A Corpus-based Analysis of Cross-cultural E-pal Projects

Nam-Sook Chung

Department English Language Education, The College of Education B/D 17
Haendang-Dong, Sungdong-Gu, Seoul 133-792, Republic of Korea
sshsm2@hanmail.net

The purpose of this study is to report the e-pal projects based on a web-based BB system, and describe how vocabulary frequencies differ among the distinct learners who took part in classroom-based cross cultural e-pal projects for seven years. WordSmith software tool and Range BNC program were used to compare the vocabulary of up to 12 different countries’ texts. The results show that (1) regardless of nationality, the most frequent top 12 words are \textit{I, the, to, and, is, in, of, a, you, my, it, and that}. (2) as for vocabulary distribution, learners from China, Taiwan and the UAE are the highest users in high frequency words. Learners from Burkina Faso, and Argentina are the highest users in low frequency words. (3) high-frequent content verbs are \textit{tell, get, go, know, think, see, make} and \textit{want}. A corpus-based research can offer more reliable and authentic guide to vocabulary teaching and grammar.

1. Introduction

The focus of second language learning and teaching has been promoting communicative ability and creative self expression through interaction in a social context. However, it is fairly difficult to find enough interactive opportunities to promote communicative ability in EFL contexts. These e-pal projects have made it possible to expand the scope of interaction to computer-based interaction. In addition, corpus-based research was possible on the basis of learner corpora. The data of all students’ electronic written texts was stored into host University data base automatically. Many corpus-based linguistic studies have been presenting a new and different perspective in terms of frequency, phraseology and collocation. Recently, the fields of using corpora are grammar-related reference books, ideology and culture, translation, stylistics, forensics and plagiarism.

The purpose of this study is to describe how vocabulary frequencies differ among the distinct language groups who took part in cross cultural e-pal projects for seven years. I divided the learners into four different language groups which are Asian, Western European, Central American and African learners. This study can give insight into grammar and vocabulary teaching and provide realistic information about high-frequency words found in the e-pal projects.

2. Literature Review

Corpus-based based studies have been conducted since the first language corpus, the Brown corpus which was begun in 1964. Representative corpora are Lancaster-Oslo/Bergen (LOB, 1978), the bank of English (Cobuild Corpus, 1980-1997), the British National Corpus (BNC,1991-1994), Longman Spoken and Written English Corpus (LSWE). Conrad (2000) argued that corpus linguistics could radically change grammar teaching in the 21st century. Mint (1996) reported that the most frequent 10 irregular verbs cover 45 % of total English irregular verbs.

Ringbom (1998) made use of seven Western European learner corpora from the International Corpus of Learner English (ICLE), comparing them with each other and with the native speaker corpus. The study shows that a seemingly simple word frequency count may provide a perspective on the general characteristics of advanced learner language. The finding is that L1 transfer and its universal features have important effects on learning a foreign language. High-frequency verbs are \textit{be, have, do, can}, and the most frequent main verbs are \textit{think, get, make, want, take, find, know, use, go, and live}. Make is the most frequent verb form to NS (native speakers), followed by use, believe, feel. On the other hand, NNSs tend to use other verbs, especially \textit{think, get, find, want, and know}.

3. Method

Range BNC program was used to find the coverage of 12 different countries’ texts by word lists Range program provides. It was designed by Prof. Paul Nation and Linguistics Studies at Victoria University in Wellington, New Zealand. There is no need to seek permission for its use or for its free distribution to others (www.lextutor.ca). WordSmith software Tool was applied to create the high frequent wordlists.

3.1 Research Questions

1) What high-frequency words are found in all nationalities?
2) How rich a vocabulary do each different country’s learners use in e-pal projects? How much coverage of each different learner corpora of the 12 base lists Range BNC program provides?

3) What is the difference of frequent content verb in each of the distinct four language groups learners, compared with native speakers’ corpus (LSWE)?

3.2 Research Environment

Data collection period is from the years 2001 to 2007. Students from 14 different countries including Y, S and H Korean University students joined in classroom-based international e-pal projects via emails. The host University was Meikai University in Japan, where an ibunka project organizer Prof. Watanabe taught. The features of the project are anonymous and password-protected group-based discussion. (Fig.1)

![Fig.1. The Bulletin board of e-pal project](http://lkte.meikai.ac.jp/wbb/login.htm)

3.3 Participating schools

**Asia Region**

**Korea:** Yongin Songdam College, Sookmyung Women's University, Hanyang University

**China:** Hubei Polytechnology University Shanghai Institute of Physical Education

**Taiwan:** National Taipei Teachers College, National Chiayi University

**Japan:** Meikai University, Yokohama National University, Obirin University

**UAE:** American University of Sharjah, Sharjah Womems College

**Indonesia:** 24 High school Bandung, 8 Senior High School at Yogyakarta SMA Negeri 8 Yogyakarta

**Central and North America Region**

**Mexico:** Colima University

**USA:** Houston University, Wilcox High school of Santa Clara

**South America Region**

**Argentina:** Federal University

**Brazil:** Universidade Federal de Minas Gerais

Western Europe Region

**Germany:** Berufsschule Schongau,

**Finland:** Malmi School of Business

Africa Region

**Burkina Faso:** YAMMAYA high school

**Namibia:** Polytechnic of Namibia

3.4 Writing Topics

1) School Life (Sep.17-Oct.13) : A short self introduction and school life

2) Daily Life (Oct.14-Nov.3) : Various aspects of students' domestic culture such as local festivals, local towns, local food.

3) Social Issues (World Peace) (Nov.4-Nov.24) : Various domestic and international issues, such as environmental issues, regional conflicts, war.

4. Result

4.1 What high-frequency words are found in all nationalities?

![Fig.2. High-frequency top 12 word](image)

Fig.2 shows the most frequent top 12 words among total corpora by the wordlist. They are all function words (pronoun, conjunction, article, and preposition) rather than content words. *I* is the most frequent word in all corpora, followed by *the*, *to*, *and*, *is*, *in*, *of*, *a*, *you*, *my*, *it*, and *that*. This implies that most students who took part in the e-pal projects were more conscious of grammar and form than of lexical meaning. *And* seems to be somewhat highly used by NNS, while *that* is used less relatively. *The* is used more than *a* by NNS.

4.2 How rich a vocabulary do each different country’s learners use in e-pal projects? How much coverage of each different learner corpora of the 12 base lists Range BNC program provides?

Fig.3 shows the percentage of vocabulary distribution for each of the 12 different countries.
The words in a text of each different country were compared with the words in 12 base lists. In the word list column, one, two, three refer to each of the base lists. One wordlist includes the most frequent 1,000 words, two wordlist includes the most frequent 2,000 words, and 12 wordlist includes the most frequent 12,000 words. The sources of these lists are *A General Service List of English Words* by Michael West (Longman, London 1953).

Fig. 3. Percentage of vocabulary distribution

One and two wordlists imply high-frequency words which are common and easy words. The higher the percentage, the higher the ratio of frequent words. The lower the percentage, the higher the ratio of low-frequency words. Students from Burkina Faso, Argentina, and Indonesia are the highest users in low-frequency words. On the other hand, students from China, Taiwan, the UAE, and Finland are the highest users in high-frequency 2,000 words. Students from Taiwan, Japan and Finland are high users in 4,000 words. Students from Korea and Brazil are high users in 5,000 words. Argentina students are high users in 6,000 words.

4.3 *What is the difference of frequent content verbs in each of the distinct four language groups learners, compared with native speakers’ corpus (LSWE)?*

Table 1 shows high-frequency top 10 content verbs out of total corpora. Learners from Asia, Western Europe and Central America used more the verb *think* than native speaker (LSWE). Asia and Central America learners used more the verb *think* than Western European learners. *Think* and *want* are overused by Asian learners. *Go*, *get* are relatively overused by Western European learners.

This shows that overused or underused words have a clear relation to the structure of the L1. The most high-frequency verb by NS is *say* while *tell* is used by NNSs from Western Europe, Central America, and Africa region. *Make* is overused by only Asian learners and NS. The words of *Live, study, visit, call, and love* may be related with the topic. All the frequent verbs except for *take* and *come* in NS corpus also occur in all learner corpora. Consequently, vocabulary teaching on the verbs of *come* and *take* in EFL context should be considered in the future.

Table 1. High-frequency content top 10 verbs

<table>
<thead>
<tr>
<th>Rank</th>
<th>Asia</th>
<th>Western Europe</th>
<th>Central America</th>
<th>Africa</th>
<th>LSWE (NS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Think</td>
<td>Go</td>
<td>Think</td>
<td>Know</td>
<td>Say</td>
</tr>
<tr>
<td>2</td>
<td>Want</td>
<td>Work</td>
<td>Know</td>
<td>Want</td>
<td>Get</td>
</tr>
<tr>
<td>3</td>
<td>Know</td>
<td>Go</td>
<td>See</td>
<td>Go</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td>Go</td>
<td>Think</td>
<td>See</td>
<td>Go</td>
<td>Know</td>
</tr>
<tr>
<td>5</td>
<td>Make</td>
<td>Get</td>
<td>Want</td>
<td>Think</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>See</td>
<td>Tell</td>
<td>Live</td>
<td>Hope</td>
<td>See</td>
</tr>
<tr>
<td>7</td>
<td>Live</td>
<td>Want</td>
<td>Get</td>
<td>Tell</td>
<td>Make</td>
</tr>
<tr>
<td>8</td>
<td>Get</td>
<td>See</td>
<td>Take</td>
<td>Live</td>
<td>Come</td>
</tr>
<tr>
<td>9</td>
<td>Study</td>
<td>Call</td>
<td>Tell</td>
<td>Learn</td>
<td>Take</td>
</tr>
<tr>
<td>10</td>
<td>Hope</td>
<td>Visit</td>
<td>Love</td>
<td>Wish</td>
<td>Want</td>
</tr>
</tbody>
</table>

5. Conclusion

E-pal projects corpora can shed more light on the characteristics of learner corpora. This study on vocabulary frequency of different countries can help a teacher decide which items emphasize in the classroom with regard to grammar. In addition, it can contribute to developing course materials and textbooks based on more accurate and empirical information about vocabulary frequency. This study expands our understanding of vocabulary acquisition process and the nature of L1 transfer itself. Therefore, finding ways to use of the L1 in interactive language learning tasks and activities in the classroom is our teachers’ challenge in the future.

The limitation of this study is the small size of each learner corpus. Thus, it would be difficult to make global generalizations about the result of the
study. Lexical density measuring the percentage of lexical words or content words or research to calculate the ratio of lexical tags to the total number of words should be examined more detail in the future research.

The internet is a useful tool for global communication and language learning, and a tremendous source for the application of corpora in applied linguistics.

6. Implications
This study can give pedagogical implications for vocabulary teaching. First, this project is helpful for Learners to expand their vocabulary knowledge and enhance vocabulary practice. Second, English learner corpora can be used to the Error Analysis. Unlike previous error analysis, natural learner data from different L1 backgrounds gives us access not only to errors but also to learners’ total inter-language.

Acknowledgments
I am deeply grateful to an organizer of project, Professor Masahito Watanabe of Yokohama National University who distributed the comprehensive record of e-pal projects. His great guidance and thoughtful support enabled me to achieve significant growth academically.

I would like to express my sincere thanks for partner teachers, Ms. Su-hsun Tsai of Taipei Municipal University of Education, Ms. Vera Menezes of Universidade Federal de Minas Gerais and Ms. Naoko Kasami of Oberlin University. WorldCALL 2008 Fukuoka happily allowed me to meet these four partner teachers in person for the first time. In addition, our meeting led me to insights and introspection which will greatly enhance my teaching progress in the future.

References