PAC3
at
JALT
2001
Conference
Proceedings

MENU
Text Version
Help & FAQ

International
Conference
Centre
Kitakyushu
JAPAN
November
22–25, 2001

Corpus Analyses of Textbook Vocabulary in Taiwan

Chia-hsing Pan
Shu-te University, Taiwan
Nai-hua Ko
Shu-te University, Taiwan

The importance of English was officially realized in primary education in Taiwan in 2001. The consecutiveness of English education throughout primary school to junior high school has become a major concern. With the use of the *CCLang* program, this study intends to investigate the vocabulary in the main reading texts of the most popular English textbooks of primary schools in Taiwan, and compare with those of junior high schools to understand if their word frequency distributions and sentence length are significantly different, and in what way. The study corpora are also compared respectively with two comparative corpora, the *Brown Corpus* and the *LOB Corpus*. The results show no significant difference between the study corpora themselves and with the comparative corpora. The study corpora are all targeted at EFL beginners. However, the junior high corpus is supposed to be more difficult in terms of sentence complexity, but is similar in word frequency distributions.

台湾で小学校英語教育が正式に始まったのは、今年の九月である。国民からの要求の声に答えるうえに、小学校六年間と中学校三年間のカリキュラムを一貫的に計画するという「九年一貫」を通して、全面的に改革しようとしているので、とうとう台湾の文部省は今年から、正式に、全面的に小学校英語教育を実行し始めた。けれども、英語教育が始まるとともに、教師の数が足りないとか、教科書の選択とか、中学校のカリキュラムとの連関性とか、さまざまな

問題や困難が浮いてきた。語学教師として、現状にある問題点を深く探求したいと思っているので、この研究を始めとして、教科書のことをピックアップして、その語彙とセンテンスの難易度について、研究してみた。研究方法については、もっとも小学校が採用しているテキストを研究対象として、その本文の会話文の内容をコンピュータにインプットして、ＣＣＬａｎｇというソフトポログラムで分析して、その語彙やセンテンスの難易度を比べてみた。また、それぞれ“Brown Corpus”と“LOB Corpus”という二つの語彙群と比べてみた。結果としては、語彙から見ると、大きな差異が見つからなかった。また、語彙の難易度もほとんど初級英語の程度にあっている。しかし、これから「九年一貫」のカリキュラムを考えると、現在中学校の教科書にある語彙の難易度はもっと検討するべきだと思われる。

Introduction

Background and Motivation

With the economic or political advantages of the spread of English, Taiwan, as well as other non-English countries, is at a juncture of incorporating more English instruction into the educational system, especially in primary education. Children's English in Taiwan has been flourishing under the belief that "the earlier, the better." More and more children have been engaged in English lessons, either in cram schools or in primary schools, long before they attend junior high school where English instruction officially begins. This year, however, English education was implemented in all primary schools in Taiwan according to the policy of a nine-year consecutive curriculum of primary and secondary education.

One issue that has been heatedly discussed is the selection of teaching materials. Textbooks are what curricular and instructional procedures are based on; therefore, they in part determine the success of language learning. Of the components in teaching materials, vocabulary acquisition is of great importance in that vocabulary acquisition may dictate success in language acquisition. This study will focus on vocabulary of the main texts in both two kinds of textbooks since vocabulary plays an essential role in the acquisition of the four language skills: listening, speaking, reading, and writing. Besides, there have been a lot of findings of frequency effects that facilitate the acquisition of vocabulary (Chuang, 1999; Forster & Chambers, 1973; Rubenstein, Carfield & Millikan, 1970). As a result, an analysis of vocabulary in English textbooks of both educational systems can be an intriguing attempt to quantitatively analyze how and why word frequency effect and length of sentence may have an impact of reading comprehension and/or learning motivation. In particular, if learners employ the English textbooks of primary and junior high schools selected in this study, would they find them appropriate for their English proficiency in each phase in terms of vocabulary and sentence length?

According to Kennedy (1998, p.94), "learners of a

language should be exposed first to reading the most frequently occurring words in the language." Therefore, word frequency distributions of the textbooks will be discussed and compared. In addition, the time and effort of processing a sentence would increase in proportion to the length of the sentence. Consequently, sentence length in textbooks of the study can be one indicator of text complexity or even difficulty. In other words, the longer a sentence is, the more difficult it may be. Categories of conjunctions, prepositions and infinitives all signify longer sentences, and therefore more complex sentence structures (Chuang, 1999, p.86).

This study is aimed at analyzing vocabulary of the most popular English textbooks for primary school students and the unified version of English textbooks in junior high schools with a corpus-based approach. It scientifically and quantitatively substantiates vocabulary learning of textbooks, a core of instruction and learning. The main texts of textbooks will serve as the study corpora to understand whether these textbooks contain great discrepancies in terms of word frequency and sentence complexity. It is hoped that this investigation, that examines vocabulary in particular, can draw more people's attention to vocabulary acquisition.

Research Questions

1. What are the most frequent words in the three study corpora and two comparative corpora? How are they different or similar?
2. Are the primary school English textbooks easier than the junior high school ones in terms of sentence length?

Review of the Literature

Related background knowledge of corpus linguistics

Corpus linguistics, like all linguistics, is concerned primarily with the description and explanation of the nature, structure and use of language, and languages and with particular matters such as language acquisition, variation, and change (Kennedy,1998) With the advances of computers, the scope of corpus linguistics has been enormously extended. Since the *Brown Corpus* in 1961, the first significant computer American English corpus, was established, a great number of corpora have been built for different purposes with the power of both computers and computer software available. Computers have permitted linguists to work with a large variety of texts and thus to seek generalizations about language and language use which can go beyond particular texts or the intuitions of particular linguists (Kennedy, 1998).

In 1961, the first significant computer corpus, the Brown Corpus, was compiled. It was a synchronic corpus of approximately one million words representative of the written English printed in the United States. Following the example of the *Brown*

Corpus, the *Lancaster-Oslo/Bergen (LOB) Corpus* was intended to be a British English counterpart to the *Brown Corpus* between 1970 and 1978, and contains 1,006,825 words. By the 1990s, the second-generation computer corpora contain 100 million words or more, such as the *Cobuild Corpus*. It was a joint achievement between Collins, a commercial publisher, and Cobuild, a research team at the English Department of the University of Birmingham. The great importance of it lies in the size of the computer corpus and the association of making corpora with a particular commercial research and development project to produce corpus-based dictionaries, grammars and language teaching courses. Other corpora of the second generation also include the British National Corpus and the International Corpus of English.

Corpora have provided a more realistic foundation for the study of language and a fruitful basis for comparing different varieties of English and for exploring the quantitative and probabilistic aspects of the language (Cited in Chuang, 1999). The essential characteristics of corpus-based analyses are (Biber, Conrad & Reppen, 1998):

1. it is empirical, analyzing the actual patterns of use in natural texts;
2. it utilizes a large and principled collection of natural texts, known as a " corpus," as the basis for analysis;
3. it makes extensive use of computers for analysis, using both automatic and interactive techniques;
4. it depends on both quantitative and qualitative analytical techniques.

The corpus-based approach should not be regarded as the unique correct approach, but as a complementary approach to more traditional approaches. For example, structural analyses, one of the traditional approaches, are related to the corpus-based approach. Moreover, lexicography and grammar can be significantly examined with it. For instance, dictionary makers can benefit from obtaining the information about the most frequent words and their uses, the collocational patterns of related words, and the contexts in which words occur most frequently by studying the lexical and grammatical associations in representative corpora (Chuang, 1999).

Research Methodology

Data Collection

Due to time limit, the author decided to go directly to qualified publishers of English textbooks proved by the Ministry of Education for their distribution lists over all the primary schools in Taiwan. Those qualified publishers for the fifth-grade include Nan-i,(NI) Zen-ling, Melody, Oxford, Prince Hall, HESS, Kwang-Fu (KF), Kan-Shan (KS), Chia-In (CI), Longman, Kid Castle (KC), Giraffe, Banana Boat, and CYC School.

According to the data, what school chose which textbook was clearly identified and recorded. Those for the sixth grade are HESS, CI, NI, KC, KF, and KS. For example, take Chang-hwa County and Chang-hwa City. The most popular textbook is KF, about 4308 copies (57 among 158 schools, about 36.7%) for the fifth grade and 5870 copies (55 among 158, about 34.8%) for the sixth grade. After an overall investigation, the most popular textbooks are HESS and KF. In the north Taiwan, HESS is the most frequently used; in the middle or south Taiwan, KF. According to the publishers, children in north Taiwan have been exposed to English earlier than those in the other parts of Taiwan. Since for most of the publishers, within such a short time only the first volumes for both grades are available, the author decided to analyze both textbooks, the two representatives to all the population.

Data Processing

The main texts of English textbooks from HESS and KF publishers were computerized to form two corpora as well as the unified version of junior high school English textbooks (only the first volume) to form another corpus, Junior High English Textbook Corpus (JHET Corpus). All the data were respectively processed again using *CCLang* so that all the data such as word frequency lists, word tagging and sentence length could be obtained for statistical significance tests and reference purposes. *CCLang*, computer software for corpus analysis developed by Cheng (1998, 1999) at the University of Illinois at Urbana-Champagne, was utilized to investigate the word frequency and the average length of sentence. Finally, the SPSS for Windows was used to perform statistical significance tests required to test the descriptive statistics for this study.

Data Analysis

On the basis of word frequencies and lexico-grammatical features, quantitative data analyses were performed to find out the descriptive and inferential statistics of HESS, KF and JHET reading texts. In the descriptive statistics, using the *CCLang* program would show word frequencies and their ranking. Furthermore, with the help of word tagging, the word tag frequencies of the reading texts were grouped into twenty grammatical categories such as nouns, adjectives, and so on. Every frequency was further analyzed by a significance test, *Pearson Product Moment Correlation*, to see if HESS, KF and JHET are significantly different. Finally, qualitative interpretation would be demonstrated to account for the results.

Results and Discussion

Findings

The *Longman Corpus*, the *HESS Corpus*, and the

KF Corpus are the most widely used primary school English textbooks in the fall semester of 2001 in the metropolitan cities in Taiwan according to the data given by the publishers. Another study corpus is the first volume of English textbooks for junior high schools. In Table 1, graphic word types and word tokens of the six corpora are presented. The numbers of those of the JHET Corpus are far larger than those of commercial ones. However, compared with the graphic word types and word tokens of the Brown Corpus and the LOB Corpus, those of the JHET Corpus and the commercial market are very small in number. Moreover, the sentence lengths of each of the three study corpora are found (See Table 2). Apparently, the JHET Corpus has the longest sentence length and is more difficult for learners, while the Longman Corpus has the shortest sentence length.

Table 1: *The Graphic Word Types and Word Tokens in the Eight Corpora*

Corpus	Graphic Word Types	Word Tokens
Longman	45	62
HESS(fifth)	81	193
HESS(sixth)	66	146
JHET	199	744
KF(fifth)	107	284
KF(sixth)	138	363
Brown	42,958	1,000,000
LOB	40,626	1,000,000

Table 2: *Sentence Length of the Six Study Corpora (only one volume for Longman and one for JHET)*

Grade	Longman	HESS	KF	JHET
Fifth grade	2.39	3.45	3.05	6.44
Sixth grade	2.39	3.04	3.29	6.44

The 20 most frequent words in the eight study corpora, including the six study corpora and the two comparative corpora, are discussed in Appendix 1. The numbers under each corpus are the frequencies and the percentages for each graphic word types that are used most frequently in the respective corpora. The reason why only the twenty most frequent words were selected is because the population of those corpora was too small and most of the words only appeared once.

It is clear that in all the six corpora the 20 most frequent words are very similar, including common pronouns, prepositions, articles, and verbs. Generally speaking, most of them are function words. Unlike content words, the number of function words is limited. Therefore, students usually enlarge their vocabulary sizes in content words. Moreover, there are several points worth noting. First, the top five graphic word

types in both primary and junior high school corpora do not correspond to those of the *Brown Corpus* and the *LOB Corpus*. However, those corpora are very similar in their first five frequently used words, most of which are pronouns. Second, "what" appears in all the six study corpora instead of the comparative corpora. It can be assumed that the six study corpora are the English textbooks especially for beginners who very often have to learn simple questions first, including wh-questions. Third, the pronoun "you" was not found in the 20 most frequent graphic word types in the Brown Corpus and the LOB Corpus but in the other corpora. This may reflect its discoursal use. The pronoun is frequently used in colloquial texts. The corpora investigated are all for young learners in the beginning level and are presented in conversations a lot. As a result, it is reasonable to see the use of "you" in them instead of the two comparative corpora of the written genre in nature. Fourth, the percentage of the definite article of the JHET Corpus was not higher than that of the other corpora. Surprisingly, the textbooks of Longman and HESS for the sixth grade and that of KF for the fifth grade did not have the definite article at all on top of the twenty frequent words. It can be seen that pronouns are used more frequently instead of "the" since it is easier for learners to identify what or who the pronouns refer to. Even for adult learners of foreign languages, "the" is still a big hurdle to them. However, it cannot be ignored in daily-life conversation. Even beginners are supposed to use the most authentic teaching materials without deliberately ignoring the existence of "the." Fifth, the coordinating conjunction *and* was found in the JHET Corpus as well as HESS for the sixth grade but was not found in the other corpora among the 20 most frequent words. It reflected the facts that the conjunction can make a sentence longer and more complex and beginners may find it difficult to construct such complex sentences. Sixth, *is* was found in the most frequent graphic word types in all corpora. Seventh, "to," the infinitive that can make a sentence longer, only appeared in the sixth-grade book of HESS. Finally No past tense was found since the past tense was harder for the beginners than the present tense and was supposed to be presented later.

In order to calculate the proportion of the 20 most frequent words that overlap in the Longman Corpus and the JHET Corpus, the HESS Corpora and the JHET Corpus and the KF Corpora and the JHET Corpus, *CCLang* is used to calculate their graphic word types. It is found that the percentage of the 20 most frequent words that overlap with those of the JHET Corpus is 55% in the Longman Corpus, 60% and 60% in the HESS Corpora, and 60% and 65% in the KF Corpora. This shows that those three corpora are all intended for English beginners; therefore, the overlapping is large.

Moreover, the results of the Independent Samples T-

test for all of the corpora are not significantly different.

Conclusion

In this study, two research questions are answered to provide insights into the word frequency distributions of vocabulary in the primary school English textbooks and the new versions of junior high school English textbooks. First, on the basis of the descriptive statistics of the graphic word types in each corpus, independent sample T-tests were conducted and a no significant differences were found. In other words, the most popular English textbooks of primary schools in three cities are not significantly different from the unified version of the junior high school English textbooks. Hence, if they are taught to different levels of students, the motivation of high-level students will be decreased. Second, the junior high school textbooks are more difficult than the primary school ones in terms of sentence length and the number of word tokens.

References

Biber, D., Conrad, S. and Reppen, R. (1998). *Corpus linguistics: investigating language structure and use*. Cambridge, U.K. Cambridge University Press

Chang, Y.L. (1995). *A reading package of a children's English program*. Unpublished manuscript. National Kaohsiung Normal University.

Chou, C.T. (1991). A study on the effectiveness of the early start in learning English as reflected in children's English competence. *English Teaching & Learning,* 15 (4), 45-54.

Christenbury, L. & Kelly, P.P. (1994). What textbooks can and cannot do. *English Journal*, March, 76-80.

Halliwell, S. (1992). *Teaching English in the primary classroom*. New York: Longman Group UK Limited.

Marks, C. B., Doctorow, M. J., & Wittrock, M.C. (1974). Word frequency and reading comprehension. *Journal of Educational Research,* 67 (6), 259-262.

Moa, L. W. (1993). Discussion of foreign language teaching at primary school. *New Horizon Bimonthly for Teachers in Taipei*, 67, 2-6.

Nadler, H. (1969). *Criteria for the selection of ESOL materials*. Paper presented at the Third Annual TESOL Convention, Chicago, 5-8.

Richards, J. C., & Rodgers, T. S. (1986). *Approaches and Methods in Language Teaching: A description and analysis.* Cambridge: Cambridge University Press.

Shih, S. C. (1992). Cognition and children English teaching. *English Teaching & Learning,*17, (2), 57-67.

Skierso, A. (1991). Textbook selection and evaluation. In Marianne Celce-Murcia (Ed.). *Teaching English as a Second or Foreign Language.* Boston: Heinle & Heinle Publishers.

Solomon, M. (1978). Textbook selection committees: What teachers can do. *Learning,* 6, (7), 43.

Sorenson, A. (1967). Multilingualism in the Northwest Amazon. *American Anthropologist,* 69, 670-684.

Stahl, S. A. & Fairbanks, M. M. (1986). The effects of vocabulary instruction: A model-based meta-analysis. *Review of Educational Research*, 56(1), 72-110.

Tsao, F. F, Wu, Y. H., & Hsieh, Y. L (1994). A continuous study of English learning of third-grade primary school students. *Educational Research & Information*, 2 (3), 111-122.

Tsao, S. H. (1994). A study on children English learning in Taipei area. *New Horizon Bimonthly for Teachers in Taipei,* 2, 49-61.

Wang, C. C. (1996). The third contact of children English and junior high school English. *Caves English Teaching,* 10, 19-24.

Appendix 1

Rank	Longman			HESS (5th)			HESS (6th)			KF (5th)			KF (6th)			Junior			Brown		LOB	
1	S	56	7.53%	YOU	6	5.88%	S	19	9.85%	YOU	12	8.22%	I	22	7.75%	IT	18	4.96%	THE	6.77%	THE	6.59%
2	A	33	4.44%	I	4	4.41%	IT	17	8.81%	I	8	5.48%	YOU	17	5.99%	WHAT	15	4.13%	OF	3.52%	OF	3.50%
3	IS	26	3.50%	YES	3	4.41%	WHAT	11	5.70%	IS	7	4.80%	S	16	5.63%	S	15	4.13%	AND	2.79%	AND	2.76%
4	YOU	21	2.82%	‘S	3	4.41%	IS	11	5.70%	WHAT	5	3.43%	IS	12	4.23%	IS	15	4.13%	TO	2.53%	TO	2.63%
5	MY	19	2.55%	IS	3	4.41%	YOU	9	4.66%	SHE	4	2.74%	M	10	3.52%	I	15	4.13%	A	2.27%	A	2.22%
6	IT	19	2.55%	ARE	2	2.94%	HOW	5	2.59%	SARAH	4	2.74%	HAVE	10	3.52%	YOU	14	3.86%	IN	2.07%	IN	2.06%
7	CAN	18	2.42%	THIS	2	2.94%	ARE	5	2.59%	S	4	2.74%	WHAT	8	2.82%	ARE	11	3.03%	THAT	1.04%	THAT	1.10%
8	THIS	17	2.29%	OK	2	2.94%	YOUR	4	2.07%	NAME	4	2.74%	ARE	7	2.47%	A	9	2.48%	IS	0.97%	IS	1.06%
9	I	16	2.15%	‘M	2	2.94%	NUMBER	4	2.07%	ARE	4	2.74%	TOMMY	6	2.11%	M	8	2.20%	WAS	0.94%	WAS	1.00%
10	WHAT	15	2.02%	HI	2	2.94%	NAME	4	2.07%	YOUR	3	2.06%	IT	6	2.11%	TIME	6	1.65%	HE	0.94%	IT	0.98%
11	GOOD	14	1.88%	HELLO	2	2.94%	M	4	2.07%	YES	3	2.06%	OLD	5	1.76%	THE	6	1.65%	FOR	0.91%	FOR	0.90%
12	TOM	12	1.61%	AM	2	2.94%	A	4	2.07%	TO	3	2.06%	MY	5	1.76%	YOUR	5	1.38%	IT	0.88%	HE	0.84%
13	ARE	12	1.61%	YUMMY	1	1.47%	WEATHER	3	1.55%	NICE	3	2.06%	HI	5	1.76%	SHE	5	1.38%	WITH	0.70%	AS	0.71%
14	YOUR	11	1.48%	YOUR	1	1.47%	THE	3	1.55%	MORNING	3	2.06%	HE	5	1.76%	ON	5	1.38%	AS	0.70%	I	0.70%
15	SHE	11	1.48%	WHAT	1	1.47%	TELEPHONE	3	1.55%	MEET	3	2.06%	YOUR	4	1.41%	MY	5	1.38%	HIS	0.67%	BE	0.70%
16	M	11	1.48%	WELCOME	1	1.47%	NO	3	1.55%	M	3	2.06%	THANK	4	1.41%	FOR	5	1.38%	ON	0.65%	WITH	0.70%
17	AND	11	1.48%	THAT	1	1.47%	MY	3	1.55%	IT	3	2.06%	SHE	4	1.41%	BOOKS	5	1.38%	BE	0.61%	ON	0.68%
18	THE	10	1.34%	THANK	1	1.47%	I	3	1.55%	GOOD	3	2.06%	NAME	4	1.41%	THESE	4	1.10%	I	0.57%	HIS	0.60%
19	THAT	10	1.34%	SWIM	1	1.47%	HAPPY	3	1.55%	DO	3	2.06%	JENNY	4	1.41%	THANK	4	1.10%	AT	0.52%	AT	0.58%
20	HI	10	1.34%	SOUNDS	1	1.47%	YES	2	1.04%	AND	3	2.06%	HOW	4	1.41%	PLAY	4	1.10%	BY	0.51%	BY	0.57%