English Language Teaching; Vol. 8, No. 8; 2015
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

An Analysis of the Application of Wikipedia Corpus on the Lexical Learning in the Second Language Acquisition

Jing Shi[1]

[1] School of English for International Business, Guangdong University of Foreign Studies, Guangzhou, Guangdong, P. R. China

Correspondence: Jing Shi, School of English for International Business, Guangdong University of Foreign Studies, Guangzhou, Guangdong, 510420, P. R. China. E-mail: sj@oamail.gdufs.edu.cn

Received: May 5, 2015 Accepted: July 22, 2015 Online Published: July 27, 2015

doi:10.5539/elt.v6n8p171 URL: http://dx.doi.org/10.5539/elt.v6n8p171

Abstract

Corpus linguistics has transformed linguistic research but has a slightly moderate impact on the ESL teaching and learning. The Wikipedia Corpus, designed by Mark Davis is introduced in this essay. The corpus allows teachers to search Wikipedia in a powerful way: they can search by word, phrase, part of speech, and synonyms. Teachers can also find collocates, and see re-sortable concordance lines for any word or phrase. The application of Wikipedia corpus is conducted in the experimental group whereas the conventional lexical teaching and learning mode with teacher imparting lexical information to students is carried out. The collected data is assessed and evaluated. The empirical evidence reveals the beneficial effects of corpus linguistics on ESL teaching and learning.

Keywords: wikipedia corpus, lexical learning, pedagogic processing of corpora

1. Introduction

ESL teaching in China is often notoriously associated with the production of "bubble English". The traditional instructional approach, with teachers as instructors and students as listeners, isolates language skills and is without appropriate contextual clues in a classroom environment where the instructor is didactic expert and students complacently follow along. As an English language instructor in a prestigious university which is renowned for its foreign language teaching and learning, I have always been keen on sorting out interactive and student-centered activities that can offer an alternative to the traditional method.

Corpus linguistics has revolutionized linguistic research and also has a moderate impact on the ESL teaching and learning; however, much work remains to be done to narrow the gap between research and practice. Corpus linguistics, with an immense potential in the field of ESL teaching and learning, is marginalized in its application. Language teachers reluctantly use or even feel resistant towards the pedagogical application of corpus in the classroom mainly because they do not have basic training in working with corpora and they feel intimidated by the corpus data.

The pedagogical application of corpus in the classroom can find its road into the mainstream only if teachers' needs could be taken into account and more user-friendly corpus resources that are already freely available online could be introduced to them. The Wikipedia Corpus, compiled and designed by Mark Davis (http://corpus.byu.edu/wiki/ (accessed: 29/03/2015)), contains the full text of the English version of Wikipedia, 1.9 billion words in more than 4.4 million articles. This corpus tool is not challenging for teachers as the interface is easy to handle. The corpus allows teachers to search Wikipedia in a powerful way: they can search by word, phrase, part of speech, and synonyms. Teachers can also find collocates, and see re-sortable concordance lines for any word or phrase.

This essay introduces the pedagogical application of Wikipedia and evaluates the effectiveness of its application via pretest and posttest. All the statistics are collected and analyzed by SPSS 21.0. It is hoped that the easy-to-use corpus tools and methods can benefit the lexical teaching of ESL in China.

2. Theoretical Framework

The pedagogical application of corpus is the direct application of corpus linguistics in the classroom setting, affecting how language is taught and learned (Romer, 2011). This essay mainly focuses on the direct application

of corpus linguistics in the L2 classroom.

The pedagogical application of corpus linguistics is a significant alternative to the conventional method as it is beneficial for ESL teaching and learning. First, observing corpus data and analyzing language information is of vital importance to the development of learners' intelligence. The process of learners' analyzing corpus data can exercise their ability of attention, memory and reasoning (Robinson, 2001). Concordance, especially the KWIC (Key-Word-In-Context) format can easily attract learners' attention and the data-observing can ensure an opportunity for learners to exercise their ability of analyzing and reasoning. Sinclair (1991), Skehan (1993) and Aston (1997) have already pointed out that the pedagogical application of corpus linguistics plays a positive role in language learners' schemata construction.

Second, the mode of autonomous learning can enhance learners' language awareness and raise their consciousness of language format. Language awareness and consciousness can help language learners to bridge consciousness gap and knowledge gap between L1 and L2. The concordance data can serve as an impetus for learners to improve their abilities in lexical acquisition because: first, the similar language patterns are displayed frequently which can be noticed by learners. The noticing can enable language patterns to be imbedded in learners' interlanguage (Schmit, 1990). Second, collocations of words are shown in a salient format which can be regarded as a lexical focus for learners.

The future development of pedagogical application of corpus lies in the following aspects: (1) learner and teacher needs are being prioritized, (2) direct uses of corpora in L2 teaching needs improvement (Romer, 2011). The introduction of Mark Davis's Wikipedia Corpus in L2 teaching can satisfy these two needs. The user-friendly corpus system is able to be used by teachers and students. This direct use of corpora can endow the L2 teaching and learning environment with more authentic language information.

3. Implementation of Wikipedia Corpus in the Lexical Teaching

48 English majors, studying at a university in mainland China for the first year participates in the research. Participants are categorized into two main groups: experimental group and control group. Participants in two groups are using the same teaching material, having the same amount of English lessons, and being evaluated on the same basis for enabling for the purpose of controlling variables tightly. The participants have been learning English for ten years at least but they have been exposed to the high-stake-exam-oriented educational system and they are weak at communicating with English native speakers. The paramount objective for the students is improving their listening, speaking, reading and writing skills of English so as to be able to communicate with English speakers.

The application of Wikipedia corpus is conducted in the experimental group whereas the conventional lexical teaching and learning mode with teacher imparting lexical information to students is carried out. The corpus tool is not too challenging for teachers who do not have a good command of corpus linguistics as the interface of the corpus tool is not so daunting which is similar to the search engine which teachers are familiar with (see Figure 1).

Students in the experimental group received a simple and short orientation of the application of Wikipedia Corpus before they begin their lexical teaching and learning. Students are divided into six groups for discussion. The word "finance" is taken as an example. First, the students are given the lexical items which they are required to acquire. Second, students type in the word "finance" and choose the display format as KWIC. Third, the concordance lines are shown in the KWIC format, as displayed in Figure 2. Fourth, students observe the concordance lines and make efforts to find out the collocations of the word. They are asked to figure out the meaning, part of speech and most commonly-used expressions of the word. Fifth, they have a heated discussion about their findings of the word and later will report their assumptions to the whole class. The learner-centered activity offers an opportunity for students to cultivate their critical thinking ability in language learning. Students are no longer receiving lexical information in a passive manner, instead, they become active "digger" of vocabulary. Seldom can you spot dozing students in the classroom as they have a lot of hand-on work to do. The simple-to-use Wikipedia Corpus keep them busy working and active learning.

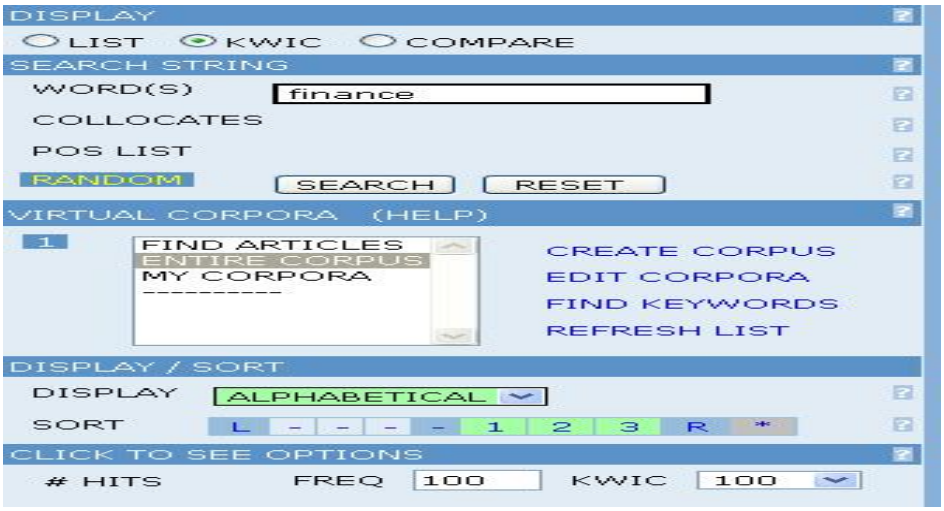

Figure 1. The interface of wikipedia corpus

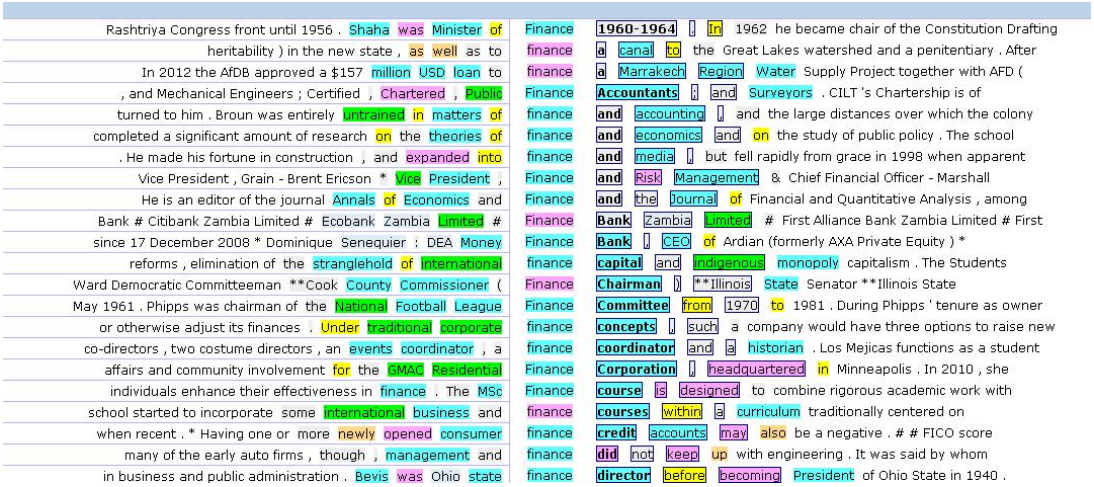

Figure 2. Concordance lines of "finance"

4. Research Method

Students in both the experimental group and the control group are required to take two simple reading comprehension tests to evaluate their capacity in lexical acquisition. The pretest and posttest are modified on the basis of Boulton's tests (2012). Pretest and posttest are of the identical pattern. The tests consist of two short articles of the similar word count, from the magazine *the Economist* (http://www.economist.com/ (accessed: 27/03/2015)). The topics of the two selected articles are not unfamiliar to students (see Appendix 1 & 2). In both pretest and posttest, students in both two groups are allowed to read the articles for 2 minutes and then the sheets of articles are collected. The answer sheets with reading comprehension questions, focusing on words' meanings and forms, are distributed to them. Students have another 2 minutes to complete all the questions. The levels of vocabulary of two articles are evaluated by the Software *Range* (http://www.victoria.ac.nz/lals/about/staff/paul-nation (accessed: 27/03/2015)) and the results are shown in Table 1 and Table 2. From the statistics, it indicates that the levels of vocabulary of two articles are similar.

Table 1. *Range*'s test results of article for pretest

WORD LIST	TOKENS/%	TYPES/%	FAMILIES
One	0/ 0.00	0/ 0.00	0
Two	28/ 7.07	22/11.28	20
Three	22/ 5.56	19/ 9.74	19
Not in the list	346/87.37	154/78.97	?????
Total	396	195	39

Table 2. *Range*'s test result of articles for posttest

WORD LIST	TOKENS/%	TYPES/%	FAMILIES
One	0/ 0.00	0/ 0.00	0
Two	41/10.70	35/15.15	32
Three	27/ 7.05	24/10.39	22
Not in the list	315/82.25	172/74.46	?????
Total	383	231	54

Students in both two groups take pretest before the implementation of the experiment and the posttest is conducted after they have received different ways of lexical teaching and learning for one semester. The test scores are displayed in Table 3.

Table 3. Test scores of experimental Group (Group E) and Control Group (Group C)

Group E	Pretest	Posttest	Group C	Pretest	Posttest
E1	15	15	C1	15	12
E2	10	13	C2	12	18
E3	15	14	C3	11	11
E4	16	11	C4	14	14
E5	17	16	C5	15	12
E6	10	10	C6	13	15
E7	16	15	C7	18	15
E8	18	12	C8	16	17
E9	15	10	C9	16	13
E10	14	13	C10	16	16
E11	15	13	C11	15	8
E12	14	14	C12	18	10
E13	17	14	C13	16	14
E14	11	11	C14	16	14
E15	16	12	C15	14	10
E16	12	13	C16	15	13
E17	13	14	C17	15	14
E18	17	17	C18	14	12
E19	16	10	C19	17	14
E20	14	15	C20	18	12
E21	16	16	C21	13	15
E22	18	13	C22	14	10
E23	14	13	C23	14	9
E24	17	14	C24	13	12
Average Score	14.6	13.1	Average Score	14.9	12.9

The raw data reveals that both the experimental group and the control group did worse in the posttest than the pretest but the average score of the experimental group's posttest is higher than that of the control groups' and the difference is statistically significant. This reveals the pedagogical application of Wikipedia functions better than the conventional method in terms of learners' lexical acquisition.

Both the experimental group and the control group are subdivided into three different levels in accordance with the scores of the pretest. Learners, scored 16-18 are in Level I; scored 14-15, Level II; scored 9-13, Level III. Within the experimental group, only the Level III learners did better in posttest than pretest. The situation is similar among the control group. (see Table 4)

Table 4. Pretest and posttest scores at three levels

	Experimental Group		Control Group	
Levels	pretest	posttest	pretest	posttest
Level I (scored 16-18)	16.73	13.64	16.78	13.89
Level II (scored 14-15)	14.43	13.86	14.5	11.4
Level III (scored 9-13)	10.83	11.83	12.4	14.2

3. Results

We can come to the following conclusions safely by analyzing the scores of pretest and posttest between the experimental group and the control group:

First, learners in Level I of two groups did worse in the posttest than the pretest, and students in the control group did better than students in the experimental group, but the differences does not meet the standard of significance. It implies that the conventional method, compared with the pedagogical use of corpus linguistics, is slightly more suitable for advanced learners (see Table 5).

Second, learners in Level II in the experimental group did better than those in the control group, and the difference is statistically significant. Medium-level learners welcome the innovative method and they find it easy to adapt to the new method. The innovative method functions best with medium-level learners probably because they have the ability to use the corpus tool and they welcome the new change in lexical teaching method (see Table 6).

Third, learners in Level III in the control group did better than those in the experimental group, and the difference is statistically significant. The new method does not function well in the low-level learners. Much work remains to be done if the pedagogical application of Wikipedia corpus is conducted among the ESL beginners (see Table 7).

Table 5. Independent-samples t test of Level I

		Independent-Samples T Test								
		Levene's Test for Equality of Variances		T-test for Equality of Means						
		F	Significance	t	df	Significance (2-tailed)	Mean Difference	S.E. Difference	95%Confi.Interval of Diff. Lower	Upper
Average Score (Posttest-Pretest) Level I	Equal Variances Assumed	.	.	.	0	.	-.25000	.	.	.
	Equal Variances Not Assumed			.	.	.	-.25000	.	.	.

Table 6. Independent-samples t test of Level II

		Independent-Samples T Test								
		Levene's Test for Equality of Variances		T-test for Equality of Means						
		F	Significance	t	df	Significance (2-tailed)	Mean Difference	S.E. Difference	95%Confi.Interval of Difference	
									Lower	Upper
Average Score (Posttest-Pretest) Level II	Equal Variances Assumed	.	.	.	0	.	2.46000	.	.	.
	Equal Variances Not Assumed			.	.	.	2.46000	.	.	.

Table 7. Results of independent-samples t test of Level III

		Independent-Samples T Test							
		Levene's Test for Equality of Variances		T-test for Equality of Means					
		F	Significance	t	df	Significance (2-tailed)	Mean Difference	S.E. Difference	95%Confi.Interval of Difference
Average Score (Posttest-Pretest) Level III	Equal Variances Assumed	.	.	.	0	.	-2.37000	.	.	.
	Equal Variances Not Assumed			.	.	.	-2.37000	.	.	.

4. Conclusion

This essay reports on the application of Wikipedia corpus and the analysis of its effects on ESL learners. The findings are complicated, indicating there is room of improvement for the pedagogical application of Wikipedia.

The user-friendly Wikipedia functions not very well as expected. In general, the innovative method is more beneficial for ESL learners but much work remains to be done. Advanced learners had similar test results in the experimental group and the control group. The new method is most welcomed by medium-level learners to improve their lexical acquisition. Low-level learners find the new method too challenging and it does not work well in that group. It is hoped that the pedagogical use of corpus linguistics will be improved to meet the unavoidable need to educates ESL learners to be more interested, curious and critical in English lexical acquisition.

Acknowledgements

This essay is supported by Foundation for Distinguished Young Talents in Higher Education of Guangdong, China; the Talent Cultivation Project of Guangdong University of Foreign Studies (Project Code:GWTP-LH-2014-01).

References

Aston, G. (1997). Enriching the learning environment: Corpora in EFL. In A. Wichmann, S. Tligelstone & T. McEnergy (Eds.). *Teaching and Language Corpora*. New York: Longman.

Boulton, A. (2012). Language awareness and medium-term benefits of corpus consultation. In G. Sanz (Ed.). *New Trends in CALL-Working Together* (pp. 39-46).

Robinson, P. (2001). Task complexity, task difficulty, and task production: Exploring interactions in a component framework. *Applied Linguistics, 22*(1), 27-57. http://dx.doi.org/10.1093/applin/22.1.27

Romer, U. (2011). Corpus research applications in second language teaching. *Annual Review of Applied Linguistics, 31*, 205-225. http://dx.doi.org/10.1017/S0267190511000055

Schmit, R. (1990). The role of consciousness in second language learning. *Applied Linguistics, 11*, 17-46.

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

Skehan, P. (1993). Second language acquisition strategies and task-based learning. In V. D. Scarpis, L. Innocenli, F. Marucci & A. Pajalich (Eds.), *Intreccie contaminazioni*. Venice: Supernova.

Appendix

Appendix 1: The Article for Pretest

Millions of Chinese have dreamed of attending Harvard University. “Harvard Girl”, a how-to manual published in 2000 by the parents of one successful applicant, was a national bestseller. Georgia Institute of Technology, a prestigious university in Atlanta, has enjoyed less name-recognition. Yet this is fast changing: the number of Chinese applicants to Georgia Tech has surged, from 33 in 2007 to 2,309 last year. Some applicants are from the best schools in China, and all are ready to pay around $ 44,000 (for yearly fees and housing costs)-the equivalent of nearly ten times the average annual disposable income of urban households.

The ambitions of Chinese students are shifting: no longer are they attracted just by the glittering names. Pursuit of education abroad is becoming an end in itself. Universities far less renowned than Georgia Tech are reaping the benefits. More than 800,000 Chinese went abroad to study at all levels in 2012 and 2013. In those two years they made up more than a quarter of the 3m who had done so since China began opening to the outside world in 1978. At the end of 2013 nearly 1.1m Chinese were studying abroad, according to the Ministry of Education-more than three times as many as a decade earlier. China has long been the largest source of foreign students enrolled in higher education globally, with its share rising steeply. Since at least 2009 China has provided the most foreign students not just to the English-speaking countries of the developed world but also to numerous others including France, Germany, Italy, Sweden, Finland, Japan and South Korea.

The boom in study in America is especially striking. More than 110,000 students from China were enrolled as undergraduates at American universities in the academic year of 2014-14, eleven times as many as in 2006-07. They now account for 30% of all foreign undergraduates. By comparison, the number of Chinese undergraduates in Britain less than doubled over the same period, to 35,000. The total number of Chinese in all types of higher education in America-274,000-was more than four times as many as in 2006-07, according to the New York-based Institute for International Education.

A fast-growing number of families are sending their children to America earlier to study (and moving with them) as well. In 2013 about 32,000 Chinese received visas for study at secondary schools in America, up from just 639 in 2005. The growth has occurred despite a steep decline since 2010 in the number of Chinese aged between 18 and 22, from 121m to 89m this year.

(Word Count: 419)

Appendix 2: The Article for Posttest

It is often described as the world’s biggest recurring movement of people: a 40-day period spanning the lunar new year (which fell on February 19th this year), during which astonishing numbers of people travel to join distant family members to celebrate the “spring festival”. Officials call this period *chunyun*, or spring transportation. The term evokes horror in the minds of many: trains so jammed that the only place to sit is on lavatory floors. This year the projected number of journeys on public transport during *chunyun*, which will end on March 15th, is nearly 2.9 billion, a 10% increase over the comparable period a year ago. Yet there are reasons to be a little less gloomy about what this entails.

The numbers suggest that despite rapid urbanization, the pull of the countryside remains strong. Many of the journeys involve *mingong*, or peasant workers, as the nearly 300m migrants from the countryside who work in urban areas are often snootily called. Their families are often divided. Children and parents stay in the villages, because a fragmented social-security system makes it difficult for migrants to enjoy subsidized education and health care in the cities. Many migrants think it a good idea that some relatives remain: the stay-behinds can help

retain land-use rights which might come in handy for the migrants if urban work dries up. The authorities themselves are keen for migrants to keep their backstop.

But migration patterns are changing. Wang Kan of the China Institute of Industrial Relations says that, during *chunyun*, trips between provinces have been declining. This is because migrants are often working closer to home, thanks to the relocation of some industries away from the coast to inland provinces where labor is cheaper. "We can see the emergence of more regional hubs, " says Mr Want. No longer is the *chunyun* rush so concentrated in the biggest and wealthiest cities.

Analysing *chunyun* data is difficult. Xiaohui Liang of Renmin University says that companies have recently begun providing private long-distance coach transport for their workers. These trips do not get counted in official statistics. Other workers, he says, get counted twice if they go by train to a regional hub and from there continue by bus to their hometowns. A single worker doing this in both directions would account for four *chunyun* journeys.

(Word Count: 385)

Appendix 3: Pretest Sheet

1. Which of the following words are used in this article?

1) A. handbook B. manual
2) A. prestigious B. esteemed
3) A. surge B. rush
4) A. non-reusable B. disposable
5) A. shift B. alter
6) A. reap B. obtain
7) A. considerable B. numerous
8) A. occur B. happen
9) A. enroll B. inscribe
10) A. glitter B. sparkle

1）-5）_____ _____ _____ _____ _____

6)-10) _____ _____ _____ _____ _____

2. Choose the best translation for the words from the article.

1) bestseller

A. 最好卖的东西 B. 畅销书 C. 很多人买的东西 D. 最受欢迎的东西

2) name-recognition

A. 识别度 B. 辨别度 C. 知名度 D. 区别度

3) equivalent

A. 平等 B. 相同 C. 相似物 D. 对等物

4) annual

A. 年均的 B. 月平均的 C. 平均一季度的 D. 日平均的

5) urban

A. 城市的 B. 农村的 C. 国家的 D. 个人的

6) household

A. 个人 B. 全家 C. 企业 D. 大家

7) renowned

A. 有声望的 B. 臭名昭著的 C. 受欢迎的 D. 受冷落的

8) steeply

A. 大幅度地 B. 险峻地 C. 惊险地 D. 飞快地

9) striking
A. 容貌出众的 B. 显著的 C. 妩媚动人的 D. 吓人的
10) boom
A. 激增 B. 繁荣 C. 减少 D. 衰落
1) -5）_____ _____ _____ _____ _____
6)-10) _____ _____ _____ _____ _____

Appendix 4: Posttest Sheet
1. Which of the following words are used in this article?
1) A. projected B. expected
2) A. span B. bridge
3) A. apart B. distant
4) A. prompt B. evoke
5) A. crush B. jam
6) A. lavatory B. toilet
7) A. comparable B. corresponding
8) A. sad B. gloomy
9) A. entail B. involve
10) A. snobbishly B. snootily
1)-5）_____ _____ _____ _____ _____
6)-10) _____ _____ _____ _____ _____
2. Choose the best translation for the words from the article.
1) recur
A. 年复一年地出现 B. 回想 C. 重现 D. 长久存在
2) astonishing
A. 使人害怕的 B. 惊人的 C. 使人伤心的 D. 使人开心的
3) fragmented
A. 芳香的 B. 片段的 C. 打碎的 D. 四分五裂的
4) subsidized
A. 有优惠政策的 B. 有补贴的 C. 发津贴的 D. 发工资的
5) authority
A. 官方 B. 当权者 C. 权威 D. 行政管理
6) keen
A. 热心的 B. 厉害的 C. 强烈的 D. 敏捷的
7) migration
A. 迁徙 B. 搬家 C. 离开 D. 移居
8) relocation
A. 重新定位 B. 重新迁移 C. 再定位 D. 搬迁
9) statistics
A. 数学 B. 统计学 C. 统计的数据 D. 统计法
10) hub
A. 中心 B. 轮轴 C. 焦点 D. 电线插孔

1)-5）_____ _____ _____ _____ _____
6)-10) _____ _____ _____ _____ _____