sometimes been familiar with the sound of a word, but not its written form. While they may still benefit from practice on this word, such instances could compromise the validity of the pre-test.

The disappointing lack of evidence for learning benefits of listening could be due to several things. It could be that listening is not beneficial or that with an average of 12 opportunities to listen per student, the study is too small to test the feature. Additionally, the words used this semester are shorter and less complex, on average, than those used in other semesters, due to the requirements imposed by the WSD study. The fact that students had requested audio and that a few students used the feature frequently suggests that research on the inclusion of audio should not be abandoned.

The surprisingly strong correlation (shown in graph 3.6.3.2) between first hearing lower quality synthesis and the chance that a student would go on to use the audio feature again also requires further research. Ideally the use of synthesized audio pronunciation would be compared to the use of recordings of natural speech. A shortcoming of the study described here is the lack of exit survey about the audio feature, which may have explained the unexpected negative correlation between quality and usage. It may be that

low quality synthesis made students curious. Nevertheless, the willingness of some students to listen, even to lower quality synthesis, indicates that they are willing to accept the technology. This is encouraging for the future of TTS in CALL.

The results of this study make several recommendations for enrichment of the student knowledge model in REAP, as well as suggestions for the design of future studies on synthesis in REAP. It shows that differences in phonology and L1 writing system are manifest in different challenges in ESL vocabulary acquisition. Students with different language backgrounds are differently affected by word characteristics. As a result of different word-processing skills, they also differ significantly in their use of the audio feature.

Acknowledgements

Michael Heilman, Maxine Eskenazi, Anagha Kulkarni, Le Zhao and Jamie Callan were all involved in the version of REAP used in the summer 2007 study. Michael was very helpful to all stages of the project. Anagha ran the simultaneous WSD study and helped design this study.

Bibliography

- Abu-Rabia, S. (1997a). Reading in Arabic orthography: The effect of vowels and context on reading accuracy of poor and skilled native Arabic readers in reading paragraphs, sentences, and isolated words. *Journal of Psycholinguistic Research*, 26, 465-482.
- Abu-Rabia, S. (1997b). Reading in Arabic orthography: The effect of vowels and context on reading accuracy of poor and skilled native Arabic readers. *Reading and Writing* 9, 65-78.
- Abu-Rabia, S. (1999). The effect of Arabic vowels on the reading comprehension of second- and sixth-grade native Arab children. *Journal of Psycholinguistic Research*, 28, 93-101.
- Akamatsu, N. (2003). The Effects of First Language Orthographic Features on Second Language Reading in Text. *Language Learning*, 53:2, 207-231.
- Black, A. & Lenzo, K. (site creators). *Unit Selection Databases*. Retrieved August 15, 2007, website; <http://festvox.org/bsv/c2641.html>.
- Chang, J. (2001). Chinese Speakers. In Swan, M. & Smith, B. (eds.), *Learner English: a teacher's guide to interference and other problems*. Cambridge, UK: Cambridge University Press.
- Collins-Thompson, K. & Callan, J. (2004). A language modeling approach to predicting reading difficulty. In *Proceedings of the HLT/NAACL 2004 Conference*. Boston, MA.
- Coxhead, A. (2000). A New Academic Word List. *TESOL Quarterly*, 34:2, 213-238.
- Dupoux, E., Kakehi, K., Hirose, Y., Pallier, C., & Mehler, J. (1999). Epenthetic vowels in Japanese: A perceptual illusion? *Journal of Experimental Psychology: Human Perception and Performance*, 25, 6, 1568-1578.
- Egan, B. K., & LaRocca, S. A. (2000). Speech recognition in language learning: A must. In *proceedings of InSTIL 2000* (pp. 4-9). Dundee, England: University of Abertay Dundee.
- Fender, M. (2003). English word recognition and word integration skills of native Arabic- and Japanese-speaking learners of English as a second language. *Applied Psycholinguistics*, 24, 289-315.

- Handley, Z. & Hamel, M. (2005). Establishing a Methodology for Benchmarking Speech Synthesis for Computer-Assisted Language Learning (CALL). *Language Learning & Technology*, 9:3, 99-119.
- Hawkins, R. (2001). *Second Language Syntax: a Generative Approach*. Malden, Massachusetts: Blackwell Publishers.
- Heilman, M., Juffs, A., & Eskenazi, M. (To Appear). "Choosing Reading Passages for Vocabulary Learning by Topic to Increase Intrinsic Motivation." *Proceedings of the 13th International Conference on Artificial Intelligence in Education*. Marina del Rey, CA. (poster).
- Heilman, M., Collins-Thompson, K., Callan, J., and Eskenazi, M. (2007). "Combining Lexical and Grammatical Features to Improve Readability Measures for First and Second Language Texts." *Proceedings of the Human Language Technology Conference*. Rochester, NY.
- Heilman, M. & Eskenazi, M. (2006). "Authentic, Individualized Practice for English as a Second Language Vocabulary." Presented at Interfaces of Intelligent Computer-Assisted Language Learning Workshop at the Ohio State University, Columbus, OH, 12/06. Unpublished.
- Heilman, M., Collins-Thompson, K., Callan, J., & Eskenazi, M. (2006). "Classroom success of an Intelligent Tutoring System for lexical practice and reading comprehension." *Proceedings of the Ninth International Conference on Spoken Language Processing*. Pittsburgh, U.S.A.
- Horst, M., Cobb, T. & Meara, P. (1998). Beyond Clockwork Orange: Acquiring second Language Vocabulary through Reading. *Reading in a Foreign Language*, 11:2, 207 – 223.
- Kabac, B. (2003). *The Perceptual Processing of Second Language Consonant Clusters*. Dissertation. University of Delaware. Retrieved June 10, 2007 website: <http://ling.uni-konstanz.de/pages/home/kabak/KabakPhDdissertation.pdf>
- Koda, K. (1988). Cognitive process in second language reading: Transfer of L1 reading skills and strategies. *Second Language Research*, 4, 133-156.
- Koda, K. (1990). The use of L1 reading strategies in L2 reading: Effects of L1 orthographic structures on L2 phonological recoding strategies. *Studies in Second Language Acquisition*, 12, 398-410.
- Koda, K. (1998). The role of phonemic awareness in second language reading. *Second Language Research*, 14, 194-215.

- Koda, K. (1999). Development of L2 Intraword Orthographic Sensitivity and Decoding Skills. *Modern Language Journal*, 83, 51-64.
- Lee, J. (2001). Korean Speakers. In Swan, M. & Smith, B. (eds.), *Learner English: a teacher's guide to interference and other problems*. Cambridge, UK: Cambridge University Press.
- Muljani, D., Koda, K. & Moates, D. (1998). The development of word recognition in a second language. *Applied Psycholinguistics*, 19, 99-113.
- Oxford-Hachette. (2003). *Oxford-Hachet French Dictionary on CD-ROM* (Verions 2.0). Oxford, England: Oxford University Press.
- Pelletreau, Timothy R. (2006). *Computer-assisted vocabulary acquisition in the ESL classroom*. Unpublished master's thesis. University of Pittsburgh, Pittsburgh, PA.
- Perfetti, C. A. & Liu, Y. (2005) Orthography to phonology and meaning: Comparisons across and within writing systems*. *Reading and Writing*, 18, 193-210.
- Smith, B. (2001). Arabic Speakers. In Swan, M. & Smith, B. (eds.), *Learner English: a teacher's guide to interference and other problems*. Cambridge, UK: Cambridge University Press.
- Sobkowiack, W. (1998). Speech in EFL CALL. In K. Cameron (Ed.), *Multimedia CALL: Theory and practice* (pp. 23-34). Exeter, England: Elm Bank.
- Tan, L.H., Sprinks, J. A., Eden, G. F., Perfetti, C. A. & Siok, W. T. (2005). Reading depends on writing, in Chinese. Retrieved June 10, 2007 from *Proceedings of the National Academy of Science*, web site:
<http://www.pnas.org/cgi/reprint/102/24/8781.pdf?ck=nck>
- Taylor, I. & Taylor, M. M. (1995). *Writing and literacy in Chinese, Korean and Japanese*, Philadelphia, PA: John Benjamins.
- Thompson, I. (2001). Japanese Speakers. In Swan, M. & Smith, B. (eds.), *Learner English: a teacher's guide to interference and other problems*. Cambridge, UK: Cambridge University Press.
- Tomoda, T. (2005). Change in Script usage in Japanese. Retrieved June 10, 2007 from *Electronic Journal of Contemporary Japanese Studies*, website:
<http://www.japanesestudies.org.uk/articles/2005/Tomoda.html>
- Treiman, R., Mullennix, J., Bijeljac-Babic, R., & Richmond-Welty, E.D. (1995). The special role of rimes in the description, use, and acquisition of English orthography. *Journal of Experimental Psychology: General*, 124, 107-136.

- Wade-Woolley, L. (1999). First Language Influences on Second Language Word Reading: All Roads Lead to Rome. *Language Learning*, 49:3, 447-471.
- Wang, M., Koda, K., & Perfetti, C.A. (2003). Alphabetic and nonalphabetic L1 effects in English word identification: a comparison of Korean and Chinese English L2 learners. *Cognition*, 87, 129-149.
- Yoon, H., Bolger, D. J., Kwon, O. & Perfetti, C. A. (2002). Subsyllabic units in Reading. In Ludo, V., Elbro, C. & Reitsma, P (eds.), *Precursors of Functional Literacy*, 130-163, Philadelphia, PA: John Benjamins.

Appendix A: Focus words and included morphological variants

appreciate appreciates appreciated appreciating
bond bonded bonding bonds
brief brevity briefed briefing briefly briefs
channel channeled channeling channels
code coded codes coding
complex complexities complexity
conceive conceives conceived conceiving
factor factored factoring factors
foundation foundations
issue issues
major majors
manual manuals
monitor monitored monitoring monitors
panel panelled panelling panels paneling paneled
principal principals principally
procedure procedures procedural
qualify qualifies qualified qualifying
suspend suspended suspending suspends
tape taped tapes taping
transmission transmissions
volume volumes
aid aided aiding aids
contract contracted contracting contracts
depress depressed depresses depressing depression
function functions
parallel paralleled parallels paralleling
pose posed poses posing
shift shifted shifting shifts
supplement supplementary supplemented supplementing supplements
trigger triggered triggering triggers

Appendix B: TTS Evaluation

Native speaker feedback, ranked by intelligibility:

word	Position in native speaker evaluation	Native speakers who correctly identified (of 7 total)	intelligibility	naturalness
panel	20	4	1.571	1.429
contract	10	5	2.000	1.333
trigger	22	3	2.286	2.143
volume	14	7	2.571	2.143
appreciate	27	7	2.714	2.286
channel	30	6	2.714	2.143
depress	13	6	2.857	1.571
manual	23	7	2.857	1.857
brief	1	3	3.000	2.143
issue	3	6	3.000	2.571
major	11	7	3.000	2.286
procedure	7	7	3.000	2.143
tape	6	6	3.000	2.571
bond	16	7	3.143	2.429
suspend	15	7	3.143	2.571
aid	18	6	3.286	2.857
complex	19	4	3.286	2.286
pose	12	5	3.286	2.857
function	9	7	3.429	2.143
code	24	6	3.714	3.429
parallel	4	7	3.714	2.714
conceive	2	6	3.857	2.571
monitor	8	7	4.000	2.571
factor	28	7	4.143	2.571
principle	26	7	4.143	2.429
qualify	5	7	4.143	2.857
supplement	29	7	4.143	3.143
foundation	17	7	4.286	3.286
transmission	25	7	4.286	3.000
shift	21	7	4.571	3.571