COMPUTING THE VOCABULARY DEMANDS OF L2 READING

Tom Cobb
Université du Québec à Montreal

Linguistic computing can make two important contributions to second language (L2) reading instruction. One is to resolve longstanding research issues that are based on an insufficiency of data for the researcher, and the other is to resolve related pedagogical problems based on insufficiency of input for the learner. The research section of the paper addresses the question of whether reading alone can give learners enough vocabulary to read. When the computer's ability to process large amounts of both learner and linguistic data is applied to this question, it becomes clear that, for the vast majority of L2 learners, free or wide reading alone is not a sufficient source of vocabulary knowledge for reading. But computer processing also points to solutions to this problem. Through its ability to reorganize and link documents, the networked computer can increase the supply of vocabulary input that is available to the learner. The development section of the paper elaborates a principled role for computing in L2 reading pedagogy, with examples, in two broad areas, computer-based text design and computational enrichment of undesigned texts.

INTRODUCTION

There is a lexical paradox at the heart of reading in a second language. On one side, after decades of guesswork, there is now widespread agreement among researchers that text comprehension depends heavily on detailed knowledge of most of the words in a text. However, it is also clear that the words that occur in texts are mainly available for learning in texts themselves. That is because the lexis (vocabulary) of texts, at least in languages like English, is far more extensive than the lexis of conversation or other non-textual media. Thus prospective readers of English must bring *to* reading the same knowledge they are intended to get *from* reading. This paradox has been known in outline for some time, but in terms loose enough to allow opposite proposals for its resolution. On one hand, Nation (e.g., 2001) argues for explicit instruction of targeted vocabulary outside the reading context itself. On the other, Krashen (e.g., 1989) believes that all the lexis needed for reading can be acquired naturally through reading itself, in a second language as in a first. It is only recently that the dimensions of this paradox could be quantified, with the application of computer text analysis to questions in language learning. What this quantification shows is the extreme unlikelihood of developing an adequate L2 reading lexicon through reading alone, even in highly favourable circumstances. This case is made in the initial research part of the paper. The subsequent development section goes on to show how the text computing that defined the lexical paradox can be re-tooled to break it, with (1) research-based design of texts and (2) lexical enrichment of undesigned texts. Empirical support for computational tools will be provided where available; all tools referred to, both analytical and pedagogical, are publicly available at the *Compleat Lexical Tutor* website (www.lextutor.ca).

DEFINING THE LEXICAL PARADOX

In applied linguistics conversations, turn-taking can involve a delay of several years. An example is Krashen's (2003) paper entitled *Free voluntary reading: Still a very good idea,* which criticizes the findings of a study by Horst, Cobb and Meara (1998) that had called into question the amount of vocabulary acquisition that normally results from free, pleasurable, meaning-oriented extensive reading. This study found that even with all the usual variables of an empirical study of extensive L2 reading controlled rather more tightly than usual, the number of new words that are learned through the experience of reading a complete, motivating, level-appropriate book of about 20,000 running words is

minimal, and does not indicate that reading itself can reasonably be seen as the only or even main source of an adult reading lexicon. The gist of Krashen's response (2003) was that such studies typically underestimate the amount of lexical growth that takes place as words are encountered and re-encountered in the course of free reading. To support this contention, he calculated an effect size from the Horst, Cobb and Meara data that he interpreted to show stronger learning than these researchers' conclusions had implied. But more importantly, beyond the data, he believes that many words and phrases are learned from reading that do not appear in the test results of this type of study, owing to the crude nature of the testing instruments employed, which typically cannot account for partial or incremental learning. According to this argument, word knowledge is bubbling invisibly under the surface as one reads, and may appear as a known item in a vocabulary test only some time later[1]. This hidden vocabulary learning from reading is seen as extensive enough to "do the entire job" (Krashen, 1989, p. 448) of acquiring a second lexicon, an idea that Waring and Nation (2004, p. 11) describe as "now entrenched" within second and foreign language teaching. Similar claims are many (e.g. Elley's belief that children graduating from a book flood approach had learned "all the vocabulary and syntax they required from repeated interactions with good stories," 1991, pp. 378-79); but clear definitions of "the entire job" are few.

Krashen has taken part in a number of conventional vocabulary-from-reading studies that use conventional measures, but these studies have not provided empirical evidence of either the extent of such hidden learning, or its sufficiency as the source of a reading lexicon. Instead, he cites the "default explanation" for the size of the adult lexicon, an account borrowed from first language (L1) theorizing (e.g., Nagy, 1988; Sternberg, 1987), whereby the lexical paradox is resolved through the sheer volume of reading time available over the course of growing up in a language. According to this explanation, a lifetime of L1 reading must eventually succeed in doing the job – even if very little measurable vocabulary knowledge is registered in any one reading event – since there is no other plausible way to account for the large number of words that adult native speakers typically know.

The extension of research assumptions and procedures from L1 to the L2 learning contexts is questionable at best[2], particularly in the absence of empirical support. But as will be shown here, both the extent and sufficiency of hidden vocabulary learning can in fact be investigated empirically within the L2 context, without recourse to default arguments. Key to this undertaking are a research instrumentation, method, and technology for measuring small increments of lexical knowledge that can be applied to sufficient numbers of words over a sufficient length of time to be plausibly commensurate with the known vocabulary sizes of learners: roughly 17,000 English word families in the case of a typical literate adult L1 lexicon (as calculated by Goulden, Nation & Read, 1990), or the 5,000 most frequent word families in the case of L2 (proposed as minimal for effective L2 reading by Hirsch & Nation, 1992). That is to say, the experimentation requires substantially more than the handful of words normally tested in this type of research (typically between 10 and 30, as discussed in Horst et al., 1998) in order to arrive at a credible estimate of "the entire job."

Claim A: The Extent of Hidden Learning

An instrument capable of measuring incremental knowledge is Wesche and Paribakht's (1996) vocabulary knowledge scale, or VKS, which asks learners to rate their knowledge of words not in binary terms (I know/I don't know what this word means) but on a five-point scale (ranging from "I don't remember having seen this word before," to "I can use this word in a sentence.") But since the VKS requires learners to also demonstrate their knowledge (e.g. by writing sentences), it is cumbersome to use in measuring changes in the knowledge of large numbers of words over time through repeated encounters, as would be needed to test the claim of extensive amounts of hidden acquisition. Therefore, Horst and Meara (1999) and Horst (2000) devised the following ratings-only version, which was suitable for adaptation to computer.

0 = I definitely don't know what this word means

1 = I am not really sure what this word means

2 = I think I know what this word means

3 = I definitely know what this word means (Horst, 2000, Chapter 7, p. 149)

Following a reading of a text, learners can efficiently rate their knowledge of a large number of its words using a computer input that employs this scale and stores the number of words rated 0, 1, 2, and 3 for each learner and each reading. But the real innovation of the adaptation is the conversion of the scale to a matrix, which allows the comparison of ratings over two (or more) readings of the same text. The matrix (shown in Figure 1) is essentially the 4-point scale in two dimensions, so that each cell represents results at both time *n* and after a subsequent reading (time *n+1*). For example, the data in the first horizontal row shows that 75 words had been rated 0 after reading n and were still rated 0 (I don't know) after reading n+1, but that 27 words had moved from 0 to 1, nine words from 0 to 2, and three words from 0 to 3. The second row shows how words rated 1 (not sure) at time n were distributed at time n+1, and so on. In other words, the cell intersections capture the numbers of words that have changed or failed to change from one knowledge state to another as a result of a subsequent reading.

	Reading n+1				
Reading n		**0**	**1**	**2**	**3**
	0	**75**	27	9	3
	1	4	**20**	20	6
	2	2	4	**4**	35
	3	0	0	0	**75**

Figure 1. From scale to matrix (Horst, 2000)

Employing a methodology of repeated readings and a computer-based testing apparatus that allowed the tracking of large numbers of words, Horst and Meara were able to trace the ups and downs of word knowledge that normally pass below the radar of conventional tests. What new information emerges from this methodology? For just this one state of the matrix Heading 3 (column 6) of Figure 1 shows that of the 300 learnable targets, 44 (or 3 + 6 +35) have moved into the "I definitely know this word" state from another knowledge state. This is new knowledge that would probably have shown up on a standard test. However, another 56 words (27 + 9 + 20) have made lesser gains (into the "not sure" and "think I know" territory) that would probably not have shown up on a standard test. These ratios change over the course of several readings, as the learning opportunities diminish, but in the first three readings there are often at least as many words moving rightward below the radar as above it, i.e., moving to knowledge-state 1 or 2 rather than 3. A surprising number of words move to the right and then back to the left for a time, presumably reflecting either a learning and forgetting cycle, or a hypothesis testing phase, or elements of both (Horst, 2000). The evidence from the matrix studies broadly shows that Krashen is right: there is more vocabulary learning from reading than most tests measure. It seems uncontroversial to generalize from Horst and Meara's data that the total amount of vocabulary learning from reading might be as much as double what the various studies using more conventional measures have typically shown.

An alternate source of evidence for substantial amounts of hidden vocabulary growth through reading is provided by Waring and Takaki (2003) using a different methodology. These researchers tested twenty-five words acquired from reading with measures at three levels of difficulty—passive recognition (that a word had been seen in the text), aided meaning selection (by a multiple choice measure), and unaided recall (through a translation test)—and found that scores were almost 2.5 times higher for multiple choice than translation, and more than 3 times higher for recognition than for translation. In other words, most of

the initial learning represented by remembering that a word had appeared in the text would not have registered on either of the other more difficult tests. This finding thus complements the matrix finding, albeit for a smaller number of items.

But can we get from here to sufficiency? Even if it is clear that more learning takes place through word encounters than most tests measure, is free reading able to provide a sufficient number of such encounters?

Claim B: The Sufficiency of Hidden Learning

Krashen's related claim, the sufficiency of hidden vocabulary growth, can also be tested empirically in an L2 context, as the following very basic experiment in corpus analysis demonstrates. But first we need some definitions.

To arrive at an operational definition of sufficiency, we might ask questions such as: How many words are enough for various purposes, such as to begin academic study in a second language, or to undertake a professional activity? Vocabulary researchers working on questions of coverage calculate the minimum number of word families needed for non-specialist reading of materials designed for native speakers to be between 3000 (Laufer, 1989) and 5000 word families (Hirsch & Nation, 1992) -- provided these are high frequency items and not just random pick-ups. How many encounters are needed for word learning to occur? The number varies with a host of individual and contextual factors, but the majority of studies (reviewed in Zahar, Cobb & Spada, 2001) find that an average of six to ten encounters are needed for stable initial word learning to occur. In Horst's (2000) matrix work, six encounters were the minimum exposure for words to travel reliably from state 0 to state 3 and stabilize. Will anything like 3,000 word families be met six times apiece through free reading?

Investigation 1

The materials assembled to answer this question were chosen to give the free reading argument optimal chances of succeeding. Thus the vocabulary size assumed to be sufficient for comprehension and learning was set as low as could be deemed plausible, at 3000 word families of written English rather than 5000. In contrast, the amount of reading a typical L2 learner would be likely to achieve was set as high as could be deemed plausible. A sample of the free reading that an ESL reader might be expected to undertake over a year or two of language study was extracted from the 1 million word Brown corpus (Kucera & Francis, 1979). This classic corpus comprises 500 text samples of roughly 2,000 words grouped into sub-corpora of various sizes (different kinds of fiction, etc., as shown in the bottom half of Figure 2). To reflect the kinds of reading learners might do, the original sub-corpora were further grouped into three broad categories (press, academic, and fiction) of roughly similar size (179,000 words, 163,000 words, and 175,000 words, respectively). It is reasonable to suppose that one of these three groupings is a plausible if optimistic representation of the amount of free reading of authentic material that learners might achieve over a year or two of language study (these word counts are roughly equivalent to 100 pages of newspaper text, six stories the size of *Alice in Wonderland*, or 17 academic studies the length of this one.)

High frequency words were extracted from the 100-million-word British National Corpus (Leech, Rayson, & Wilson, 2001) and grouped into families and then into thousand-family lists by Nation (2006, available at http://www.lextutor.ca/vp/bnc/). The first three of Nation's lists (i.e. the 3000 most frequent word families) represent the current best estimate of the basic learner lexicon of English. A random item-from-wordlist generator (available at http://www.lextutor.ca/rand_words/) produced 20 sets of three 10-word samples from the 1000, 2000, and 3000 British National Corpus (BNC) lists. One of these sets was selected randomly for use as sample learning targets in the investigation[3].

A computer program calculated the number of occurrences of each sample word family that a learner would encounter in each of the Brown sub-corpora. This computer program called *Range* (Heatley &

Nation, 1994) was adapted for Internet by the author and is available at http://www.lextutor.ca/range/. Figure 2 shows the distribution of a word, phrase, or family throughout a set of texts. The original version of the program allowed users to specify their own texts; the online version shown in Figure 2 provides a set of standard texts, namely the 15 original sub-corpora or the three larger groupings of the Brown corpus already mentioned. In the present experiment, word families as opposed to individual words were the search units. This was achieved by entering a stem form plus apostrophe for each item as appropriate (*abandon'* finds *abandons, abandoning, abandonment,* as shown in Figure 2). Since it cannot be taken for granted that learners will recognize family members as being related (Schmitt & Zimmerman, 2002), incorporating whole families in the analysis is likely to provide a generous estimate of the learning opportunities in the text sample. A similarly generous assumption is that learners have perfect memory for encountered items over extended time and text.

The distribution of the BNC 4000-level word family *abandon'* throughout the three major divisions of the Brown corpus is requested in Figure 2; the output of the Range search is shown in Figure 3. The point to notice is that while this item appears in all three samples, it appears more than six times in only one of them (press writing).

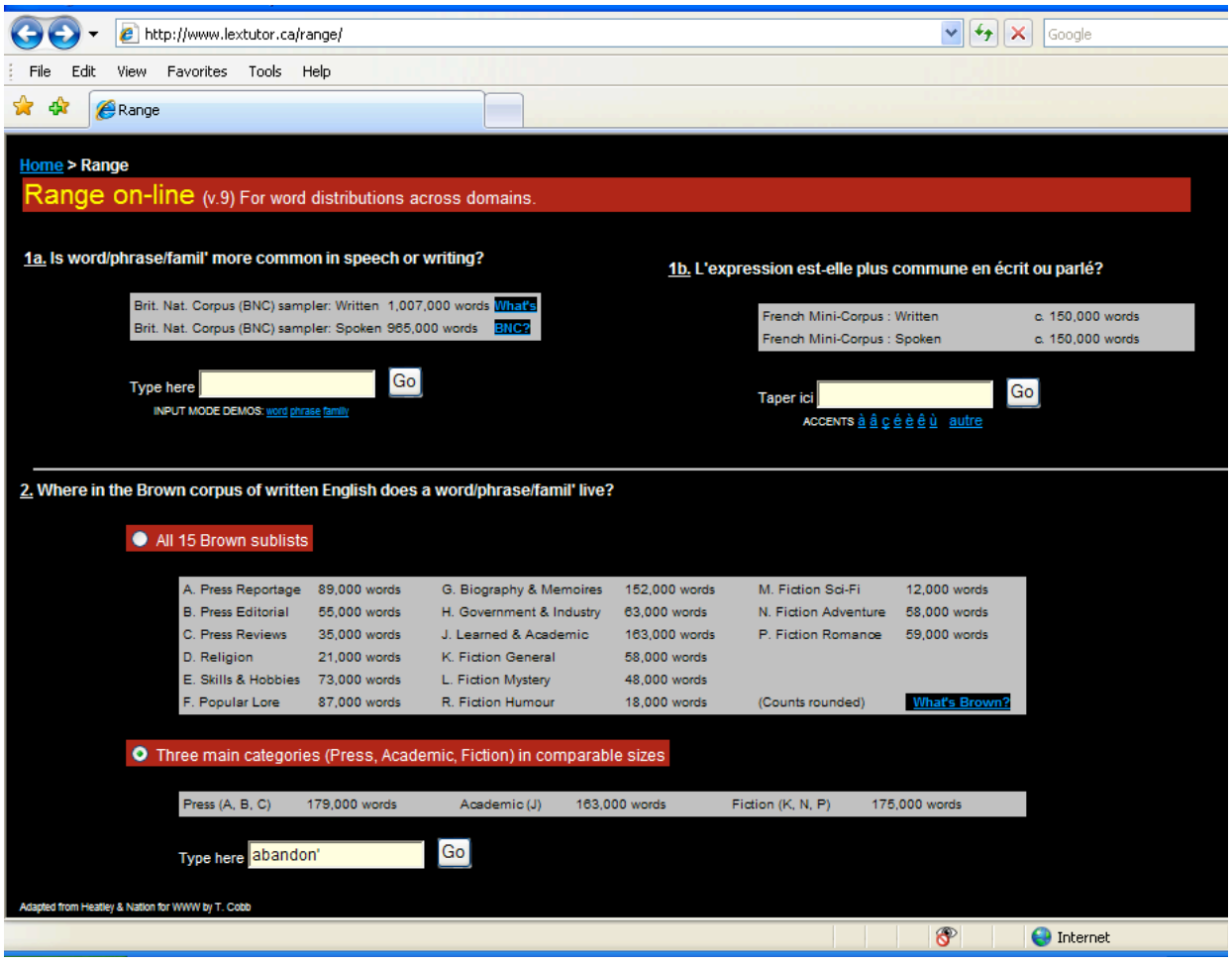

Figure 2. Range requesting word distributions of *abandon'* family

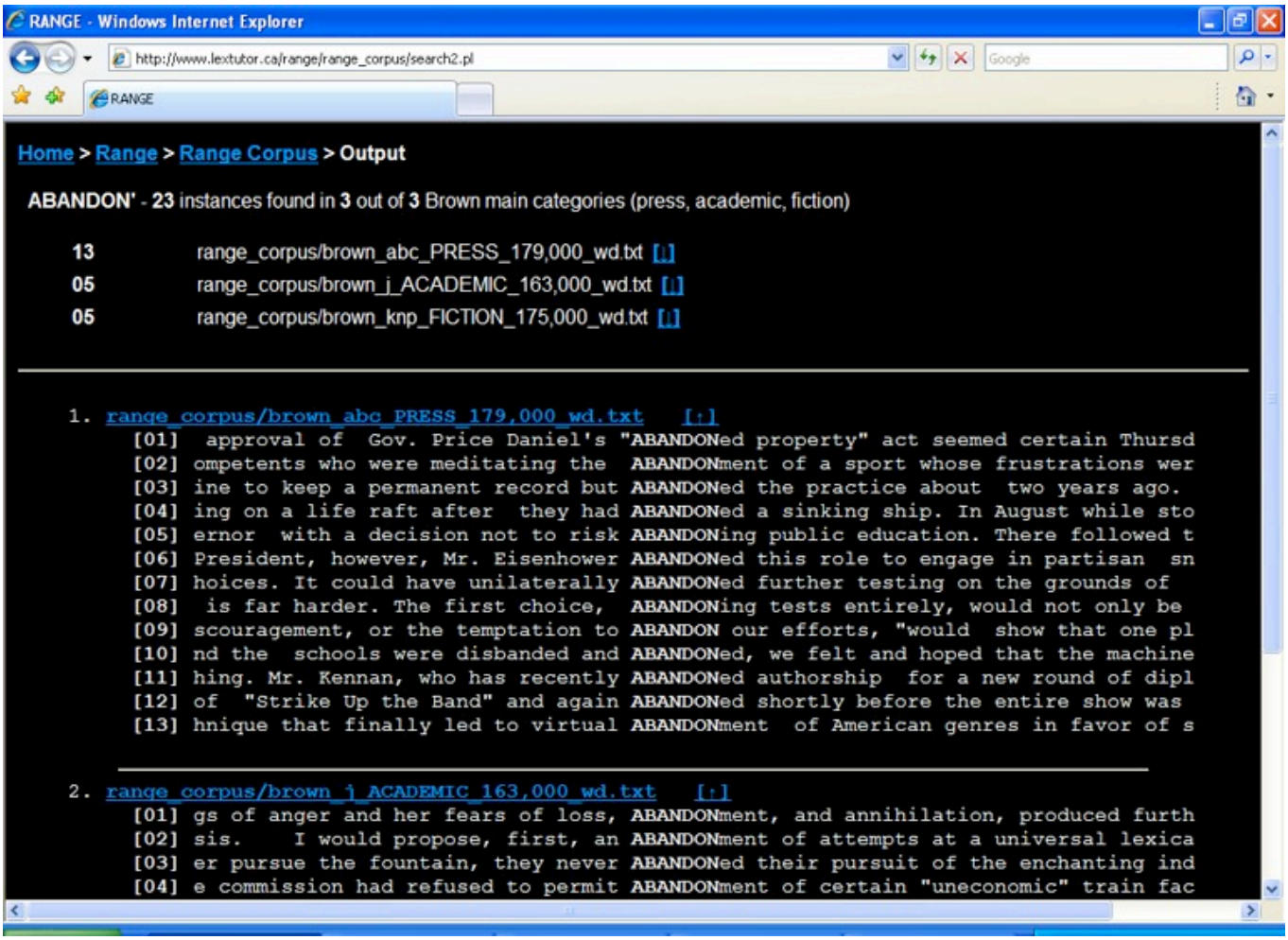

Figure 3. Range for word distributions – output

The overall and perhaps unexpected finding from this analysis is that after the most frequent 1000 items, family ranks tend to thin quite rapidly, and with them the learning opportunities. Table 1 shows the distributions in the three Brown samples for the ten target word families from each of the three most frequent BNC levels. For each target word family, the total number of occurrences in each sub-corpus is shown; at the bottom of each column, the number of targets appearing more than six times in each sub-corpus is shown. As can be seen, all 1000-level word families will be met more than six times in press writing, all except *bus* more than six times in academic writing, and all except *bus* and *associat'* in fiction. However, five 2000-level families (*persua', technolog', wire', analy'*, and *sue*) will dip below six encounters in one or more areas. And none of the 3000-level families will be encountered six times in all three areas, and half or more are not met six times in any area. (No member of the *irritat'* family is met in 163,000 words of academic text!) A sideline finding is that fiction writing, once the usual focus of free reading programs for learners in the process of acquiring 1000 and 2000-level vocabulary, does not present the strongest learning opportunities in either of these zones. Fiction does, however, seem to be a reasonable source of 3000-level items, providing six occurrences for five of its 10 words, as compared to four for press and three for academic writing. It is therefore worth looking at the vocabulary growing opportunities of fiction reading more closely.

Table 1. Decreasing Likelihood of Meeting Words in Grouped Sub-Corpora

1000 level				**2000 level**				**3000 level**			
Word family	Press	Aca-demic	Fic-tion	Word family	Press	Aca-demic	Fic-tion	Word Family	Press	Aca-demic	Fic-tion
lead'	200	64	45	persua'	17	3	7	irritat'	3	0	6
point'	106	194	61	grade'	14	25	8	millimeter'	0	0	0
bus	15	1	1	technolog'	9	8	0	urgen'	7	1	7
associat'	66	49	4	moon'	6	27	31	transmi'	5	9	1
press'	66	100	38	wire'	3	5	20	chew'	0	0	3
creat'	61	51	21	Maintain'	16	49	6	naked'	2	1	18
real'	153	86	172	analy'	12	129	4	civiliz'	5	12	12
other'	383	355	273	drama'	40	14	8	contest'	15	1	4
special'	90	52	20	depress'	14	7	9	charm'	19	1	12
final'	62	66	57	sue	8	7	1	prompt'	6	4	5
MEAN (SD)	120.20 (106.17)	101.80 (101.97)	69.20 (86.61)		13.9 (10.21)	27.40 (38.41)	9.40 (9.35)		6.20 (6.23)	2.9 (4.23)	6.8 (5.63)
Words/10 with 6+	10	9	8		9	8	7		4	3	5

Investigation 2

For a complementary investigation, the sufficiency of a generous diet of free fiction reading as the sole or main source of vocabulary growth for 3000-level families is now examined. At the same time, the reading sample is changed from a corpus sample of texts produced by many writers to a sample of texts produced by a single author, where the vocabulary learning opportunities are arguably greater (through characteristic themes, repetitions, etc.). A corpus of just under 300,000 words was assembled from seven Jack London stories (including school favorites *Call of the Wild* (1903) and *White Fang* (1906) all offered free of cost at http://london.sonoma.edu/Writings/) as a second plausible representation of a heavy diet of free reading. Would a learner who read all these stories meet most of the 3000-level families six times apiece?

The computational tool used in this analysis is lexical frequency profiling, in this case the BNC version of VocabProfile (available at http://www.lextutor.ca/vp/bnc/, illustrated below in another context), which breaks any English text into its frequency levels according to the thousand-levels scheme already employed. The results of this analysis are as follows: The full collection of London adventure stories was shown to contain 817 word families at the 3000 level; however, only 469 of them are met six times or more, while 348 are met five times or less (181 of them twice or less). In other words, fewer than half will be met enough times for reliable learning to occur. Interestingly, this result is similar to that shown in Table 1, where half the 3000-level words appeared six times or more in the fiction sub-corpus.

Conclusion

Together, these projections indicate that even the largest plausible amounts of free reading will not take the learner very far into the 3000-family zone. It is thus somewhat redundant to raise the matter that even words met more than six times are not necessarily learned. New word meanings are normally inferable in environments containing no more than one unknown item per 20 known items, (Laufer, 1989; Liu Na & Nation, 1985). However, VocabProfile analysis of one of the best known of the London stories (*Call of the Wild*, comprising 31,473 words) shows that 10% of the text's words (not including proper nouns) come from frequency zones beyond the 3000 level itself, sometimes well beyond it. This means that many of the novel's 3000-level items will be met in environments of 1 unknown item per 10 words, or double the density that research has shown learners able to enjoy or learn from[4].

To summarize, this analysis is based on the most generous conditions possible: a 3000 word size requirement rather than 5000; six occurrences for learning rather than ten; a one in twenty new word density; a larger and broader diet of input than many learners will provide for themselves; an assumption that family members are usually recognized; and an assumption of minimal forgetting between reading encounters. Even then, fewer than half the 3000 level words present themselves sufficiently for reliable learning to occur. Further, the situation only gets worse for word families at the 4000 and 5000 levels and beyond. Thus, while there may well be more word learning from random encounters in free extensive reading than meets the eye, the fact is interesting but irrelevant, since most post-2000 words simply will not be encountered at all in a year or two of reading. Therefore free reading alone is not sufficient to "do the entire job" of building a functional second lexicon in any typical time frame of L2 learning.

To refute this finding, sufficiency proponents would need to define what the "entire job" of reading in an L2 is, and then show, either empirically or in principle, how this job can be done through reading alone, given the learning rate, learning conditions, and lexical profile findings outlined above and elsewhere. Until then, the common finding that many ESL learners tend to plateau with usable knowledge of about 2000 words families or less (leaving them poorly equipped to comprehend most texts) remains entirely explicable (Cobb, 2003).

BREAKING THE LEXICAL PARADOX

The findings presented thus far present a basis in text analysis for what many studies have shown empirically in the past 20 years (e.g. Alderson, 1984; Bernhardt, 2005), that L2 reading is "a problem" and that the main problem is lexis. This longstanding awareness, in the research if not in the teaching community, has produced many proposals to supplement vocabulary growth from reading with other and more direct approaches to vocabulary learning. Examples include Paribakht and Wesche's (1997) reading-plus (plus vocabulary activities) scheme and various vocabulary course supplements (e.g., Barnard, 1972; Redman & Ellis, 1991; Schmitt & Schmitt, 2004).

But there are problems in principle with the supplement solution, all of which rely to some degree on separating learning the words for reading from the act of reading. One problem is that lexical knowledge does not necessarily transfer well from vocabulary exercise and dictionary look-up to text comprehension (Cobb, 1997; Krashen, 2003; Mezynski, 1983), especially when there is a delay between the two. Second, the number of words to be met and recycled typically proliferates the vocabulary supplements to sets of several volumes (e.g., Barnard's five volumes; Redman & Ellis' four volumes), diverting a large amount of instruction time away from reading itself. The reading-plus approach is reading-based in that the target words are drawn from a text just read, but it has the disadvantage that this work must be prepared by a teacher with a text and vocabulary items that have been selected in advance and so can only be developed for a small handful of texts.

What is missing from either supplement scenario is some way of focusing attention on and proliferating encounters with new words at any level within the act of reading, or shortly after reading, for any type of text, and for lots of texts. The following section of this paper will look at several concrete proposals for doing this, with reference to empirical validation where available. The goal is to use computing to preserve the free in free reading. Two broad approaches will be described and illustrated. The first is computer-based text design, and the second is computer-aided enrichment of undesigned texts.

Computer-Aided Text Design: The Case for Home-made Simplified Materials

In principle, simplified or graded texts can meet some of the word learning requirements outlined above. Texts can be written to a particular vocabulary knowledge level, with words beyond that level introduced in environments that meet the '1 unknown word in 20' ratio mentioned earlier as the criterion for reliable guessing from context. New words can be recycled the desired number of times, in a process extending over a series of texts, until a vocabulary target, whether 3000 or 5000 frequent word families, has been

met. Doing such re-writing well is clearly a difficult and expensive job. Perhaps for this reason, there is no set of graded readers in English that explicitly attempts to do it all. The arguably best designed of the graded reader sets available (e.g., the Longman *Penguin* series or Oxford's *Bookworm*), while useful, share a number of limitations that are readily evident without the help of detailed text analysis. As noted by Hill (1997), these texts are almost exclusively based on just one text genre, narrative fiction (either classics safely out of copyright or custom written originals). They employ a variety of unspecified frequency classification systems and offer no method of matching learner level to text level other than self-selection. And they make no claims about how many stories at each level a learner would have to read to achieve mastery at that level, or what coverage this mastery would provide with respect to real texts (although researchers like Nation & Wang (1999) have looked at some of these questions).

Computer text analysis can add two further limitations. One is that no series of graded readers proceeds systematically beyond the 3000-families level, and even those that get this far do not cover it particularly well. This is shown in a VocabProfile analysis of a whole set of graded readers similar to the analysis of the Jack London stories above. If a learner read all 54 stories at six levels in the *Bookworm* series (a total of 377,576 words), he or she would indeed meet 931 of the thousand word families at the 3000 frequency level, but would meet just over half of them (511 families) six times or more. This analysis is remarkably similar to the two above which also showed only half of the 3000-level families appearing six times or more. A difference, however, is the overall known-word density of the contexts the words will be met in, owing to the unequal presence of low frequency (post-3000-level) items in the two collections (10% for *Call of the Wild* and a more manageable 5% in the higher-end *Bookworms*). Still, the problem of providing exposure to all or most of the 3000-level families in a series of graded readers has not really been cracked. Indeed it may be impossible to do so within the constraints of publishing costs, school budgets, the number of stories it would take to provide full coverage, and so on.

A further limitation is that fictional text whether simplified or not may be inherently limiting as a source of vocabulary growth given most learners' reading goals. Taking up a concern expressed by Venezky (1982), Gardner (2004) performed a text analysis on fiction vs. expository L1 school reading materials and found that the lexis of expository text (i.e., the language used in academic and professional settings) differs from the lexis of fiction in substantial ways. Expository text basically comprises more words, different words, and more difficult words, in addition to unfamiliar discourse patterns that are not simple mirrors of real-life time sequences. Gardner questions the suitability of using fiction texts as preparation for reading expository texts, as is common practice.

There is clearly a case for producing graded materials beyond those presently available. In addition to materials that systematically target vocabulary growth beyond the 2000-level, there is also a need for expository materials to supplement the fictional. If publishers are unlikely to produce a complete range of research-indicated, graded materials, then these can be produced by institutions or by individuals, but presumably with some difficulty. However, two complementary types of text computing can make the task of in-house text simplification feasible, if not simple. Frequency profiling software can be used to find, adapt or create texts to a pre-specified lexical profile and coverage; and text comparison software can be used to establish degree of lexical recycling over a series of texts.

Writing Words Out: Lexical Frequency Profiling

An example of profiling software is VocabProfile (available at www.lextutor.ca/vp/), which categorizes the lexis of texts according to frequency. Users can select either Nation's original frequency scheme of General Service List and Academic Word list (Coxhead, 2000) or 20 BNC thousand-family lists (Nation's 14 BNC-based lists mentioned above were recently expanded by the author to 20. Viewable versions of the 20 lists are linked from BNC Vocabprofile's entry page at http://www.lextutor.ca/vp/bnc/.)[5]. Figure 4 shows the input to the BNC version of VocabProfile for a text by a Canadian journalist, Rex Murphy, and Figure 5 shows the output for the same text.

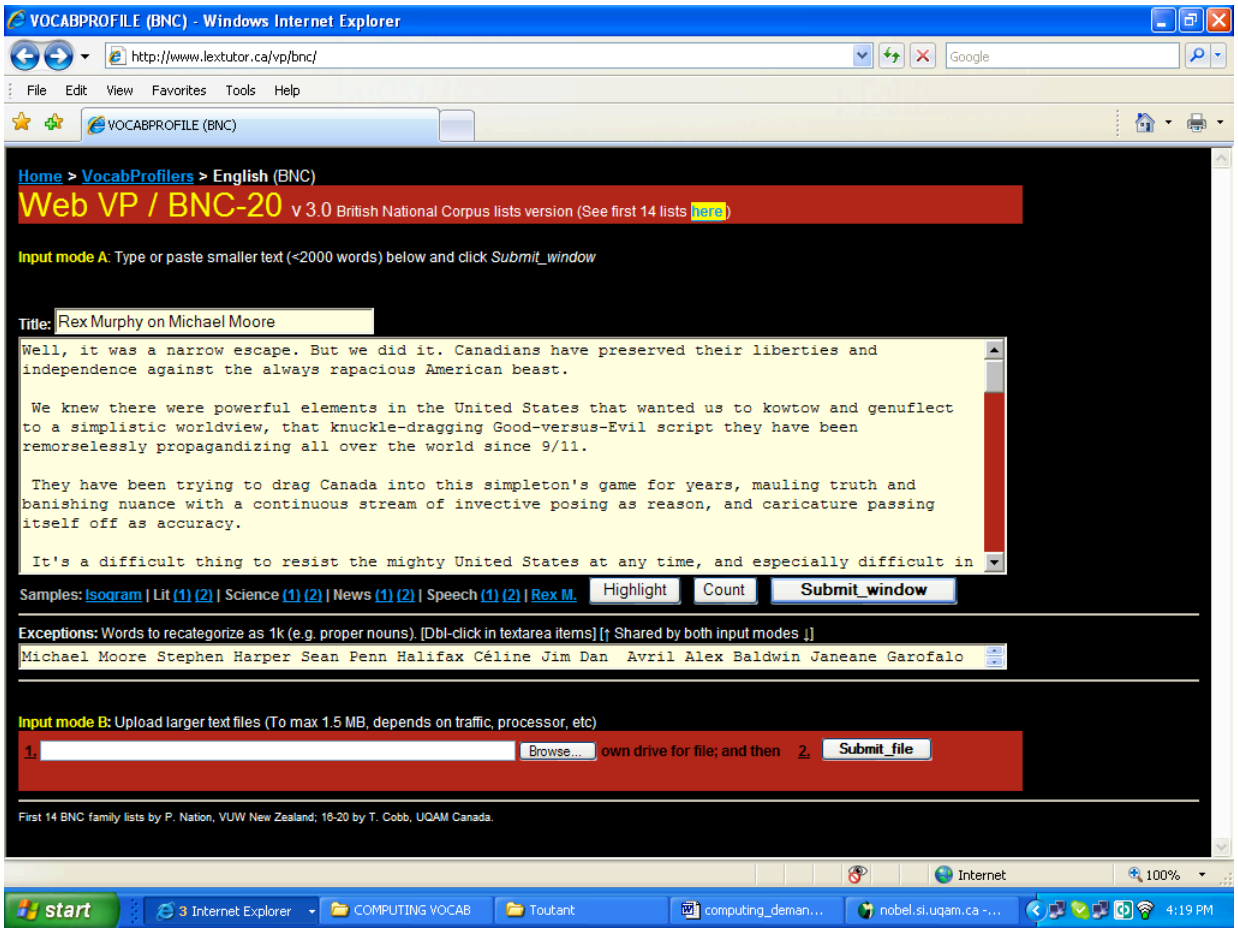

Figure 4. VocabProfile input

It is a fairly simple matter to use VocabProfile interactively to modify the lexical level of a sizeable text. VocabProfile will identify the words of learning interest for a particular group, say those between 5,000 and 10,000 frequency level for advanced learners, and thereby indicate the words that need to be written out so that target items occur in suitable known-to-unknown ratios. In the Murphy text that would mean writing out about 20 items. Using the window entry mode, the editor can go back and forth, editing and checking in iterations. This work is easier if the learners' approximate level is known using the same testing framework that the software employs; this is the case with many of the measures available at http://www.lextutor.ca/tests/.

Writing Words In: Text Comparison Software

As already mentioned, research indicates that the average number of encounters needed for reliable retention of a novel lexical item is between six and ten. There are sub-dimensions to this basic learning condition, such as the spacing between encounters (Mondria & Wit-de Boer, 1993) and the properties of the contexts surrounding the items (Cobb, 1999; Mondria & Wit-de Boer, 1991); but as shown above, just ensuring six encounters of any kind for a significant proportion of any post-2000 word list is not simple.

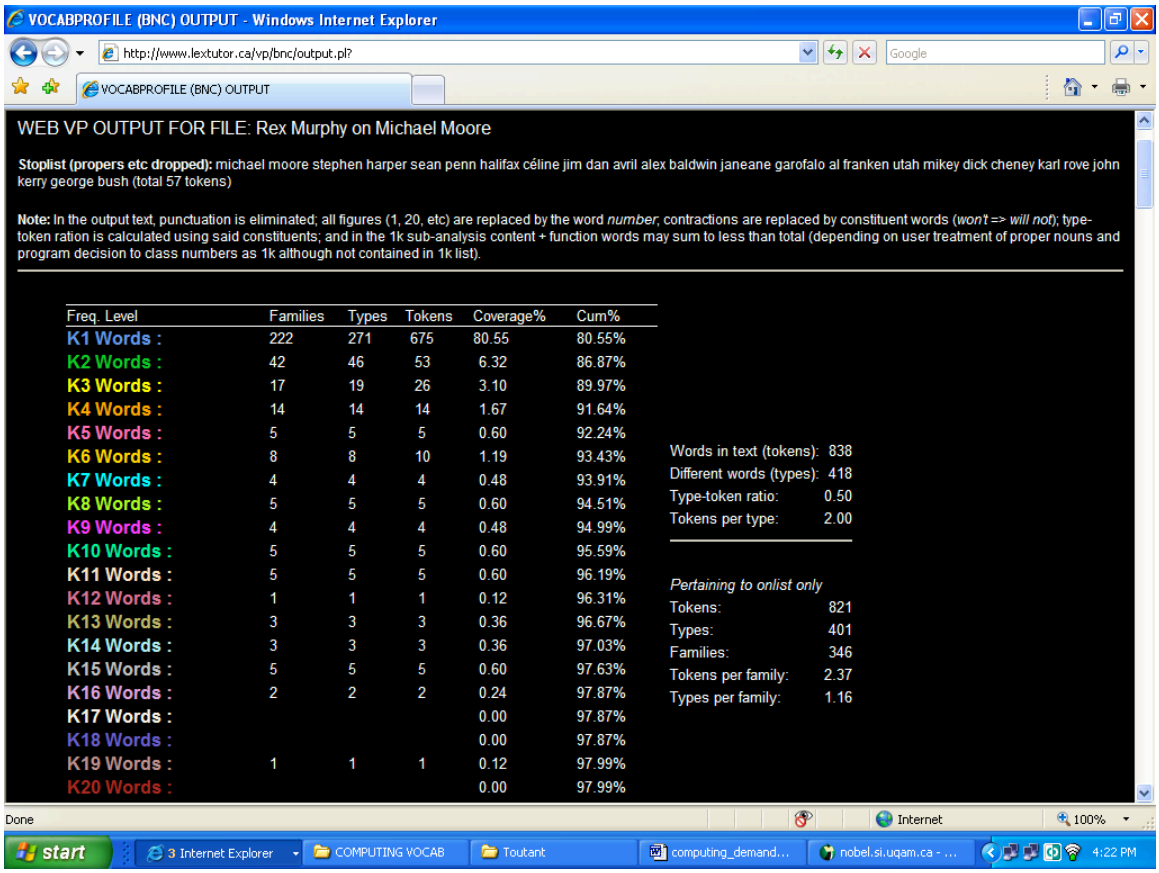

Figure 5. VocabProfile output

Interesting schemes have been proposed for finding existing texts with high degrees of repeated lexis, for example by following one topic through a number of related news stories (Wang & Nation, 1989) or through narrow reading (Schmitt & Carter, 2000). Such schemes have proven able to ensure high degrees of recycling, but only for relatively small sets of words. It seems likely that found texts would have to be supplemented by designed texts to ensure systematic opportunities for vocabulary expansion on a larger scale. A way of testing the amount of lexical repetition in found texts, or creating it through interactive modification, is to use text comparison software that can track large numbers of words through several successive texts. Such a program is TextLexCompare (available at www.lextutor.ca/text_lex_compare), which takes two or more texts as input and gives numbers of repeated and unrepeated words as output. Figure 6 shows two related texts by the same author ready for analysis in the program's dual input windows, namely the first two chapters of the aforementioned *Call of the Wild* by Jack London; Figure 7 shows the output.

The software also provides an experimental *recycling index* (recycled words/total words in the second text), which is currently being calibrated to establish norms of repetition. Initial indications (from the four demonstration texts available on the entry screen) are that the degree of repetition between two unrelated texts by different authors is about 40% of word tokens (largely function words); between unrelated texts by the same author about 60%; and between related or sequential texts by the same author about 70%.

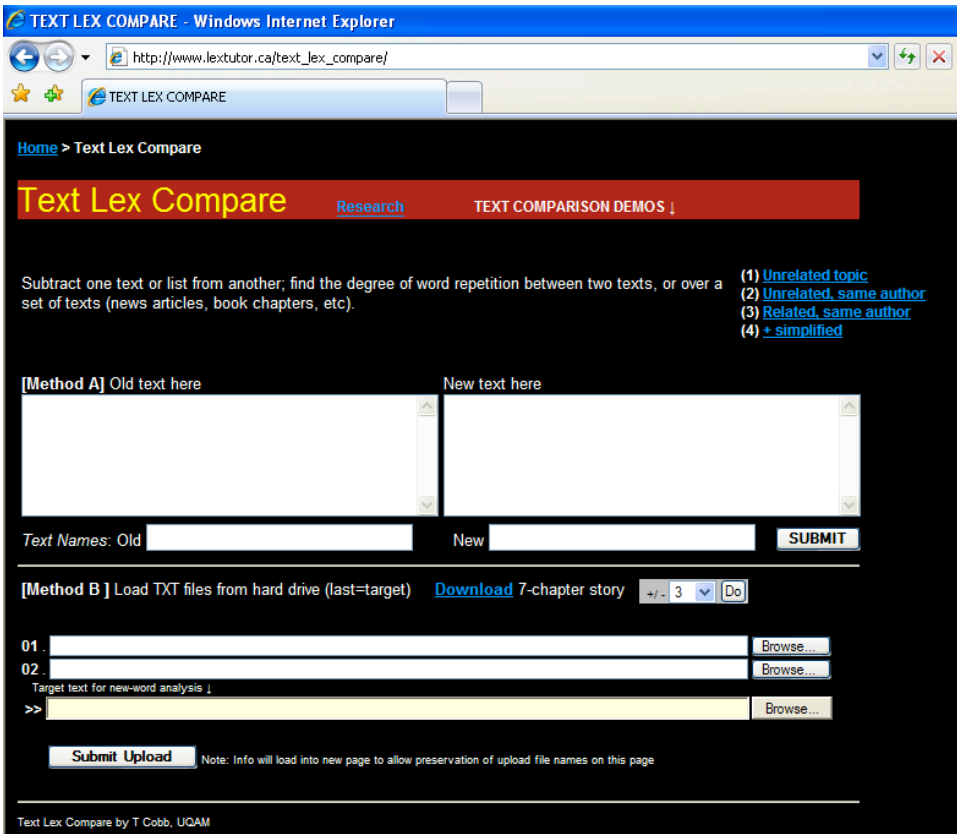

Figure 6. *TextLexCompare* input

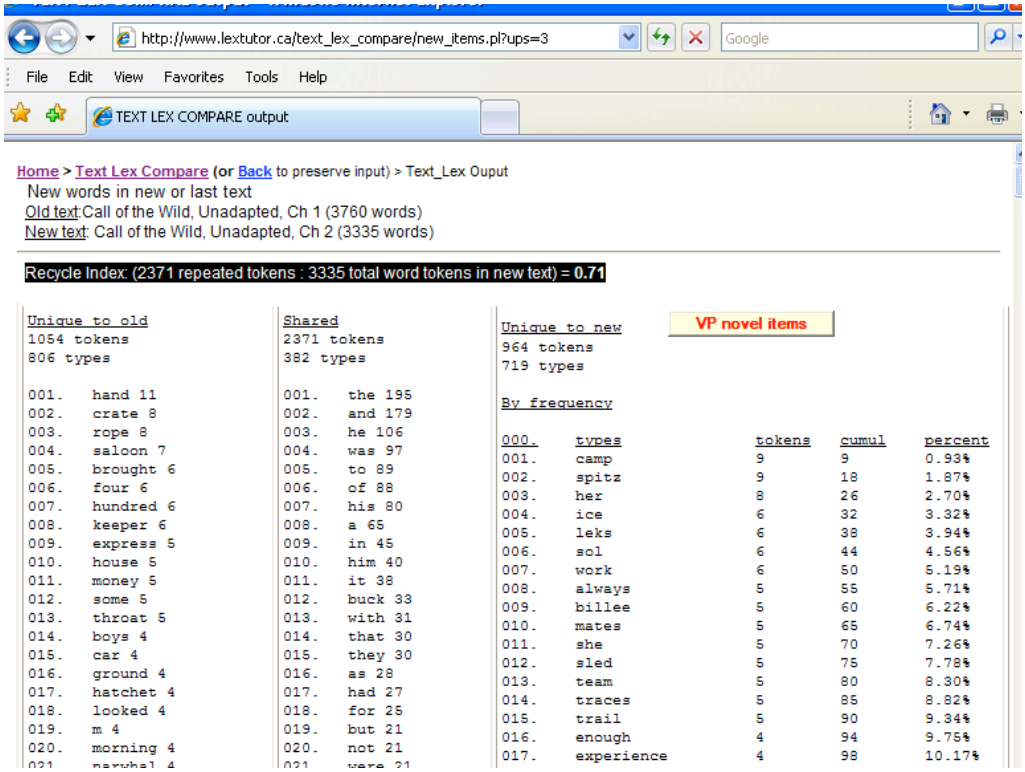

Figure 7. TextLexCompare output

The output in the sample analysis shown in Figure 7 shows that of 3,335 total word tokens in the second chapter of the book, 2,371 are repeated. In other words, a reader will have already met about 70% of the running items in the previous chapter (and about 30% will be 'new'). From a vocabulary learning perspective, this is probably a low proportion of repeated items, as will be outlined below. The provenance of the unrepeated items in frequency terms can be further investigated by clicking 'VP Novel Items' at the top right of the output screen (Figure 7), which is a direct link to VocabProfile with the novel items as input. The VP analysis shows that for these texts 36% of the unrepeated lexis is drawn from the 4,000 to 19,000 frequency zones.

In a narrative text, the rate of recycling should logically increase as the story proceeds. How much does it increase in *Call of the Wild*? To answer this question, tokens in each new chapter were matched against tokens in the combined preceding chapters using the multi-text input feature of TextLexCompare (see bottom half of Figure 6). Results of the analysis for the seven chapters of the London novel are shown in Table 2. The point to notice is that the recycling index never goes above 90% for any chapter. This means that many or most words throughout the story are being met in density environments of one unknown word in 10. This means that many or most words throughout the story are being met in density environments of one unknown word in 10 (double the density that learners can handle, according to Laufer, 1989), and, further, that this situation persists right to the end of the novel. In other words, as it stands this is not a very useful learning text for many L2 readers. However, the text could be modified to become a useful learning text by systematically reducing the flow of novel lexis to a particular level and then increasing the repetition of what remains.

Table 2. Recycling over Several Chapters of Ungraded and Graded Texts

	Call of the Wild - Original				**Call of the Wild - Graded**			
Chapter	Word tokens	New words	Recycled words	Recycle Index (%)	Word tokens	New words	Recycled words	Recycle Index (%)
Ch 1	3727				880			
Ch 2	3275	965	2310	70.53	860	192	668	77.67
Ch 3	5107	946	4161	81.48	1553	189	1364	87.83
Ch 4	3227	383	2844	88.13	1254	92	1162	92.66
Ch 5	5300	846	4454	84.04	1165	107	1058	90.82
Ch 6	4660	698	3962	85.02	1542	135	1407	91.25
Ch 7	6141	744	5397	87.88	1760	105	1655	94.03
MEAN	4491	763.67	3854.67	82.85	1287.71	136.67	1219.00	89.04
SD	1113.95	214.55	1118.11	6.53	347.24	44.00	340.36	5.95

That is what has been done by Longman's writers for its Penguin graded version of *Call of the Wild.* To calculate the success of their reworking of the story, the first seven chapters of the simplified version were fed into TextLexCompare, as was done in the analysis of the original. The results are shown on the right side of Table 2. As can be seen, the recycling index is not only higher overall in the graded than in the original story (89.04% for graded against 82.85% ungraded, (t (16.9), p<.001)), but also the index rises over the course of the story so that in the final chapter the learner is actually meeting new words in an environment of almost 95% previously met words – previously met at least once, that is, with the actual number of repetitions recoverable from the type of data shown in Figure 7.

TextLexCompare or similar software can be used, then, either as an inspection tool to verify the degree of recycling in sequences of found texts, or (in conjunction with VocabProfile) as an interactive aid to principled modification. To summarize, a fairly simple computer-aided in-house procedure for turning sequences of natural texts into sequences of learning texts is as follows: use diagnostic testing to

determine students' growth zone (or i+1) in terms of families; find text sequences that have a high proportion of words from this frequency zone; use VocabProfile to *write out* as many words as possible that fall beyond this level; and use TextLexCompare to *write in* more of the same or other words from this level, with the goal of reaching a recycling ratio of 95% well before the end of the story.

This procedure clearly presupposes that candidate texts are available in machine readable format, as indeed they increasingly are. This format also offers two further opportunities. First, as noted above, there is a dearth of graded materials in the world of ESL (and no doubt other languages) generally, and of both expository materials and post-3000-focused materials in particular. Using the scheme of frequency based tests and tools outlined above, it is possible in principle to organize a large online repository of graded text materials categorized by size, text type, target level, and recycling schedule.

Computer-aided Enrichment of Undesigned Texts

More than 10 years ago Cobb and Stevens (1996) argued that the flood of text then poised to go online should be a boon for L2 learners. They proposed a number of ways that computers would be able to deliver an expanded supply of text, and also enhance learning through various kinds of processing of the text before, during, and after delivery. It now seems safe to say that the amount, quality, diversity, and availability of such text have exceeded expectations. It seems only a little less safe to say that the vast majority of reading undertaken by people who have grown up in this period now takes place on a computer screen. And yet it is not clear that the computer, for its part, is as of yet serving as more than the delivery vehicle. This is unfortunate, because just as the text was more than expected, so are the opportunities for computers to do more than simply download, display, and distribute texts.

The Internet in its current form contains dictionaries, pictures, routines that turn text into speech, and many more resources that can, in principle, be usefully linked to learners' texts. In the bewildering array of opportunities, the challenge is to evolve rationales for selecting and sequencing resources and to provide smooth, principled access to them. The remainder of this paper presents a number of worked-out ideas for integrating texts with Internet based text and text-based resources. As before, these centre on ways of increasing the number of encounters with words met in texts, but here the approach is external rather than internal. That is, word learning opportunities are enhanced through networking the text to resources beyond the text rather than through modifying the text itself. The earlier discussion of text modifications dealt with the "before" aspects; let us turn now to "during" and "after."

Solution 1: Link Texts to Other Texts

Suppose a learner reads a text containing a number of new words that occur only once (Horst's, 2000, study involved learning 300 singly occurring items from a short literary work). The chances of the L2 reader retaining many of these, unless participating in a repeated readings experiment, are minimal. But what if the occurrence of each of these words could be expanded during or after reading by connecting it to the same word as it appears in a range of other texts, i.e. in a corpus?

In principle, a collection of texts is able to disclose patterns in language that go beyond what can normally be perceived in more naturalistic situations in the short term. A pattern of relevance to the current discussion is the broader meaning of a lexical item, which can often be disclosed only over several occurrences.

The usability of corpus information by language learners is a topic of interest in language learning research (e.g., Cobb, 1997; Cobb, 1999a; Gaskell & Cobb, 2004). A potentially useful finding in this research is that even beginning learners, provided with a clear learning objective and usable concordance interface, appear to be able to use this information as a replacement for the process of meeting words repeatedly in texts over the course of a great deal of reading. Research by Cobb (1999b) showed that learners who used concordance information in a dictionary building project were more able to transfer new word learning to novel contexts than controls who had used glossary information.

A way of using the findings of this research is to link concordances to learner texts as a learning resource. For example, a learner could click on a word and get a concordance of additional examples or conceptualizations of the word. This is the approach in an experimental series of programs available on Lextutor at www.lextutor.ca/hypertext. The learner pastes the text she is reading into a text box, and the program transforms it into a text with every word click-linked to a concordance from the Brown or other corpus and from there to an online dictionary. The proposed interaction is shown in Figure 8, where the learner has clicked on a key word from the passage, usage, and been given 14 further instances of the word from Brown. From the concordance, there is a further link to a choice of dictionaries (in this case, the online version of Princeton's Wordnet).

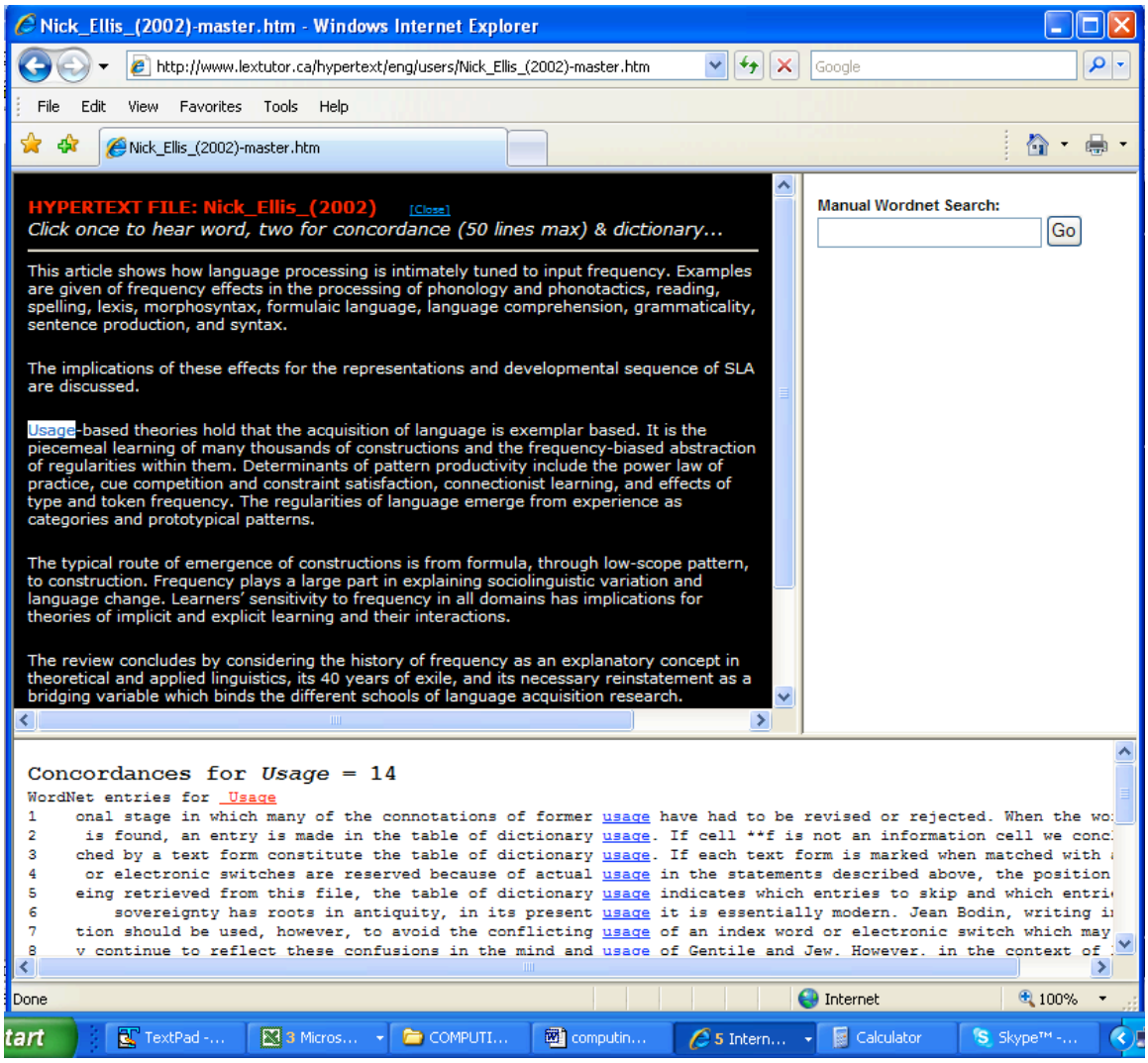

Figure 8. A concordance-linked text

Figure 8 shows this concordance-linked idea in its most basic form. Variations on the theme can be made available along the following four dimensions: (1) when the corpus information will be consulted -- during or after reading; (2) how actively the corpus information will be processed; (3) whether a general or a text-related corpus provides the best learning opportunities; and (4) whether learners are able to build their own concordances from the 'corpus' of their reading assignments. Experiments in implementing several of these ideas can be seen in Figure 9, which shows a resource enriched version of a text discussed

earlier, Chapter 1 of *Call of the Wild*. (This original version was described previously as unsuitable for learners without modification or enrichment[6].)

During or after?

The reader can double-click for a concordance while reading, or Alt-click the word (click while holding the keyboard's Alt-key) to place it in the "Wordbox" for further attention later. There is no need to interrupt the reading to avoid losing the word, a common problem for lexically alert L2 readers. Once in the Wordbox, the words stored after a reading session can be saved or used in a number of ways.

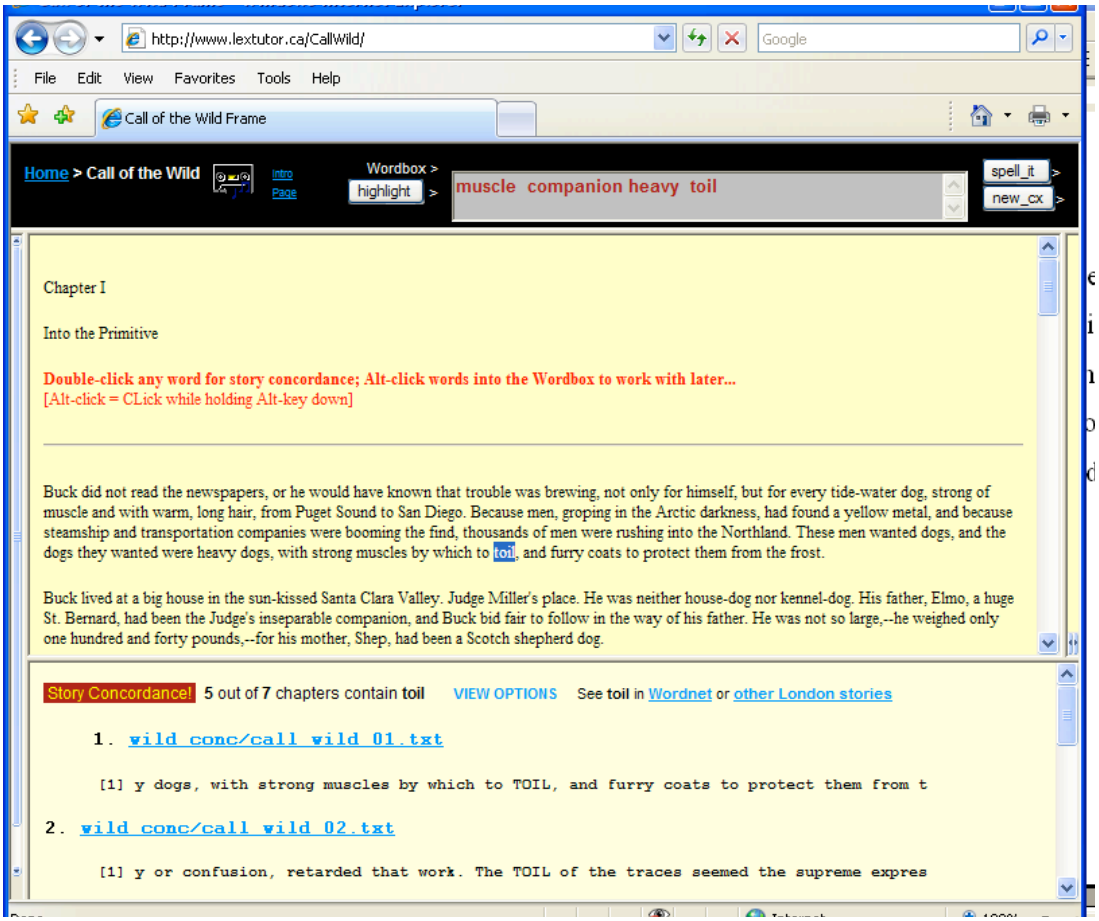

Figure 9. Story concordance 1

Active processing

The words in the Wordbox can be sent via a click of the *new_cx* (new contexts) button to a concordance transfer exercise (shown in Figure 10). The words are integrated behind the scenes by a Multiconcordance routine (at www.lextutor.ca/multi_conc) into a multiple concordance quiz, where each of the words must be returned to the correct Brown or other corpus lines from which the program has removed them. This type of post-reading vocabulary activity is comparable to Wesche and Paribakht's (1996) reading-plus activity except that it is learner generated and can be devised instantly from any machine readable text.

Figure 10. Transferring words to novel contexts

Targeted corpora

Figure 9 also shows the learner's other option, which is to double click for a concordance. However, the corpus used to produce the concordance in this case is simply the rest of the text the learner is currently reading broken down by chapters. For example, the word *toil* is shown as appearing in five further chapters of the story for a total of 15 occurrences, giving the learner some idea of how much investment it is worth. From this interface a dictionary can also be accessed, as well as concordance lines from a further corpus comprising several further works by the same writer. Figure 11 shows another interesting word, *progeny,* as it appears in other London stories.

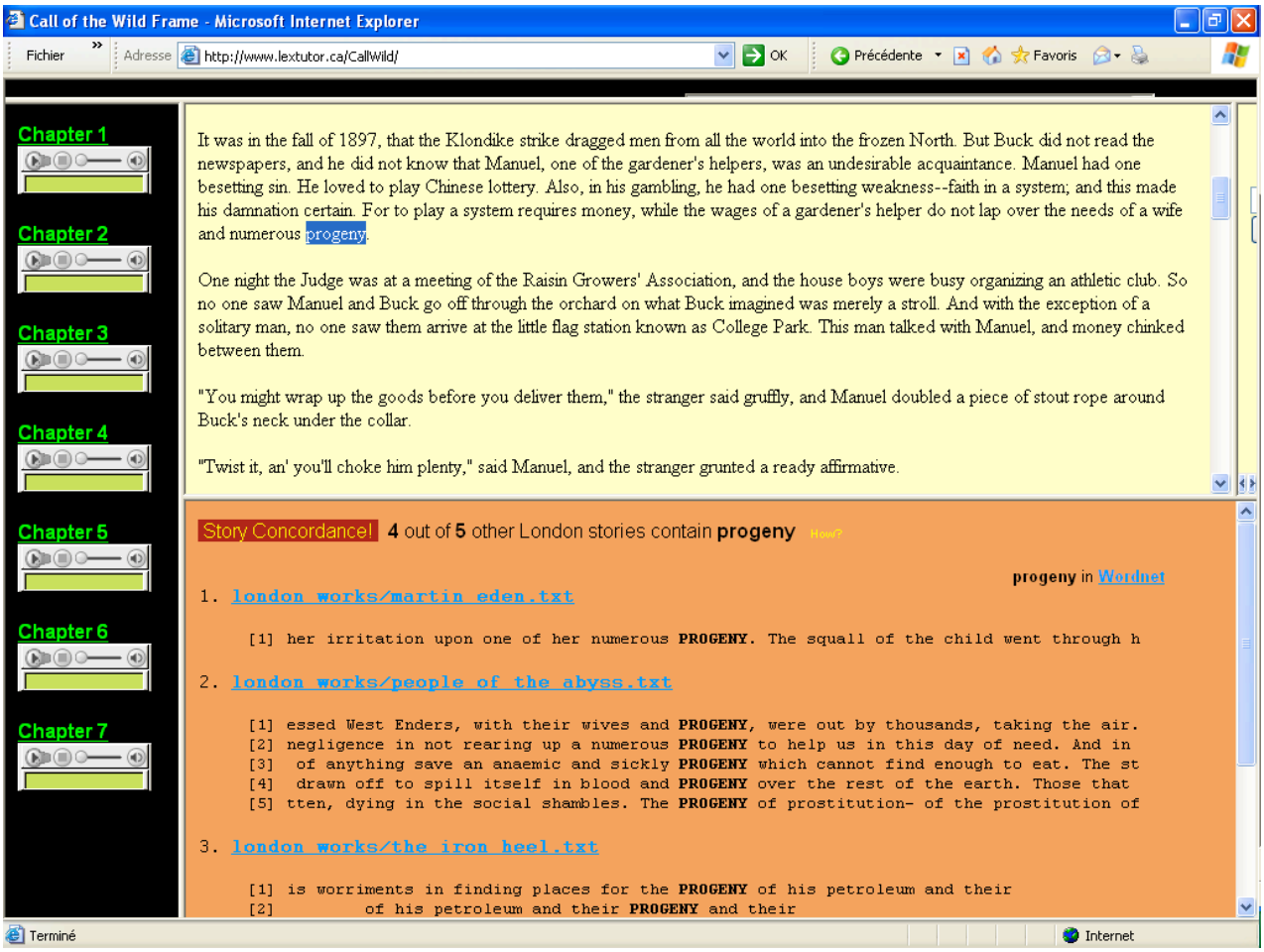

Figure 11. Story concordance 2

It should be noted that while these examples come from a literary text, any text broken into chapters or sections could be employed using the same technology. Further context and empirical validation of concordance-linked reading can be found in Cobb, Greaves and Horst (2001) and in Cobb (2006).

Learner-generated concordances

Another approach to the target concordances idea involves letting groups of learners develop their own concordances for words in a text or set of texts. This can be achieved through the use of a Group Lexical Database (shown in Figure 12, and online at www.lextutor.ca/group_lex/demo/).

The database allows learners to put together their lexical acquisitions, again as a way of overcoming the slow pace of natural occurrences on an individual basis. Learners enter examples and definitions for words from whatever they are reading, independently or as directed by a teacher. The teacher can specify the number of examples to be provided for each word, although learners often provide several examples spontaneously (as in the case of *inspect, intermingling,* and *invigorated* in Figure 12). The program also provides opportunities for active processing in the form of a quiz option, for learner selected items, in which key words must be replaced in the example sentences provided by another learner (as shown in Figure 13) or from a corpus (accessed via the 'Tougher Quiz' option). Positive research on the group-lex concept is reported in Horst, Cobb and Nicolae (2005).

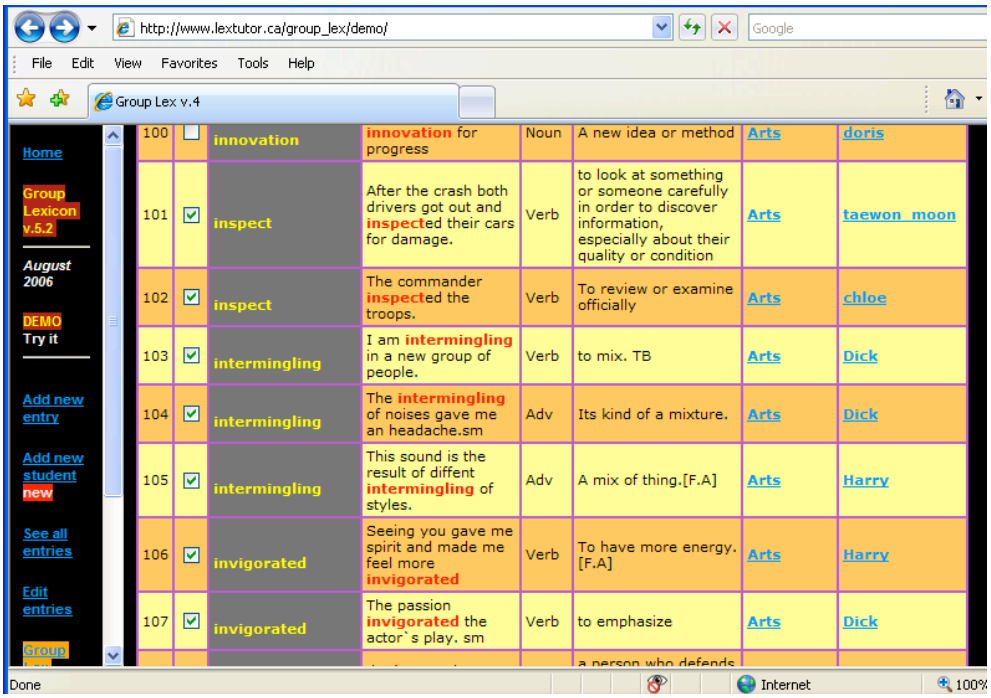

Figure 12. Learner generated concordances

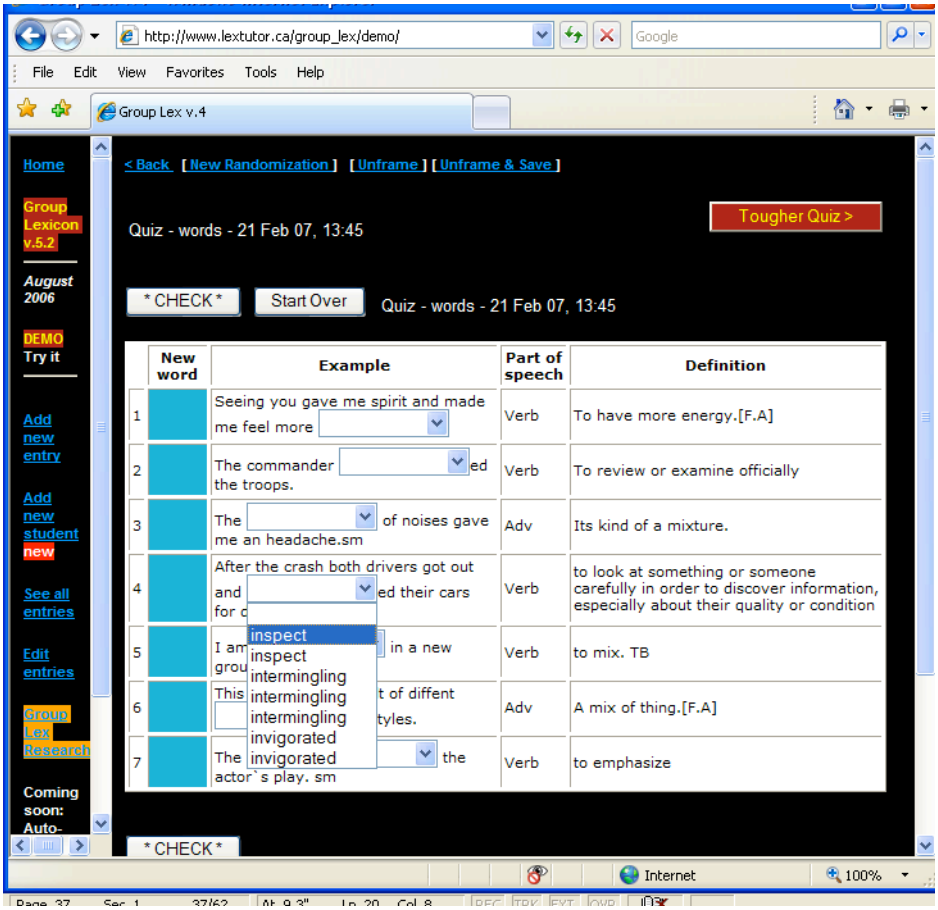

Figure 13. Learner generated quizzes

To summarize, these are concrete suggestions for using networked computing to link learners' texts to the broader world of online text and to overcome some of the limitations inherent in natural text input.

Solution 2: Link Texts to Speech

In addition to the world of text, the world of online speech is a rich resource that can be used to multiply learners' exposure to new words.

As has already been shown, the occurrences of words in texts beyond the 2000 frequency level thin out rapidly such that ensuring six encounters becomes problematic. It is also well known that words in the medium and lower frequency ranges appear mainly in text rather than in speech (e.g., Stanovich & Cunningham, 1992). However, it is not the case that lower frequency words *never* appear in speech, and in fact it is quite likely that language learners are both reading and hearing some of the same medium frequency words. But are they aware of it? It is probably safe to say that learners do not always recognize words in speech that they have met once or twice in texts, and hence that they miss learning opportunities that are actually on offer in the speech environment. How many opportunities might they miss?

The present analysis again involves the program Range, but this time we are using the part of the program (see Figure 2, top left) that compares distributions of lexical items between written and spoken corpora. The corpora for this analysis are the roughly equal (about 1 million words apiece) written and spoken BNC sampler sub-corpora (Oxford University Press, 1998). The 36 sample items to be tested in these corpora are sets of six randomly selected words from each of the following six frequency levels: the 100 most frequent content words of English, and then the 1000, 2000, 3000, 4000, and 5,000 BNC frequency lists. Numbers of written and spoken occurrences of sample items in the first four of these lists are shown in Table 3.

It is to be expected that very high frequency words will be more common in speech than writing, and conversely that lower frequency items will be more common in writing than speech, in rough proportion to frequency. This is what Table 3 shows, as well as some interesting crossover points.

Table 3. Occurrences in Speech vs. Writing of Words at Different Frequency Levels

First 100 word families, per million			**1000 level word families, per million**			**2000 level word families, per million**			**3000 level word families, per million**		
Word family	in writing	in speech	Word family	in writing	in speech	Word family	in writing	in speech	Word family	in writing	in speech
look'	1336	2124	welcom'	122	53	accus'	64	7	glor'	62	20
think'	638	3703	sun'	403	104	accustom'	42	2	stain'	17	2
people'	1211	1757	produc'	768	334	ach'	5	5	collar	14	13
out	1880	2635	worth'	160	224	admir'	61	14	widow'	42	3
up	2087	3074	Try'	195	342	afford'	43	78	chew'	8	12
back'	1224	1562	somebody	50	358	Alike	20	4	appal'	17	10
MEAN	1396.00	2475.83		283.00	235.83		39.17	18.33		26.67	10.00
SD	474.99	748.85		242.46	120.26		21.02	26.95		19.06	6.14
% total	36.05%	63.95%		54.55%	45.45%		68.12%	31.87%		72.72%	27.27%

A provisional pattern can be generalized from the Table 3 means. Very frequent (first 100) English words are almost twice as common in speech as they are in writing. Frequent words (at the 1000 level) are more or less equally common in speech and writing. Medium frequency (2000-level) words are over twice as common in writing. 3000-level items are 2.5 times as common in writing, with the predominance of writing occurrences increasing with each thousand-level (not shown) until the pattern stabilizes at about five written per spoken occurrence after the 5000-word level. [7]

What this means for the learner who is listening and reading in equal amounts at the same content level, working on a post-2000 lexicon, and not consistently mapping phonemes to morphemes, is that numerous

occurrences of the same word are being lost as learning opportunities. For 2000-level words, the loss could be as much as one-third; for 3000-level words as much as one-quarter; and so on.

The lost opportunities could be even greater in learning contexts where exposure to sophisticated oral text is high; an example is university ESL learners listening to academic lectures in content courses. Testing of the sample items used in the comparison above with an academic spoken corpus bears this out. The items shown in Table 3 occurred almost twice as often per million in the Michigan Corpus of Academic Spoken English (Simpson, Briggs, Ovens, & Swales, 2003) than they did in the more general BNC spoken corpus.

Just from the perspective of added exposure to words in context, it is clear that most learners would benefit from being able to recognize the words they are meeting in texts when they hear them spoken. How could this happen? A low-disruption way of integrating speech information into learners' online texts is through text-to-speech (TTS) technology. This technology is now able to provide acceptable pronunciations for 99% of words and is available to Web-based documents. Further, TTS can be scripted into these documents in several different ways.

The most basic implementation of TTS on Lextutor simply transforms a learner's text into a speaking text. This is achieved by pasting any text into a window at www.lextutor.ca/tts/ and clicking *Submit*. Lextutor uses the cost-free DirectExtras speech plug-in and voices from www.speaksforitself.com, which users must download onto their own or institutional machines in order to hear their texts spoken. Once they have done this, they can hear any word they click on. But TTS can be integrated with other resources in more interesting ways. For example, in the *Call of the Wild* reading activity shown in Figure 11, two clicks on a word produces the story concordance, while a single click produces the speech rendition. TTS is integrated as a resource option into several Lextutor routines, including those shown above (Group Lex and Hypertext) and others.

Another approach to joining text and sound is to attach a text file to a recorded sound file. The Internet has many sources of such multimedia texts. One of the Hypertext routines (http://www.lextutor.ca/hypertext/eng_2) is capable of making the link, provided the user has been able to locate the address of the sound file. The dictionary-linked reading page that results is shown in Figure 14, one-click linked to the TTS rendition for individual words, two-clicks linked to the *Cambridge Advanced Learner's Dictionary*, with the sound file running through the player at top right of the screen.

In Figure 14 there is again a Wordbox, which can either just store learners' words for their private purposes or send them to a reading-plus activity. The *Meaning* button sends them to a meaning-based Multiconcordance activity, as already described, but the *Spelling* button sends them to a TTS-based spelling activity called Dictator. This routine transforms any set of learner selected words into a TTS based spelling dictation activity, in either practice or test formats. In Figure 15, a learner has created a training exercise to practice spelling the words he or she hears. The learner clicks a word to hear it, tries to spell it, and is given help with any errors. The help is provided by the resident Guidespell tutor (Cobb, 1997), which does not give away the correct answer but rather tells the student how many letters were correct in the attempt to spell *accompany*. When ready, the learner can enter the same words into a Test version of the program, which gives no help but simply a score when all words have been entered. Learners can access Dictator directly (at http://www.lextutor.ca/dictator/) or via Wordboxes in several Lextutor routines. (Since the time of writing, the two Reading-Plus activities accessed from the Wordbox have been joined by a third, a practice activity for timed lexical access.)

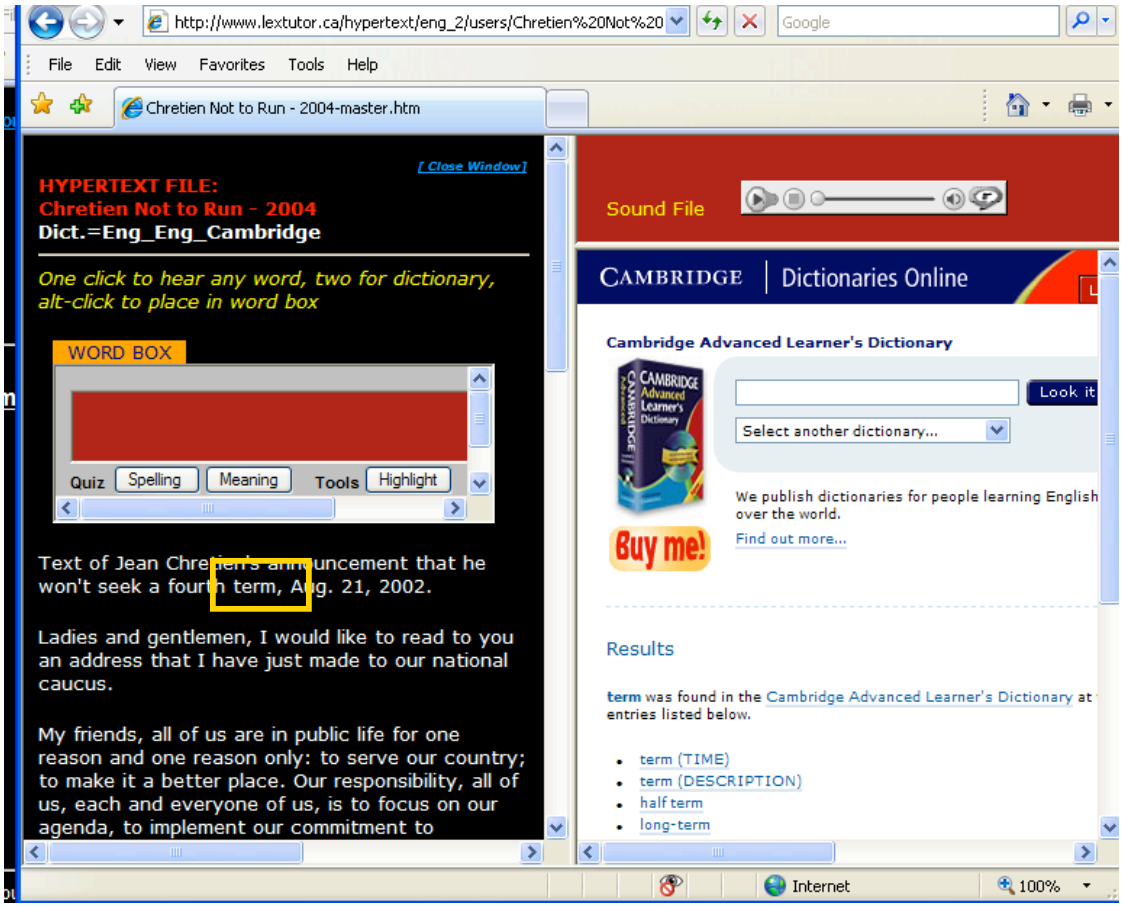

Figure 14. 2-way sound

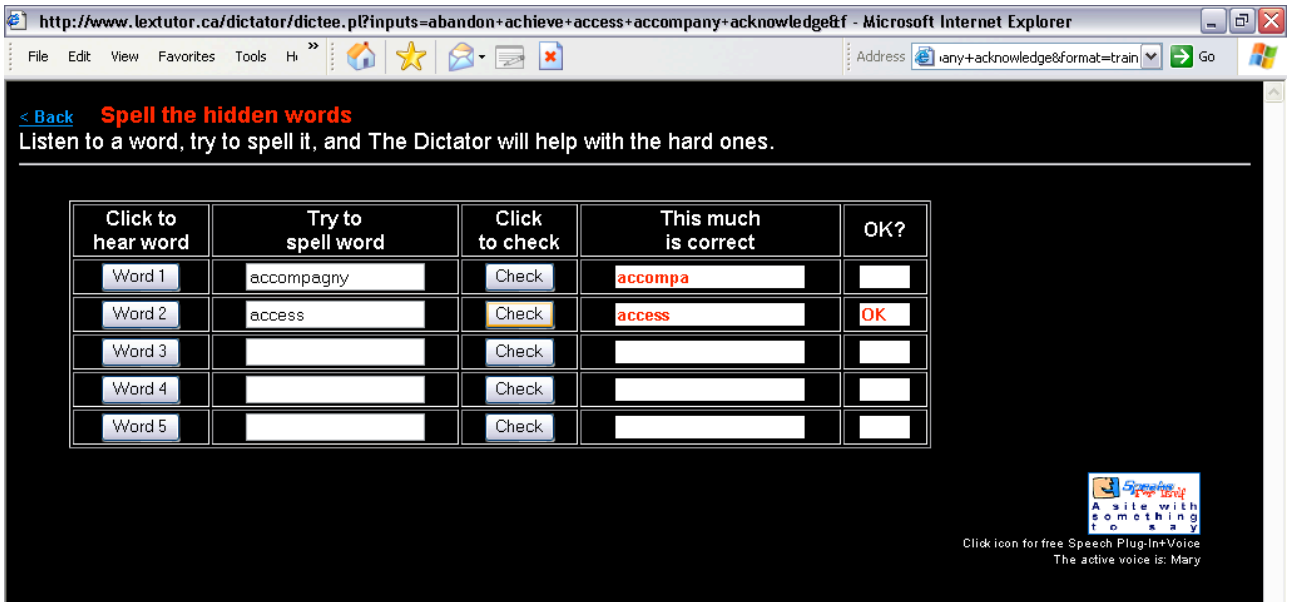

Figure 15. Dictator training activity under way

To summarize, this is only a small sample of the ways the interlinked universe of texts can be employed to multiply the quantity and quality of input for L2 vocabulary learning.

CONCLUSION

The key problems of learning through extensive reading are clear. Corpus analysis shows that words beyond the 2000 most frequent are unlikely to be encountered in natural reading in sufficient numbers for consistent learning to occur. Lexical profile analysis shows that the amount of new vocabulary in natural texts is at odds with both the lexical level and learning capacity of most learners. Text comparison analysis further shows that the rate of new word introduction in a text designed for native speakers is far more than most L2 learners will be able to cope with. And yet these same tools can also be employed positively to multiply learning opportunities, whether by facilitating the adaptation of texts that learners can read and learn from, or by habilitating unadapted texts through external resourcing.

In a classic paper on vocabulary growth from reading, Krashen (1989) remarked that a number of books can be purchased for the price of one computer, thereby implying that the books were the wiser choice. In 2007 books and computers are less a choice than a partnership. I hope I have convinced the reader that the role of computing in L2 reading instruction can and should go well beyond the functions of delivery, distribution, and printing.

NOTES

1. For a review of the approaches to vocabulary testing that Krashen may be referring to, see Read (2000) or visit www.lextutor.ca/tests/ for computer versions of several of these tests.

2. See Cobb (in press), Gardner (2004), and Grabe (1991), among others, on the limitations of applying L1-based reading research to L2 contexts.

3. Readers can create similar sets and replicate this experiment for themselves at http://www.lextutor.ca/rand_words/.

4. The full text of this novel and its frequency profile analysis can be seen at http://www.lextutor.ca/callwild/.

5. These 20,000 families arguably define the complete non-specialist lexicon of English.

6. This text and the linked resources are available at www.lextutor.ca/CallWild/.

7. This is admittedly, a rough picture. The standard deviations for speech and text means are large at all levels, probably lending support to Biber's (1988, p. 162-3) view that "there is no single dimension of orality versus literacy" in the language at large.

ACKNOWLEDGMENTS

The research and development work discussed in this paper was generously supported by the Quebec government through the Centre for the Study of Learning and Performance at Concordia University in Montreal. Many of the software routines elaborated on the Lextutor website are based on programs and ideas developed by Paul Nation and his students and colleagues at Victoria University of Wellington in New Zealand. Welcome assistance with the writing was provided by LLT reviewers and editors.

ABOUT THE AUTHOR

Tom Cobb has designed, taught, and coordinated almost every type of English reading and writing course possible in a career spanning 20 years and five continents. He was convinced quite early that whatever the

target skill, there would never be enough time for language learners to get very far with it, but that well instructed computers could radically increase the effectiveness of the time available. He now consults in language program development internationally, supplies learning and research tools to the profession through his website, The Compleat Lexical Tutor (http://www.lextutor.ca), and helps young Montreal ESL teachers get the most out of computers in their classrooms.

Email: cobb.tom@uqam.ca

REFERENCES

Alderson, C. (1984). Reading in a foreign language: A reading problem or a language problem? In J. Alderson & A. Urquhart (Eds.), *Reading in a Foreign Language* (pp. 1-24). London: Longman.

Barnard, H. (1972). *Advanced English Vocabulary: Workbooks*. Rowley, MA: Newbury House.

Bernhardt, E. (2005). Progress and procrastination in second language reading. In M. McGroarty (Ed.), *Annual Review of Applied Linguistics, 25* (pp. 133-150). New York: Cambridge University Press.

Biber, D. (1988). *Variation across speech and writing.* London: Cambridge University Press.

Cobb, T. (1997). Is there any measurable learning from hands-on concordancing? *System 25*(3), 301-31.

Cobb, T. (1999a). Breadth and depth of vocabulary acquisition with hands-on concordancing. *Computer Assisted Language Learning 12*, 345 - 360.

Cobb, T. (1999b). Applying constructivism: A test for the learner-as-scientist. *Educational Technology Research & Development 47*(3), 15-33.

Cobb, T. (2003). Analyzing late interlanguage with learner corpora: Quebec replications of three European studies. *Canadian Modern Language Review*, 59(3), 393-423.

Cobb, T. (2006). Internet and literacy in the developing world: Delivering the teacher with the text. *Educational Technology Research & Development 54*(6), 627-645.

Cobb, T. (In press.) Necessary or nice? The role of computers in L2 reading. In Z. Han & N. Anderson (Eds.), *Learning to Read & Reading to Learn* (tentative title), for TESOL.

Cobb, T., Greaves, C., & Horst, M. (2001). Can the rate of lexical acquisition from reading be increased? An experiment in reading French with a suite of on-line resources. In P. Raymond & C. Cornaire (Eds.), *Regards sur la didactique des langues secondes* (pp. 133-153). Montréal: Éditions logique.

Cobb, T., & Stevens, V. (1996) A principled consideration of computers and reading in a second language. In M. Pennington (Ed.), *The power of CALL* (pp. 115-136). Houston: Athelstan.

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

Elley, W. (1991). Acquiring literacy in a second language: The effect of book-based programs. *Language Learning 41*(3), 375-411.

Gardner, D. (2004) Vocabulary input through extensive reading: A comparison of words found in children's narrative and expository reading materials. *Applied Linguistics 25*(1), 1-37.

Gaskell, D., & Cobb, T. (2004). Can learners use concordance feedback for writing errors? *System 32*(3), 301-319.

Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics 11*(4), 341-358.

Grabe, W. (1991). Current developments in second language reading research. *TESOL Quarterly 25*, 375-406.

Heatley, A. & Nation, P. (1994). *Range*. Victoria University of Wellington, NZ. [Computer program, available at http://www.vuw.ac.nz/lals/.]

Hill, D. (1997). Setting Up An Extensive Reading Programme: Practical Tips. The Language Teacher Online. Retrieved September 21, 2007, from http://www.jalt-publications.org/tlt/files/97/may/hill.html.

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

Horst, M. (2000*). Text encounters of the frequent kind: Learning L2 vocabulary from reading.* University of Wales (UK), Swansea: Unpublished PhD dissertation.

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.

Horst, M., Cobb, T., & Nicolae, I. (2005). Expanding academic vocabulary with a collaborative on-line database. *Language Learning & Technology, 9*(2), 90-110.

Horst, M., & Meara, P. (1999). Test of a model for predicting second language lexical growth through reading. *Canadian Modern Language Review, 56*(2), 308-328.

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

Krashen, S. (2003). *Explorations in language acquisition and use: The Taipei lectures*. Portsmouth, NH: Heinemann.

Kucera, H., & Francis, W. (1979). *A Standard Corpus of Present-Day Edited American English, for use with Digital Computers* (Revised and amplified from 1967 version). Providence, RI: Brown University Press.

Laufer, B. (1989). What percentage of text-lexis is essential for comprehension? In C. Lauren & M. Nordman (Eds.), *Special language: From humans thinking to thinking machines* (pp. 316-323). Clevedon, UK: Multilingual Matters.

Leech, G., Rayson, P., & Wilson, A. (2001). *Word frequencies in written and spoken English.* London: Longman.

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

London, J. (1903) *Call of the Wild.* Serialized in *The Saturday Evening Post*, June 20-July 18.

London, J. (1906) *White Fang.* Serialized in *The Outing Magazine*, May-Oct.

Mezynski, K. (1983). Issues concerning the acquisition of knowledge: Effects of vocabulary training on reading comprehension. *Review of Educational Research, 53*, 253-279.

Mondria, J-A., & Wit-De Boer, M. (1991). Guessability and the retention of words in a foreign language. *Applied Linguistics, 12*(3), 249-263.

Mondria, J-A. & Wit-De Boer, M. (1993). Efficiently memorizing words with the help of word cards and 'hand computer': Theory and applications. *System 22*, 47-57.

Nagy, W. (1988). *Teaching vocabulary to improve reading comprehension.* Newark, DE: International Reading Association.

Nation, P. (2001). *Learning vocabulary in another language.* Cambridge: Cambridge University Press.

Nation, P. (2006). How large a vocabulary is needed for reading and listening? In M. Horst & T. Cobb (Eds.), *Second special vocabulary edition of Canadian Modern Language Review*, *63*(1), 1-12 .

Nation, P., & Wang, K. (1999). Graded readers and vocabulary. *Reading in a Foreign Language, 12*(2). Retrieved January 30, 2007, from http://nflrc.hawaii.edu/rfl/PastIssues/rfl122nation.pdf.

Oxford University Computing Services. (1995). *The British National Corpus.* Oxford: Oxford University Press.

Paribakht, T.S., & Wesche, M. (1997). Vocabulary enhancement activities and reading for meaning in second language vocabulary acquisition. In J. Coady & T. Huckin (Eds.). *Second language vocabulary acquisition: A rationale for pedagogy.* Cambridge: Cambridge University Press.

Penguin Readers. London: Longman. Information from http://www.penguinreaders.com/.

Read. J. (2000). *Assessing vocabulary.* London: Cambridge University Press.

Redman, S., & Ellis, R. (1991). *A way with words: Vocabulary development activities for learners of English, Books 1, 2, 3.* Cambridge: Cambridge University Press.

Simpson, R., Briggs, S., Ovens, J., & Swales, J. M. (2003). *The Michigan Corpus of Academic Spoken English.* Ann Arbor, MI: The Regents of the University of Michigan. Retrieved February 07, 2007, from http://www.hti.umich.edu/m/micase/

Schmitt, N., & Carter, R. (2000). The lexical advantages of narrow reading for second language learners. *TESOL Quarterly 9*(1), 4-9.

Schmitt, N., & Schmitt, D. (2004). *Focus on academic vocabulary: Word study from the Academic Word List.* New York: Longman.

Schmitt, N., & Zimmerman, C. (2002). Derivative word forms: What do learners know? *TESOL Quarterly 36*(2), 145-171.

Stanovich, K.E., & Cunningham, A.E. (1992). Studying the consequences of literacy within a literate society: The cognitive correlates of print exposure. *Memory & Cognition, 20,* 51-68.

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

Venezky, R. (1982). The origins of the present-day chasm between adult literacy needs and school literacy instruction. *Visible Language 16*(2), 113-136. Reprinted (2000) in *Scientific Studies of Reading, 4*(1), 19-39.

Waring, R., & Takaki, M. (2003). At what rate do learners learn and retain new vocabulary from reading a graded reader? *Reading in a Foreign Language, 15*(2). Retrieved January 30, 2007, from http://nflrc.hawaii.edu/rfl/October2003/Waring/waring.html

Waring, R., & Nation, P. (2004). Second language reading and incidental vocabulary learning. *Angles on the English Speaking World, 4*, 11-23.

Wesche, M., & Paribakht, S. (1996), Assessing vocabulary knowledge: Depth vs. breadth. *Canadian Modern Language Review*, 53(1), 13-40.

Zahar, R., Cobb, T., & Spada, N. (2001). Acquiring vocabulary through reading: Effects of frequency and contextual richness. *Canadian Modern Language Review*, 57(4), 541-572.