January 2002, Vol. 6, Num. 1
pp. 147-164

PROVIDING CONTROLLED EXPOSURE TO TARGET VOCABULARY THROUGH THE SCREENING AND ARRANGING OF TEXTS

Sina Ghadirian
McGill University, Montreal

ABSTRACT

This article considers the problem of how to bring foreign language students with a limited vocabulary knowledge, consisting mainly of high-frequency words, to the point where they are able to adequately comprehend authentic texts in a target domain or genre. It proposes bridging the vocabulary gap by first determining which word families account for 95% of the target domain's running words, and then having students learn these word families by reading texts in an order that allows for the incremental introduction of target vocabulary. This is made possible by a recently developed computer program that sorts through a collection of texts and a) finds texts with a suitably high proportion of target words, b) ensures that over the course of these texts, most or all target words are encountered five or more times, and c) creates an order for reading these texts, such that each new text contains a reasonably small number of new target words and a maximum number of familiar words. A computer-based study, involving the sorting of 293 *Voice of America* news texts, resulted in the finding that a) the introduction of new target vocabulary in each text could be kept to a reasonably small amount for the majority of texts, and b) the number of target vocabulary items occurring fewer than five times could be kept to a minimum when the list of target vocabulary accounted for 96% of the domain's running words, rather than 95%.

THE PROBLEM: L1 VERSUS L2 VOCABULARY ACQUISITION

There is considerable evidence that L1 learners acquire a large amount of their vocabulary through guessing from context (Nagy & Herman, 1987; Sternberg, 1987). The frequency at which the L1 learner encounters words, and the variety of contexts in which words are encountered, ensure that the learner will eventually come across most new words in a context where the word is guessable. Research suggests, however, that foreign language students do not undergo the same rich and varied exposure to vocabulary (Singleton, 1999). As a result, although EFL elementary-level students quickly learn many of the high-frequency words that occur in teaching materials, they experience a breakdown in their ability to guess from context when faced with the much lower frequency words found in unsimplified texts. This is because the low-frequency words found in unsimplified texts make up too large a proportion of those texts. In other words, since there are not enough familiar words in the text for the learner to use as clues, guessing unfamiliar words from context becomes extremely difficult or impossible.

The problem, then, is how to expand a student's vocabulary knowledge to the point where he or she recognizes enough of the words in unsimplified texts to be able to guess unfamiliar words from context. Put another way: what is needed is a strategy for bridging the gap between a knowledge of the kinds of high-frequency words found in elementary texts, and a knowledge of the words necessary for the student to be able to resume incidental vocabulary learning. The problem can be broken into two parts: a) Which words are needed in order to bridge this gap? b) Which methods should be used to teach these words quickly and effectively?

Which Words

Carroll, Davies, and Richman (1971) pointed out, nearly three decades ago, that about 80% of the running words (tokens) in any English text are accounted for by the 2,000 most frequent word families[1] of English. Nation (1990) has drawn to our attention the importance of knowing these word families to reading comprehension. A reader who is familiar with 80% of the tokens in a text, however, is still not able to adequately comprehend the text. Studies by Liu & Nation (1985) and Laufer (1989) point toward 95% as the amount of coverage required in order for a reader to adequately understand a text and guess new words from context. Finding a reasonably-sized vocabulary list that accounts for 95% of the tokens of all unsimplified texts, however, has proven difficult. Instead, it may be more feasible to focus on moving the student from elementary-level texts to texts in a specific domain or genre.

How to Get 95 % Coverage in Academic Texts

Researchers interested in vocabulary acquisition by students enrolled in ESP (English for Specific Purposes) courses point out that just over 90% of the running words in academic texts can be accounted for by two word-lists, West's General Service List (GSL; 1953) -- which includes the 2,000 most frequent word families of English -- and Xue & Nation's University Word List (UWL; 1984) -- which is made up of words frequently found in academic texts (Nation & Hwang, 1995).[2] In addition, academic texts contain a number of word families specific to the academic domain that is the subject of the text (Sutarsyah, Nation, & Kennedy, 1994). In one study, researchers working with an economics textbook found that word families from the GSL and UWL accounted for over 91% of tokens in the text, and estimated the number of domain-specific word families at 460 (Sutarsyah et al., 1994). Unpublished research by the author suggests that domain-specific word families (defined by their greater frequency of occurrence in a narrow range of texts circumscribed by the domain) account for more than 4% of academic economics texts' tokens, thus bringing the total to 95%.[3] If we assume that this figure holds true for other academic domains, we can conclude that for academic texts it *is* possible to come up with a reasonably-sized combination of word lists (GSL at 2,300 word families + UWL at 800 word families + economics domain list at 460 word families) that accounts for 95% coverage of the text. Knowing these word families should allow learners to comprehend the texts and attempt to guess the remaining 5% of tokens from context.

Which Method

Once a word list or combination of word lists accounting for 95% of tokens in the target domain has been found, the next question to consider is which method is best suited to acquainting students with the word families on this list quickly and effectively. Some interesting solutions to this problem have been suggested by researchers interested in the problem of ESP vocabulary acquisition. At issue for these researchers is how to integrate the speed of explicit instruction with the traditional benefits of reading-based vocabulary acquisition. In response to this problem, a number of instructional strategies have been devised which attempt to teach target vocabulary items quickly, while ensuring that each item is supplied with some form of meaningful context.

One strategy that has shown much promise over the last few years is vocabulary instruction via computer-based concordancing. First, a computer-based corpus is created by scanning texts in the students' target domain into a computer. Subsequently, any word that exists in the corpus can be viewed by the student surrounded by its immediate context (or *contexts*, as there are usually multiple instances of the word in the corpus). Cobb and Horst (1999) argue that a concordance-based tutor has three advantages over incidental reading-based and traditional word list learning strategies: a) computer concordancing conserves the efficiency of list targeting while allowing for exposure to the new word in multiple contexts, b) it allows for a way to ensure that each word is encountered a minimum of five times, and c) the learner can choose among the example sentences generated by the concordancer for one that makes sense to him or her (Cobb & Horst, 2001). Note that, relevant to the second argument, a study by Saragi, Nation, & Meister (1978) has shown that a word needs to be encountered at least five times in order to be well retained.[4]

Other computer-based lexical tutors have been drawing attention in recent years. Of note is a tutor developed by Peter Groot (2000) named CAVOCA (Computer Assisted VOCabulary Acquisition). CAVOCA is designed to operationalize current theories about how lexical storage works. Hence, students using CAVOCA are introduced to a word by having to guess the word from context, think about correct versus incorrect usage of the word, read the word in the context of example sentences, and finally produce the word in a CLOZE exercise. According to Groot, this kind of rigorous involvement with the word should encourage deeper processing and longer-term retention than traditional learning strategies like bilingual word list memorization.

PROVIDING CONTROLLED EXPOSURE TO TARGET VOCABULARY THROUGH THE SCREENING AND ARRANGING OF TEXTS

The strategies mentioned above offer alternatives to reading-based incidental vocabulary learning, which, as both researchers point out, is not necessarily best-suited to ESP purposes. The three major complaints about reading-based vocabulary acquisition are that a) it is an inefficient strategy for learning target words (readers must wade through many other words, in haphazard fashion, before they come across a target word), b) even if a target word is encountered during reading, there is no guarantee it will be encountered five or more times, and c) even if a and b were not problems, the high proportion of unfamiliar words in unsimplified texts ensures that for L2 learners with a limited vocabulary of high-frequency words, guessing new target words from context is difficult or impossible.

If these three problems could somehow be resolved, however, there may be good reason for encouraging reading-based vocabulary acquisition over non-reading-based strategies. Krashen (1989) has argued vigorously that extensive reading is the only strategy that provides the learner with complete and non-superficial knowledge of a word. The pleasure that many learners experience when reading a whole text is also an important factor to consider, since, ideally, it creates the motivation to read more (and hence, learn more words). Finally, important reading skills are exercised during the reading of whole texts that are not exercised during the reading of example sentences (making predictions, recognizing genre, etc.). Developing these skills may be crucial to further reading (and again, further vocabulary learning).

My contention is that the three problems mentioned above -- a) the difficulty of finding texts with a high proportion of target words, b) the difficulty of knowing whether a reader has encountered a word five or more times, and c) the breakdown in learning new words that occurs because students do not recognize enough familiar words in the text -- can be resolved. They can be resolved by carefully selecting, screening, and arranging texts by means of a recently developed computer program (Ghadirian, 2000). Specifically, this program a) finds texts with a suitably high proportion of target words, b) ensures that over the course of these texts most or all target words are encountered five or more times, and c) creates an order for reading these texts such that each new text contains a minimum number of new words and a maximum number of familiar words.

Such a program is designed, in effect, to improve the reading-based acquisition of target vocabulary by applying careful control over what words are contained in the texts, how often they occur throughout the texts, and how many are introduced to the reader in each text, that is, to create the conditions for what we might call "controlled" or "optimized" reading-based vocabulary acquisition. (Note that this article does not attempt to prove that using the program *does* result in measurable vocabulary acquisition. Rather, it presents the program's strategy for providing controlled exposure to target vocabulary as a plausible means of bringing this acquisition about.)

Of course, this is not the first attempt at regulating texts' vocabulary content for the purpose of optimizing vocabulary acquisition and reading comprehension. Graded reading schemes, which attempt to move learners through a sequence of vocabulary levels by having them read texts suited to the levels, have been around for quite a while. However, there are important differences between learning vocabulary from graded readers and learning vocabulary from texts arranged by the program mentioned above. These differences involve the amount of new vocabulary required to bridge the gap between texts at consecutive levels, as I discuss elsewhere in the article.[5]

Providing Controlled Exposure to Target Vocabulary: The Process of Selecting and Arranging Texts

So far, I have talked about a program that provides controlled exposure to target words by supplying the reader with a sequence of texts in which a) target words make up a large proportion of each text, b) five or more instances of each target word occur throughout the texts, and c) the texts are ordered in such a way that each text contains a maximum number of familiar words and a minimum number of new target words. What I wish to do at this point, is explain the criteria used by TextLadder (the computer program briefly described above; Ghadirian, 2000) as it screens texts and sorts them into this sequence. To assist my explanation, I am going to draw on a hypothetical ESP teaching situation involving students with specific vocabulary needs.

Imagine a classroom located in a country where English is not the first language and made up of students who are beginning a master's degree program in Economics. All the students have received bachelor degrees in economics; however their experience with economics-related texts in English varies from limited exposure to no exposure whatsoever. They have been informed that over the course of their master's degree program texts written in English will be among the readings assigned. Their immediate need, then, is to be able to read English academic texts related to economics as quickly as possible. As mentioned previously, various studies suggest that students need to be familiar with 95% of the tokens in a text in order to adequately understand that text. As well, I have already said that researchers have been able to identify three word lists that, together, can provide 95% token coverage of academic texts related to economics. Therefore, from the point of view of the students' vocabulary needs, learning the word families from these three lists (the GSL, the UWL, and the economics specialty-words list) takes priority over learning any other kind of vocabulary.

The first step, then, before TextLadder even becomes involved, is for the ESP instructor to find a large number of relatively short texts (article or news story-sized would be ideal) that conceivably contain words from these lists in a high proportion. A good choice would probably be news stories related to economics or business, and academic articles suitable for the undergraduate level. Note that articles considered graduate-level reading are *not* a good choice. An important criterion of text selection at this stage is that if the texts were somehow translated into the students' L1 they would be comprehensible. As the students are at the *beginning* of their graduate program, this criterion is not met for graduate-level texts.

Once the texts have been collected by the ESP instructor, the next step is to scan these texts onto the computer. Once that is done, TextLadder takes over. TextLadder's first task is to check each text to see whether 95% of the tokens in the text are accounted for by the three word lists. (Optionally, proper nouns can be included in this 95%, but whether or not proper nouns qualify as "familiar vocabulary" is a complicated question that is discussed in more detail later) Texts that do not make the 95% cut-off are dropped. The criterion at work here is this: If the reader knows all the words from the three lists found in that text, he/she should be able to understand the text. As we shall see, the incremental nature of the program's sorting process ensures that the reader *is* familiar with the list words found in that text by the time he or she reaches it.

If the word list or combination of word lists really does consistently account for 95% of tokens in texts from the target domain, then only a small number of texts will in fact be eliminated. The texts that remain will be representative of the target domain: 95% of their tokens (or slightly less, if proper nouns are allowed) will be words found in the GSL, UWL, and economics word lists.

Once the elimination process is completed, Textladder moves on to the sorting process. The program sifts through the collection of texts and finds the text with the smallest number of unfamiliar words. It does this by comparing the words of each text with a pool of "familiar words." This familiar words pool consists, initially, of high-frequency words of the kind taught in elementary-level EFL texts. To be more precise, it consists of the first 176 words of the GSL (i.e., the 176 words most frequently encountered in English language texts) and an assortment of other words that are found in a broad range of elementary texts: basic numbers, basic colours, days of the week, and so forth.[6] Once the text with the fewest number of unfamiliar words is found, the text is placed first on the "sequence list" (the list which describes the order in which the texts should be read), and its unfamiliar words are added to the familiar words pool. (Note that, of course, only the unfamiliar words that are also GSL, UWL, and economics list words are added. Proper nouns and other words not on the lists are not added.) The program now repeats the process of looking through the texts, trying to find the text with the smallest number of unfamiliar words. Again, it compares the words of each text with the words of the (now slightly enlarged) familiar words pool. Once this text is found, it is placed *second* on the sequence list, and its words added to the familiar words pool. The program repeats this process over and over until no more unfamiliar list words remain in any of the texts.

Notice that there will always be a certain number of unfamiliar words for each new text. The number of unfamiliar words per text varies depending on a) the size of the text itself; b) the number of texts that were ultimately selected to undergo sorting (the greater the number of texts, the fewer unfamiliar words per text); and c) the text's position on the sequence list (in general, the number of unfamiliar words for the first few texts will be high). The unfamiliar list words found in each text are displayed prior to that text. Before reading the text, the student is asked to acquaint him/herself with these words through the use of a bilingual dictionary, preferably a computer-based one accessible either on CD-ROM or via the Internet. (A study by Hulstijn, Hollander, & Greidanus [1996] has highlighted the reluctance students often demonstrate toward using paper-based dictionaries in reading situations.) The rationale for this kind of pre-reading dictionary activity is that the resulting superficial knowledge of the word attended to will be reinforced during the actual reading of the text and further reinforced by encounters with the word in new contexts over the course of succeeding texts.

Notice too that, over the course of the preceding paragraphs, I have not explained how the five-encounters-per-word criterion is satisfied. Let me now elaborate. TextLadder keeps track of the number of times a word is encountered over the course of all the texts with which it is dealing. If, by the end of the whole sorting process, there remain words that were encountered fewer than five times, TextLadder informs the instructor of these words, as well as the number of times each of these words was encountered. The instructor then has a number of options concerning how to make further encounters with the words possible (e.g., direct teaching or the construction of customized texts). In the following section, a fuller discussion is provided on how the number of target words occurring fewer than five times can be kept to a minimum.

A third note, in the previous discussion I said that TextLadder judges a word in the text to be "familiar" if it matches a word in the familiar words pool. The meaning of "matches" needs to be clarified. I do not mean to say that the word in the text has to be an exact replica of the word in the pool. The word in the text can vary in certain allowable ways: it can be pluralized, have a different verb tense, or be a derived form of the same base word. In short, the general criterion for a "match" is that the two words belong to the same word family (i.e., Level 6 of Bauer & Nation [1993], with certain differences).[7] In some cases, TextLadder comes across a word that is apparently derived from a familiar base word but in fact has a completely different meaning (e.g., "homely"). In such cases, TextLadder does not consider the two words a match. (TextLadder is not be able to catch all these cases but it should be able to catch most of them.)

There is one more note to consider before we wrap up our example. TextLadder does not ensure that all the words on all three lists (GSL, UWL, and economics word list) are encountered. Rather, it ensures that all the words from these lists that occur in *all the texts being sorted through* are encountered. If the number of texts scanned onto the computer is large enough, then conceivably all the words in all the lists would be encountered. However, this is not a prerequisite, a fact that allows the instructor to use a word list, or combination of word lists, which is larger than is strictly necessary.

At the end, the instructor has a list of texts throughout which all the words from the three lists necessary for comprehending the *complete* set of texts (including those not on the sequence list) are encountered. The majority of them have been encountered five or more times. The texts on the list are ordered in such a way that reading and comprehending each new text should not be a struggle for the student: The number of words that have to be learned (via the pre-reading dictionary exercise) in order to understand each text is at a minimum, so that the student is not overwhelmed by these words during the reading. The number of texts is substantial (probably somewhere between 150 and 300 news story-length texts), and the instructor is confronted with some difficult decisions concerning pace. (I estimate that the students must read one text a day over the course of an average two-semester course, or two a day for a one semester course, in order to finish. Whether reliable learning can actually take place at this pace is a separate question that is not dealt with in this paper.) The heavy workload is hopefully offset for the students in the class by the satisfaction derived from understanding each text and seeing recently encountered words appear in subsequent texts.

FIVE POTENTIAL PROBLEMS

Grammar

An important issue to consider once the sequence list has been produced is how to guarantee that the student has the necessary grammatical knowledge to tackle each text as s/he comes to it. This is not an easy problem to solve. There is, of course, no guarantee that the texts, sequenced for incremental vocabulary acquisition, will also be sequenced for incremental introduction of grammatical usage. In many cases, students will come across structures in the text (e.g., the present perfect continuous tense, the passive voice, etc.) before they learn them in the classroom. A partial answer to this problem is that the instructor only focus on the structures and tenses that most commonly occur in texts from the target domain/genre. An important article by Flowerdew (1993) describes how computer concordancing can be used to focus on which grammatical structures, notional areas, and discursive formations are most used in a given text or group of texts. This may allow the instructor to streamline his/her instruction so that the material most relevant to immediately comprehending the texts can be taught earlier and later expanded on over the course of subsequent classes.

Multi-Word Units

A second problem concerns multi-word units. When TextLadder comes across the phrasal verb "blow up," it does not recognize the entire unit but only the individual words of which it is made. In fact, not only is TextLadder unable to recognize phrasal verbs, it is unable to recognize any multi-word unit, be it a compound noun (e.g., "home run") or an idiomatic expression (e.g., "make a run for it"). This is obviously problematic since students may understand the individual words making up a multi-word unit without comprehending the unit itself. Ideally, TextLadder should recognize these units, include them in the pre-reading activity, and factor their presence into the overall comprehensibility of the text. It is conceivable that TextLadder could be modified in a future version to allow for this. In the meantime, concordancing, in the context described by Flowerdew (1993) could be useful in determining which phrasal verbs and other multi-word units are most common in the texts.

Proper Nouns

Proper nouns present us with another problem. As mentioned, various studies indicate that a student must be familiar with 95% of the tokens in a text in order to be able to adequately understand the text. But do proper nouns qualify as "familiar"? Hirsh and Nation (1992) have discussed this problem in the context of a specific kind of text: the simplified novel. They present us with two arguments for why proper nouns in simplified novels should be considered as words that do not require previous learning: a) the text reveals what we need to know about the proper nouns as the story progresses, and b) the initial capitalized letter of the proper noun informs the reader that it *is* a proper noun, which is already an important piece of information. As well, Hirsh and Nation pointed out that the number of proper nouns was small and their frequency of occurrence high in the simplified novels under study. This reasoning appears sound, and it is difficult to see why proper nouns in simplified novels that have a low number of high-frequency proper nouns (all of which adequately introduced by the text) should not qualify as part of the 95%. Still, the situation will be different for every domain and every genre of text. Some texts may assume a prior acquaintance with proper nouns that the reader does not in fact have, and they may contain a large number of these proper nouns scattered over the text, each one mentioned only once or twice. This is, of course, a worst-case scenario. Most news stories are careful to introduce proper nouns that are names of people (although place names can be more problematic). Ultimately, the choice of whether or not to include proper nouns in the 95% is left to the instructor, who must consider the kinds of texts being read and decide whether or not to select the "include proper nouns" option in TextLadder. Not selecting the option may mean spending excessive amounts of time locating texts and scanning them into the computer (since the number of texts that can pass the 95% test will obviously be much smaller if proper nouns are not included). Selecting the option risks the possibility that students may not be able to comprehend the texts (since, when the text's proper nouns are completely unfamiliar or unexplained, the percentage of tokens that the student is in fact familiar with will be below 95%). Hopefully, future studies will clarify the question of exactly how proper nouns factor into readers' comprehension of a text.

Low-Encounter Words

A fourth problem concerns the effect that "familiar" words that have been encountered very few times have on comprehension. To explain, I will draw on a hypothetical situation. A student comes across a word for the first time (we'll say the word is "bargain") in one of the texts. The word, of course, has been seen by the student in the pre-reading activity at the beginning of the text. However, over the course of the text, the word is encountered only once. Three or four texts later, the student comes across "bargain" once again. Because TextLadder no longer considers it to be an unfamiliar word, "bargain" does not appear in the pre-reading activity. However, as the student has only encountered the word once (twice, if we include the original dictionary activity), there is a good chance s/he does not remember it well, thus making comprehension of the text more difficult. The question, then, is how to factor the number of times familiar words have been encountered into the sorting process. In a version of TextLadder on which I am currently working, I have come up with a provisional answer. When TextLadder is deciding which text has the fewest unfamiliar words, it penalizes texts that have familiar words that have been encountered fewer than five times. A familiar word that has been encountered once will incur the greatest penalty, while one that has been encountered four times will incur the smallest penalty, and words that have been encountered five times or more incur no penalty at all. Because of this penalty system, a text that has eight unfamiliar words might be placed on the sequence list before a text that has seven unfamiliar words, if the latter text contained more low-encounter familiar words.[8]

Homographs

A final problem concerns homographs. Say, for example, the word *tear* shows up in a pre-reading dictionary activity. Which meaning should a student focus on? This problem is partly resolved by the fact that if the student is using a CD-ROM or Internet dictionary, s/he will be able to click on the word during the reading of the text and figure out from context which meaning is appropriate. However, there remains a problem. The next time a student encounters *tear* in a text, TextLadder will already consider the word to be a familiar word, even though this *tear* may bear a different meaning from the first *tear*. How do we get around this problem? One solution, which has not been implemented yet, is to have TextLadder run all the texts through a part-of-speech tagging program. This way, although TextLadder will not be able to differentiate between two homographs which are also the same parts-of-speech, it will be able to differentiate between homographs which are different parts-of-speech, which would effectively allow it to tell *tear* (as in "to rip") from *tear* (as in "what comes out of your eye when you cry").

PERFORMANCE ANALYSIS: TEXTLADDER AND *VOICE OF AMERICA* SIMPLIFIED NEWS TEXTS

In order to obtain a quantitative measure of TextLadder's ability to bridge the gap between a limited vocabulary of high-frequency words and an expanded vocabulary consisting of words from a target domain or genre, a genre was chosen and relevant texts amassed, so that TextLadder could be put through a test-run. The genre (or rather, sub-genre) chosen for the study was *Voice of America* simplified news texts. By "simplified news texts," I am referring to news stories that have been specifically written for an audience that consists of EFL students. Texts from this genre are appearing in increasing numbers on the Internet. Some examples of simplified news texts include *Voice of America* Special English news stories, *Key News Reader Newspaper* stories, KXTV abridged news stories, and CNN abridged news stories (the latter two are also directed at adult L1 students in literacy programs). My decision to work specifically with *Voice of America* simplified news texts was made for the following reasons: a) much of the vocabulary in *Voice of America* simplified news texts is drawn from a list of words called the "Special English" word list, a fact that makes finding the combination of word lists necessary for 95% coverage much easier, and b) *Voice of America* has a large EFL reader/listener base, which suggests that these are the kinds of texts EFL students find useful and of interest.

The goal of the *Voice of America* study was to see whether TextLadder would be able to ensure two things: 1) that each text selected for reading contained a reasonably small number of unfamiliar list words, if read in the appropriate order, and 2) that all, or most, of the list words encountered in the selected texts were encountered five times or more. The first step in the study was the compiling of a word list that accounted for 95% of the tokens in *Voice of America* simplified news texts. It was found that the Special English word list by itself did not provide 95% coverage. Therefore, the list was supplemented by words from the GSL, UWL, initial "familiar words pool" list, and the Longman Defining Dictionary (LDD) word list. (The LDD and Special English word lists are accessible in a list by Rick Harrison called *Vital English Vocabulary* [Harrison, 1997].) It was found that the new list, in combination with proper nouns, consistently accounted for 95% of the tokens in *Voice of America* texts.[9]

Next, 293 *Voice of America* simplified news texts were downloaded onto a computer. The size of the texts varied between 300 and 1,500 words. All parts of the text that were not part of the news story itself were edited out.

TextLadder was then run and 266 texts made the 95% cut-off and underwent the sorting process. At the end of this process, two pieces of output were produced: 1) a list of the names of the 253 texts selected (i.e. the "sequence list" mentioned in the last section), with corresponding figures describing the amount of new target vocabulary introduced in each text; and 2) a list of all the target vocabulary encountered over the course of the 253 selected texts, each word accompanied by a figure stating the number of times the word had been encountered.

1) Amount of Unfamiliar Target Vocabulary Introduced in Each Text

Figure 1 describes how the amount of new target vocabulary introduced per text changes over the course of the 253 texts placed on the sequence list. Note that "amount of unfamiliar target vocabulary per text" is synonymous with "percentage of unfamiliar list-word tokens in each text." The latter is determined by dividing the number of unfamiliar list-word tokens in the text by the total number of tokens in the text.

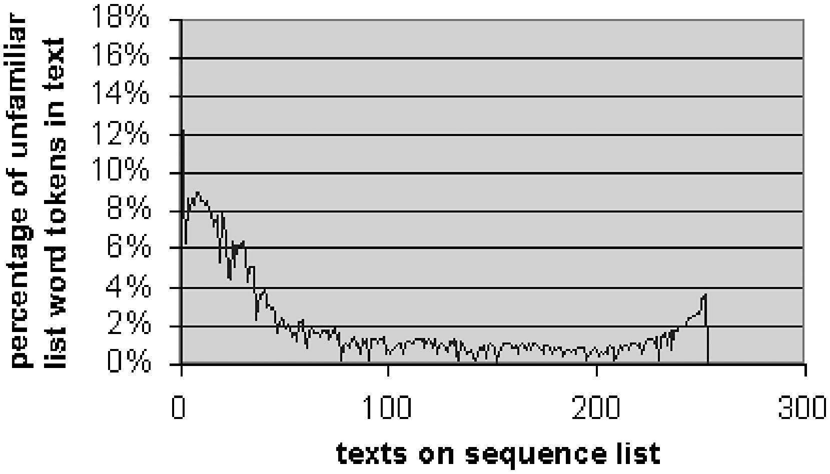

Figure 1. Amount of new target vocabulary introduced per text

As Figure 1 makes clear, the percentage of unfamiliar list words in each text is quite high at the beginning. However, after only 80 texts, it has already dipped below 1%. In the following 140 texts, unfamiliar list-words regularly make up less than 1% of the tokens in each text. It is only in the last 35 texts that the percentage begins rising back up toward 4%.

The question that arises from this result is whether students should invest time and effort in reading those 140 texts (over 55% of the total number of texts on the sequence list), when new words make up such a small percentage of those texts (see Appendix B for the actual number of words). If an accelerated pace of vocabulary learning is the aim, then the answer is probably no.

The problem lies with TextLadder's sorting process: For each slot in the sequence list, TextLadder looks through the texts and finds the one with the smallest number of unfamiliar words. Although this sorting strategy is desirable at the beginning when the lowest percentage of unfamiliar words is well over 10%, it becomes less desirable later in the sequence.

As a result of these findings, the TextLadder program was modified. The program was allowed to follow the old sorting strategy for the first few texts on the sequence list. However, when the lowest percentage of unfamiliar words dipped below 5%, the sorting strategy was reversed: TextLadder now began looking for texts with the *highest* number of unfamiliar list words. The results proceeding from the adjusted TextLadder program are shown in Figure 2.

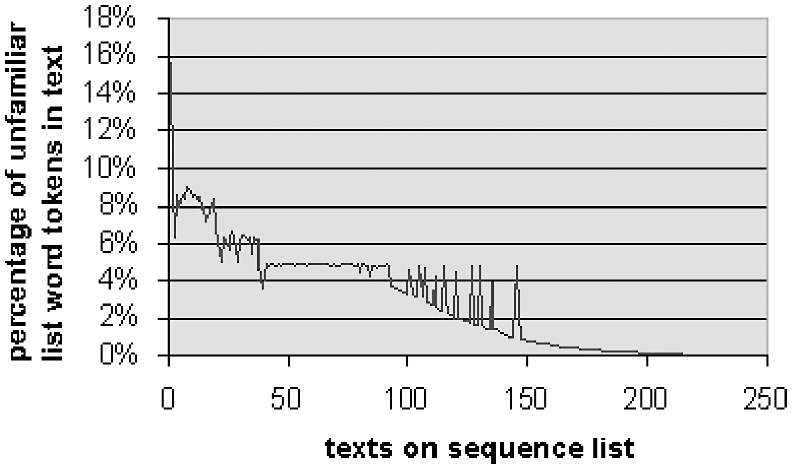

Figure 2. Amount of new target vocabulary introduced per text (Adjusted)

Note that rather than making a steep drop down to 1% within the first 80 texts, the percentage of unfamiliar words now decreases at a much slower rate, not dipping below 1% until the 144th text, and hitting zero after the 216th text. The number of texts in which new list words make up less than 1% of tokens is therefore only 73 texts (those between the 144th and 216th), significantly fewer than in the unadjusted version of TextLadder. Note also that all the list words in the 266 texts have been encountered after only 216 texts. (In other words, students do not need to read any other texts besides these 216 in order to encounter all the list words in all the texts.) The new version of TextLadder thus created shorter sequence lists than the unadjusted version, which required 253 texts in order to encounter all the list words.[10]

2) Number of List Words Occurring Fewer than Five Times Throughout the Selected Texts

TextLadder produced, as output, a list of all the target vocabulary encountered over the course of the selected texts and the number of times each word[11] was encountered. The words on this list were then divided into two categories: those occurring four times or fewer throughout the selected texts and those occurring five times or more.[12] Figure 3 displays the difference between these two categories.

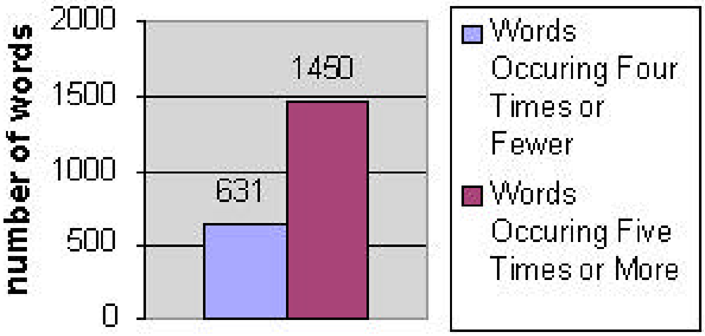

Figure 3. Words occurring four times or fewer vs. words occurring five times or more

As the above chart shows, a substantial number of list words (631) occurred four times or fewer over the course of the selected texts. This is clearly a problem, as there are serious doubts as to whether a student reading through the selected texts would effectively learn these words. As I have noted before, TextLadder always informs the instructor of the identity of words occurring four times or fewer so that the instructor can take steps to provide the students with exposure to the words through alternate means.

However, 631 words (30% of the total number of words encountered) is clearly too large a load for the instructor to handle.

As a result of this finding, two modifications were made: one to the compiled list of target words providing 95% coverage of *Voice of America* texts, and the other to TextLadder itself. The word list was altered (or rather, supplemented) in order to allow for 96% token coverage of *Voice of America* simplified news texts, rather than 95%. This was done by adding a new list to the combination of lists that already provided 95% coverage. This new list was obtained by selecting all the words occurring 1,000 times or more over the course of four years of *The Guardian* newspaper, minus proper nouns. (A word frequency-based analysis of *The Guardian* produced by Mike Scott [1997], served as the basis for this selection.) Once this new list was added to the previous ones, the resulting combination of lists was found to consistently provide 96% coverage of *Voice of America* simplified news texts (if proper nouns are included in the 96%).[13]

The advantage to having 96% list word/proper noun coverage, rather than 95%, is this: A student only needs to be familiar with 95% of the tokens in the text in order to comprehend that text. If list words account for 96% of a text's tokens, then the student does not in fact need to know all the list words encountered in the text. Indeed, up to 1% of the of the text's tokens can be made up of list words that the student can ignore (by "ignore," I mean that it is not necessary for the student to pre-acquaint him/herself with these words through the pre-reading dictionary exercise or through prior exposure).

Which list words should be ignored? The best candidates would be the ones that occur four times or fewer over the course of all the selected texts.

There is a problem, however. List words that occur four times or fewer over the course of all the selected texts (we shall call them "low-frequency words" for convenience) still tend to make up more than 1% of the tokens in individual texts. Therefore, some of the low-frequency words do need to be learned, in order for comprehension to take place. TextLadder's strategy for selecting which low-frequency words should be learned is to focus on those texts where ignored low-frequency words make up more than 1% of tokens. From each of these texts, low-frequency words are selected for learning until the number of ignored low-frequency words makes up less than 1% of tokens in the text. This is done in a way that ensures that higher-frequency low-frequency words (e.g., words that occur four times throughout the selected texts) are selected before lower-frequency low-frequency words (e.g., words that occur once throughout the selected texts).[14] Once the ignored low-frequency words make up less than 1% of the tokens in all the selected texts, TextLadder stops. At this point, it has determined exactly which low-frequency words still have to be learned in order for a reader to comprehend each of the selected texts.

Figure 4 shows the results of running the adjusted TextLadder with a list that provides 96% coverage (including proper nouns) of *Voice of America* texts.

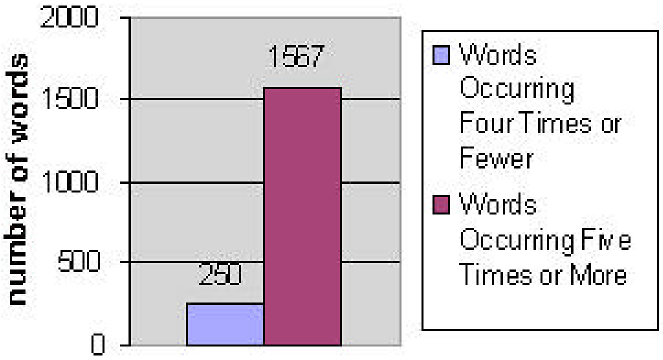

Figure 4. Words occurring four times or fewer vs. words occurring five times or more (96% list)

As is clear from Figure 4, a significantly smaller number of low-frequency list words is required for comprehension, when the newly-adjusted TextLadder is used in conjunction with a list providing 96% coverage. Presumably, the number should be even smaller with a list providing 97% coverage, although this was not tested. Note that TextLadder produces, as output, a list identifying these 250 words and specifying the number of times they have occurred over the selected texts. The instructor can use this information to provide the students with further encounters with the words, either through direct vocabulary instruction or through customized text creation. (For example, the 250 words above could conceivably be presented to students -- with the appropriate number of repetitions -- over the course of 10 or so texts designed and written by an instructor expressly for that purpose. These texts could then be read by the students once they had finished all the texts on the sequence list.)

Figure 5 shows the amount of new vocabulary introduced per text, when the 96%-coverage list is used.

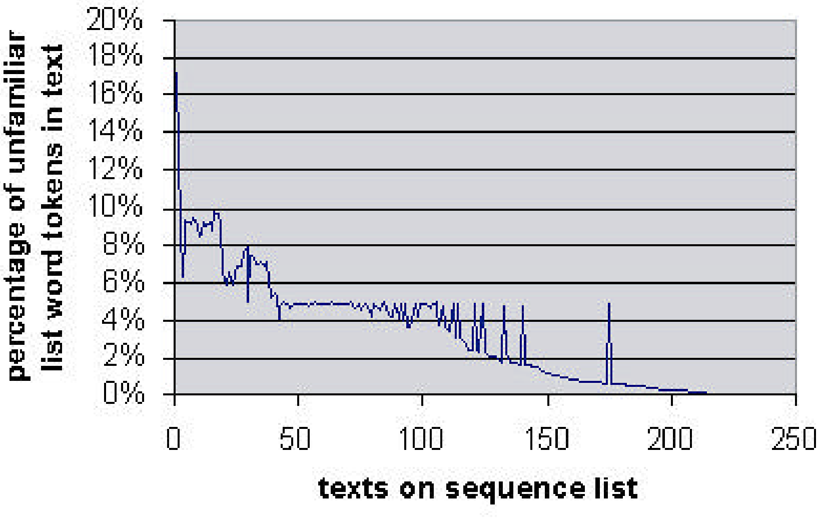

Figure 5. Amount of new target vocabulary introduced per text (96% list)

Note that the results displayed in Figure 5 resemble, to a great extent, the results displayed in Figure 2. This suggests that using a word list that provides 96% coverage instead of using a word list that provides 95% coverage does not greatly alter the distribution of new vocabulary over the course of the texts on the sequence list. Based on this finding, we can conclude that the 96% coverage list is the better list to use, since it a) allows for a similar amount of new vocabulary to be introduced in each text over the course of a similar number of texts, and b) ensures that a significantly smaller percentage of the target vocabulary that is encountered fewer than five times needs to be learned.

CONCLUSION

The results of this study show that an adjusted version of TextLadder can ensure that a reasonable amount[15] of new vocabulary is introduced in each text, for the majority of the texts on the sequence list. The study also indicates that of the list words occurring four times or fewer throughout the selected texts a significantly smaller number need to be learned when a list providing 96% coverage of the target domain can be obtained.

These results suggest that two major obstacles to the reading-based acquisition of target vocabulary -- namely, the high number of unfamiliar words in each text, and the insufficient number of repetitions of a target word -- can likely[16] be overcome through computer-driven intervention. Two important issues, however, have not been addressed by this study.

First, the study does not deal with how a reader might learn all the words on a given word list, but only those list-words encountered over the course of a miniature corpus consisting of 293 texts. This leaves unanswered the questions of how big a corpus would have to be to include all the words on the list, and how many texts would have to be read in order to encounter all these words.

Second, although the TextLadder program has been written with an eye to taking advantage of recent findings in vocabulary acquisition studies, no attempt has been made to determine whether it in fact results in measurable learning.

It is hoped that future studies will address these questions, so that TextLadder's feasibility as a method for optimizing reading-based vocabulary acquisition can be more definitively determined.

TextLadder is available at http://www.readingenglish.net.

TextLadder is free for research purposes.

APPENDIX A. NUMBER OF LIST WORDS OCCURRING FEWER THAN SIX TIMES THROUGHOUT THE SELECTED TEXTS

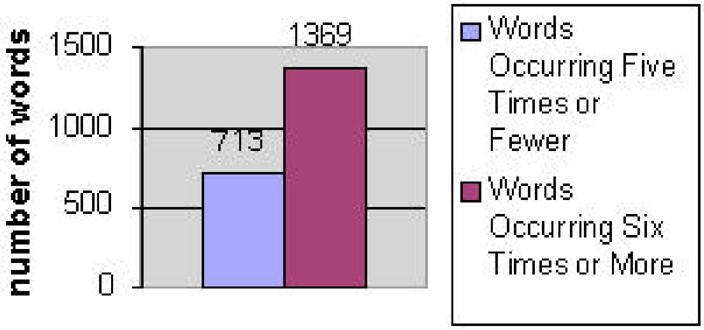

Figure 6. Words occurring five times or fewer vs. words occurring six times or more (95%)

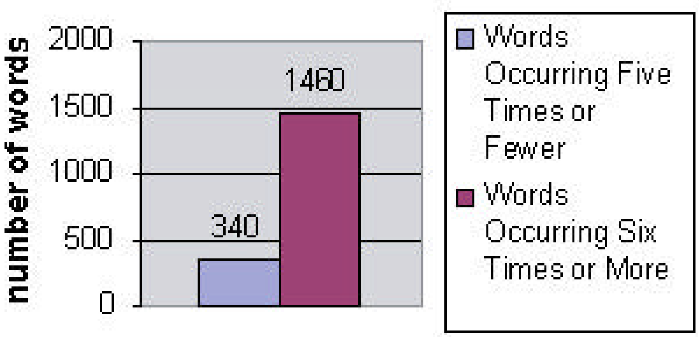

Figure 7. Words occurring five times or fewer vs. words occurring six times or more (96% list)

APPENDIX B. NEW LIST WORDS INTRODUCED PER TEXT (EXPRESSED AS NUMBERS RATHER THAN AS PERCENTAGES)

Note: Figures 8, 9, and 10 correspond with Figures 1, 2, and 5, respectively.

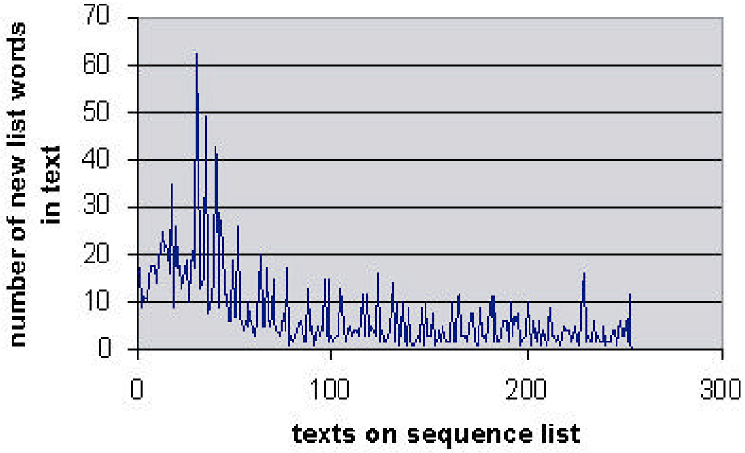

Figure 8. Number of new list words introduced per text

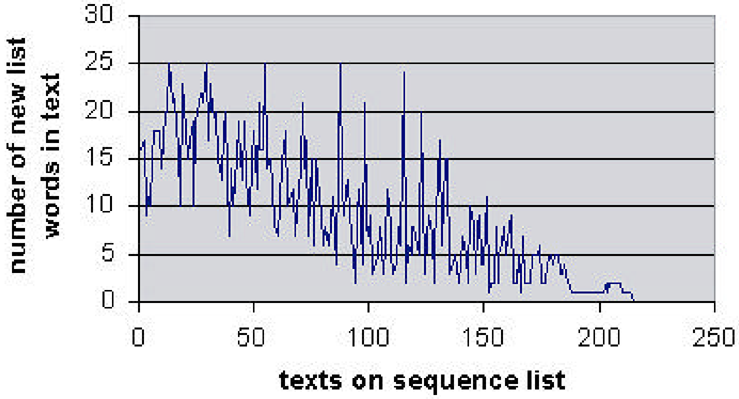

Figure 9. Number of new list words introduced per text

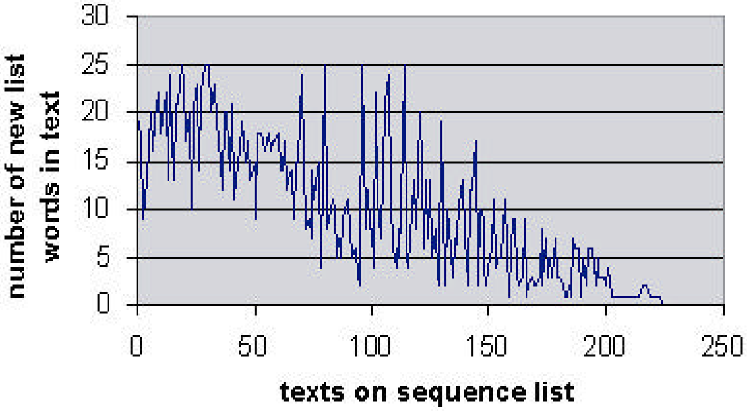

Figure 10. Number of new list worlds introduced per text (96% list)

NOTES

1. A word family consists of the base form of a word and all the inflected and derived forms of it that can be understood by a learner without having to learn them separately. (This will depend to a great extent on the learner's command of prefixes and suffixes.) For example, *connect, disconnect,* and *connectable* are subsumed by a single word family (Bauer & Nation, 1993).

2. A new list, the Academic Word List (AWL), has recently been developed (Coxhead, 2000). This list is shorter than the UWL and provides better token coverage for academic texts (10% as opposed to the UWL's 8.5%). There is also evidence that the GSL may not be the best general service vocabulary list available. A study by Nation and Hwang (1995) suggests that a 1,945-word family list consisting of the various overlaps between three word lists -- the GSL, a frequency list based on the Brown corpus, and a frequency list based on the LOB corpus -- provides better token coverage (83.4%) than the GSL itself (82.3%).

3. It is still uncertain whether or not the percentage remains above 4% when frequently occurring proper nouns are removed from the list of domain-specific words.

4. The number of encounters required for learning is controversial. This issue is brought up again later in the article, and discussed in detail in Note #12.

5. According to an analysis by Nation & Wang (1999) of a set of popular graded readers, the word lists for four out of six levels did not provide 95% coverage of readers at subsequent levels, even when proper nouns were allowed to be part of that coverage. The difference between word list sizes of any two consecutive levels (i.e., the "vocabulary gap" between texts at those levels) averaged out to about 400 words. In contrast, the number of words needed to bridge the vocabulary gap between any two consecutive texts arranged by the computer program varies between 1 and 25 words.
   Of course, the graded readers are much longer than the texts used by the program, which are news story-sized. However, this size difference is precisely the point: Smaller texts ensure that new vocabulary can be incrementally introduced in smaller, more manageable chunks.
   The use of shorter texts to sequence the introduction of new vocabulary *has* been attempted in a recent study (Worthington & Nation, 1996), and found to be problematic due to the varying rate of vocabulary introduction: a very large amount of vocabulary is introduced over the first few texts, while texts later in the sequence contain very little new vocabulary. The last section of the present study attempts to demonstrate, however, that applying careful control over the amount of new vocabulary introduced in each text can help mitigate this problem.

6. The choice of the first 176 words of the GSL along with basic colours, numbers, and so forth, as the basis for the initial familiar words pool, is somewhat arbitrary. In the future, I hope to conduct an extensive word-frequency analysis of elementary-level ESL/EFL texts, in order to better determine which words a post-elementary-level student can be expected to be familiar with.

7. Many of the prefixes and a number of suffixes for Levels 5 and 6, for example, are not considered when determining a match.

8. Although this approach ensures maximum comprehensibility for each new text, it may also create new problems by prolonging the interval between encounters of low-encounter words, thereby making it more likely that the student will forget the word in that interval. Such a situation might make acquisition of the word less likely. Clearly, then, there is a tension between the need for increased comprehensibility, which requires that the interval between exposures to low-encounter words be increased, and the need to reduce the likelihood of forgetting, which requires that the interval between exposures to low-encounter words be decreased. Ultimately, testing will have to be done in order to determine which of the factors takes precedence. If it is found that students are irrecoverably forgetting words in the prolonged interval between exposures, then two options remain:

a) simply remove the penalty system, or b) actually reverse the penalty system, so that texts with high numbers of low-encounter words are "rewarded" and placed earlier on the sequence list rather than later.

9. The list provided 95% token coverage (including proper nouns) for 91% of domain texts tested.

10. Also, note that TextLadder was adjusted in this second version to *not* place texts on the sequence list if they contained more than 25 unfamiliar words, regardless of the percentage. This restriction had the effect of saving longer texts for last.

11. Note that only one word per word family appears on the list.

12. To explain why a threshold of five encounters was chosen, I refer back to the Saragi, et al. (1978) study which showed a significant increase in learning when a new word was encountered five times, and an even greater increase for six encounters. A more recent study by Rott (1999) showed a significant increase in learning for students exposed to a word six times, compared to students exposed to the word two or four times. These results, like those of the Saragi et al. study, suggest a threshold somewhere in the vicinity of five or six encounters. Other studies have placed this threshold higher, at eight exposures (e.g., Horst, Cobb, & Meara, 1998), or even twenty (Herman, Anderson, Pearson, & Nagy, 1987). However, a very recent study by Zahar, Cobb, & Spada (2001) suggests that the minimum number of exposures necessary for learning is dependent on the student's prior vocabulary size, the reasoning being that if the reader is familiar with the words surrounding the word in question, then the exposure will lead to better acquisition.
    TextLadder's incremental approach to vocabulary introduction and its system of privileging high-encounter words attempt to ensure that students will already be familiar with most of the words that form the context of low-encounter words. For this reason, a threshold of five encounters was provisionally chosen, which TextLadder allows to be increased to six. A summary of the results obtained with a threshold of six, rather than five, is available in Appendix A.
    It should not be taken for granted, however, that encountering a word five or six times over an extended period of time (e.g., one or two semesters) will result in acquisition of the word. The studies mentioned above all involved studying vocabulary acquisition over a relatively short time period. The question of whether five or six encounters will suffice for acquisition, for TextLadder-processed texts read over an extended period of time, will have to be determined by future research. See Note #8 for details on how TextLadder might be modified to decrease the interval between word encounters and thereby perhaps increase the likelihood of acquisition.

13. The new combination of lists provided 96% token coverage (including proper nouns) for 88% of domain texts tested.

14. The process of selecting low-frequency (list) words for learning is described here in more detail. As mentioned, TextLadder focuses on texts where ignored low-frequency words up make up more than 1% of tokens. From those texts, it selects words that occur four times throughout the selected texts for learning. It then checks to see whether any texts remain in which ignored low-frequency words make up more than 1% of tokens. If so, it focuses on those texts, selecting low-frequency words that occur three times throughout the selected texts. Once again, it checks to see how many +1% texts remain, then focuses on those texts and selects low-frequency words that occur two times throughout the selected texts. It checks again, then repeats, selecting words that occur once throughout the texts. At any point, if TextLadder finds no remaining texts in which ignored low-frequency words make up more than 1% of tokens, the selection process is stopped.

15. i.e., greater than 1% and less than 10%

16. This assumes that acquisition of a word after 5 or 6 encounters is possible when extended time intervals between encounters are involved. See Note #12.

ABOUT THE AUTHOR

Sina Ghadirian holds a BA in English and a certificate in TESL. He has taught ESL in Canada and Thailand, and is currently completing a MA in Second Language Education at McGill University (Montreal, Canada). His research interests include vocabulary acquisition and computer-assisted language learning (CALL).

E-mail: sina.ghadirian@mail.mcgill.ca

REFERENCES

Bauer, L., & Nation, I. S. P. (1993). Word families. *International Journal of Lexicography, 6*(4), 253-279.

Carroll, J. B., Davies, P., & Richman, H. (1971). *Word frequency book*. New York: Houghton Mifflin.

Cobb, T., & Horst, M. (2001). Reading academic English: Carrying learners across the lexical threshold. In J. Flowerdew, & M. Peacock (Eds.), *Research perspectives on English for academic purposes* (pp. 315-329). Cambridge, UK: Cambridge University Press.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly, 34*(2), 213-238.

Flowerdew, J. (1993). Concordancing as a tool in course design. *System, 21*(2), 231-244.

Ghadirian, S. (2000). TextLadder [Computer software]. Available from http://www.readingenglish.net/.

Groot, P. (2000). Computer assisted second language vocabulary acquisition. *Language Learning & Technology, 4*(1), 60-81. Retrieved February 27, 2001, from http://llt.msu.edu/vol4num1/groot/default.html.

Harrison, R. (1997). Vital English Vocabulary word list. Retrieved February 27, 2001, from http://www.rick.harrison.net/langlab/vitaleng.html.

Herman, P. A., Anderson, R. C., Pearson, P .D., & Nagy, W. E. (1987). Incidental acquisition of word meaning from expositions with varied text features. *Reading Research Quarterly, 22*(3), 263-284.

Hirsh, D., & Nation, I. S. P. (1992). What vocabulary size is needed to read unsimplified texts for pleasure? *Reading in a Foreign Language, 8*(2), 689-696.

Horst, M., Cobb, T., & Meara, P. (1998). Beyond a clockwork orange: Acquiring second language vocabulary through reading. *Reading in a Foreign Language, 11*(2), 207-223.

Hulstijn, J., Hollander, M., & Greidanus, T. (1996). Incidental vocabulary learning by advanced foreign language students: The influence of marginal glosses, dictionary use, and reoccurrence of unknown words. *The Modern Language Journal, 80,* 327-339.

Krashen, S. D. (1989). We acquire vocabulary and spelling by reading: Additional evidence for the input hypothesis. *The Modern Language Journal, 73,* 440-464.

Laufer, B. (1989). What percentage of lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machine* (pp. 69-75). Clevedon, UK: Multilingual Matters.

Liu, N., & Nation, I. S. P. (1985). Factors affecting guessing vocabulary in context. *RELC Journal, 16*(1), 33-42.

Nagy, W. E., & Herman, P .A. (1987). Breadth and depth of vocabulary knowledge: Implications for acquisition and instruction. In M. G. McKeown & M. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 19-35). Hillsdale, NJ: Erlbaum.

Nation, I. S. P. (1990). *Teaching and learning vocabulary*. New York: Newbury House.

Nation, I. S. P., & Hwang, K. (1995). Where would general service vocabulary stop and special purposes vocabulary begin? *System, 23*(1), 35-41.

Nation, I. S. P, & Wang, K. (1999). Graded readers and vocabulary. *Reading in a Foreign Language, 12*(2), 34-50.

Rott, S. (1999) The effect of exposure frequency on intermediate language learners' incidental vocabulary acquisition and retention through reading. *Studies in Second Language Acquisition*, *21*(4), 589-620.

Saragi, T., Nation, I. S .P., & Meister, G. F. (1978). Vocabulary learning and reading. *System, 6,* 72-78.

Scott, M. (1997). Guardian Word List [word frequency list based on nearly all of *The Guardian* newspaper text from 1991-1994]. Available at http://www.liv.ac.uk/~ms2928/

Singleton, D. (1999). *Exploring the second language mental lexicon.* Cambridge, UK: Cambridge University Press.

Sternberg, R. J. (1987). Most vocabulary is learned from context. In M. G. McKeown, & M. Curtis (Eds.), *The nature of vocabulary acquisition* (pp. 89-105). Hillsdale, NJ: Lawrence Erlbaum Associates.

Sutarsyah, C., Nation, I. S. P., & Kennedy, G. (1994). How useful is EAP vocabulary for ESP? A corpus based case study. *RELC Journal, 25*(2), 34-50.

West, M. (1953). *A general service list of English words.* London: Longman, Green & Co.

Worthington, D., & Nation, P. (1996). Using texts to sequence the introduction of new vocabulary in an EAP course. *RELC Journal*, *27*(2), 1-11.

Xue, G., & Nation, I. S. P. (1984). A university word list. *Language Learning and Communication, 32,* 215-219.

Zahar, R., Cobb, T., & Spada N. (2001). Acquiring vocabulary through reading: Effects of frequency and contextual richness. *Canadian Modern Language Review, 57*(3), 541-572.