Vocabulary sizes of some City University students.

By Tom Cobb & Marlise Horst, Division of Language Studies, City University of Hong Kong

REF: Cobb, Tom and Marlise E. Horst (1999). Vocabulary Sizes of some City University Students. *Journal of the Division of Language Studies of City University of Hong Kong*, 1 (1), 59-68.

ABSTRACT

Two groups of City University of Hong Kong (Division of Language Studies) students, one first year and one second year, were tested on their ability to recognize the meanings of high-frequency English vocabulary. All subjects scored high on tests of the 2000 and 3000 most common words, but lower scores on the test of the University Word List suggest that some students may not have the word knowledge they need either to read authentic texts efficiently or to infer the meanings of the new words they encounter. This impression was confirmed by the finding that second year students scored only slightly higher than first year students on the test at the 5000-word level. It was further confirmed by the finding that no vocabulary growth occurred during a six month period in the first year group.

Deux groupes d'étudiants en langues de la City University à Hong Kong ont participé à une évaluation de leur capacité de reconnaître des mots de vocabulaire de grande fréquence. Les étudiants des deux groupes-dont un de première année et l'autre de deuxième année-ont bien réussi quand il s'agissait des items figurant parmi les 2000 or 3000 mots les plus communs par ordre de fréquence, mais certains performaient moins bien quand il s'agissait des mots faisait partie de la University Word List, la liste des mots jugés nécessaires à la réussite universitaires. Ce résultat nous fait croire que certains étudiants n'ont pas suffisamment de mots de vocabulaire pour lire et comprendre les textes qui font partie de leur programme d'études. Cette impression est confirmée par le fait que les résultats des étudiants de deuxième année étaient à peine supérieur aux résultats des étudiants de première année quand il s'agissait de leur connaissance des mots qui figuraient au niveau des 5000 par ordre de fréquence, autrement dit, les mots beaucoup moins fréquents. Une autre confirmation de cette tendance vient du fait que nous avons trouvé un manque de croissance dans le vocabulaire des étudiants de première année sur une période de six mois.

INTRODUCTION

How many English words does a university student in Hong Kong need to know? There are several possible ways to answer this question. One is to look at what the students themselves aspire to. Hong Kong university students appear to have very high expectations, and in our classes some have expressed hopes of reaching native-like levels of proficiency. One of the subjects in the study noted in a journal entry that he hoped to "reach the high standard and compare to the native speaker." Goulden, Nation and Read (1990) have determined that the vocabulary size of an average native-English-speaking university student is about 17,000 word families (a word family being a baseword together with its derived forms, e.g. happy, unhappy, happiness), or as many as 40,000 different word types. Clearly our Hong Kong learners' hopes are very high indeed.

Another way of answering the question might be to consider the number of words needed to speak English at a native-like level. Many Hong Kong students appear to recognize the importance of speaking and seek opportunities for oral practice. For instance, in the sample of university students discussed below, over half expressed appreciation for speaking activities and/or a desire for more of these in course evaluations. Many courses emphasize speaking skills, and the oral presentation appears to be a fixture in many classrooms. No doubt this is in an attempt to redress the scarcity of opportunities for oral expression in secondary schools where classes are large, instruction is teacher-centered, and opportunities to speak are limited, as Johnson and Yau (1996) have observed. But spoken language draws on a small supply of words. In one study, native speaker teachers in classrooms were found to rarely stray beyond the 2500 most common basewords of English (Meara, Lightbown & Halter, 1997) in their speech, and ordinary native-speaker conversations are also likely to stay within a limited range. It is clear that if the goal is to have enough vocabulary to attain native-like speech, then the number of words students need to know is much smaller than the 17,000 figure above.

Yet another way to answer the question is to assume that Hong Kong students need to know enough words to be able to read authentic, university-level textbooks in English. Using a reading criterion makes sense for two reasons. One is that the use of course texts designed for native speakers appears to be the norm at Hong Kong universities. Admittedly, help with difficult texts is often available in the form of pre-teaching in Cantonese or predigested point-form notes (Johnson & Yau, 1996), but nonetheless, it is reasonable to assume that most Hong Kong students, sooner or later, face the task of reading academic texts in English.

A second reason for thinking in terms of reading is the fact that encountering new words in written text is the main way new vocabulary is acquired. People who read more know more words (West & Stanovich, 1991), and encountering words in texts plays a major role in first language vocabulary development, accounting for the acquisition of thousands of new words each year by school-age children (Nagy, Hermann & Anderson, 1985). Krashen (1989) and others argue that reading is crucial to second language vocabulary acquisition as well. So in order to access this major avenue for L2 vocabulary growth (and attain the high levels of proficiency they so desire) students clearly need to be able to read successfully.

HOW MANY WORDS ARE NEEDED TO READ ACADEMIC TEXTS

One of the revelations of corpus analysis of English texts is that a small number of words makes up a large part of the texts we read. Table 1 below indicates that a learner who knows the 1000 most frequent words of English will already be able to understand half of the words that occur in an average text, and a learner who knows 1000 more understands a quite a large proportion of the text indeed (81%). Nation (1990) has argued for familiarizing L2 learners with the 2000 most common words of English through direct classroom instruction, an idea that has gained currency with the appearance of frequency-based course books like the Collins COBUILD English Course.

Table 1. Word frequencies, based on a count of 5 million running words

Different words	Percent of average text
86,741	100 %
43,831	99.0
5,000	89.4
3,000	85.2
2,000	81.3
1,000	49.0
10	23.7

(Carroll, Davies, and Richman, 1971, cited in Nation, 1990, p.17).

But how far does having control over 80 percent of the words of a text really take a learner? It might be supposed that if 80 percent of the words are familiar (i.e. four out of five), the meaning of the rest of the words can be worked out from context or looked up in a dictionary. But in fact, trying to read at a density of one unknown word in five is an arduous task. The experience of this density can be roughly simulated for native speakers if they try to make sense of the following text with eight in forty, or one in five, words gapped (in a text assuming schema knowledge of New Zealand forestry, a topic about as familiar to most language experts as many of the texts we give our students):

```
If _____ planting rates are _____ with planting _____ satisfied in each _____ and the forests milled at the _____ opportunity, the _____ wood
supplies could further _____ to about 36 million cubic meters _____ in the period 2001-2015. (Adapted from Nation, 1990, p. 242.)
```

Intuitively, a text demanding this much effort seems likely to become fatiguing in half a page or less. Research confirms the impression; studies by both Laufer (1989) and Hirsh & Nation (1992) find that knowing 80 percent of the words in a text is a very unreliable base for either comprehension or further inference. In fact, they argue that the need for direct vocabulary instruction is not satisfied until learners know more like 95 percent of the words in their text, or 19 tokens in 20. But Table 1 suggests that to reach this crucial level at which learners would find themselves in command of their reading assignments, they would need to know an unspecified number of words lying somewhere in the limitless expanse between 5,000 and 40,000 word types. The size of the task for student and teacher alike appears to be unmanageable.

But surely the difference between 90 percent--which could be attained by learning a more manageable 3000 words--and 95 percent is trivial? The difference may seem small but it is not, in either importance or attainability. With 90 percent of tokens known, one word in ten remains to be inferred from context, but with 95 percent known, only one in twenty remains. In other words, the supporting context for inferences is two times as rich at 95 percent as at 90 percent. Hirsh and Nation (1992) dramatize this doubling effect as one unknown word every line of printed text versus one every two lines.

The difference can be experienced directly by readers as the effort needed to conjure semantic entities for gaps at a ratio of 1:10 and then again at 1:20. Here is the forestry text with a gap ratio of 4:40, or 1:10 (90 percent of context words known):

```
If _____ planting rates are maintained with planting targets satisfied in each _____ and the forests milled at the earliest opportunity, the _____
wood supplies could further _____ to about 36 million cubic meters annually in the period 2001-2015.
```

And here is the text with a gap ratio of 2:40, or 1:20 (95 percent of the context words known):

```
If current planting rates are maintained with planting targets satisfied in each _____ and the forests milled at the earliest opportunity, the
available wood supplies could further _____ to about 36 million cubic meters annually in the period 2001-2015.
```

The gaps feel loosely constrained at 90 percent, tightly at 95. A missing word in 20 can either be guessed adequately enough (the first gap is something about a place, the second something about an increase), or is constrained to the point where dictionary look-up is straightforward. And of course the one unknown word in 20 is never likely to go away for the language learner, because every interesting text contains low-frequency items that the learner is unlikely to know, and which, as Kucera (1982) pointed out, typically carry a disproportionate share of its information load.

SHORTCUTTING TO HIGHER KNOWN-WORD DENSITIES

Given the unlikelihood of second language learners acquiring 95 percent of their word tokens by inching toward the native speaker's 40,000 level, several researchers in the last 25 years have tried on behalf of their students to find ways for them to shortcut their way to higher known-word densities. Their main tool has been corpus analysis. The basic idea is that different zones of discourse may have pools of high-frequency items well beyond the 2000 and 3000 levels which, if known, would give learners reading competence within those zones. Such zones have been sought and found in both general and subject-specific categories.

Praninskas pioneered the general approach at the American University of Beirut in the 1970s. She noticed that her Arab students who had mastered the 2000 level struggled to read academic texts, and wondered if there might not be a core list of academic words that could be identified and taught to these students. In a largely by-hand analysis, she used every tenth page from ten of her students' first-year academic texts, to produce a corpus of 272,466 running words to arrive at a list of high-frequency items. After subtracting out West's (1953) 2000 list, she was left with a high-frequency residue of 507 words occurring frequently across all ten texts. Praninskas' 507 words were published as the American University Word List (1972) and became the focus of a successful introductory vocabulary course in Beirut.

Several similar academic vocabulary lists were produced using similar methodologies by other teacher-researchers at roughly the same time in other parts of the developing world. Eventually Xue and Nation (1984) combined four of these lists to produce an integrated list of just over 800 words which they called the University Word List (UWL).

A further shortcut to high levels of known word density has been found by considering the lexis of specific disciplines. Nation, Sutarsyah, and Kennedy (1994), for example, analyzed a corpus of economics texts and found that with the 2000-level and UWL words subtracted out, a relatively small number of economics terms remained which recurred a good deal throughout the corpus, a number that could be clearly identified and directly taught to students who needed to read economics texts. Indeed, these researchers found that West's 2000 list, plus the University Word List, plus the subject-specific word list, accounted for 95 percent of tokens in an academic text -- in addition to being a manageable number of words for direct instruction (less than 3500 total). In view of this information, they argue, any English for Specific Purposes course should endeavor to identify and teach a core lexis to its learners directly, and thereafter introduce them to strategies of contextual inference.

QUESTIONS AND DESIGN OF THE STUDY

The main question that this study asks is the following: Where do Hong Kong students stand in relation to this apparently crucial 95 percent figure? In other words, are students reading in a known-word-density zone that enables them to understand authentic academic texts and to learn new words from them? To answer this question, Nation's (1990) Levels Test was administered to get a sense of the number of words entry-level students know. Second year students were also tested to determine whether there was any evidence of vocabulary growth over the period of a year of academic study in English. Briefly stated the research questions were as follows:

1. How many high-frequency word families do subjects know?
2. Do subjects who have been studying in English longer know more words?

SUBJECTS AND INSTRUMENTS

There were two groups of subjects, one entry level (n = 21) and the other second-year (n = 28). All the subjects were students in the Higher Diploma in English for Professional Communications program of the Division of Language Studies at the City University of Hong Kong. To qualify for entrance into the diploma program, students need a mark of C on the Hong Kong Certificate of Education Exam (HKCEE). The subjects' knowledge of English can be roughly classified as high intermediate.

Subjects took Nation's (1990) Levels Test at the 2000-word, 3000-word, 5000-word, and University Word List (UWL) level. The test measures receptive word knowledge words at five frequency levels in a multiple-choice format that asks students to identify short definitions of target items. A question from the UWL section of the test is shown below (Nation, 1990):

1. affluence	
2. axis	_4_ introduction of a new thing
3. episode	___ one event in a series
4. innovation	___ wealth
5. precision	
6. tissue	

Subjects were tested during the first week of October 1996. Also, the first-year group was re-tested in early April, 1997.

HOW MANY WORDS DO STUDENTS KNOW?

Test results indicated that virtually all of the students knew virtually all of the 2000 most basic word families of English, much as might be expected of these Hong Kong learners who have studied English throughout their secondary education. The high degree of word knowledge all subjects display at these levels is a credit to the language instruction they have received in their secondary education in Hong Kong, particularly in view of the general lack of attention to lexis at the 2000-3000 level (and beyond) discovered in typical commercial language learning materials by Meara (1993) and Nation (1994), and at the 2000 level itself by Cobb (1994).

Generally, performance at the 3000 and 5000 levels was also high as shown in Table 2 below. Nation (1990) judges scores of 80 percent or more as indicative of mastery of a particular level, so students are clearly in full control of the 3000 level with scores averaging over 90 percent in both groups.

Table 2. Lexical profile of entry level and year 2 subjects, Oct. 96

	Year 1 (n=21)	**Year 2 (n=28)**	**Difference (by t-test)**
3000	91.7% (SD 6.6)	91.0% (SD 9.3)	t(47)=1.85 p<.05 s.d
5000	64.4 (21.4)	73.4 (12.2)	t(47)=1.85 p<.05 s.d
UWL	72.2 (16.4)	72.0 (16.3)	t(47)=0.06 p>.05 n.s.d.

More interesting scores occur at the 5000 level with first year students at a slightly lower mean score than second year students. A t-test for independent sample showed this difference to be significant, but in terms of numbers of words the difference is not great. Assuming that a subject knows all words at the 2000 and 3000 levels, the average score of 64 percent in the first-year group represents receptive knowledge of about 4300 of the 5000 most common words of English, while the 73 percent figure in the second year group represents about 4500 words. So both groups of subjects are near the 5000 mark that should give them control over 90 percent of the words they encounter in general reading (see Table 1), but still a large, unspecified number of words away from the 95% that should make authentic texts readily comprehensible and foster efficient learning of new words.

Results at the UWL level, which is relevant to academic reading, are somewhat less encouraging. Results in the two groups are practically identical with a mean score of 72 percent which represents receptive knowledge of about 580 of the 800 UWL words.

ARE STUDENTS LEARNING NEW WORDS?

The overall picture presented in Table 2 is worrying. Lexical growth for the second-year students over the course of their first year of studies appears to be limited. A t-test for independent samples indicated that the small difference between first and second-year students on the 5000-level measure was significant, but as mentioned above, this amounts to an increase in vocabulary size of just 200 words. For learners studying in an (officially) English-medium environment over the period of a whole year, this does not seem impressive, given that the average rate of vocabulary growth for European secondary students learning English (a few hours a week?) is typically 275 words per six-month term (Milton & Meara, 1995).

In the case of the 2000 and 3000 levels, the lack of change is due to a ceiling effect - little remains at these levels to learn. But it is the lack of a clear difference at the University Word List level, as well as the large variance, that is most troubling. Only one-third of the students in either class have control at this level, one-third scoring less than 60 percent, and this does not change over a year of study. By contrast, in a lexical growth study by Laufer (1994) with Hebrew and Arabic-speaking students in Israel, the UWL level was the one level that did grow over a year of reading and study.

It is possible that the tests simply did not capture vocabulary growth that actually occurred in these subjects. For instance, they may have learned many new words at levels beyond the 5000 most-frequent category - it is clear that any further research with these subjects should also include the 10,000 level part of the Levels Test to avoid ceiling effects. However, it is also possible that the lack of a strong growth finding is due to the cross-sectional design; that is, a better-than-usual first year group may have been compared to a poorer-than-usual second year group. For this reason it was decided to test the same first-year group again six months later.

The results of the second test do not give cause for optimism. Post-test scores were very similar to pre-test results, suggesting that no change in mean vocabulary size occurred during the six-month period (see Table 3 below). In fact, mean performance was marginally lower at the second sitting on parts of the test (e.g. UWL). A t-test for matched samples indicated that differences between pre- and post-test scores were not significant, so the original impression that students are not learning many new words through academic study in English appears to be confirmed.

Table 3. Lexical profile of year 1 subjects, Oct. 1996 and April 1997 (n=14).

	Pre-Test	**Post-Test**	**Difference (by t-test)**
3000	92.1% (SD 6.9)	91.4% (SD 6.2)	t(13)=0.38 p>.05 n.s.d
5000	70.5 (19.7)	71.8 (12.2)	t(13)=0.30 p>.05 n.s.d
UWL	76.8 (14.0)	72.3 (17.7)	t(13)=1.16 p>.05 n.s.d.

The observant reader will have noticed that the number of first-year students in this comparison (14) is smaller than in the earlier comparison (21) with second year students. This is partly due to absenteeism, but there is a second reason: three students have left the course. Before concluding, it is interesting to take a closer look at the vocabulary profiles of these three.

THREE SPECIAL CASES

Like their peers, all three scored high on the tests of knowledge of the 2000 and 3000 most frequent words of English. But is at the 5000 and UWL levels that the profiles of these three learners become interesting (Table 4 below).

Table 4. Lexical profiles of three drop-outs, year 1

	2000	**3000**	**5000**	**UWL**
Subject A	94%	94%	44%	61%
Subject B	100	83	11	55
Subject C	94	94	55	55

Given their large deficits at the UWL level, it is clear that reading academic texts must have been laborious for these students, an exercise of filling in the gaps with not enough known words in the context to make good guesses, and time consuming references to dictionaries. It hard not to think to think that inability to read efficiently played some role in their decision to leave the course.

CONCLUSION

What can be done to make sure that Hong Kong university students are able to read English academic texts effectively, and to understand and eventually learn the new words they encounter? Knowing more words to start with appears to be part of the answer, though obviously not all of it. Reading effectively is much more than knowing meanings of a large proportion of the words that occur in a text, and learning a particular list of words cannot magically enable second language learners to achieve native-like reading proficiency. Clarifying the connection between vocabulary knowledge and successful reading remains a challenge to researchers. In the meantime, it is clear that some Hong Kong students have some identifiable vocabulary deficiencies that may be hindering their academic progress. Fortunately, there is an increasing body of knowledge about words and their frequencies that can help teachers identify the vocabulary that students need to know to achieve their goals.

REFERENCES

Cobb, T. (1994). *Which course prepares students for the PET?* Research Report. Muscat: Sultan Qaboos University.

Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics, 11* (4), 341-363.

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

Johnson, R. K., & Yau So Ngor, A. (1996). Coping with second language texts: The development of lexically-based reading strategies. In D. A. Watkins & J. B. Biggs (Eds.), *The Chinese learner.* Hong Kong: Hong Kong University

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

Kucera, H. (1982). The mathematics of language. In *The American Heritage Dictionary.* Second College Edition. Boston: Houghton Mifflin.

Laufer, B. (1994). The lexical profile of second language writing: Does it change over time? *RELC Journal, 25* (2), 21-33.

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.

Lightbown, P.M., Halter, R., & Meara, P. (1997). Classrooms as lexical environments. *Language Teaching Research, 1* (1).

Meara, P. (1993). Tintin and the world service: A look at lexical environments. *IATEFL: Annual Conference Report,* 32-37.

Milton, J., & Meara, P. (1995). How periods abroad affect vocabulary growth in a foreign language. *ITL Review of Applied Linguistics, 107/108,* 17-34.

Nagy, W. E., Herman, P.A., & Anderson, R. C. (1985). Learning words from context. *Reading Research Quarterly, 20* (2), 233-253.

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

Nation, P. (1994). Review of four vocabulary coursebooks. *System, 22* (2), 283-287.

Praninskas, J. (1972). *American University Word List.* London: Longman.

Sutarsyah, C., Nation, 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.

West, R. F., & Stanovich, K. E. (1991). The incidental acquisition of information from reading. *Psychological Science, 2* (5), 325-329.

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