Learning Vocabulary for University Entrance Exams: A Word Frequency Study

Paul ROSS

Abstract

This paper reports on a frequency analysis of the English vocabulary items that university-bound students in Japan are expected to master. The items are compared with the three frequency bands that current corpus-driven research shows to offer the widest range of coverage for the minimal amount of learning investment. The comparison shows that students in Japan are being exposed to a large amount of vocabulary that is not in any of these three bands. Discussion of the problems with corpus-driven research and some of the pedagogical issues raised by the study is included.

Introduction

Language education in Japan is often criticized for being driven by university entrance exams. From the time students begin studying English in junior high school, they are thought to spend an inordinate amount of time memorizing arcane points of grammar and long lists of esoteric vocabulary items to prepare for the exams. However, aside perhaps from the school teachers and cram school instructors responsible for preparing students to pass the entrance exams, few people involved in English language education in Japan know much about the language that students are exposed to or how they are trained to acquire it. This paper attempts to fill in the gaps in our knowledge of one of these areas, focusing on the vocabulary items that university-bound students in Japan are expected to master.

Specifically, I will report on a frequency analysis of the vocabulary found in three exam preparation self-study manuals. I am especially interested in finding out the degree to which the items in these manuals fit into the three basic frequency bands of vocabulary that corpus-driven research shows to offer the widest amount of coverage. This will shed some light on the nature of the vocabulary that Japanese EFL students are exposed to, and it will offer some empirical evidence in answer to the question of whether or not vocabulary instruction is in fact inefficient and ineffective.

1. Study Description and Results

To determine which lexical items university-bound students are expected to master, the following three self-study manuals published by major cram schools were chosen: *Shisutemu Eitango* (*システム英単語*), *Nyūshi Eitango no Ōdō* (*入試英単語の王道*), and *Sokudoku Eitango, Vol. 1* (*速読英単語, Vol. 1*). The items covered in the books were fed into an open-source, web-based vocabulary profiler (see The Compleat Lexical Tutor, http://www.er.uqam.ca/nobel/r21270/textools/web_vp.html). The profiler provides an analysis of where the items fit into the bands that recent corpus-based vocabulary research (e.g. Coxhead 1999 and Nation 2002) shows to offer the widest range of coverage with the fewest number of lexical items.[1]

1.1 Results: *Shisutemu Eitango* (*システム英単語*)

	Families	Types/Tokens	Percent
1st K	384	417	20.27
2nd K	430	444	21.53
AWL	404	425	20.61
Off List	?	775	37.58

Figure 1 Number of headwords: 2,061
On-list items: 1,286
On-list families: 1,218

Results of the lexical analysis of the items in *Shisutemu Eitango* are summarized in Figure 1. To the left of the grid are the frequency bands: the first and second thousand most basic words (based on West 1953, cited in Nation 2002), the words on Coxhead's (1999) new academic word list, and the off-list band for words that are not on any of the previous three bands. Moving to the grid itself, we see that each band is analyzed according to three categories: the number of word families, the number of types and tokens, and the overall percentage of items in each band. The information under the grid gives the number of headwords found in the text, the number of those items that are on one of the first three frequency bands, and finally how many word families are represented in that number.

Looking at Figure 1, we find that *Shisutemu Eitango* contains 2,061 headwords. Approximately 60 percent of these are on one of the three basic frequency bands, and slightly less than 40 percent are not found on any of those three lists. Of the on-list items, representation in each band is fairly consistent (approximately 20 percent each). Notice that the type/token ration is 1:1, meaning that the headwords receive unique mention. Notice also that the number of word families (1,218) and the number of on-list items (1,286) are close, meaning that the learning burden is relatively high.

1.2 *Nyūshi Eitango no Ōdō* (入試英単語の王道)

	Families	Types/Tokens	Percent
1st K	299	312	15.22
2nd K	408	424	20.68
AWL	394	414	20.20
Off List	?	900	43.90

Figure 2 Number of headwords: 2,050
On-list items: 1,150
On-list families: 1,101

Results of the analysis of the items in *Nyūshi Eitango no Ōdō* (Figure 2) are presented in the same way. The number of headwords is similar (2,050 versus 2,061), as are the percentage of words from the second thousand and AWL bands (as above, approximately 20 percent each). Notice, however, that coverage of words in the first thousand band is approximately 5 percent lower (15.22 percent versus 20.27 percent), while off-list items are approximately 5 percent higher (43.90 percent versus 37.58 percent). Again, a high learning burden is suggested by the nearly 1:1 ratio of on-list families to on-list words.

1.3 *Sokudoku Eitango, Vol. 1* (速読英単語，*Vol. 1*)

	Families	Types/Tokens	Percent
1st K	593	921/928	30.21
2nd K	536	750/755	24.58
AWL	388	624/625	20.35
Off List	?	762/764	24.87

Figure 3 Number of headwords: 1,857
Total # of words covered: 3,057
On-list items: 2,295
On-list families: 1,517

A quick look at Figure 3 shows a considerable difference in this text as compared with the two others in the first band (30.21 percent versus 20.27 percent and 15.22 percent) and the AWL band (24.87 percent versus 37.58 percent and 43.90 percent). Readers will also have noticed several differences in the way that the information for this text has been presented. First, there is a difference in the type/token ratios. This is due to the coverage of hyphenated items that the text analyzer doesn’t recognize. For example, the text includes the items *self*, *self-confidence* and *self-esteem*, yielding three tokens of the same type (*self*), with *confidence* and *esteem* analyzed as separate items. There are fifteen such items in the text, and while this inevitably leads to a slight skewing of the data, the number of instances is fairly low.

Also notice that a distinction is made in Figure 3 between the number of headwords and the total number of words contained in the text. All of the texts under consideration include a large number of items in addition to their main entries. These include members of the same word family, antonyms, collocations, and formulaic phrases. *Shisutemu Eitango* and *Nyūshi Eitango no Ōdō* present these items as 'extra' and encourage students to be aware of them, but *Sokudoku Eitango* is explicit about the need to master these items as well. The point to keep in mind is that the number of items that the learners are exposed to in these texts is quite a bit higher than the number of headwords the texts claim to contain.

2. Discussion

The first issue that needs to be explored is the question of how these texts determine which lexical items to cover. Given that the books are intended to prepare students for university entrance exams, it should come as no surprise that decisions are based on an analysis of the lexical items that appear in the reading passages of university exams. The texts refer to the universities they include by referring to either the general level or specific name of the institution. They also mention how many years of exams their analysis covers or state the number of running words that their corpus consists of. Only one book (*Sokudoku Eitango*) is explicit about its use of a computer-based corpus analysis, while the others claim more simply that they have carried out a 'careful frequency analysis' to ensure that the items they cover offer the most effective and efficient investment of time and energy on the part of learners while providing them with the maximal return on that investment.

2.1 On the frequency bands

The data presented in Section 1 raises several important questions. First, notice that items from the first two frequency bands (i.e. from either the first or second thousand words) make up between a low of approximately 30% of the words covered in a single text (*Shisutemu Eitango*) and a high of 55% (*Sokudoku Eitango*). How are we to interpret these figures? I would argue that the relatively high numbers of basic vocabulary is evidence that learners preparing to enter university are not felt to have gained sufficient mastery of basic vocabulary items. It is worth noting that nowhere in these manuals is mentioned made of reviewing or consolidating work on vocabulary from high school, suggesting that the writers have little confidence that the learners have been exposed to the basics.

At the same time, a large percentage of the items in the manuals are from either the third and lowest frequency band (i.e. the academic word list) or else are entirely off list (i.e. items not on any of the three frequency bands). Items from the academic word list make up approximately 20% of the items presented in each manual, and off-list items account for between a low of approximately 25% (*Sokudoku Eitango*) and a high of approximately 44% (*Nyūshi Eitango*). Given that the reading passages of entrance exams are traditionally — and notoriously — 'challenging', the large number of such words does not come as a surprise.

However, it would be a mistake to assume that items from the academic word list or even those that are off list are necessarily esoteric. The following are samples from these bands taken at random from two of the books under discussion:

Academic word list

recover aware encounter attain access document reject principle approach capability compatible export status rely unique accommodate community period devote welfare

(from *Nyūshi Eitango no Ōdō*)

Off-List Words

absorb bacteria calorie datum Egyptian facial gaze habitat jewelry linguist malice nonverbal oblige pace recall score tablet underground van

(from *Sokudoku Eitango, Vol. 1*)

Notice that while these lists do contain words that strike us as beyond what any Japanese high school student might reasonably be expected to know (e.g. *malice, compatible*), many are within the realm of the reasonable, and a number of them are likely to be familiar to learners as loanwords (e.g. *approach, unique, community, bacteria, calorie, jewelry,* and *score*). It is of course possible to find a number of examples of items that seem hard to justify exposing learners to (for example, *folly, forsake, grumble* and *impudent* are all found in *Sokudoku Eitango*), but we should resist the assumption that once we move outside the core vocabulary of the first two thousand items we are necessarily placing an undue learning burden on our students.

2.2 Should frequency counts drive vocabulary teaching and learning?

It is hard to argue against the pedagogical merits of grounding vocabulary teaching and learning on frequency counts, and few in the ELT field would be likely to do so. The problem, of course, is that frequency counts are based on specific corpora, and the results of those counts will depend on what corpora are used to provide them. As this study shows, for many university-bound learners in Japan, vocabulary learning is being driven by corpora made up of entrance exam reading passages. Most in the ELT field would argue that a much broader, more representative and more authentic corpora should be driving the frequency counts, and the consensus seems to be that the three frequency bands used in this study provide the most useful framework currently available.

However, until consensus at all levels of language education can be reached, we are likely to be stuck in a situation in which competing corpora make it difficult to reach the level of standardization we would like to see. A clear example of the problem can be found in the widely different ways current dictionaries handle frequency counts. According to *Kenkyusha's New English-Japanese Dictionary*, the 'essential vocabulary' for university-level English study consists of at least 4,000 words.[2] *Genius*, on the other hand, puts the number at 9,600.

The confusion only increases when we compare the various frequency bands adopted by major dictionaries and examine which bands specific words fit into in. Figure 4 presents a comparison of four popular dictionaries: *Kenkyusha's New English-Japanese Dictionary* (*KNEJD*), *Genius*, *Longman's Dictionary of Contemporary English* (*LDCE*), and *Collins Cobuild English Dictionary for Advanced Learners* (*CCEDAL*).

	Band 1	**Band 2**	**Band 3**	**Band 4**	**Band 5**
KNEJD	1,000 (JHS)	1,000 (HS)	2,000 (univ. entrance exams+ study)	3,000 (next most basic)	
Genius	1,000 (JHS)	3,400 (HS)	5,100 (univ. and beyond)		
LDCE	1st 1,000	2nd 1,000	3rd 1,000		
CCEDAL	1st 680	Next 1,040	Next 1,580	Next 3,200	Next 8,100

Figure 4

Notice that the dictionaries have between three and five frequency bands, with varying numbers of items in each band. The explanation of the two Japanese-English dictionaries contains information about each band in terms of level of education (junior high school through university and beyond). The monolingual English dictionaries present the number of items in each band, with *LDCE* emphasizing that its three bands comprise the most ‘basic’ or ‘core’ vocabulary in descending order of frequency. *CCEDAL*, on the other hand, presents a total of five bands, explaining that the first four bands comprise the most ‘basic’ or ‘core’ vocabulary, again in descending order of frequency.

Since the bands in these four dictionaries are broken down so differently, learners looking up the same word in them will come away with rather different information regarding word frequency. The chart below summarizes the information about the (putative) frequency of three words: *ability, fear, and commit.*[3)]

	ability (1st K)	***fear*** (1st K)	***commit*** (AWL)
KNEJD	Band 2 (HS)	Band 2 (HS)	Band 3 (University)
Genius	Band 1 (JHS)	Band 1 (JHS)	Band 3 (University +)
LDCE	Bands 1/2 (S/W)	Bands 3/1 (S/W)	Bands 2/3 (S/W)
CCEDAL	Band 2 (1st 1,720)	Band 1 (First 680)	Band 2 (First 1,720)

Figure 5

Note: LDCE analyzes words according to frequency in spoken (S) and written (W) registers. For example, *ability* is in Band 1 in the spoken register and Band 2 in the written register.

The top row of the chart includes the three words and their location within the three major frequency bands. Both *ability* and *fear* are on the first thousand list, while *commit* is on the academic word list. Notice, however, that the frequency band of a given word may differ by dictionary. To take just one example, *KNEJD* places *ability* in its second band (i.e. among the second thousand items, appropriate for the high school level), while *Genius*, the other Japanese-English dictionary, places it in the first band (i.e. among the first thousand items, appropriate for the junior high school level). In the monolingual learners’ dictionaries, *ability* is in *LDCE’*s Band 1 for spoken register (i.e. among the first thousand items), and in Band 2 in the written register (i.e. in the second thousand words), while *CCEDAL* places *ability* in its second band (i.e. among the first 1,720 words). A look at the other two items will show similar differences among the dictionaries, offering support for the claim made earlier that the current lack of standardization means that the dream of corpus-driven vocabulary teaching and learning is far from being realized.

3. Concluding Remarks

The present study offers some support for the traditional view that Japanese learners of English spend considerable time and energy on low-frequency linguistic items.

Unfortunately, in the case of vocabulary at least, this expenditure of time and energy apparently does not come after mastering the basics. As mentioned above, a major portion of the items covered in the self-study manuals are words from the two highest bands of frequency (i.e. the first two thousand words), suggesting that learners are not assumed to have covered these items in either junior or senior high school language study. A study by Matsui et. al. (2004) gives further evidence of the gaps in command of basic vocabulary, with findings suggesting that first and second year university students were familiar with only between 40-70 percent of the core vocabulary used to define words in three major learners' dictionaries.[4] One possible explanation, of course, is that learners are spreading themselves thin by concentrating on the lower frequency items before they have full command of the higher frequency items.

This, of course, is precisely the problem that corpus-driven research in applied linguistics aims to address. The attraction of this research to those involved with language education is undeniable: not only will it help ensure that the language items presented to learners are more accurate reflections of language as it is really used, but it offers the tantalizing prospect of boosting the effectiveness and efficiency of language instruction by offering reliable guidelines based on empirical evidence that can help determine what items to include on the syllabus.

However, as this study makes clear, obstacles remain before the benefits of corpus-driven research and pedagogy can be fully realized. As suggested above, the most basic issue in need of resolution is how — indeed, if — standardization of the benchmark corpora can be achieved. We have seen that university entrance exams provide the corpus that propels vocabulary acquisition for many learners in Japan. It is also clear that the frequency counts of lexical items culled from this corpus differ greatly from the frequency bands posited by the most recent research in applied linguistics. We are witnessing a clear-cut case of the 'washback effect', and the only chance of negating this effect is if universities in Japan recognize the pedagogical implications of what they are doing and consider it to be in their interests to change their approach.

However, as we have also seen, the problem is not limited to the entrance exam system. Major reference materials, both bilingual dictionaries published in Japan and monolingual learners' dictionaries published abroad, base their frequency analysis of lexical items on different corpora and have different conceptions of how those bands are best broken down.

For all the excitement that corpus-driven research is generating in ELT circles, we are still only at the beginning stages of applying the findings stemming from that research. We will no doubt continue to see an increase in classroom materials such as McCarthy et. al. (2005) that choose and order the language items it presents based on findings from corpus-driven research. It is also clear that corpus-driven research will continue to be of great help to teachers in presenting linguistic features that are truly representative of the way language is actually used. However, perhaps the time has come to worry less about increasing the depth and breadth of the corpora we use, and turn our attention to exploring the possibility of reaching a consensus on which corpora will be used as the benchmark for both research and pedagogical purposes in the ELT field.

Notes

1) See Appendix 1 for a more detailed consideration of the items presented in each book.

2) 4,000 items is the minimum; depending on how you interpret their explanation it could be as high as 7,000. Either way, these numbers are quite a bit different from the 9,600 claimed by *Genius*. It is also interesting to note that the number of 'essential' items in the vocabulary self-study manuals referred to in this study is considerably lower. This also shows that the corpus that the cram schools use is not the same one used by the dictionary publishers.

3) These are some of the words used in *Kenkyusha's New English-Japanese Dictionary* to illustrate its own use of word frequency bands.

4) As Matsui et. al. (2004) point out, these figures strongly suggest that university students are unprepared to use monolingual learners' dictionaries. They suggest that a major goal of a university language program should be to focus on giving students the vocabulary they need to use these kinds of dictionaries. I agree with them that setting up such concrete goals has pedagogical advantages related to motivation; it also makes sense from a general vocabulary acquisition standpoint since the defining vocabulary of any learners dictionary will rely heavily on high frequency items. For example, a vocabulary profile I carried out on the *LDCE* found that approximately 90% of its defining vocabulary is from that dictionary's top three frequency bands (See Appendix 3). However, the question about the pedagogical benefits of monolingual dictionaries is far from settled (see Ross 2001).

References

Books and articles

Coxhead, A. (1999). A new academic word list. *TESOL Quarterly*, 34, 213-238.

Emoto, Y., Shimada, H., Yoneyama, T., Fukuzaki, G., and Gill, T. (2004). *Nyūshi Eitango no Ōdō*. Tokyo: Kawai Press.

Kazahaya, H. (2003). *Sokudoku Eitango, Vol. 1* (4th edition). Tokyo: Z-Kai Press.

Nation, I.S.P. (2002). *Learning Vocabulary in Another Language.* Cambridge: CUP.

Matsui, S., Okada, T., Ishihara, K., and Pavloska, S. (2004). Toward the use of monolingual dictionaries: Building knowledge of defining vocabulary. *Doshisha Studies in Language and Culture*, 7(1), 83-112.

McCarthy, M., McCarten, J., and Sandiford, H. (2005) *Touchstone*. Cambridge: Cambridge University Press.

Ross, Paul (2001). Dictionaries and the language learner. *The Journal of the Institute for Language and Culture*, Konan University, 3, 51-65.

Tone, M., and Shimo, Y. *Shisutemu Eitango*. Tokyo: Sundai Bunko.

West, M. (1953). *A General Service List of English Words.* London: Longman, Green, and Co.

Dictionaries

Collins Cobuild English Dictionary for Advanced Learner, 3rd edition (2001) (electronic edition).
Genius Japanese-English Dictionary, 3rd edition (2001). Taishukan (electronic edition).
Kenkyusha's New College English-Japanese Dictionary, 6th edition (1994).
Longman Dictionary of Contemporary English, 3rd edition (1995).

Websites

The Compleat Lexical Tutor: http://www.er.uqam.ca/nobel/r21270/textools/web_vp.html

Appendix 1

Section 1 is limited to an overall summary of the frequencies of the items found in the three vocabulary self-study manuals. In the manuals themselves, the items are broken down into several levels (e.g. 'basic', 'standard', and 'advanced'), and this section presents data on the items found in each of those levels.

a. Shisutemu Eitango (***システム英単語***)

Level 1: Basic

	Families	Types/Tokens	Percent
1st K	211	215	35.77
2nd K	159	160	26.62
AWL	172	173	28.79
Off List	?	53	8.82

Number of headwords: 601
On-list items: 548
On-list families: 542

Level 2: Essential

	Families	Types/Tokens	Percent
1st K	50	53	8.79
2nd K	173	174	28.86
AWL	150	153	25.37
Off List	?	223	36.98

Number of headwords: 603
On-list items: 380
On-list families: 373

Level 3: Advanced

	Families	Types/Tokens	Percent
1st K	11	11	2.70
2nd K	71	71	17.44
AWL	63	63	15.48
Off List	?	262	64.37

Number of headwords: 407
On-list items: 145
On-list families: 145

Level 4: Final

	Families	Types/Tokens	Percent
1st K	1	1	0.37
2nd K	14	14	5.22
AWL	21	21	7.84
Off List	?	232	86.57

Number of headwords: 268
On-list items: 36
On-list families: 36

Level: Brush Up (Common words with multiple meanings)

	Families	Types/Tokens	Percent
1st K	136	138	75.41
2nd K	25	25	13.66
AWL	14	14	7.65
Off List	?	6	3.28

Number of headwords: 183
On-list items: 177
On-list families: 175

All

	Families	Types/Tokens	Percent
1st K	384	417	20.27
2nd K	430	444	21.53
AWL	404	425	20.61
Off List	?	775	37.58

Number of headwords: 2,061
On-list items: 1,286
On-list families: 1,218

b. Nyūshi Eitango no Ōdō (入試英単語の王道)

Level: Basic

	Families	Types/Tokens	Percent
1st K	111	115	38.33
2nd K	96	99	33.00
AWL	44	44	14.67
Off List	?	900	43.90

Number of headwords: 300
On-list items: 258
On-list families: 251

Level Two

	Families	Types/Tokens	Percent
1st K	149	155	12.92
2nd K	296	304	25.33
AWL	273	287	23.92
Off List	?	454	37.83

Number of headwords: 1,200
On-list items: 746
On-list families: 718

Level 3

	Families	Types/Tokens	Percent
1st K	3	3	1.00
2nd K	14	14	4.67
AWL	54	55	18.33
Off List	?	228	76.00

Number of headwords: 300
On-list items: 72
On-list families: 71

Level 4

	Families	Types/Tokens	Percent
1st K	0	0	0.0
2nd K	1	1	0.50
AWL	23	23	11.50
Off List	?	176	88.00

Number of headwords: 200
On-list items: 24
On-list families: 24

All

	Families	Types/Tokens	Percent
1st K	299	312	15.22
2nd K	408	424	20.68
AWL	394	414	20.20
Off List	?	900	43.90

Number of headwords: 2,050
On-list items: 1,150
On-list families: 1,101

***c. Sokudoku Eitango, Vol. 1* (*速読英単語，Vol. 1*)**

Group 1

	Families	Types/Tokens	Percent
1st K	522	659/660	41.93
2nd K	316	375/375	23.82
AWL	268	326/326	20.71
Off List	?	213/213	13.53

Number of headwords: 1,574
On-list items: 1,360
On-list families: 1,106

Group 2

	Families	Types/Tokens	Percent
1st K	146	164/164	16.95
2nd K	255	288/289	29.89
AWL	175	204/204	21.10
Off List	?	310/310	32.06

Number of headwords: 967
On-list items: 656
On-list families: 576

Group 3

	Families	Types/Tokens	Percent
1st K	95	104/104	19.59
2nd K	81	91/91	17.14
AWL	79	95/95	17.89
Off List	?	240/241	45.39

Number of headwords: 531
On-list items: 290
On-list families: 255

Appendix 2

Below are examples of items from the first and second thousand frequency bands taken at random from *Shisutemu Eitango*:

First thousand

allow base consider decide expect follow gain honor include join limit manufacture notice offer provide remain suggest trust unite value wonder youth

Second thousand

argue behavior compare determine encourage frequently government hurt improve lean mistake narrow own prefer qualify refer search tend upset violence worry

Appendix 3

Below is a frequency analysis of the lexical items used in defining entries in the *LDCE*. As the table shows, approximately 88 percent of the items are from the first two bands of frequency (i.e. first and second thousand words), with approximately 1.5 percent from the academic word list. It should be noted that approximately 145 tokens of a total of 312 of the off-list words consist of items such as abbreviations needed to use the dictionary (e.g. *Av, NV, AD,*) and a list of affixes (e.g. -*ation*, -*ment*, and -*ness*)

	Families	Types/Tokens	Percent
1st K	908	1157/1766	59.70
2nd K	699	833/835	28.23
AWL	41	45/45	1.52
Off List	?	237/312	10.55

Number of headwords: 2,958
Number of types: 2,272
On-list items: 2,035
On-list families: 1,648