English Language Teaching; Vol. 8, No. 10; 2015
ISSN 1916-4742 E-ISSN 1916-4750
Published by Canadian Center of Science and Education

Corpora and Collocations in Chinese-English Dictionaries for Chinese Users

Lixin Xia[1]

[1] Center for Lexicographical Studies, Guangdong University of Foreign Studies, Guanghzou, China

Correspondence: Lixin Xia, Center for Lexicographical Studies, Guangdong University of Foreign Studies, Guanghzou, China. E-mail: cdhuiyi@aliyun.com

Received: August 13, 2015 Accepted: September 19, 2015 Online Published: September 21, 2015

doi:10.5539/elt.v8n10p162 URL: http://dx.doi.org/10.5539/elt.v8n10p162

Abstract

The paper identifies the major problems of the Chinese-English dictionary in representing collocational information after an extensive survey of nine dictionaries popular among Chinese users. It is found that the Chinese-English dictionary only provides the collocation types of *v+n* and *v+n*, but completely ignores those of *v+adv*, *adj+n* and *adv+adj.* And as a common practice, this kind of dictionary doesn't give different collocates of synonymous equivalents. Besides, it provides collocational information on the basis of the headword, but not the equivalents. This leads to the suggestions for a new way of representing collocational information in Chinese-English dictionaries. The most important thing is that the Chinese-English dictionary should provide collocational information about the equivalents instead of the headword. All the five types of collocation are given the same status in a dictionary. Moreover, with corpus data, the selective restrictions of the English equivalents in meanings, frequency and semantic prosody should be explicitly represented to dictionary users by glosses, illustrative examples, collocation columns, etc. It is argued that by doing so, the dictionary will better meet the needs of the users.

Keywords: corpus, collocation, Chinese-English dictionary, lexicography

1. Introduction

Collocation refers to a group of words that habitually appear together and thereby convey meaning by association. According to Lewis, collocation is "the readily observable phenomenon whereby certain words co-occur in natural text with greater than random frequency (Lewis, 1997)." Collocations can be further classified into the following types: grammatical collocation, semantic collocation and lexical collocation.

A grammatical collocation is the co-occurrence of a dominant word such as a verb, noun, or adjective, with a function word or grammatical structure such as a preposition, infinitive or clause in a syntactic pattern. Therefore, "*hungry for*" and "*a pleasure to do*" are grammatical collocations. A semantic collocation is the semantic valency of the words. In other words, some words go together only with particular words as their subjects or complements, as "*dogs bark*", and "*rain heavily*". Therefore, a semantic collocation specifies the rules under which a verb collocates with a noun, or a verb with an adverb. A lexical collocation is a type of construction where a noun, adjective or adverb forms a predictable connection with another word, as in "*pure chance*" and "*completely satisfied*".

Due to the limited space, the present paper will focus on semantic and lexical collocations in a Chinese-English dictionary, excluding grammatical ones. However, there is only limited literature on the subject. Siepmann (2005, 2006) critically reviews different approaches to collocation, and tries to synthesize recent advances in collocational theory into a coherent framework for lexicological theory and lexicographic practice. He concludes that "the traditional dictionary-making process should be turned on its head: rather than starting from an alphabetical framework it should proceed from a bilingual or multilingual onomasiological research base (Siepmann, 2006)." Svensén (2009) discusses in more details the approaches to collocational information in a L1-L2 dictionary. He argues that in a L1-L2 dictionary, "all types of relevant collocations need to be included, transparent or not, directly translatable or not (Svensén, 2009)."

Atkins and Varantola (1997) made an empirical study on dictionary use, and found that dictionary users looked up collocations of L2 words (11%) more frequently than they looked up grammar of L2 words (4%). He (2008) did a similar test on the use of Chinese-English dictionaries, and reached the same conclusion that dictionary users look up collocational information more frequently than they look up grammatical information. This shows the importance of collocation in a Chinese-English dictionary. However, no other literature is found to discuss collocational information in a Chinese-English dictionary. Therefore, the present paper attempts to answer the following questions:

1) How is collocational information treated in current Chinese-English dictionaries?

2) What types of collocation are provided in the dictionaries?

3) How can corpus data about collocation be explicitly represented in the Chinese-English dictionary?

2. Methods

In order to answer the above research questions, current Chinese-English dictionaries will be chosen as the research object. A Chinese-English dictionary can be classified into the dictionary for domestic users and that for foreign users. The working definition for a Chinese-English dictionary for Chinese users (CEDCU) is that a Chinese-English dictionary is one made specifically for the users whose native language is Chinese.

From tens of published Chinese-English dictionaries, we finally chose nine of them. They are the first edition of "*A Chinese-English Dictionary (CED1)*" by Wu Jingrong in 1980, the second edition of "*A Chinese-English Dictionary (CED2)*" by Wei Dongya in 1995, the third edition of "*A Chinese-English Dictionary (CED3)*" by Yao Xiaoping in 2010, "*Concise English-Chinese Chinese-English Dictionary (CECCED)*" by Feng Juehua in 2004, "*A Modern Pocket Chinese-English Dictionary (MPCED)*" by Chen Hongan in 1990, the third edition of "*The Chinese-English Dictionary (TCED)*" by Wu Guanghua in 2010, "*A Practical Chinese-English Dictionary for Translation (PCEDT)*" by Wu Wenzhi in 2001, "*A New Century Chinese-English Dictionary (NCCED)*" by Hui Yu in 2003, and "*New Age Chinese-English Dictionary (NACED)*" by Wu Jingrong in 2000.

They were chosen because they were all published after China adopted its opening policies in 1978, and that they are the most popular ones among Chinese users. For example, the CED series enjoy the greatest number of users in China, and the CED1 is the most influential in China. The TCED includes the greatest number of entries. The NCCED and NACED are said to have the greatest number of collocation in them. The CECCED and MPCED stand for the dictionary of pocket size, and the PCEDT for the dictionary for translation.

After the dictionaries were chosen, all types of collocation that are explicitly represented in the nine dictionaries were collected. That is to say, the collocations in the glosses are included, but those represented implicitly in the illustrative examples are excluded. They are the following five types of collocation: 1) v+n, as *make a decision*; 2) n+v, as *lions roar;* 3) adj+n, as *sharp increase;* 4) v+adv, as *rain heavily,* 5) adv+adj, as *bitterly disappointed* (*cf* Svensén 2009).

After the collocations were collected, data are analyzed based on corpus data. The corpus chosen is BNC because it is the most representative one of its kind containing contemporary British English. And the *word sketch* is used to analyze the data.

3. Results

All the data is shown in Table 1.

From Table 1, we can see that all the dictionaries provide certain types of collocation of the equivalents except the pocket *MPCED*. This may suggest that the editors of the dictionaries are fully aware of the importance of collocation in the Chinese-English dictionaries. The *CED1* was the first one of this kind of dictionary published after China adopted the open policy in 1970's. Its editor included some information about collocation in it, and its revised editions included more.

Table 1. Types of collocation in Chinese-English dictionaries

Title	v+n	n+v	adj+n	v+adv	adv+adj
CED1	+	+	-	-	-
CED2	+	+	-	-	-
CED3	+	+	-	-	-
TCED	+	+	-	-	-
CECCED	+	+	-	-	-
MPCED	-	-	-	-	-
PCEDT	+	+	-	-	-
NACED	+	+	-	-	-
NCCED	+	+	-	-	-

However, these dictionaries only provide two types of collocation, i.e. "*v*+*n*" and "*n*+*v*" in the form of glosses either in English or in Chinese.

Ex. 1 【磕磕绊绊】kēke-bànbàn <形>❶(of a road) bumpy; rough:…. ❷(of a person) limping; jerky: …from *NCCED*

In the above example, the collocates of the equivalents are given in the glosses in English.

Ex. 2 **上课** [- -] ① (学生听课) attend class; go to class…② (教师讲课) conduct a class; give a lesson… from *TCED*

In Ex. 2, the collocates of the equivalents are included in the Chinese glosses. However, a close look at the dictionary shows that the Chinese glosses are actually definitions of the headword. That is to say, the *TCED* defines the headword both in Chinese and English. Sometimes the Chinese definition may contain the collocates. But in most cases it doesn't. Thereby, we may conclude that the *TCED* doesn't provide collocational information about the equivalents of the headword purposely.

4. Discussion

Although most of the Chinese-English dictionaries investigated provide some types of collocation, they have some common problems as shown below.

1) Insufficient number of collocation and a bias in representing different types of collocation

In the examination of the nine Chinese-English dictionaries, we find that they only supply a very limited number of collocations. Besides, they fail to give collocation information in a systematic way, which results in a lack of collocation information about many equivalents.

Moreover, all the nine dictionaries fail to provide the following three types of collocation: *v*+*adv*, *adj*+*n* and *adv*+*adj*. They are frequently used in the English text, and are as important as the collocation types of *v*+*n* and *v*+*n*. But they are not treated in the same way as the latter ones.

2) Not giving different collocates of the synonymous equivalents

In theory, a L1-L2 dictionary should tell its users the differences between the equivalents. However, in practice, it just lists the equivalents without differentiating between them. It is the same case with current Chinese-English dictionaries.

Ex. 3 【宣传】…<动>publicize; propagate; disseminate; preach: ~交通法规 publicize traffic regulations || ~政策 publicize a policy; give publicity to a policy || ~政治观点 propagandize political ideology || 进行~carry out/conduct propaganda… from *NCCED*

In Ex. 3, the dictionary lists four English counterparts of the Chinese headword without further explaining the use of and the difference between them. A user would feel confused facing the four synonymous equivalents. Although the dictionary gives the collocates of "publicize" in the examples part, no collocation information are given about the other three equivalents. Due to the space limit in a dictionary, it is not likely to list all the

collocates of the equivalents by illustrative examples. Besides, it's demanding for a dictionary user to extract and sum up all the use and collocation structure of the equivalents from the examples given in the examples.

3) Providing collocation information on the basis of the headword

For a dictionary user, he expects not only an equivalent in the dictionary, but also use of the equivalent, including collocation information about it. According to Atkins (1997), a dictionary user looks up collocational information about the equivalents more often (10%) than grammatical information about the equivalents (5%) when he looks up in a L1-L2 dictionary. This shows the importance of collocational information in a L1-L2 dictionary.

However, current Chinese-English dictionaries don't provide the needed collocation information about the English equivalents of the Chinese headword. In most cases, if the headword has a collocation structure, the dictionary will provide the counterpart in English. If there is a corresponding collocation structure in English, no problems occur. But this correspondence is rare. In most situations, there is no exact correspondence between them. It is more likely that a Chinese headword has several equivalents of different collocation structures, and the dictionary doesn't further illustrate the collocation patterns of them. That is to say, a Chinese-English dictionary provides collocation on the basis of the headword in the source language rather than the equivalents in the target language. A possible outcome of this treatment is that dictionary users may misuse the equivalents in an incorrect collocation.

Corpus data is widely used in the compilation of the dictionaries, especially in the English learner's dictionaries (Xia 2011). But none of the investigated Chinese-English dictionaries were made on the basis of corpora. With the use of large corpora and software in the compilation of the dictionary, dictionary writers can better represent linguistic information at various levels, including collocation information.

1) Highlighting collocation information based on the design features of the Chinese-English dictionary

Collocation plays an active role in the production of a text, and it is the key to the "naturalness" of it (Xia 2009). Chen (2008) argues that a scientific and systematic representation of collocation in a learner's dictionary will be the chief design feature of the next generation dictionaries. The aim of a Chinese-English dictionary is to help its users produce an English text. Therefore, it should show its users the collocational behaviors of the equivalents in a systematic way.

First, all the collocation types discussed above should be given in a Chinese-English dictionary. That is to say, the types of *v+adv*, *adj+n* and *adv+adj* should be given the same status as the types of *v+n* and *v+n*.

In addition, because of the cultural difference, Chinese collocational structures are quite different from English ones. Therefore, the Chinese-English dictionary should tell its users the difference between them in various ways, such as in the form of collocational columns, error warnings, and illustrative examples.

2) Explicit use of corpus data to show frequency restrictions of the equivalents

Large corpora and software break through the limitation of traditional research on collocation. Now we can count the frequency of the words in a corpus. For example, the *word sketch* not only gives the frequency of particular collocates, but also the significant degree of the collocates.

Corpus data analysis has helped dictionary writers make such decisions on headword selection, sense differentiation, sense ordering, definition, etc. In treating collocational information in a Chinese-English dictionary, dictionary editors should make full use of the corpus data, and highlight collocational structures according to their frequency, including the selection of collocates and the ordering of the collocation structures.

In Ex. 1, the editor gave the collocational structures of the English equivalents, such as "(of a road) bumpy; rough", and "(of a person) limping; jerky", but no frequency data were given about them. This would make the dictionary users think that there is no difference between the two collocation structures, and they can use them at random.

However, through data retrieval by the *word sketch*, we find that the following words tend to go together with *bumpy*: *ride* (22), *road* (11), and *track* (8); and the following words with *rough*: *ground* (60), *ride* (53), *track* (40), *sea* (34), and *grass* (22). In addition, the frequency of *bumpy* is 156, and that of *rough* is 3291. We can thereby conclude that *bumpy* is a low-frequency word mainly used to describe the state of an uneven road, and *rough* has a greater frequency mainly used to describe the state of an uneven ground, a sea surface or a grass land.

As to the collocates of *jerky* and *limping*, they also tend to co-occur with different groups of words: *jerky* is mainly used to describe an unsteady movement (23) or motion (2) while *limping* is mainly used to describe a tottering people (2) or walk (2). And the frequency of jerky (90) is greater than that of limping (35). Based on the corpus data, we revise the entry in Ex. 1 as below:

Ex. 4 **磕磕绊绊** kēke-bànbàn <形>❶(of a ground, ride, grass, etc.) **rough**; (of a ride, road, track, etc.) **bumpy**:…. ❷(of a movement, etc.) **jerky**; (of a person，walk, etc.) **limping**:…

In Ex. 4, we give separate glosses for the two equivalents in sense 1 and 2. Careful users would notice the difference between them. Compared with Ex. 1, we reorder the English equivalents according to their frequency. And the collocates are also listed according to their frequency.

3) Representing semantic prosody of the equivalents

The notion of semantic prosody arises from corpus linguistics. Louw (1993) defines it as "a consistent aura of meaning with which a form is imbued by its collocates". As Rundell (1998) notes that in English certain words tend to co-occur with the words with positive or negative associations. For examples, *event* always collocates with historical, important and other words with positive associations. But *incident* goes together with serious, unfortunate, unpleasant and other words with negative associations.

As it is important for language learners, lexicographers have made efforts to present information about semantic prosody in their dictionaries, and Partington (1998) notes that English monolingual dictionaries based on corpora provide more information about semantic prosody than those not based on corpora. But Chinese-English haven't provided any information about semantic prosody of the English equivalents.

In Ex. 3, after data retrieval by the *word sketch*, we find that *preach* usually co-occurs with the words with positive associations, such as religion, sermon, socialism. *Disseminate* often goes together with the words with neutral associations, such as knowledge, idea, information, etc. *propagate* is often used with contraction, plant in the technical field. Finally, *publicize* collocates with words with negative associations, such as crime, grievance, dispute, etc. Based on the above corpus data, Ex.3 is revised as below:

Ex. 5 **宣传** xuānchuán<动> **preach (**a sermon/gospel/system, etc.**)**; **propagate** (an idea, knowledge, etc.); <正式> **disseminate** (information, a result/ finding/message, etc.); <贬>**publicize** (crime/grievance, a dispute, etc.)**:** 宣传教义 preach a sermon||宣传知识 propagate knowledge||他们的研究发现已经广为宣传。Their research findings have been widely disseminated….

In Ex. 5, we list *preach* as the first equivalent because it has the greatest frequency. Besides, a label <贬> (disapproving) is placed in front of the equivalent *publicize* to present explicitly the semantic prosody of the collocations.

5. Conclusions

In sum, a Chinese-English dictionary should present the collocational behaviors of the English equivalents rather than those of the Chinese headwords as the users need them in their production of an English text. In particular, when there is a difference between the Chinese collocational structures and the English ones, the dictionary should show its users the difference. Finally, in the treatment of collocational information, we should use corpus data explicitly to present the users the selection rules of English equivalents in meanings, frequency and semantic prosody through various means, such as glosses, illustrative examples, collocational columns, error warnings, notes, etc. to better meet the users' needs.

References

Atkins, B. T. S., & Varantola, K. (1997). Monitoring dictionary use. *International Journal of Lexicography, 1*, 1-45. http://dx.doi.org/10.1093/ijl/10.1.1

Chen, G. & Tian, B. (2008). The design features of the next-generation English learner's dictionaries (xia yidai yingyu xuex cidian de sheji tezh). *Foreign Language Teaching and Research (waiyu jiaoxue yu yanjiu), 3*, 224-233.

He, J. (2008). *Towards a Model of a Collegiate Chinese-English Learner's Dictionary.* Beijing: Science Press.

Lewis, M. (1997). *Implementing the Lexical Approach: Putting Theory into Practice*. Hove, England: Language Teaching Publications.

Louw, B. (1993). Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and Technology: In Honour of John Sinclair*. (pp. 157-76). Philadelphia/Amsterdam: John Benjamins. http://dx.doi.org/10.1075/z.64.11lou

Partington, A. (1998). *Patterns and Meaning: Using Corpora for English Language Research and Teaching*. Amsterdam & Philadelphia: John Benjamins. http://dx.doi.org/10.1075/scl.2

Rundell, M. (1998). Recent trends in English pedagogical lexicography. *International Journal of Lexicography, 4*, 315-342. http://dx.doi.org/10.1093/ijl/11.4.315

Siepmann, D. (2005). Collocation, colligation and encoding dictionaries. Part I: Lexicological aspects. *International Journal of Lexicography, 18*(4), 409-43. http://dx.doi.org/10.1093/ijl/eci042

Siepmann, D. (2006). Collocation, colligation and encoding dictionaries. Part II: Lexicographical aspects. *International Journal of Lexicography, 19*(1), 1-39. http://dx.doi.org/10.1093/ijl/eci051

Svensén, B. (2009). *A Handbook of Lexicography: The Theory and Practice of Dictionary-Making*. Cambridge: Cambridge University Press.

Xia, L. (2009). Recent developments and trends in corpus lexicography (Part I/Part II) (yuliaoku cidianxue de z`uixin fazhan he weilei qushi (shang/xia)). *Journal of Lexicographical Studie*s, 3/4, 71-78/81-91.

Xia, L. (2011). Some reflections on the compilation and publication of the Chinese-English dictionary (dui hanying yuwen cidian bianzuan he chuban de yixie sikao). *Publishing Science (Chuban Kexue), 2*, 23-27.