An Empirical Study on Corpus-driven English Vocabulary Learning in China

Jiao Binkai

Associate Professor, School of Foreign Languages of Xuchang University, China

E-mail: binkaisir@126.com

Abstract

This paper focuses on the autonomous English vocabulary learning in corpus-based contexts. Language teaching practice is becoming more learner-centered in the field of language teaching, learner autonomy has been an ongoing concern of foreign language educators in china. As an assistant tool in language learning, corpus makes an easy and quick analysis of the greatest amounts of linguistic data possible, and the learners are provided with a new approach to learn a language independently. The empirical study explored the aspects in which corpus can be applied in English vocabulary learning were investigated and illustrated. At the end of the research, a questionnaire was designed and answered to test the effectiveness of the learning model.

Kewwords: Corpus, Vocabulary learning, Learner autonomy, Effectiveness

1. Background of the Study

The development of individualized study methods and the autonomous learning ability on part of the students is an important indicator of the successful reform of the teaching model in the past decade in China. As one of the major components of language system, vocabulary is especially essential for English as Foreign Language (henceforth EFL) learners (Lewis, 2000). Chinese college students generally believe that the learning of large amount of vocabulary is one of the most challenging as well as necessary tasks. Among various approaches that contribute to vocabulary acquisition, corpus is one of the latest and most enlightening, for a corpus makes an easy and quick analysis of great amounts of linguistic data possible(Sinclair, 2003).

The application of the computer corpus in the EFL teaching and learning has marked a fundamental shift both in methodology and in ideology with regard to linguistic studies and language learning. With constructivism as the guidance, the vocabulary teaching method based on corpus helps to develop students' autonomous learning ability. As an assistant tool in language education, corpus is a new study field in applied linguistics in China. Based on a large collection of authentic materials and advanced soft wares, corpus makes an easy and quick analysis of the greatest amounts of linguistic data possible, and the learners are provided with a new approach to learn a language independently. This is an empirical study into the application of corpus-driven approach in vocabulary teaching and learning based on a conduct in two classes, and it is intended to tap into the promotion of the new autonomous learning model.

Likewise, many non-native learners of English admit that once they have exceeded the initial stages of language acquisition, they often regard vocabulary learning as "their greatest single source of problems" (Halliday, 2004)). For the Chinese learners of English, vocabulary has always been a challenging as well as necessary task in their study. According to a survey on Chinese college students, 67.8% of them regard vocabulary as the most difficult part in language learning, and 97.46% think their progress in vocabulary during middle school is rather limited (Jiao 2008). Therefore, how to select the words to learn, how to teach and learn vocabulary efficiently and accurately and how to avoid making mistakes in vocabulary constitute a question that language teachers and learners must consider and study carefully. In all, the application of the corpus in vocabulary learning which I described in the following is part of the outcomes of the empirical research.

2. Software and Corpora Involved in the Study

Program used in the study mainly is Wordsmith Version 4.0. Wordsmith was developed by Mike Scott (1996).

Wordsmith Tools is an integrated suite of programs for looking at how words behave in texts. You will be able to use tools to find out how words are used in your own texts, or those of others. The Wordlist tool lets you see a list of all the words or word-clusters in a text, set out in alphabetical or frequency order. The concordancer, Concord, gives you a chance to see any word or phrase in context--so that you can see what sort of company it keeps. With key words, you can find the key words in a text. Concordances can be produced in a number of formats. The most usual form is KWIC concordance which puts the search word right at the center of the screen. By looking at both sides of the keyword, we will know quite easily what vocabulary and grammatical patterns it is used in and how strong collocability is for it. If the program is a bit more sophisticated, it might also provide its user with lists of collocates or frequency lists.

The corpus that I will focus on in this study is *New College English Corpus* (henceforth NCEC) and *BNC*. NCEC is adopted in the teaching and learning practice by teachers and students of our university. New College English (henceforth NCE) is an EFL coursebook series specially developed for non-English major students in China. NCE is designed to conform to the requirements set forth by *the National College English Teaching Syllabus*, and NCE corpus is a resource and a learning tool for the coursebook. The corpus contains approximately 1.7 million words and the texts are mainly taken from the series of coursebooks, that is, band 1 to 6. Besides, the corpus can be classified into six categories: listening typescript, new words, work book, teaching notes and translation, reading and writing, and additional material available on website only.

The NCE concordancer possesses very useful functions for language analyses and can find out any existing word or phrase in the course book with their occurrence contexts. The range and the length of the context can be adjusted according to the requirement of users. If you like, contexts of the form of the search word can be displayed together on the screen. The BNC is very useful for a very wide variety of research purposes, in fields as distinct as lexicography, artificial intelligence, speech recognition and synthesis, literary studies, and all varieties of linguistics. It provides a unique and authoritative view of the state of the English language today, with carefully balanced representation of as many different varieties of English as possible. It can be used to exercise Natural Language Processing systems of all kinds, as a fertile source of real life examples for language learners, or simply to explore the way the language is currently used.

3. Research Procedures

The New College English Book is the major material used in the class. The book consists of three passages in each unit, focusing on the same topic. The first one passage is supposed to be learned more intensively. Since we only have six hours for each unit, it is really a hard task to learn 70 to 80 words thoroughly in the two texts. As a result, a study was conducted based on the *NCEC* and *BNC*.

In the first class, I just briefly introduced to the students the present situation of vocabulary learning in college and emphasized the importance of vocabulary learning. Then, the application of the corpus was demonstrated as an aid to vocabulary learning. Finally, the students were divided into groups of four with one in each group in charge to practice the new approach.

The Internet-based classroom is open to students in weeks evening, so before the study of a new unit, the groups were asked to choose the words or phrases appeared in the unit, usually, ten to twenty words were chosen. The groups then got access to the corpus and ran the concordancer to find out the senses of meaning, colligation and collocation of these words. They were supposed to work collaboratively to fulfill the tasks.

When these words were actually learned in the classroom, some of the groups were picked out to report their findings and shared them with their peers. The teacher would be present to answer their questions and give them guidance. The teacher could employ the citations as examples when explaining a word or adapt them for exercises. If a student made a mistake, the teacher could also ask him or her to refer to the concordance. Sometimes, if the number of citations was too big or time was limited, the teacher would choose part of the concordancing lines and present to the class as handouts. After class, the students could resort to the corpus when they had an urge to consult the usages of the words they were interested in or confused about.

Although corpus in language teaching and learning is still a new field to explore, its potentials empirical have already been demonstrated by some researches. This thesis is a tentative and empirical research and practice into the vocabulary teaching and learning, based on the NCE Corpus and BNC adopted in my class. Research is mainly conducted in the following six aspects.

3.1 Word Frequency

All the students are motivated to find out the frequency of some word in the unit, then the figures are gathered and the frequency list of each unit takes shape. The list helps the students have a sense of the distribution of the word in

the whole vocabulary of the course books, and acts as an important guide in their vocabulary learning.

3.2 Deductive Approach

Through a deductive approach, the definition and explanations of a word can be tested by running a concordancer. The teacher and the students are then exposed to a large number of contexts of the same word to confirm what is learned so as to refine the original generalizations. This, therefore, assists the student in exploring the language in great detail and thereby gaining further insights into its vocabulary.

3.3 Inductive Approach

During the teaching practice, the study of some key words can be conducted through this approach. The teacher can ask the students to observe and discuss the concordance before explaining the word. Working with corpora does not necessarily require students sitting at computer terminals. The printed results of concordance searches can also be provided in classroom teaching, and the concordancing lines can be chosen by the teacher so as to avoid students' confusion by a large amount of information. Anyway, with guidance from the teacher, and through the inductive approach, the students are motivated in their learning of English and at the same time, learn some important skills that would make their learning less intimidating and keep them in regular contact with English.

3.4 Colligation, Collocation and Prosody

Colligation and collocation are related item, but the former is on a grammatical level while the latter is more on a vocabulary level, to explain simply. Nation (2001) holds that "fluent and appropriate language requires colloational knowledge." The computerized corpus, with sufficient, authentic and typical data, enables users to conduct comprehensive and overall research into collocational patterns. It is assumed that, once patterns of non-native deviance have been discovered, students can be explicitly made aware of these patterns, and that, given time, motivation and the opportunity to practice, they will eventually be able to modify their linguistic behavior into a more native-like direction.

3.5 Synonyms

The use and differentiation of synonyms is usually a problem for many Chinese learners of English. If the students only know the Chinese equivalents of the synonyms, they still have difficulty grasping the shades of their meanings and usages. The best way to help them to deal with this problem is exposing them to ample examples of real English, which can be obtained easily via a large English corpus. In this aspect, concordances can play a unique role since no description of differences between words.

3.6 Misuse of Words

The purpose of error analysis is to seek the causes of the errors and the law of learning a language, but finding out the representative problems or errors requires the amount of the learner's output be as large as possible. Hence, the analysis of the learners' interlanguage would be much more efficient if the outputs are transferred into an electronic corpus accumulatively from learners and then explored via a computer program. By observing the development of a learner's interlanguage or making error analysis, the teacher and the students are able to improve the efficiency of their teaching and learning practice respectively.

4. Questionnaire and Analysis

In order to test the effectiveness of the corpus-driven vocabulary learning and provide a guide for the future study, a questionnaire was designed and answered by the subjects at the end of the semester.

4.1 Questionnaire Design

The questionnaire was employed to test the effectiveness of the corpus-driven vocabulary learning and get more information, epically the attitudes, from the subjects. Participants were asked to rate each of the treatments on a scale of choices. Responses from the questionnaires were analyzed and converted into percentages to indicate the usefulness of the vocabulary learning model. The questionnaire consists of 19 closed-multiple-choice questions and one open-ended question concerning the following aspects:

1. the students' opinion on the new vocabulary leaning approach and the concordance

2. the influence of the new approach on student' learning habits and methods

3. the students' assessment on the effectiveness of the learning methods.

4.2 Data Collection

At the end of the semester, the students were asked to complete a questionnaire. The students were informed that the survey was only for the purpose of research, and not to be concluded to evaluate their performance. The questions

were explained so that the students would have no difficulty understanding them. 89 students answered the questionnaire. Among the 89 collected copies, 2 were invalid (incomplete or misunderstood), so the final valid copies were 87. For the sake of accuracy and efficiency, the responses to the questionnaires were input into the computer and analyzed with SPSS.

4.3 Results and Analysis

4.3.1 Research Question 1

What are the advantages of the new vocabulary learning model to foster learner autonomy of Chinese EFL students?

On the whole, the students' attitudes toward corpus-assisted vocabulary learning are positive. 58 out of 87 students, that is 66.7% of them say they like the new vocabulary learning model and 26 (29.8%) of them like it very much. compared with the traditional learning model, 47(54%) of the students think the corpus-driven method is helpful and 21(24.2%) of them think there is no difference between the two. That shows most of them accept the new learning model. As far as the concordancing software is concerned, only 4 students are dissatisfied with it (4.6%) while most of them are satisfied or think it is all right.

It is revealed in the questionnaire that to search for the usage of a word with a question in mind is likely to motivate the learners and stimulate their interest in exploring, since nearly half (45.7%) of the subjects find it interesting. Moreover, the vast amount of information is regarded as an advantage of the learning model according to the subjects.

As far as the influence on the students' learning habit is concerned, the result is also positive. Data analysis indicates that two-thirds of the students (59/87=68%) can finish the task assigned before class while only less than 10% rarely do so. Most of the subjects claim that language learners should be responsible for the outcomes of their learning. In answering the question "I evaluate my learning to find the weak point and the ways to improve", 42.5% (37/87) of them choose "sometimes", and 39% (34/87) choose "often" or "always". Furthermore, in analyzing the reasons for their success or failure in vocabulary learning, a large percentage of them attribute to one's own efforts for both questions (56/87=64.4% and 69/87=79.3% respectively). All these show that the students are fully aware of the importance of autonomous learning.

However, as to the question "Have you found a suitable approach to learn English vocabulary", only 23(26.4%) of them choose "yes", while 27 (31%) choose "no" and 37 (42.6%) choose "it's hard to say", Consequently, when introducing a new learning model to the students, it is of great importance to bear in mind that they need to learn how to learn. In other words, the teacher should try to help them develop an effective learning method.

When confronting a vocabulary problem, over half of the students will turn to the dictionary (52.9%) and 20.7% will ask their classmates for help. It is noticeable that 48.3% of the subjects will also turn to the corpus for a solution, which indicates that some of them have taken the corpus as one of the learning aids.

As to the teacher's guidance in corpus-driven vocabulary learning; most of the subjects admit they need it (66/87=75.9%) or sometimes need it (19/87=21.8%). That shows the learners have realized the negative influence of the new learning model, and they need the teacher's assistance to cope with the large amount of information. In other words, the learners view the new approach as enhancing, instead of replacing the classroom-based instruction. Therefore, the function of the corpus in vocabulary learning cannot be exaggerated.

According to the final statistics, corpus is taken as an effective way to build up vocabulary knowledge and most students accept the fact that using corpus is beneficial for language acquisition. 72 out of 87, that is 94.2% of them, report their vocabulary knowledge has improved to some extent, and 5 (5.8%) say it has improved a little. As to the improved aspects, most of them choose vocabulary (83/87=95.4%) and/or collocation (74/87=85.5%), which is consistent with my judgment and observation. Moreover, a large percentage of them think their studying ability has improved (greatly improved: 25/87=28.7%; to some extent: 49/87=56.3%; a little: 13/87=15%).

4.3.2 Research Question 2

What problems exist with the application of the corpus-driven vocabulary learning model?

Some problems (disadvantages) are also revealed from the questionnaire. About one-third of the subjects cannot fulfill the assigned task every time. That may due to their limited time (55/87=63% of them mentioned it) or to the abundant materials that frustrate their initiative to learn (43/87=49.4%). Besides, the concordancing software needs to be improved, as nearly one-third of the subjects think it is not quite satisfactory (26/87=29.9%). As mentioned above, only a small number of students claim they have found a suitable approach to learn English, which calls for the teacher's appropriate guidance.

Moreover, in open-ended question part, some objects gave more ideas about the learning model, the main one is:

concordancing plus dictionary is better. The following statements are examples made by the objects. "Concordancing plus dictionary is better, because concordances plus dictionary explains to me exactly what the word means, it gives us more choice. I could use the two (concordancing and dictionary) together to help me remember it clearer. We can know the meaning clearly form the dictionary when we can't understand the examples quite well." Participants enumerated a number of features of each method. The following are examples of the participants' answers. "Concordancing list is hard to catch the word's meaning, it's good for us to learn a new word meaning but it's better for us to understand the usage of a new word. Concordancing plus dictionary is good. It's easy to know word's meaning and leant it further". "The words in dictionary are easier for us to learn, but keep in our brains for short time, Concordancing plus dictionary is on the contrary." "Dictionary: clear but lack context; concordancing: just the opposite; concordancing plus dictionary: very good."

In spite of the problems, the corpus-driven vocabulary learning approach is favorably rated and proves to be an effective aid.

4.3.3 Main Findings

The thesis mainly discusses and exemplifies the concrete application of corpus in vocabulary teaching and learning. It also encourages the teachers and students to explore and discover the laws of vocabulary use under actual language environment so that they can grasp the correct usage of English vocabulary.

Corpus can be applied in various aspects of vocabulary learning. In the first place, by counting the word frequencies, the learner can obtain a clear sense of the relative importance of the vocabulary. The empirical study reveals that the words listed between 10 and 25 in each unit prove to be the relatively key words (since most of the words at the very beginning of the list are already familiar to the students).

In the second place, the large number of examples retrieved from the corpus provides the learners with ample information about the meanings and usage of the target words, which is a distinct advantage over tradition vocabulary learning.

In the third place, concordancing lines are of great help in discovering the collocation, colligation and prosody of the search word. This was also echoed by the students when they answered the questionnaire.

In the last place, the corpus-driven vocabulary learning method can be applied in discriminating synonyms or correcting the learner's vocabulary mistakes.

The research and the empirical study in the thesis prove that the corpus-driven approach is helpful in vocabulary learning and can contribute to autonomous learning at the same time. That can be shown in the following aspects. Firstly, corpus and concordance provide a new medium for vocabulary learning. For the students, a more active approach emphasizing discovering instead of memorizing can make learning more engaging and less intimidating. Secondly, students have more opportunity to control their own learning, so their learning awareness is promoted. Thirdly, the large amounts of data in the corpus enrich the teaching contents, and more importantly, the way the data is presented facilitates the discovery of some vocabulary knowledge, such as word frequency, collocation etc by individual users. Lastly, through the process of discovering, the students developed a sense of achievement and an impression that English is at least "learnable". The awareness that words can be learned by observing and testing hypothesis are developed and fostered.

4.3.4 Reflection on the Main Findings

Types of corpus-driven vocabulary learning activities haven't been explored extensively. It is true that some types of corpus-driven vocabulary learning activities have been used during the research, but the types are too limited and the students' enthusiasm haven't been fully inspired, sometimes some students even felt dull and lost the interest. Various corpus-driven exercises for students should have been designed and employed to encourage students to learn vocabulary autonomously (e.g.: cloze, colligation and extending context, using keywords and making courseware.)

5. Implications

This is just an empirical research which calls for improvement and further study in implementing such a model, many factors should be taken into consideration and certain cautions are essential.

Students engaging in corpus-driven learning environment must first familiarize themselves with the use of the computer and the concordancing program. Therefore, the teacher should provide an initial tutorial for the students. Apart from the basic skills, the teacher needs to explain to the class why and when it is appropriate for them to use the corpus rather than regular textbooks so that the students understand the purpose and the basic mode of this learning approach.

The value of the corpus is real and definite, but whether a learning method is successful or not is not determined by how much data is used but by careful planning (Kennedy, 1998). Therefore, how to integrate corpora into English language teaching and learning constitutes a challenge. Here, it is suggested to introduce learners to this process in controlled stages to ensure success and continued enthusiasm. For instance, a sensible, workable number of examples can be selected by the teacher and then introduced to the students in printed handouts with carefully thought-out procedures to follow. That is because the students may get somewhat bewildered when faced with pages of concordancing materials, especially in the initial stage of their learning. Later, they can be allowed to work with the concordance on the computer and begin to formulate their own questions.

In a learner-centered classroom, students are seen as being able to assume a more active and participatory role than is usual in traditional approaches. However, one point must be made clear: the corpus-driven language learning does not mean isolated learning; interaction and collaboration, either among learners or between learners and their teachers, is the "driving force to individuals' cognitive development" (Bruffee 1986). According to some students, working with a partner opens a channel for positive interaction and made the learning experience more rewarding. In addition, small group work relieves the stress of having to take on a difficult task oneself, which is particularly significant to students who lack confidence in their own proficiency.

In the corpus-driven teaching, the need for teachers will not decrease, but their roles, and the role of teaching in the learning process, will change. Teachers should not only be viewed as instructors, but as helpers and facilitators (Dickinson 1995). Therefore, it is crucial for the teacher to establish a good relationship with students, supporting and guiding them in their learning. This can best take place when teachers come to meet learners' specific needs and demands. Besides language problems, these include getting to know students well enough to understand what they need and what they are able to do.

Since most students in China's EFL context have depended so much on teachers, they will have more difficulty adopting autonomous learning at the very beginning. In this case, if they start with collaborative work with their peers under the guidance of their teacher, they will be more likely to arrive at a greater degree of interdependence (Blin 1999).

References

Barnbrook, G. (1996). *Language and Computers: A practical introduction to the computer analysis of language*. Edinburgh: Edinburgh University Press.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpous Linguistic*. Cambridge: Cambridge University Press.

Blin, F. (1999). CALL and the development of learner autonomy. In R. Debski & M. Levy (Eds.), *World CALL: Global perspectives on computer-assisted language learning* (pp.169-184). The Netherlands: Swets &Aeitlinger Publishers.

Bruffee, K. A. (1986). Social construction, language and the authority of knowledge: A bibliographical essay. *College English*, *48*, 773-790. http://dx.doi.org/10.2307/376723

Coady, J., & Juckin, T. (Eds.). (1997). *Second Language Vocabulary Acquisition*. Cambridge, UK: Cambridge University Press.

Dam, L. (1995). Learner Autonomy 3: *From Theory to Classroom Practice*. Dublin: Authentik.

Dickinson, L., & Carver, D. (1980). Learning how to learn: Steps towards self-direction in foreign language teaching in schools. *ELT Journal*, *3*, S(1), 1-7.

Dickinson L. (1993). Talking shop: aspects of autonomous learning. An interview with Leslie.

Dickinson, L. (1995). Autonomy and Motivation: A Literature Review. *System*, *23*(2), pp. 165-174. http://dx.doi.org/10.1016/0346-251X(95)00005-5

Fosnot, C. T. (1996). Constructivism: A psychological theory of learning. In. Fosnot, C. (Ed.), *Constructivism: Theory, perspective and practice* (pp.8-33). New York: Teachers College Press.

Gardner, D., & Miller, L. (1999). *Establishing Self-access from Theory to Practice*. Cambridge: Cambridge University Press.

Gremmo, Marie-Jose, & Riley, Philip. (1995). Autonomy, self-direction and self-access in language teaching and learning: the history of an idea. *System, 23*(2), 151-164.

Gross, D. (1991). *A Practical Handbook of Learning Teaching*. London: Cassell.

Halliday, M. A. K. (2004). *Lexicology and Corpus Linguistics*. London/New York: Continuum.

James, Cx. (1996). *English in Computer Science*. Hong Kong: Longman.

Jiao Binkai. (2006). Study on corpus-driven English vocabulary learning. *Education Exploration*, *4*, 89-91.

JI Kangli. (2001). The Possibilities of Fostering Learner Autonomy Through Course Design. Papers Presented at the 3rd International Symposium on ELT in China. Foreign Language Teaching and Research Press. (June, 2003)

Johns, A. (1997). *Text, role and context.* Cambridge, UK: Cambridge University Press.

Jonassen, D. (1999). *Learning with technology*: *A constructive perspective*. NJ: Prentice-Hall.

Kennedy, Cx. (1998). *An Introduction to Corpus Linguistics*. London: Longman.

Lee, L. (1998). Supporting Greater Autonomy in Language Learning. *ELT*, *52*(3), 212-218.

Lewis, M. (2000). *Teaching Collocation*. London: Commercial Color Press plc.

Littlewood, William. (1996). Autonomy in Language Learning. *System, 24*, 427-435.

Littlewood, William. (1999). Defining and Developing Autonomy in East Asian Contexts. *Applied Linguistics*, *20*(1), 71-94.

Louw, B. (1993). Irony in the Text or Insincerity in the Writer? The Diagnostic Potential of Semantic Prosodies. In M. Baker, G. Francis and E. Tognini-Bonelli (eds.), *Text and Technology: In Honour of John Sinclair*. Amsterdam: John Benjamins.

Nation, L. S. P. (2001). *Learning vocabulary in Another Language*. Cambridge: Cambridge University Press.

Partington, A. (1996). *Patterns and Meanings: Using Corpora for Research and Teaching*. Philadelphia: John Benjamins Publishing Company.

Pennycook, A. (1997). Cultural Alternatives and Autonomy. P. Benson &P. Voller (ed.), *Autonomy and Independence in Language Learning* (pp. 35-53). New York: Longman.

Susan Hunston. (2001). *Corpora in Applied Lingustics*. Oxford: Oxford University Press.

Sheerin, S. (1997). An exploration of the relationship between self-access and independent learning. In Benson, P. and Voller, P., (eds.), *Autonomy and Independence in Language Learning.* London and New York: Addison Wesley Longman.

Sinclair, J. (1991a). *Corpus, Concordance, Collocation.* Oxford: Oxford University Press.

Sinclair, J. (2003). *Reading Concordances*. London: Longman.

Taylor, L., Leech, Cx., &Flingelstone, S. (1991). A survey of English machine-readable corpora. In Johansson, S. & Stenstr (eds), *Manual of Information to Accompany the Lancaster-Oslo-Bergen Corpus of British English, for Use with Digital Computers.* London/New York: Longman.1991. pp. 319-354.

Tognini-Bonelli, E. (Ed.). (2001). *Corpus Linguistics at Work.* Amsterdam/Philadelphia: John Benjamins Publishing Company.

Tribble, C. (1990). Concordancing and an EAP Writing Programme. *CAELL Journal*, *1*(2), 10-15.