VOCABULARY CONSTRAINT ON READING MATERIALS

Cucu Sutarsyah

FKIP Universitas Lampung, Jl. Prof. Dr. Sumantri Brojonegoro No. 1
e-mail: csutarsyah@yahoo.com

Abstract: The aim of the study is to identify and describe the vocabulary in reading materials and to seek if the texts are useful for reading skill development. A descriptive qualitative design was applied to obtain the data. Some available computer programs were used to find the description of the vocabulary in the texts. It was found that the texts are dominated by low frequency words which compose 16.97% of the words in the texts. In terms of high frequency words occurring in the texts, function words dominate the texts. In the case of word levels, it was found that the texts being used have very limited number of words from GSL (West, 1953). The proportion of the first 1,000 words of GSL only comprises 44.6%. The data also show that the texts contain too large proportion of words which are not in the three levels (the first 2,000 and UWL). These words constitute account for 26.44% of the running words in the texts. It is believed that the constraints are due to the selection of the texts which are made of a series of short-unrelated texts. This kind of text is subject to the accumulation of low frequency words especially those of content words and limited number of words from GSL. It could also impede the development of students' reading skills and vocabulary enrichment.

Keywords: vocabulary, low frequency words, high frequency words, word list, word level.

In Indonesia, English is learned by a large number of students who will almost never have the opportunity of conversing with native speakers, but who will have access to the literature and periodicals, or scientific and technical journals written in the language they are learning (Rivers, 1981). Thus, reading skill is virtually more important in this context than other skills (Huda, 1992). In fact, it is not enough for a student to have reading skills in their first language.

However, despite this specific need for a foreign language, very frequently, students reading in a foreign language seem to read with less understanding than one might expect them to have (Alderson, 1984). It is observable that many students are still not able to comprehend texts when they enter the tertiary level. Day and Bamford (2000) also claim that, in general, learners learning to read English as a foreign language find it a difficult process, and as a result, they do not enjoy it.

Of the English skills, the ability to read today is the most realistic in terms of language use. Indonesian language is used in all schools and colleges. Students are expected to be able to read academic texts in English and have to master a total of 4,000 vocabulary items as set by the 1984 curriculum. For some reasons, in the 1994 curriculum, the number of vocabulary items was reduced to about 2,500 items for Social and Science Program and 3,000 for Language Program (Depdikbud, 1993:1). This vocabulary size is actually sufficient for a student in order to read academic texts. Nevertheless, these objectives are not well achieved.

However, some studies show that the students have low achievement in reading skills and vocabulary gain. It is predicted that the SMU graduates have only mastered about 1,000 words. Results of some research also support the view that reading in a language which is not the learners' first language is a source of considerable difficulty (Alderson, 1984, Brown, 2001).

Many efforts have been made to improve the learners' reading ability. Some of them are shown by the existence of books on theories of reading and some books on reading exercises or learning to read. However, as stated by some experts, e.g., Sadtono (1995), the fact shows that many learners still encounter problems or difficulties when trying to understand an English text.

To the writer's knowledge, there is no thorough investigation on reading materials especially on vocabulary analysis. Most studies deal with the output

of learning-teaching, such as, reading achievement and vocabulary gain, learners' test scores (Sadtono, 1995; Nurweni, 1995; Samhati, 2006). Some others deal with the relationship between independent variables and a dependent variable in reading comprehension such as that of Chandrasegaran's (1981). Some concentrate on the input, such as syllabus, teaching materials, learners' behavior and characteristics.

If we want to identify what makes for success and what causes failure in learning to read, we will find a very long list of variables that can be drawn up. However, some experts agree that there are three factors which are very dominant, that is, the learners, the teacher, and the materials. In spite of the interdependency of these three factors, each can be considered independent of the others. Dakin (1973) has an interesting idea about these three variables. She explains that the materials have their own structure because they are written by different people. The material writers, again, have their own perception about the learners. Thus, the materials have their own plan of development, and their own weaknesses, which both teacher and learners must overcome.

The three factors have relationships and interdependency that are set up in the classroom. The teacher has his or her own method of teaching and using the materials and his or her own perception about the learners. The learners develop their own methods of learning to respond to both the teacher and to the materials. Finally, the materials are written by authors who have a particular teacher and particular learners in mind.

Success or failure in learning to read depends on what is learned, not on what the teacher and materials set out to teach. The job of the teacher and the materials is to promote in the learners successful technique of learning (Dakin, 1973). Thus, even though our attention is paid to the factors, we still have too many variables to look at. Based on the illustration above, this research is only focused on the reading materials used by EFL learners.

The vocabulary in the texts is the aspect that needs to identify. It is claimed that the condition of the words in a text has a great influence on readers' comprehension. It is also commonly believed that comprehension depends on the extent to which the words in a text are familiar to the readers (Nation, 1997). On the other hand, the familiarity of words implies that the readers have already met the words for several times and have stored them in their long-term memory. These words usually belong to high frequency words and occur in wide ranges of texts and most of them belong to the most frequent words of General Service List of English Words (West, 1953). In fact, word frequencies are important in planning word lists for language teaching (Richards, 2002:7). In contrast, the words that are infrequent are usually considered difficult for most readers. Some of these words are of technical words and are not in the three *Base Word List* of Nation's (1997). Thus, an ideal text is the text that has enough high frequency words and not too many low frequency words.

Furthermore, an ideal text or course book has a good density index. Density index of a passage or a lesson or a course book is the proportion of different words to the total number of words or running words in it (Nation, 1990). If this proportion is high, reading is relatively easy. In other words, in order to get high-density index, many of the different words must be frequently repeated. This also implies that the text containing too many low frequency words is relatively difficult. According to Nation (1990), the density index of an English course book is usually of the ratio of 1:2.4. This means that, on the average, each word is repeated between two or three times. An efficient course book for beginning students should have density index of around 1:20 (Nation, 1990). In later years the density index should be decreased to about 1:10 or 1:12. A course book containing a large number of words that occur less than five or six times can not be efficient. This course book usually has density index of 1:2.4 or 1:4.

This case study was carried out at the Undergraduate Program, English Departement, State University of Malang (UM). The students come from high school at the age of 18. This undergraduate lasts about four years. In this Department, English is used as a medium of instruction for most subjects taught. It prepares English teachers to teach English as a foreign language in schools where reading receives the most emphasis. The main goals of the Department are, first, to improve the students' command of English for the purpose of teaching the language, and, second, to enhance their knowledge in the field of language teaching.

Based on the discussion above, the texts being used for a reading class at the English Department of UM are believed to contain some constraints that make them difficult for the students to read. The study was guided by one main research question: to what extent can the texts being used for a reading class support students' reading skills? Specifically, the study was focused on the research question: to

what extent is the vocabulary in the texts useful for reading skill development?

METHOD

The reading materials being investigated were the ones used by the students in the class (Reading Comprehension I) and the ones used for the tests. A computer-based calculation was used by using available program in the computer. To do this, all the texts were retyped to be computer readable data. During the semester, 20 texts were used in both the classroom and the tests. All information indicates the complexity of a text and is used to judge the condition of a text. As it has been discussed, the text condition can influence students' reading performance. The investigation on the texts was made on vocabulary condition. This vocabulary examination was divided into two sections. The first one was dealing with the examination of high and low frequency words occurring in the texts. The second one dealt with the classification of words in the texts into three levels.

Computer programs called *WORD* and *Vocaprofile* were used to analyze words in these two investigations. These computer programs were prepared and designed in Computing Service Center, Victoria University of Wellington, New Zealand (Nation, 1997). Word program simply uses an input file of written materials and analyzes it to find the frequency of occurrence of each word. The output of this program is a series of word lists arranged alphabetically with their frequency of occurrences and rank order of word frequency list with their proportion. From this rank order list, we can see the most frequent words up to the least frequent words. Both word lists were produced in the form of word types or word forms.

It has been claimed that word frequency count is important for a text evaluation and to find the most important words to learn, because word selection is needed since it is impossible to learn all the words at once. This investigation is useful to see if the texts are appropriate to help develop students' reading skills.

Other investigation on vocabulary is based on word levels. By using the computer program, *Vocaprofile*, the words in the texts were grouped into three levels (three Base Word Lists). The *Vocaprofile* provides a series of selected three Base Word Lists that are used to run the program. The program processes an input file (written materials) and automatically classifies words in the text into three levels based on these three Base Word Lists. The output is a series of word lists arranged alphabetically with their frequency of occurrences.

The main purpose of this analysis is to see what and how many words in the texts are and are not in the lists. By having this, we can see how many words in the texts are among the high frequency words of English (Level I and II or Base Word List I and II) and how many in the University Word List or Level III (Xue and Nation, 1984). This investigation is also useful to see to what extent the texts with the vocabulary they contain can help the learners develop their reading skills.

The words in Base Word List one and Base Word List two consist of the first 2,000 most frequent words taken from *A General Service List of English Words* (West, 1953), called GSL for short. According to West (1953, vii), the words represent a list of 2,000 General Service Words and considered suitable as the basis of vocabulary for learning English as a second language. Thus, Base Word List one consists of the first 1,000 most frequent of GSL and Base Word List two consists of the next 1,000 most frequent words of GSL. Base Word List Three consist of words not in any of these lists, the first 2,000 most frequent words, but which are frequent in university texts from wide range of disciplines. These words were originally from the University Word List prepared by Xue and Nation (1984).

The selection of words for Base Word List one and two is based on word frequency. Words with frequency of 332 or more occurrences per five million running words in the GSL were put in Base Word List One (1^{st} 1,000 most frequent words). Words with frequency of 331 or less per five million running words in the GSL were put in the Base Word List two, 2^{nd} 1,000 most frequent words.

It should be taken into account that the classification of words in the text is mainly based on word families. It means that words with the same family are considered one word family. Take, for example, the root *relate* with its derivational words, was divided into several word families as follows:

* relate – relates – related – relatedness – relating - unrelated
* relation – relations
* relationship
* relative – relatively - relativity

Thus, if all these words occur in a text, they are considered as 4 word families, *relate, relation, relationship*, and *relative* whose frequency of occurrence is 6, 2, 1, and 3 respectively. This word

family classification is based on derivational affixation of level 1 up to 4 of Bauer and Nation (1993). The idea is that when a student knows the word *relate*, it is assumed that he knows the word *relates*, *related*, relatedness, *relating*, and *unrelated*. Therefore, these words are considered one word family. It is also assumed that he cannot recognize the words *relative* and *relationship* based on the base word *relate*. Therefore, they are considered different word families. In addition, the number of running words and different words or word types in the texts are presented. These data can be used to find the density index of the text being used

RESULTS AND DISCUSSION

As the aim of the study, the investigation on the reading materials is limited to the condition of vocabulary in the texts. This aspect is usually used as the basis to judge the acceptability of a reading materials. The discussion of vocabulary is divided into two subsections, that is, the discussion of high and low frequency words in the texts and the classifications of word levels.

High and Low Frequency Words

The study identifies the proportion of high and low frequency words that occur in the texts. As has been discussed, the computer program reveals that the texts consist of 7,945 running words. Table 1 lists the distribution of word occurrence beginning from the most frequent to the least frequent. It was found that the most frequent word occurred 515 times and this accounts for 6.49% of the words in the text. There is only one item (the) found with this frequency. The next most frequent word occurs 238 times and it covers 3.01% of the words. The cumulative percentage of this word together with the previous one is 9.50% and so on.

Looking at the low frequency words, the condition is somewhat different. The table shows that the least frequent words, the words occurring once, dominate the texts. There are 1,346 words or 16.97% of the words of this type found in the texts. This means that in the average of ten words, about one to two words occur once. Thus, having met these words for the first time, the reader would not meet again. Obviously, no repetition for these words is found by the reader working through the texts.

Table 1. The Distribution of High and Low Frequency Words in the Texts

No	Frequency/ range	Number of words	Cumulative percentage (%)	Percentage of words (%)
1	515	1	6.49	6.49
2	238	1	9.50	3.01
3	229	1	12.38	2.88
4	211	1	15.04	2.66
5	201	1	17.58	2.54
6	187	1	19.94	2.36
7	117	1	21.41	1.47
8	102	1	22.70	1.29
9	66-87	4	26.39	3.69
10	46-65	7	31.29	4.90
11	40-45	5	34.02	2.73
12	28-39	8	37.38	3.36
13	25-27	6	39.36	1.98
14	22-24	6	41.11	1.75
15	19-21	7	42.89	1.78
16	17-18	9	44.85	1.96
17	14-15	13	47.24	2.39
18	12-13	13	49.27	2.03
19	10-11	27	52.86	3.59
20	9	9	53.88	1.02
21	8	12	55.09	1.21
22	7	26	57.39	2.30
23	6	42	60.57	3.18
24	5	56	64.10	3.53
25	4	93	68.79	4.69
26	3	153	74.58	5.79
27	2	335	83.03	8.45
29	1	1346	100.00	16.97

Furthermore, the number of words whose frequency is 4 and less is 2,847 words or 35.90% of the running words in the texts; whereas, the occurrences of the four most frequent words are, 515, 238, 229, 211 which account for 6.49%, 3.01%, 2.88%, and 2.66% respectively. These, altogether, account for only 15.04% of the words. This condition indicates that the proportion of low frequency words in the texts is greater than that of the high frequency words.

In the case of vocabulary learning through reading, too much proportion of low frequency words is of a little value. This suggests that students working their way through all the texts would be continuing meeting words that they would not meet again or would meet only a few times in the texts. Ideally, the texts do not consist of too many low frequency words but they have more high frequency words, especially those of content words. At this point, long continued texts of the same topic might be more valuable than a series of unrelated texts. This kind of text can provide a lot of repetition that is important for optimal learning to occur. Meanwhile, a series of unrelated texts must use a bigger proportion of low frequency words.

Furthermore, it was also found that most of the high frequency words in the texts consist of function words. Of the first 50 most frequent words, there are only two content words, i.e., *people* and *said;* the rest are, of course, function words. In a study done by Sutarsyah *et al.* (1994), there are 18 content words in the 50 most frequent words from economic corpus.

From the list, we can also identify some function words from these first 50 most frequent words successively, i.e., *the, to, of, a, and, in, he, was, it*, and *had.* In the second 50 most frequent words, more content words are found, such as *read, book, like, now, come, long, make, eyes, look, money, manager, see, top*, etc. This indicates that many function words dominate high frequency words in the texts. The texts being used are not rich enough as to develop students' reading skills because there are too many function words and limited content words which occur in the texts. Many content words are of low frequency in the texts. This means that the students have a little opportunity to learn and enrich their vocabulary as to increase their reading skills because they are not provided with enough repetitions of these words. This text condition is of a little value to the students who learn to read.

To conclude, we can have some points that can be taken into account. The proportion of low frequency words found in the texts is much greater than that of the high frequency words. In written texts, the writer has more processing time. Because of desire to be precise in writing, and simply because of formal convention of writing, low frequency words often appear (Brown, 2001: 305). These low frequency words can present stumbling blocks to learners. Having too many low-frequency words, the readers tend to forget them when meeting for the first time in the series of unrelated texts.

The function words dominate the texts with considerably high frequency. Even though the number of function words is smaller than that of content words, function words are used more frequently in any text (Bolinger, 1975). These words belong to classes that are relatively closed. Content words, on the other hand, which are of a great value to reading learners, are very limited in the texts, and yet their occurrences are considered low. This implies that learning content words are more important than function words because function words do not develop. Learning function words does not give much contribution to increase students' reading performance.

Word Levels

As previously mentioned, this study also classifies words in the texts into three levels based on the three base word lists provided in the computer program. Table 2 shows that most words belong to the first level. There are 6515 tokens or running words (82%) which account for 1147 word types or 54%. This makes up 706 word families with the coverage of 44.65%. In other words, less than a half of the word families in the texts are included in the most frequent 1,000 words of GSL.

The number of word families at level two is almost a half as many as the number of word families at level one (the first 1,000 words), that is, 335 word families or 21.19%. These two groups of words belong to the first 2.000 words GSL, that is 65.84% of the word families in the texts. In fact, with this proportion of high frequency words in the texts, it seems rather difficult for the students to read the texts. Ideally, to read optimally, a reader is expected to know 90% coverage of words in the texts. These words are mostly available in the first 1,000 words and some in the second 1,000 words. In other words, the first 2,000-word level is the basic vocabulary needed for every learner of English as a second/foreign language to read English text and to have other skills as well. Numerous studies, for example, Hwang, 1989; Hwang and Nation, 1995; Sutarsyah *et al.*, 1994, have confirmed that this high

frequency vocabulary accounts for around 80% of the total running words (tokens) of the texts.

Furthermore, the number of the words at level three or University Word List in the texts is very limited. It consists of 122 word families or 7.72% of word families. In fact, this group of words is important for the learners to develop their reading skills in academic study.

Studies on word analysis with two corpora by Sutarsyah *et al.* (1994) and Sutarsyah (1993) show a different result from the present study. The studies identify the vocabulary in the two texts of different features, economics and academic vocabulary. The two texts consist of roughly the same length, that is, 300,000 running words. These differences can be seen in Table 3 at all levels.

The table shows that the proportions of word families at level one in both corpora are 77.72% and 74.11% and at level two, 4.78% and 4.32%. Thus, compared to these two texts, the texts in the present study are significantly dominated by words from the second 1,000 words level. The proportion of words from the second level (2^{nd} 1,000) in this study is much greater than that of the words of the same level in the other two studies. In fact, the texts need more words from the first 1,000-word level so that the students can read optimally. The table also shows that the other two word analyses are comparable in that they have almost the same proportion. However, the proportions of words in the texts in the present study are quite different from those other two.

Another constraint is that a group of words not in any of these levels is too large, that is, 26.44% of the total word families in the texts. Most of the words in this group are not very useful for student's reading development. These words are uncommon in most texts. Some of them are the name of people, places or countries such as *Africa, America, Burma, China*, etc. The words on the name of people were even found to have high frequency of occurrences such as *Della* (7), *Cheseborough* (11), *Henry* (15), *Jim* (14), etc. Moreover, it was also found that a number of words belong to technical or subtechnical words which are usually difficult for learners. Take, for example the words *apothecaries, audible, catastrophes, casmo, damning, drought, embankments, extinguish*, etc. This seems to be particular to and useful for a specific area of knowledge. On the other hand, in this group we can find many complicated words such as *apologetically, drifting, complacency, disastrous, engenders,* etc. Therefore, readers whose experience is limited in such specific area will find the texts difficult.

In terms of density index, the text or the course book being used is considered rather difficult for students. It is found that, based on the data in Table 2, the density index of the text is 1: 3.77 (2107: 7945), which means that on an average each word is repeated between three to four times. This density index is still considered low. In terms of vocabulary learning and reading development for the first year students the text with this index is inefficient. As has been mentioned, an efficient textbook for this group of students has density index of about 1:20 with a low number of "one timers". In fact, most modern English textbooks have density index of 1: 2.4 containing 40% of words of one occurrence (Nation, 1990).

It seems that the unfavorable aspect of the texts is mainly due to the choice of texts which are made of a series of unrelated texts. This kind of texts can increase a vocabulary load of the course enormously because the texts consist of many low frequency content words and technical words. Thus, it seems that the texts cannot provide enough repetition for content words which are valuable for skill development of the learners.

Table 2. The Number of Words and Percentage of Coverage in the Texts

Word Level	Tokens	Types	Families
The first 1,000	6515 / 82.0%	1147 / 54.40%	706 / 44.65%
The second 1,000	608 / 7.7%	402 / 19.1%	335 / 21.19%
UWL	185 / 2.3%	140 / 6.6%	122 / 7.72%
Others	637 / 8.0%	418 / 19.8%	418 / 26.44 %
Total	7945/ 100.00%	2107 / 100.00%	1581 /100.00%

Table 3. The Comparison of Word Coverage of the Present Study and the Study of Sutarsyah *et al.* (1994)

Word Level	Present study	Sutarsyah *et al* (1994) and Sutarsyah (1993)	
		Economic corpus	Academic corpus
The first 1,000	44.65%	77.72%	74.11%
The second 1,000	21.19%	4.78%	4.32%
UWL	7.72%	8.74%	8.40%
Others	26.44%	8.77%	13.16%

This finding is in line with the study done by Sutarsyah *et al.* (1994) which compared one coherent, continuous text by one writer on one topic and a series of unrelated texts made up of a variety of academic texts by many writers. The study reveals that a coherent text by a single writer on a single broad topic uses a very much smaller vocabulary than a series of unrelated texts. This means that vocabulary load increases when learners are provided with reading materials containing many different texts of unrelated topics.

CONCLUSION AND SUGGESTION

Conclusion

This paper has discussed an investigation of vocabulary on reading materials used by the students and it yields some interesting points. The study has examined the vocabulary used in the materials in terms of frequency of occurrences and word levels. It reveals that function words dominate high frequency words in the texts. The texts are also dominated by low frequency content words which occur outside the three levels. This condition is not favorable for reading skill development because the students have a little opportunity to learn these important words. In other words, the texts cannot provide enough repetition of content words for students to learn and thus, the text is difficult to read.

In terms of word levels, the texts consist of too many words outside the three levels. This implies that the students are faced with vocabulary learning load because the texts contain a large number of words of this type. In addition, the texts provide limited words from the first 1.000 words of GSL so that the students have difficulty reading them. At the same time, the proportion of the words in the first 2,000-word level is also inadequate to make the students read optimally. One reason for this constraint is that the texts were made of a series of unrelated texts with different topics or themes. Because of the different topics, the texts or the course book tend to have many different words and they are usually of low frequency of occurrences.

Suggestion

The teachers and course designers are suggested to avoid the use of reading materials consisting of a series of unrelated texts, even though it is quite impossible to have the whole reading materials made of related texts for one semester. So we can consider making a course which consists of a few themes so that the texts within a theme bear more relationship to each other. The texts, thus, will need the use of a smaller amount of vocabulary.

REFERENCES

Alderson, J.C. 1984. Reading in a Foreign Language: A Reading Problem or a Language Problem? In J.C. Alderson & A.H. Urquhart (Eds.), *Reading in a Foreign Language* (pp. 221-235). London: Longman.

Bauer, L. & Nation, I.S.P. 1993. Word Families. *International Journal of Lexicography*, 64: 253-279.

Bolinger, D. 1975. *Aspect of Language*. New York: Harcourt Brace Jovanovich.

Brown, H.D. 2001. *Teaching by Principles: An Interactive Approach to Language Pedagogy*. New York: Addison Wesley Longman.

Chandrasegaran, A.1981. *Problems of Learning English as a Second Language*. Singapore: SEAMEO Regional Language Center.

Dakin, J. 1973. The teaching of Reading. In H. Fraser and W.R. O'Donnell (Eds.), *Applied Linguistics and the Teaching of English* (pp. 118-131). London: Longman.

Day, R.R. & Bamford, J. 2000. Reaching Reluctant Readers. *English Teaching Forum*, 38 (3): 12-17.

Depdikbud, 1993. *Garis-garis Besar Program Pengajaran (GBPP) Mata Pelajaran Bahasa Inggeris Sekolah Menengah Umum (SMU) tahun 1994*. Jakarta: Depdikbud.

Huda, N. 1992. The 1994 English Syllabus for Secondary Schools: Issue and Problems. *Bahasa dan Seni*, 20 (51): 2-14.

Hwang, K. 1989. *Reading Newspaper for the Improvement of Vocabulary and Reading Skills*. Unpublished MA Thesis. Wellington: English Language Institute, Victoria University of Wellington, New Zealand.

Hwang, K. & Nation, I.S.P. 1995. Where would General Service Vocabulary Stop and Special Purposes Vocabulary Begin? *System*, 23 (1): 35-41.