Compilation of Modern Chinese Loanword Dictionary Based on Big Data

Xiaoyan Zeng
School of Foreign Languages,
China University of Petroleum- Beijing
E-mail: xiaoyanzeng0313@163.com

Abstract
A few scholars have put forward ideas for compiling a modern Chinese loanword dictionary based on big data, but they have not conducted an in-depth discussion on how to compile a Chinese loanword dictionary based on big data. This article analyses the current situation of the compilation of Chinese loanword lexicon. It has some problems as the lack of Chinese corpus data support, the lack of epochal, the disconnection between the practice and research of loanword dictionary compilation, and the imbalance of loanwords included in dictionaries. Based on the guidance of big data theory, this paper summarizes the characteristics of modern Chinese loanword data, such as source complexity, times, dynamic development, interdisciplinary and type diversity. There are four inspirations: 1. Based on Compilation of a large amount of real corpus. 2. Compilation based on analysis of usage behaviour. 3. Compilation combined with Chinese loanword big data. 4. Compilation based on real-time revision of complex dynamic data.
Keywords*:* modern Chinese; loanwords; database; loanword dictionary; big data

1. Introduction

The dictionary not only reflects the new achievements and new understandings of lexical research, but also more comprehensively reflects the new appearance of the Chinese vocabulary system in the new era. It reflects the profound changes in society from multiple angles and is full of strong atmosphere of the times. The content of a dictionary records language communication and social development. It is circumstantial evidence of the contact of two or more languages. The authenticity of the corpus is the main guarantee for the compilation of the dictionary. The many problems existing in the loanword dictionary reflect the inconsistency and unsystematicness in the theory and research methods of Chinese loanwords. The compilation of modern Chinese loanword lexicon has problems such as lack of Chinese corpus support, lack of timeliness, and the inability to balance the included content. In addition, the existing Chinese dictionaries have been unable to meet the needs of loanword research to a large extent due to the limitations of their age, interpretation, matching examples, style, and part of speech. Both the study of Chinese loanwords and the compilation of Chinese dictionaries need the support of big data. The definition of loanword in modern Chinese is "when the Chinese system is difficult to carry the practical meaning of new things, loanwords emerge as the times require, and they enter the modern Chinese system in various means and ways. Specifically speaking, loanwords depend on the interaction between the communicative subject and the available language environment (composed of social factors, cultural factors, political factors, economic factors, educational factors, policy/system factors and Internet factors) in the process of modern Chinese contacting with the languages of other countries or ethnic minorities in China Word borrowing is the result of the interaction of knowledge, frequency and Chinese characteristics, and the comprehensive effect of borrowing and evolution modes such as epochal, competitive, complementary, adaptive, normative and economic, and these words borrow the sounds (pronunciation), form (vocabulary, writing form, word formation, syntax) and meaning (semantics) of foreign vocabulary within or outside the Chinese nation

In terms of arrangement, they are in line with the characteristics of Chinese words."[11] This definition re-examines loanwords from the perspective of complex dynamic system, breaks through the traditional linear and static view of loanword research, and takes into account the actual situation of modern Chinese, English and Japanese contact, as well as the cross relationship between the two major languages borrowed by Chinese loanwords. This paper analyses the present situation of modern Chinese loanword dictionary, the big data theory of modern Chinese loanword dictionary compilation, and the characteristics of loanword data based on big data, and then discusses the compilation principles of Chinese loanword dictionary based on big data.

2. The Current Status of the Compilation of Modern Chinese Loanword Dictionaries

The current situation of the compilation of modern Chinese loanword lexicon is mainly summarized in four aspects: first, the compilation of the dictionary lacks the support of language corpus; second, the practice of loanword lexicon compilation is out of touch with the study of Chinese loanwords; third, the Chinese loanword dictionary lacks modernity; fourth, the dictionary cannot balance the loanwords included in it.

2.1. The compilation of dictionaries lacks Chinese corpus support

From the 1960s to the early 1990s, due to the lack of strong support from big data technology and computer-aided conditions in the compilation of dictionaries, many entries did not give the used or earliest documentary evidence, even though some dictionaries provided the entries. The earliest time and documentary evidence that appeared, but after investigation and verification, it was found that it was earlier than the time marked in the book. Take the term "白领"（*bailing*; meaning *white colla*r) for example, the earliest documentary evidence for this entry provided by the Modern Chinese New Word Dictionary was in 1987, and after checking through the Oriental Magazine full-text search database and the People's Daily graphic database, it was discovered that the full-text search database of Oriental Magazine had been recorded as early as 1933. The time stamp of this entry is half a century away, which will affect the investigation of the diachronic evolution of loanwords to a certain extent.

The Neologism Dictionary also contains some loanwords, but these loanwords have not appeared in the mainstream media. The reasons are as follows: 1)The vitality is weak; 2)There are variants, such as "卡拉喔开" and "卡拉哦开" (both meaning *Karaoke*); 3)It is a translation preference, and has a certain degree of regulation. For example, "卡拉OK" (Karaoke) is included in the Modern Chinese Dictionary as a standardized loanwords; 4)The communicative subject adopts ways like transliteration, paraphrase, euphony, homophone, adding or subtracting words at will for the convenience of communication or the pursuit of coordination, humor or other effects, such as "食肉男" (*eating man*) or "食肉男子" (*eating man*) . These are also valuable materials for the study of loanwords, but they have not attracted people's attention. In short, there is still a lot of research space for compiling a dictionary of loanwords based on a large corpus. The research of modern Chinese loanwords based on one or two dictionaries will affect the determination of the scope and object of loanword research to a certain extent. Only based on large-scale modern Chinese big data can the research materials of modern Chinese loanwords be scientific and accurate.

[11] Zeng Xiaoyan. The Definition of Modern Chinese Loanwords: A Complex Dynamic System Theory. *Overseas Chinese education*, 2017(5).

2.2. Loanword dictionary compilation practice is out of touch with Chinese loanword research

There have been controversies over the definition of "borrowed words" or "loanwords", and scholars have also debated the attribution of free translation words and Japanese loanwords. In order to avoid these entanglements, the editor simply changed the research object, and no longer paid special attention to loanwords, but investigated the development and changes of Chinese vocabulary from the perspective of neologisms or etymology. However, academic works on loanwords research (such as Cui Yin et al., 2013; Qiao Yan et al., 2011) and doctoral dissertations (such as Li Yanjie, 2006; Fang Xinxin, 2008; Jin Xiyong, 2011, etc.) have not stagnated, yet their theoretical results have not been used to guide the practical activities of lexicography, and eventually lead to the lack of cohesion and systemicity between the theoretical research and practical application of Chinese loanwords.

2.3. Chinese loanword dictionary lacks temporal spirit

Since the Chinese Loanword Dictionary was published in the early nineteenth century, no new edition has been issued for about 20 years, and many words reflecting social development and changes and social phenomena cannot be preserved. For example, the Languages Dictionary compiled by Hu Xingzhi in 1936, the Chinese Loanword Dictionary compiled by Liu Zhengyi in 1984, the Mandarin Daily Foreign Language Dictionary edited by the compilation group of Mandarin Times Publishing Department in 1985, and the Chinese Loanword Dictionary compiled by Cen Qixiang in 1990 cannot be updated in time. The dictionaries are outdated, the entries are outdated, and the orthodox and the variant coexist, which cannot reflect the development and changes of the society in time, and its circulation in language communication is low. In addition, because the loanword dictionary was not updated in time, the letter words that emerged in the 21st century were not included in the loanword dictionary in time.

2.4. The dictionary cannot balance the loanwords included

The Dictionary of Chinese Loanwords does not indicate the time, the authors and the names of works which have reference value. If added, it will further enhance its research value. The Dictionary of Chinese Loanwords does not have widely collected examples of important reference value such as "琵琶" (*pipa*; meaning *lute*; loanwords from the Han Dynasty) and "袈裟" (*jiasha*; meaning *robes worn by monks*; loanwords borrowed from Sanskrit after the Eastern Jin Dynasty). Have the received words really entered the Chinese language? Do some individual words exist in Chinese? Is there any phenomenon of partial receipt? Or have any words been missed? Wang Enwei verified the Chinese loanwords originating from Russian in the *Chinese Loanword Dictionary*, and found out some phenomena: more technical terms and less general vocabulary; more written words and less oral words; more common words and less dialect words.[12] There are two ways of absorbing loanwords: one is through oral communication, and the other is through written documents. The latter is lacking in collecting loanwords from the corpus of people's real verbal communication activities.

Interdisciplinary research is the general trend of scientific research in the 21st century, which is conducive to broadening the horizon and grasping the overall situation in an all-round and three-dimensional manner. The borrowing mode of loanwords has extraordinary significance in terms of the content, basis, methods and means of dictionary selection. According to the characteristics of the corpus data, it examines the compilation rationale

[12] Wang Enwei. A Glimpse of "Chinese Loan Words Dictionary"——Comments on the Chinese Loan Words of Russian Origin in the Dictionary. *Dictionaries Research,* 1987(2): 115- 121.

and principles of the dictionary of loanwords, and analyzes the real-time revision function of the dictionary of loanwords. The purpose of compiling a dictionary of Chinese loanwords is to serve global Chinese teaching and related research. It is hoped that the compilers of loanword dictionary can be inspired and provide some reference for them.

3. Big data thinking in the compilation of Loanwords in modern Chinese

3.1. Big data thinking

Big data thinking has subverted people's traditional ideas, has new guidance on the traditional database design and construction methods, and has higher requirements in the number, scale and depth of corpus. Studying massive corpus data is to study complete language communication. It is the only source for building a database of Chinese loanwords.

Big data has the characteristics of large capacity, fast speed, full type and uncertainty, which provides new ideas and methods for solving various database problems. In response to the development needs of the information age, big data has collected a large number of corpus resources for the study of loanwords. The Internet era will be an era of interconnection and sharing connecting everything. According to research, English language products are widely distributed on the Internet and are closely related to language resources in other countries. Chinese language products are relatively narrowly distributed on the Internet, and have very little contact with language resources in other countries, showing a trend of independent development. Therefore, the construction of Chinese big database should follow the development trend of the Internet in the era of big data. The modern Chinese loanword research big database will usher in an era of "incorporating user characteristics", an era of "always updated word network", an era of "data mining technology development", and an era of "interconnection and sharing".

3.2. The connection idea of big data

The Internet big data currently under construction in the world can be roughly divided into three categories: 1) Internet-connected databases; 2) Internet as a database; 3) Internet-based databases[13]. The construction of the Chinese loanword database in the era of big data can be achieved in three ways: sharing, connection and self-production. The ways are as follows: 1) Sharing is to directly collect the established databases on the Internet, such as various types of Chinese corpora, Oriental Magazine, R.China Newspapers and People's Daily, etc.; 2) Connections are network data links realized through search software such as Google, Baidu, Aol, Bing, Yahoo, Sogou, etc.; 3) Self-production refers to a large amount of corpus generated through forums and chat software such as Wechat, QQ, Skype, etc. 4) Self-built is to collect a large number of loanword dictionaries, new word dictionaries, etymology dictionaries and alphabetic word dictionaries from 1912 to 2020, and respectively use the whole and selection methods for manually entering the data of loanwords, to build a information big data of loanwords.

3.3. Predicting the use behavior of users

The use behavior of users helps the research and development of loanwords dictionaries to make choices in the selection of dictionary content, and also points out the direction of research and development for loanword dictionaries to adapt to the changes of the times. Use-oriented dictionary compilation is actually an interconnected feedback loop, mainly centered on users' usage data, forming an interconnected feedback loop with three types of groups: managers, developers, and users. The analysis of user behavior includes both

[13] Zheng Tongtao, Zeng Xiaoyan. The construction of Chinese interlanguage corpus in the era of big data. *Journal of Xiamen University (Philosophy and Social Sciences Edition),* 2016 (2): 53-63.

descriptive analysis and predictive analysis[14]. Descriptive analysis can answer questions for dictionary developers, such as "What did you borrow and from where?" "What is the reason for borrowing?" "What is the result of borrowing?", etc. Using behavior analysis can also help dictionary developers make predictions, that is, predictive analysis, such as "Which type of words are easy to borrow?" "What will happen if this trend continues?" "What is the impact on Chinese vocabulary system?" "What is the most likely problem in the process of borrowing loanwords?", etc. All the data that has occurred or is happening seems to have nothing to do with the development of the loanword dictionary, but it is these unrelated data that tell us what problems have occurred or what problems will occur. The modernity of society is a typical feature of dictionary research and development. The loanword big data can effectively improve the interaction between developers and users. Based on this big data, developers can design a more practical loanword dictionary, find potential problems, and correct them in time.

4. Characteristics of modern Chinese loanword big data

4.1. Source complexity of the data

The loanword data comes from a wide range of disciplines, such as newspapers, news, science, novels, literary works, magazines, interviews, television, radio, microblog, forum, website, etc.

4.2. Temporal spirit of the data

The loanwords have strong characteristics of the times, and corresponding loanwords will be produced in different periods. They not only record the development of Chinese language, but also record China's development and changes in society, politics, economy, foreign trade, education, art, technology and other fields. These loanwords are recorded and preserved in the form of a dictionary, providing reference value for sociology and lexical research. These loanwords have different characteristics based on the differences in the contents recorded in different periods and locations. For example, before and after the May Fourth Movement, they mainly absorbed words related to science, industry, and medicine. During the Reform and Opening period, they mainly absorbed words related to economy, society and education, etc.

4.3. Type diversity of the data

The characteristic of type diversity of Chinese loanwords are mainly reflected in the following three aspects: First, from the perspective of type, it includes vocabularies borrowed from Chinese minority languages such as Uyghur, Tibetan and foreign languages such as English, Japanese and Italian as well as the new form of vocabulary-foreign letter words; second, from the perspective of borrowing methods, it includes: 1) pure transliteration words, such as "巧克力" (*qiaokeli*; meaning *chocolate*), "比基尼" (*bijini*; meaning *bikin*i)), "沙发" (*shafa*; meaning *sofa*), etc.; 2) Homophonic transliteration words, such as "香波" (*xiangbo*;meaning *shampoo*), "酷" (*ku*; meaning *cool*), "可口可乐" (*kekou kele*; meaning *coca cola*), "脱口秀" (*tuokouxiu*; meaning *talk show*) etc.; 3) transliteration plus annotated words, such as "酒吧" (*jiuba*; meaning *bar*)), "高尔夫" (*gao er fu*; meaning *golf*), "啤酒" (*pijiu*;meaning *beer*), etc.; 4) borrowed words or translated words, such as "放送" (*fangsong*; meaning *broadcast*), "写真" (*xiezhen*; meaning *portray*), "人气"(*renqi*; meaning *popularity*), TV, CEO, ATM, BBS, IT, etc.; 5) imitation translation (Calque) , Such as "黑板" (*heiban*; meaning *blackboard*), "热狗" (*regou*; meaning *hot dog*), "快餐" (*kuaican*; meaning *fast food*), etc.; 6) paraphrase words, such as "电脑" (*diannao*; meaning

[14] May T A. Analytics, University 3.0, and the future of information technology. EDUCAUSE Review, [EB/OL]. http://net.educause.edu/ir/library/pdf/ERM1159.pdf. 2011 -09-01.

computer), "吉祥物" (*jixiangwu*; meaning *mascot*), "峰会" (*fenghui*; meaning *summit*), etc.; third, from the formal perspective , there are two main types of dictionary features with foreign letters: one is letter words composed of pure English letters, such as VIP, CEO, ATM and BBS; the other is mixed words, such as English letters and Chinese characters Combined, such as "E时代"," IT 业" and "VIP 卡".

4.4. Dynamic development of the data

With the changes of the times, the loanwords will show varying degrees of changes. The change of each loanword in the big data will inevitably cause a series of fluctuations or drastic changes in the loanword system, prompting the redistribution and recombination of the vocabulary system or language system. With the development of society, loanwords may produce new cultural connotations, and their original connotations are either discarded, or they coexist with the new connotations for a period of time and then gradually fade until they disappear. According to the earliest time when each loanword appeared according to the earliest documentary evidence or the earliest time when it was included in the dictionary, it was made into a dynamic loanword research big data to observe the changes in the content of loanwords at each stage. It is easier for students to understand, master and use loanwords.

4.5. Interdisciplinarity of the data

The interdisciplinarity refers to digging out the corpus of different subject areas and collecting corpus related to the language. The corpus of different subjects is the most complete language product and characteristic description of human beings, which can reflect the whole picture of Chinese language. People's Daily mainly involves politics, culture, law, society, economy, education, science and technology, health, environment, news and other fields. Oriental Magazine has 15 different categories, including social theory, decree, internal affairs, military, diplomacy, education, finance, industry, transportation, commerce, religion, miscellaneous things, novels, series of talks, and new book introductions. It's the most influential encyclopedia journal encompassing all information. Ge Gongzhen (2003: 128) believes that Oriental Magazine is "always working hard and has the longest period". [15]Oriental Magazine has gone through various important historical periods of modern Chinese history such as the late Qing Dynasty, the Revolution of 1911, the May Fourth Movement, the War of Resistance against Japan, and the War of Liberation. It closely followed social changes and faithfully recorded the trajectory of the modern Chinese process, and comprehensively, accurately and timely reflected the latest information on politics, economy, culture, society, education, military and science at home and abroad at that time. Under the dual influence of the great social changes in China and the introduction of western learning, complex language contact practices have continuously spawned batches of new loanwords.

5. Enlightenment on the compilation of Chinese loanword dictionary

The enlightenment of the database-based dynamic compilation of loanword dictionary includes: real corpus compilation; compilation based on usage behavior analysis; compilation combined with Chinese loanword big data; compilation based on real-time revision of complex dynamic data.

5.1. Compilation based on real corpus

The meaning of the term must be confirmed one by one by searching the text information in the big data, and the phenomenon of polysemy and collocation errors should be avoided as much as possible. In the search process, in order to accurately determine when the loanwords were introduced and the correct collocation, the composition or collocation of

[15] Ge Gongzhen. History of newspaper studies in China. Shanghai: Ancient books publishing house, 2003:128. (In Chinese)

each loanword must be carefully analyzed to avoid mistaken judgments caused by carelessness. It is a common retrieval phenomenon that two adjacent Chinese characters in a text are easily "forced" to combine. For example, when searching for "热身" (*reshen*; meaning *pre-match preparation*), it will show irrelevant contents such as "I can't stand the hot weather", "I can't warm my body". For another example, searching for "恶所"(*e suo*; no meaning) will show irrelevant contents such as "determined by personal likes and dislikes", "intimidated by the enemy's extreme evil" and so on.

In the case where the usage time of the term given in some dictionaries is inconsistent with the time used for big data retrieval, the earliest occurrence shall prevail. For example, the term" 白领 " (*white collar*), Modern Chinese New Word Dictionary provided the documentary evidence that the entry appeared in 1987, but through the Oriental Magazine full-text search database and the People's Daily graphic database to check the word separately later, it was discovered that the term had been recorded in 1933, and the time stamp of the term was about half a century apart. This will affect the investigation of the development and evolution of loanwords to a certain extent. In the case that some dictionaries do not provide the usage time of certain entries, it is recommended to use the earliest usage time retrieved from the big data as the time when the word enters the Chinese system.

The Chinese Loanword Dictionary determines the normal and variant forms of Chinese loanwords from the perspective of standardization. "巧克力"(pronounced as *qiao ke li*) was determined to be a normal body, while "巧古力(*qiao gu li*)", "巧格力(*qiao ge li*)", "朱古力(*zhu gu li*)", "朱古律"(*zhu gu lü*), "查古列"(*cha gu lie*), and "勺古力(shao gu li)" were determined to be variant forms. The included variants are also very precious materials. These variants are a true and objective reflection of social phenomena at that time. The lexicographer should conduct rigorous research on the etymology of loanwords, so that the dictionary of loanwords can be compiled in a standardized manner.

5.2. Compilation based on usage behavior analysis

We can make statistics on users' expectations and application data, and develop dictionaries that meet users' use accordingly; we can also investigate dictionary requirements based on usage behavior analysis, decompose research and development goals, and then make choices about dictionary content. Use behavior analysis includes predicting user behavior, suggesting or guiding users to use dictionary resources. There are three aspects to the research on the needs of loanword dictionaries: First, collect Chinese corpus of communicative subjects in different languages, and count high-frequency words. Second, track users' language communication patterns, collect language behavior data on social networks and new media, integrate and count frequently-used words, and further filter out loanwords. In this process, research and development as a team can make up for the limitations of material selection; inviting language psychologists to participate can enhance the accuracy of the analysis of usage behavior. Third, the purpose of use is gradually decomposed and the corpus resources are reasonably allocated to ensure the effectiveness of the use process. A large amount of data not only retains the characteristics of users, but also records the formation process of all spoken and written language products, as well as the reasons, processes and results that affect the borrowing of loanwords.

5.3. Compilation combined with Chinese loanword big data

The compilation of loanword lexicon should integrate the big data of loanwords, and implement dynamic management of data, so that the loanwords collected can change with the times, so as to improve the quality of dictionary compilation. The compilation of

loanword lexicons should be based on the difficulty of use, and the collected loanwords cannot be combined at will. The needs of users and the background of users should be investigated. The comprehensibility function in database design facilitates the compilation of loanword dictionaries. In the process, Chinese keyword analysis tool such as "SEO Word-Searching Net"[16] can be used for the analysis and mining of Chinese keywords.

Compiling a dictionary of loanwords based on the Chinese loanword big data helps to solve the following problems: First, it is necessary to clarify the object and scope of the collection, complete the collection, and increase the collection of loanwords. Second, it fits the examples and interpretations, and involves the rationale for the entry. Third, the matching rules combine cultural knowledge to make them highly practical. There are three types of matching examples in entries: word examples, language examples, and sentence examples. Good matching examples should explain the function and fully reflect the meaning of the word, so that the meaning of the word can be instantiated, concrete, and contextualized. Fourth, uniformly mark the parts of speech, earlier use time and documentary materials.

5.4. Compilation with Statistic Analysis of Typical Loanword Patterns

The user's use process is a complex and dynamic process. With the interaction of social, cultural, policy, economic, political, diplomatic, and educational factors, the user's behavior and purpose are constantly changing. The compiling of Chinese loanword dictionary cannot ignore the importance of real-time revision function. The borrowing mode of loanwords is complex and dynamic. It is integrated with the big data to make the dictionary revision with real-time function. It can extract the required data from the big data anytime and anywhere from multiple angles and revise the previous version of the dictionary. Real-time dictionary revisions greatly increase the frequency of loanwords.

Compilation of more flexible and multi-level dynamic dictionaries can maximize the use value of dictionaries, either based on paper dictionaries, based on electronic dictionaries, or based on online dictionary interactive platforms. The real-time revision of the dictionary is based on the data-oriented adaptive compilation of the dictionary of loanwords: 1)The compilation of the dictionary is based on the user's usage status; 2)Starting from the user's difficulty in using it, analyzing the difficulty of using different types of users; 3)Finding the use intervene in time when the problem arises and modify the dictionary to adapt to these users; 4) Develop and adopt more effective editing ways and methods through matching with editing tools. The development and improvement of dictionaries should not only be based on user feedback, but also on data-driven analysis that includes users' difficulty in the using and dictionary entries.

6. Conclusion

The big data provides sufficient research materials for loanword lexicographers and linguistics researchers, which has changed the lack of data in the past in linguistics research, and also challenged traditional linguistic research methods. The development of linguistics and applied linguistics supported by new technologies such as Mobile Internet and Big Data have brought new opportunities and challenges. The computing speed of computers and the Cloud Era have brought terabytes of storage space, which provides technical support for big database development. The rapid development of network and electronic publications has provided a wealth of language data for database construction, and has increased the

[16] "SEO" is the abbreviation of "Search Engine Optimization", Chinese translated as "search engine optimization", similar to "word tracker", can be used for the analysis and mining of Chinese keywords.

speed of data collection and corpus entry. In terms of the scope of corpus collection, it can provide complete loanword materials; in terms of corpus content, it can provide loanword data for interdisciplinary research; in terms of research theory and methods, it can continuously improve and develop borrowing mode of loanwords using interdisciplinary research results; in terms of corpus format, it can provide text corpus and phonetic corpus of loanwords. The loanword dictionary compiled based on the Chinese loanword big data has improved the application value of the dictionary to a certain extent.

Acknowledgment
This work was supported by the National Social Science Fund of China (No.18CYY027), the Science Foundation of China University of Petroleum-Beijing (No.2462020YJRC002), and the Science Foundation of China University of Petroleum-Beijing (No.2462020YXZZ010).

References

[1] Zeng Xiaoyan. The Definition of Modern Chinese Loanwords: A Complex Dynamic System Theory. *Overseas Chinese education*, 2017(5). (In Chinese)
[2]Fang Xinxin. *Three states and two processes in language contact.* Doctoral Dissertation of Central China Normal University, 2004. (In Chinese)
[3]Ge Gongzhen. *History of newspaper studies in China*. Shanghai: Ancient books publishing house, 2003:128. (In Chinese)
[4]Jin Xiyong. *A sociolinguistic study of Loanwords in modern Chinese*. Doctoral Dissertation of Fudan University, 2011. (In Chinese)
[5]Li Yanjie. *On the development of Loanwords in modern Chinese*. Doctoral Dissertation of Shandong University, 2006. (In Chinese)
[6] Wang Enwei. A Glimpse of "Chinese Loan Words Dictionary"——Comments on the Chinese Loan Words of Russian Origin in the Dictionary. *Dictionaries Research*, 1987(2): 115- 121. (In Chinese)
[7]Zheng Tongtao, Zeng Xiaoyan. The construction of Chinese interlanguage corpus in the era of big data. J*ournal of Xiamen University (Philosophy and Social Sciences Edition)*, 2016 (2): 53-63. (In Chinese)
[8] Zheng Tongtao, Zeng Xiaoyan. Study on development of Chinese country-specific teaching materials based on big data. *Overseas Chinese education*, 2016(3).
[9]Giora, Rachel., Understanding figurative and literal language: The graded salience hypothesis. *Cognitive Linguistics*, 1997, 8(3):183-206.
[10]Kaszubski P., *"Enhancing a writing textbook: a national perspective"*, In Granger (ed.) Learner English on Computer. London: Addison Wesley Longman. 1998:172-185.
[11]Mukherjee J. & Rohrbach J., *"Rethinking applied corpus linguistics from a language-pedagogical perspective: New departures in learner corpus research"* in B. Kettemann and G. Marko (eds.) Planning, Gluing and Painting Corpora: Inside the Applied Corpus Linguist's Workshop, Frankfurt: Peter Lang, 2006:205-232.
[12]Tono Y. Using learner corpora for L2 lexicography, *LEXIKOS*. 1996, (6): 116-132.
[13]Tono Y., *Research on Dictionary Use in the Context of Foreign Language Learning*, Tübingen: Max Niemeyer Verlag. 2001,
[14]Giora, Rachel., Understanding figurative and literal language: The graded salience hypothesis. *Cognitive Linguistics*, 1997, 8(3):183-206.
[15]May T A. Analytics, University 3.0, and the future of information technology. EDUCAUSE Review, [EB/OL]. http://net.educause.edu/ir/library/pdf/ERM1159.pdf. 2011-09-01