Data-Driven ESP Vocabulary Learning

Ping LIU

College of Foreign Languages, Huazhong Agricultural University, Wuhan, Hubei, 430070, China
Email: liup@mail.hzau.edu.cn

Keywords: Data-Driven Learning (DDL), Corpus-based, ESP Vocabulary, Language Teaching.

Abstract. This empirical study is to test the feasibility and effectiveness of Data-Driven Learning. It first introduces the concept of DDL (data-driven learning) and probes the rationality and advantages of data-driven ESP vocabulary learning. It points out that DDL is more motivating since it supplies authentic language materials and it is student-centered and majored-related learning. One semester’s ESP course applied self-constructed Agriculture English Paper Corpus and its CQPweb (web-based Corpus Query Processor) to ESP vocabulary learning as tools and resource. The results of empirical study indicates that corpus application into ESP vocabulary learning helped students in many ways, such as knowing high frequency words, grasping collocation and colligation, understanding prosody of words, etc. Vocabulary test and coverage (range) test indicate that DDL effectively improves the accuracy of ESP words use and helps students to acquire more ESP words.

Introduction

In contrast to EGP (English for General Purposes), ESP (English for Specific Purposes) refers to English used in certain profession or discipline such as computer English, science English, tourism English, medical English, agriculture English, etc. With rapid development of economy and fantastic spurt of science and technology, ESP proficiency of college students needs to be improved. Gradual transition from traditional teaching objectives of EGP to ESP or EAP (English for Academic purposes) is one of the college English teaching reform tasks. Previous studies indicate that students expect to be improved both at language and their specialties through ESP teaching and learning and get prepared for doing scientific research in English to meet the demands of internationalization and globalization of modern society. The greatest challenge of ESP teaching is ESP vocabulary. However, vocabulary is the foundation for language learning.

The development of corpus linguistics greatly facilitates ESP vocabulary teaching. Studies of Corpus-based ESP vocabulary teaching and learning in western countries mainly focus on generating AWL (academic word list) (Coxhead, 1998), discrimination of vocabulary meaning and usage, etc. Corpus-based ESP teaching in China focuses on the application of ESP corpus into ESP teaching. Generally speaking, studies of data-driven ESP teaching in China are still at the beginning stage. Most of them are descriptive research, and few empirical studies have been conducted. Considering this, a case study of the application of Agriculture English Corpus into the agriculture vocabulary acquisition will be conducted to test the feasibility and effectiveness of corpus-based data-driven ESP vocabulary learning.

Rationality of Data–Driven ESP Vocabulary Learning

The concept of Data Driven Learning (DDL) was first put forward by Tim John in 1991. As Zeng Fengcao (2005, 19) pointed out, DDL is learner-centered emphasizing developing students’ learning autonomy. The rules and regularities of language use are explored and generalized by learners themselves based on their observation of corpus data rather than imparted by teachers in a traditional sense. DDL is regarded as discovery learning, exploration learning, and autonomous learning with corpus acting as tool and resources and learners acting as language detectives and researchers.
DDL conforms to major principles of second language acquisition theory and constructivism learning theory. Corpus-based Data-driven ESP vocabulary acquisition has the following advantages: 1) corpus supplies an authentic language learning environment and input natural language; 2) Vocabulary learning converted from traditional teacher-centered to student-centered and automatic learning–centered is beneficial to exerting learners’ initiative and autonomy, which in turn helps the internalization and long memory of knowledge explored by learners themselves; 3) DDL combine the language learning and specialty learning. Therefore, learners are more motivated to acquire major-oriented ESP vocabulary through DDL. As a result, corpus-based ESP vocabulary acquisition can arouse and maintain students learning interests. Tim John (1991) classified DDL into 3 categories: free DDL, guided DDL, and controlled DDL. In this study, ESP vocabulary acquisition emphasizes after-class self-learning behavior. Therefore, free DDL and guided DDL are adopted. Guided DDL in this study refers to teacher’s introduction to corpus linguistics theory and application, and offering the technical training of corpus search so that students learn to use corpus as tools and resource in their ESP vocabulary learning.

**Experiment of Corpus-based Data-Driven ESP Vocabulary Learning**

This experiment is aimed to find the answers to the following questions:
(1) How do learners use corpus tools and resource in the ESP vocabulary learning?
(2) Whether is data driven ESP vocabulary learning effective or not?

**Research Method**

**Participant**

Both the experiment group (33 students) and control group (32 students) are from two natural classes of international discipline contest. The students in two groups are from different colleges majoring in different disciplines.

**Experiment Design**

Both groups took ESP course to learn to write English paper for international discipline contests such as mathematic modeling contest, biology contest, etc. ESP course lasted one semester. ESP course of two groups was taught by the same teacher who used the same textbook and other teaching materials. The same periods of lectures and the same teaching content were offered except that the experimental group received technical training of corpus search and they were acquired to use corpus in their ESP vocabulary acquisition process and paper writing process. The main purpose of ESP course is to teach students to write English science paper since their final goal is to write and submit English science paper for international contest in their own fields. ESP vocabulary acquisition is one of the major objectives of ESP course. The final contest paper produced by both groups were collected and analyzed to examine the actual use of ESP vocabulary. All the experiment group members were required to submit a written report on how they used Agriculture English Corpus to solve their vocabulary puzzlements at the end of ESP course learning. The students were required to present two vocabulary questions raised by themselves, corpus search results, and answers they work out on their own in their written report.

**Test of Experiment Effectiveness**

The effect of data driven ESP vocabulary learning is tested in two ways. One is the vocabulary test with full score of 30 test items being 30. The first 10 items test collocation and colligation of vocabulary in forms of gap filling. The second 10 items test the understanding of ESP words meaning in forms of words matching. The last 10 items test synonyms discrimination in forms of multiple choices. Vocabulary test is to examine the accuracy of word use. Following Nation, P.’s (1983) vocabulary testing and study method, the other way is to test the coverage of ESP words in students’
papers by using corpus tools Range. In Range, the most frequently used two 1000 EGP (English for general purpose) words were listed respectively in Baseword 1 and 2 (See Tab. 3) and 570 Academic word families were listed in base word 3(from AWL, Academic wordlist made by Coxhead, 1998).

Experiment Procedures.

Construction of Online Corpus Platform

Funded by National Social Science Foundation Project of China, an Agriculture Academic English Paper Corpus with 7,382,395 words and 1235texts was established to offer tools and resource for the ESP learning. This Agriculture English Corpus contain 3 layers of sub-corpora: The 1st layer is journal article sub-corpus with its data produced by English native experts. The 2nd layer is learner paper sub-corpus with data produced by Chinese native advanced students majoring in agriculture. The third layer is BAWE (British Academic Writing of English) whose data are produced by English native advanced learners. These three corpus data represent English native experts, Chinese native learners, English native learners respectively suitable for across-corpus comparison. All the corpus data were uploaded to campus-network-based CQPweb (online Corpus Query Processor), which makes online corpus tools and resource accessible to authorized users (http://211.69.132.28).

Guided DDL Strategies and Technical Training of Corpus Search

ESP course teacher will introduce the concept of DDL and exemplify the application of corpus in language learning to experiment group. According to Tim John(1991), DDL process includes three steps: Problem identification; Classification of corpus search data; and Generalization of laws of language use. Students in experiment group were encouraged to come up with their questions on ESP vocabulary, and then they were guided to observe search results and work out the answers to their puzzlements on language use. In technical training, students are taught to write regular expression of search target and are required to do some corpus search exercises until all of them grasp techniques of online corpus research. In independent or free DDL stage, students are encouraged to use corpus for their ESP vocabulary acquisition.

Results and Discussion

The Application of Corpus to ESP Vocabulary Acquisition

At the end of ESP course learning, experiment group’s written reports on corpus application were collected. Based on the questions or problems raised in written reports, the actual application of corpus in ESP vocabulary acquisition is categorized and summarized as follows.

Employing Corpus to Know High Frequency Vocabulary

There are a large number of synonyms in ESP to express the similar meaning. Some of them are more frequently used than others. One type of questions is about the frequency of synonyms. For example, there are so many verbs in science paper to report experiment results, such as “indicate, demonstrate, illustrate, signify, reveal, reflect, etc.” As for this point, students asked the following questions: Which verbs are used more frequently than others? What is the frequency ranking of these verbs in science papers produced by English native writers? How about frequency ranking by Chinese native writers? Similar questions asked by many students about the frequency of synonyms involves “affect, influence, and impact”, “example, case, and instances”, “raise, present, propose, come up with, and put forward”, “treat, deal with, tackle, handle, cope with, deal with, dispose”, etc. Knowledge of synonym frequency can help to grasp the genre features of science English and improve the accuracy of diction in language use.
Table 1. Frequency of six verbs in two corpora.

<table>
<thead>
<tr>
<th>corpora</th>
<th>indicate</th>
<th>demonstrate</th>
<th>reveal</th>
<th>reflect</th>
<th>illustrate</th>
<th>signify</th>
</tr>
</thead>
<tbody>
<tr>
<td>Journal Articles</td>
<td>3903</td>
<td>42.20 %</td>
<td>2182</td>
<td>24%</td>
<td>1724</td>
<td>18.64%</td>
</tr>
<tr>
<td>Learner papers</td>
<td>1892</td>
<td>52.34 %</td>
<td>583</td>
<td>16.13 %</td>
<td>658</td>
<td>18.21%</td>
</tr>
</tbody>
</table>

Tab. 1 reveals not only the frequency ranking of these synonyms but also the different diction preference between English native experts and Chinese native learners. For example, English native speaker use more “demonstrate” than “reveal”. With English natives’ diction as target language, the awareness of this difference will bridge the gap of diction between English natives and non natives.

**Employing corpus to know the collocation and colligation of Words**

Learning words involves learning the usage of words. Collocation refers to the co-occurrence between different words. Colligation refers to the co-occurrence between words and grammatical units. There are many questions about collocation and colligation in written report. For example, some students ask such collocation questions as “Is ‘method’ followed by prep. ‘of’ or ‘for’?”, “When ‘approach’ is used as a noun, which preposition is collocated with it, ‘of’, ‘for’, or ‘to’?”, “What is the difference between ‘die of’ and ‘die from’?”, “What is the difference between the two preposition ‘through’ and ‘by’ in science paper? The last question actually is the question involving colligation since these two prepositions can’t be distinguished semantically. Colligation is the way to distinguish them.

Corpus statistics indicate that the answer to the last question lies in: 1) ‘by’ (5430.63 instances per million words) is much more frequently used than “through” (671.47 instances per million words) in science paper; 2) ‘by’ can be followed by both noun and gerund while ‘through’ can only be followed by noun. Maybe colligation accounts for why ‘by’ has a much higher frequency in science paper than ‘through’ (See Fig1.and Fig. 2). In light of space limitation, the answers to other questions are omitted.

Corpus offers the context in which words appear. Context can exhibit collocation and colligation. Observation of context deepens the understanding of vocabulary usage.

![Figure 1. The colligation of “through”](image1)

![Figure 2. The colligation of “by”](image2)
Employing Corpus to Know Prosody of Synonyms

Learning words means knowing deep meaning of words. Synonym discrimination often perplexes vocabulary learners. Observations of words accompanying nodes will reveal more information that a dictionary offers. Some words accompany those words with negative meaning, some with positive meaning, while others with neutral meaning. This is defined as prosody in corpus linguistics, respectively known as positive, negative, and neutral prosody. Corpus approach provides a new perspective for synonym discrimination. Written report also revealed that students work out slight difference in prosody between synonyms by observing concordance lines. Questions on “What are the differences between ‘absolutely’ and ‘utterly’?”, “…between ‘happen’ and ‘occur’?” etc. fall into this category. The search results reveal that ‘occur’ is more likely to appear together with words with negative meaning. As can be seen in Fig. 3, the words accompanying finite verb ‘occur’ are ‘erosion’ ‘partially impaired’, ‘loss of DNA methylation’, ‘contamination’, ‘foodborne illness’, etc. In contrast, “happen” tends to have a neutral prosody. In the same way, ‘utterly’ is found to have negative prosody, while ‘absolutely’ has a neutral or positive prosody.

Prosody is generalized by observing corpus data. Thus corpus approach to prosody of words helps to improve students’ ability of observation and analysis.

| some erosion of DMR methylation does occur | (Ficz et al., 2013) |
| machinery and increased hydroxylation, which may occur in the context of a partially impaired |
| Loss of DNA methylation through 5hmC will occur via hemi-methylated intermediates since there is no |
| 2003). Microbial contamination can occur during any of the steps in the |
| estimates 11-13 million cases of foodborne illnesses occur in Canada every year. In the |
| 1998). This contamination can occur either pre- or post-harvest (Beuchat. |

Figure 3. The prosody of “occur”.

The written reports also show that students use other functions of corpus to solve their puzzlements about vocabulary use. For example, some students employed the distribution function to display uneven dispersion of certain words or phrases in the different sections of science paper or in different disciplines, or in different corpora. Other students made high frequency word list in their own disciplines by using frequency list function. The establishment of ESP wordlist is aimed to improve the efficiency and effectiveness of ESP vocabulary acquisition.

Effectiveness Tests

Vocabulary Test Result

Pre-test result of Independent samples Test indicates that the control and experimental groups are at the same proficiency level (Class1 representing experimental group, Class 2 representing control group. Significant level is defined as p <0.05). There is no significant difference in vocabulary test scores between groups (t=1.079, df=63, p>0.05) (See Tab. 2). Post-test result displays that vocabulary proficiency level of experimental group is significantly higher than that of control group (t=2.519, df=63, p=<0.05). Since other variables are effectively controlled. The differences in vocabulary proficiency can be attributed to the use of corpus tools and resource.
Table 2. Independent Samples Test of Vocabulary Test.

<table>
<thead>
<tr>
<th></th>
<th>Levene's Test for Equality of Variances</th>
<th>t-test for Equality of Means</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>F</td>
<td>Sig.</td>
</tr>
<tr>
<td>Pre-test</td>
<td>1.111</td>
<td>.740</td>
</tr>
<tr>
<td>Post-test</td>
<td>1.562</td>
<td>.216</td>
</tr>
</tbody>
</table>

Coverage Test of ESP Vocabulary

Log-likelihood values indicate that the number of first 2000 general service words (Baseword 1 + Baseword 2) used by experimental group significantly smaller than that by control group respectively (Baseword 1: LL = 17.71, P < 0.05, Baseword 2: LL = 169.38, P < 0.05), and that the number of academic words used by experimental group is significantly bigger than that by control group (Baseword 3: LL = 46.94, p < 0.05). Range test reveals that the academic words has a larger coverage in science paper written by experiment group than that by control group. This indicates that corpus tools and resource are effective in helping students learning more academic vocabulary.

Table 3. Range Test.

<table>
<thead>
<tr>
<th></th>
<th>Baseword 1</th>
<th>Base word2</th>
<th>Baseword3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Experiment Group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Token</td>
<td>92124(52.69%)</td>
<td>7393(4.23%)</td>
<td>17255(9.87%)</td>
</tr>
<tr>
<td>Type</td>
<td>71277(53.81%)</td>
<td>6964(5.26%)</td>
<td>12054(9.10%)</td>
</tr>
<tr>
<td>Families</td>
<td>748</td>
<td>369</td>
<td>472</td>
</tr>
<tr>
<td>Control Group</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Token</td>
<td>60717(55.04%)</td>
<td>5456/(4.95%)</td>
<td>10389/(9.42%)</td>
</tr>
<tr>
<td>Type</td>
<td>1585/(16%)</td>
<td>666/(6.8%)</td>
<td>1083/(11.1%)</td>
</tr>
<tr>
<td>Families</td>
<td>749</td>
<td>386</td>
<td>441</td>
</tr>
<tr>
<td>Log-likelihood</td>
<td>17.71</td>
<td>169.38</td>
<td>46.94</td>
</tr>
<tr>
<td>Sig.</td>
<td>0.000 ***-</td>
<td>0.000 ***-</td>
<td>0.000 ***+</td>
</tr>
</tbody>
</table>

Summary

Corpus–based data-driven language learning provides a new perspective for vocabulary learning. The key word concordance and context co-occurrence function of corpus lessens the obstacle of ESP vocabulary learning. Many advantages such as large size of data, authentic language material, and rapid response of search make corpus an irreplaceable tool and method in language education and linguistics studies. Vocabulary test indicates application of corpus tool and resource to ESP vocabulary acquisition is effective in improving accuracy of ESP vocabulary use. Range test reveals that corpus application can help learners to acquire more academic words. This empirical study is aimed to offer references for the application of corpus in language teaching and learning.

Acknowledgement

This research was financially supported by: 1. National Social Science Foundation Project of China (Grant NO. 14BYY162); 2. High Quality Research Paper Development Project funded by the Fundamental Research Funds for the Central Universities (Grant NO.TD2662015PY193); 3. Key
Courses Construction Project sponsored by Huazhong Agriculture University; 4. Team Building Project funded by College of Foreign Languages, Huazhong Agriculture University.

References


