Exploring the Differentiation of Near-Synonyms in Smart-Technologies Framework

Juan Li
School of Chinese Language and Literature
Beijing Normal University
Beijing, China
lijuan@mail.bnu.edu.cn

***Abstract*—Near-synonyms discrimination is one of the most difficult problems in the area of teaching Chinese as a foreign language. This paper draws on the existing analysis framework and proposes a framework for obtaining near-synonym discrimination resources based on the technologies of automatic collocation extraction and semantic similarity calculation. In this paper, the case study and the method research in the application proved the effectiveness of the proposed framework. In addition, visualization technology is used in this paper to present the discrimination data of near-synonyms, which is helpful to obtain the nuance from the complicated data and can reduce people's cognitive burden, and then promote the effect of Chinese near-synonyms teaching.**

Keywords—near-synonyms discrimination, collocational resources, visualization

I. INTRODUCTION

The existing near-synonyms discrimination dictionaries predominantly illustrate the distinctions between near-synonyms through their definitions, collocation examples, and sample sentences. Fu (2006) highlighted that it's challenging to discern the subtleties of near-synonyms solely from the content of explanatory notes, which may even lead to misunderstandings. Fu (2000) proposed that the similarities and differences among near-synonyms, encompassing lexical meanings, alterations in application-based meaning, and usage disparities, are all exhibited in combination. Thus, it's essential to analyze near-synonyms in conjunction. In other words, merely observing the meaning of near-synonyms cannot adequately reveal their similarities and differences.

Chu (2021) highlighted that the primary perspectives for observing and analyzing synonym differences are form, semantics, and pragmatic context, encompassing aspects such as collocation, distribution, semantics, emotional color, style, among others. The nuances in semantic meaning and usage of near-synonyms are predominantly exhibited through collocation and distribution.

Currently, apart from Firth's (1957) renowned dictum, "You shall know a word by the company it keeps," and the list of collocation examples in synonym discrimination dictionaries, existing thesauri remain deficient in the collocation knowledge of key words, with no thesaurus providing a detailed and comprehensive depiction of synonym collocation knowledge.

Lin (1994) categorized word collocation into two types: (1) habitual collocation, which defies justification; and (2) event collocation, where a certain collocation is reasonable but another is not, with word event collocation belonging to the semantic system and not to be mixed with the cognitive or logical systems.

For near-synonyms discrimination, clarifying the collocation differences among a group of near-synonyms can aid Chinese second language learners in grasping their usage disparities. To address near-synonym discrimination issues, it's crucial to delve into the syntactic and semantic discrepancies in near-synonym collocation and distribution. Furthermore, a set of operable mining methods and a concrete description framework for near-synonym collocation knowledge are essential.

II. RELATED WORK

A. Exploring the nuances: A dictionary defines the nuance between '爱惜(aixi) ' and '珍惜(zhenxi) '

As shown in Table 1, the *Modern Chinese Dictionary*(the 7th edition) provides the definitions for ' 爱惜(aixi) ' and '珍惜(zhenxi)' as follows:

TABLE I. THE DEFINITIONS OF '爱惜(AIXI) ' AND '珍惜(ZHENXI)' IN MODERN CHINESE DICTIONARY (7TH EDITION)

Near-synonyms	Meaning
爱惜(aixi)	Avoid squandering due to appreciation and cherish. Cherish： cherish time， Cherish the national treasures.
珍惜(zhenxi)	Cherish and treasure: cherish time.

It can be observed that the Modern Chinese Dictionary (7th edition) employs the method of mutual interpretation of near-synonyms to elucidate the meaning of a group of near-synonyms. However, this interpretation does not appear to assist second language learners of Chinese in discerning the differences between these two words.

Comparison of the Usage of 1700 Pairs of Near-*Synonyms* is a learner's dictionary oriented towards export. It highlights that the collocation of '爱惜(aixi) ' can be something that either concrete or abstract. Conversely, the collocation of '珍惜(zhenxi) ' are predominantly abstract things. Nevertheless, for Chinese second language learners, this dictionary only offers 8 collocations, without specifying which are abstract things and which are

concrete things. For low-level international students who have not yet established the Chinese abstract and concrete noun system, it remains unclear which collocation objects are appropriate. Moreover, utilizing abstract and concrete general words proves insufficient in aiding them to comprehend the differences between collocation objects.

B. *Analysis framework of near-synonyms—— 'Nuance of Dimensionality + Hierarchy of Words'*

Fu (2006) argues that the analysis model of 'Nuance of Dimensionality + Hierarchy of Words' should be adopted to distinguish synonymy nouns. Species difference can be "form, structure, function, cause, time, space, quantity, degree, evaluation", etc., and Hierarchy of Words are the upper and lower hierarchy of words.

We compare the semantic categories of nouns in Modern Chinese Semantic Dictionary and HowNet. Modern Chinese Semantic Dictionary (the former) is a machine-readable semantic dictionary for Chinese-English machine translation developed by Computational Languages of Peking University. As shown in Figure 1, the semantics of nouns fall into four main categories: things (concrete things, abstract things), processes, time, and space. HowNet is known as the network, the noun is divided into two categories of events and entities, entities include everything, time, space and components of four parts. Among them, all things include social units, matter (living, non-living) and spirit (psychological characteristics, occurrences. The red content is best placed at the bottom of Figure 1.

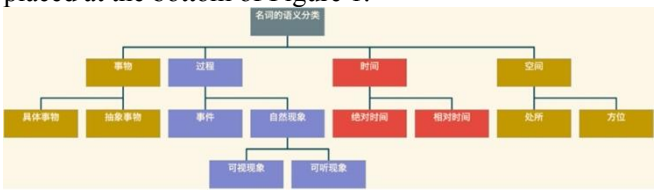

Fig. 1. Noun semantic classification in the Modern Chinese Semantic Dictionary

The analysis shows that the semantic classification of specific nouns by these two semantic resources is basically the same, but the number of abstract nouns included in these two semantic dictionaries is limited, and the semantic class of abstract nouns is difficult to be determined by these two semantic resources. In actual discrimination, researchers need to summarize more detailed semantic features according to the "word meaning component - word meaning formation pattern" of collocation words. Therefore, collocation words can be divided into different categories, which is convenient for learners to memorize and construct a complete lexical semantic network.

C. *Cluster analysis*

Clustering involves grouping objects with similar characteristics into a single category based on specific features. Zhou (2016) pointed out that "clustering attempts to divide the samples in a data set into several subsets, which are usually disjoint, and each subset is called a 'cluster'". By this division, each cluster may correspond to some underlying concept (category), such as different categories of melons can be classified as "light melons", "dark melons", "seeds melons", "seedless melons" and so on.

In recent years, many scholars have applied cluster analysis to synchronic and diachronic studies of semantics and grammar. Different categories of argumentative structure of verbs reflect their unique semantic features. The category differences of near-sense verbs in argumentative structure, such as the number, semantic features and semantic restrictions embodied by argumentative roles, can help us find the difference between near-synonyms. In identifying near-synonyms, we often show the difference between two near-synonyms verbs by listing the different categories of objects to which the verbs govern or apply.

This paper hopes to use semantic clustering to help determine the semantic category of objects in verb-object collocations, so that we can know which specific noun classes can enter this argument position, so that Chinese second language learners can clearly compare which objects can be collocated, which objects can not be collocated, and what semantic classes the collocated objects belong to.

III. CORPUS AND COLLOCATION DATA SET

Firstly, we downloaded all the corpora containing the two keywords '爱惜(aixi) ' and '珍惜(zhenxi)' in the CCL corpus. There were 4897 corpora containing '珍惜(zhenxi) and 840 corpora containing '爱惜(aixi) ' in the CCL corpus.

Secondly, based on the synonym collocation extraction method proposed in this paper, we utilize the dependency parser pyltp to extract collocation instances from the corpus, and construct a "cherish" and "appreciate" collocation database based on authentic corpus data.

Table 2 is a list of verb-object collocation with "珍惜(zhenxi)" in the collocation database and the collocation frequency is greater than or equal to 3.

TABLE II. LISTS THE VERB-OBJECT COLLOCATION RELATIONSHIP WITH "珍惜(ZHENXI)" IN THE DATABASE AND THE COLLOCATION FREQUENCY IS GREATER THAN OR EQUAL TO 3

id	word1	word2	搭配关系	搭配频率
1	珍惜	机会	动宾	310
2	珍惜	生命	动宾	182
3	珍惜	机遇	动宾	131
4	珍惜	友谊	动宾	129
5	珍惜	荣誉	动宾	92
6	珍惜	时间	动宾	91
7	珍惜	成果	动宾	77
8	珍惜	生活	动宾	73
9	珍惜	土地	动宾	63

https://pypi.org/project/pyltp/

10	珍惜	时光	动宾	62
11	珍惜	和平	动宾	54
12	珍惜	资源	动宾	38
13	珍惜	形势	动宾	34
14	珍惜	一切	动宾	33
15	珍惜	感情	动宾	27
16	珍惜	工作	动宾	24
17	珍惜	今天	动宾	24
18	珍惜	权力	动宾	24
19	珍惜	年华	动宾	23
20	珍惜	称号	动宾	20
21	珍惜	良机	动宾	3
22	珍惜	人力	动宾	3

Table 3 is a list of the collocation database that has a neutral collocation relationship with "珍惜(zhenxi)" and the collocation frequency is greater than or equal to 3.

TABLE III. COLLOCATION RELATIONSHIP WITH "珍惜(ZHENXI)" AND THE COLLOCATION FREQUENCY IS GREATER THAN OR EQUAL TO 3 LIST

id	word1	word2	搭配关系	搭配频率
1	要	珍惜	状中	625
2	十分	珍惜	状中	468
3	应该	珍惜	状中	144
4	非常	珍惜	状中	143
5	倍加	珍惜	状中	141
6	不	珍惜	状中	140
7	更加	珍惜	状中	140
8	会	珍惜	状中	132
9	很	珍惜	状中	128
10	对	珍惜	状中	102
11	都	珍惜	状中	90
12	应	珍惜	状中	86
13	更	珍惜	状中	79
14	在	珍惜	状中	79
15	好好	珍惜	状中	71
16	也	珍惜	状中	66
17	应当	珍惜	状中	66
18	格外	珍惜	状中	65
19	因此	珍惜	状中	57
20	必须	珍惜	状中	55
21	所以	珍惜	状中	54
22	特别	珍惜	状中	50
23	加倍	珍惜	状中	45
24	能	珍惜	状中	44
25	将	珍惜	状中	41
26	百倍	珍惜	状中	37
27	最	珍惜	状中	33
28	共同	珍惜	状中	30
29	因为	珍惜	状中	27
30	就	珍惜	状中	25
31	而	珍惜	状中	22
32	如何	珍惜	状中	22
33	一定	珍惜	状中	22
34	却	珍惜	状中	21
35	永远	珍惜	状中	20
36	没有	珍惜	状中	19
37	比较	珍惜	状中	3
38	曾	珍惜	状中	3

Table 4 is a list of subject-predicate collocation relationships with "珍惜(zhenxi)" in the collocation database, and the collocation frequency is greater than or equal to 3.

TABLE IV. COLLOCATION WITH "珍惜(ZHENXI)" IN THE DATABASE AND COLLOCATION FREQUENCY IS GREATER THAN OR EQUAL TO 3

id	word1	word2	搭配关系	搭配频率
1	我们	珍惜	主谓	346
2	我	珍惜	主谓	247
3	他	珍惜	主谓	115
4	你	珍惜	主谓	72
5	人民	珍惜	主谓	71
6	他们	珍惜	主谓	64
7	人	珍惜	主谓	58
8	她	珍惜	主谓	54
9	政府	珍惜	主谓	53
10	双方	珍惜	主谓	46
11	大家	珍惜	主谓	39
12	人们	珍惜	主谓	28
13	你们	珍惜	主谓	23
14	中国	珍惜	主谓	21
15	部门	珍惜	主谓	3
16	大学生	珍惜	主谓	3
17	党委	珍惜	主谓	3

Table 5 is a list of pairs in the collocation database that have a right attachment collocation relationship with "珍惜(zhenxi)" and the collocation frequency is greater than or equal to 3.

TABLE V. LIST OF THE DATABASE THAT HAS A RIGHT ATTACHMENT COLLOCATION RELATIONSHIP WITH "珍惜(ZHENXI)" AND THE COLLOCATION FREQUENCY IS GREATER THAN OR EQUAL TO 3

id	word1	word2	搭配关系	搭配频率
1	珍惜	的	右附加	256
2	珍惜	了	右附加	26

Table 6 is a list of collocation relationships with "珍惜(zhenxi)" in the collocation database and collocation frequency is greater than or equal to 3.

TABLE VI. IN THE DATABASE AND "珍惜(ZHENXI)" HAS A COLLOCATION RELATIONSHIP AND COLLOCATION FREQUENCY IS GREATER THAN OR EQUAL TO 3 LIST

id	word1	word2	搭配关系	搭配频率
1	尊重	珍惜	并列	22

Table 7 is a list of "爱惜(aixi)" collocation frequency greater than or equal to 3 in the collocation database.

TABLE VII. LIST OF MATCHES WITH "爱惜(AIXI)" IN THE DATABASE WITH A FREQUENCY GREATER THAN OR EQUAL TO 3

id	word1	word2	搭配关系	搭配频率
1	爱惜	的	右附加	80
2	要	爱惜	状中	51
3	爱惜	人才	动宾	46
4	他	爱惜	主谓	35
5	爱惜	生命	动宾	34
6	爱惜	自己	动宾	29
7	爱惜	羽毛	动宾	27
8	我	爱惜	主谓	26
9	很	爱惜	状中	23
10	也	爱惜	状中	22
11	爱惜	身体	动宾	21
12	爱惜	东西	动宾	19
13	不	爱惜	状中	19
14	爱惜	财产	动宾	3
15	爱惜	公物	动宾	3
16	爱惜	皮毛	动宾	3
17	爱惜	人力	动宾	3

IV. THE ACQUISITION OF DISCRIMINATION KNOWLEDGE OF NEAR-SYNONYMS -- TAKING '爱惜(AIXI)' AND '珍惜(ZHENXI)'AS EXAMPLES

Before discriminating near-synonyms, we should first look for the "similarity ", find out the situations in which near-synonyms are interchangeable, and summarize them from semantic, syntactic or pragmatic perspectives. Then look for "difference" (especially "slight difference"), "difference" must not be interchangeable (Chu, 2021).

A. Words that can be used interchangeably with '爱惜(aixi)' AND '珍惜(zhenxi)'

As shown in Table 8, through comparative analysis, we find that there are both abstract nouns and concrete nouns in the category of words that can be collocated with '爱惜(aixi)' and '珍惜(zhenxi)'.

TABLE VIII. THE SEMANTIC CLASS OF WORDS THAT CAN BE COMBINED WITH '爱惜(AIXI)' AND '珍惜(ZHENXI)'

宾语	与爱惜搭配	与珍惜搭配	北大语义分类体系	HOWNET 语义分类体系
人才	√	√	/	实体-万物-物质-群体
时间	√	√	时间	实体-时间
粮食	√	√	具体事物-非生物-人工物-食物	实体-万物-物质-无生物-人工物-食物

B. Cases of words that can only be combined with "爱惜(aixi)"

Table 9 is a list of words that can only be combined with "爱惜(aixi)".

TABLE IX. THE SEMANTIC CLASS OF WORDS THAT CAN ONLY BE PAIRED WITH "爱惜(AIXI)"

宾语	跟爱惜搭配	跟珍惜搭配	北大语义分类体系	HowNet 语义分类体系
东西	√	×	具体事物-非生物-人工物-器具	实体-万物-物质-无生物-人工物-器皿
公物	√	×	具体事物-非生物-人工物-器具	实体-万物-物质-无生物-人工物-器皿
财产	√	×	具体事物-非生物-人工物-钱财	实体-万物-物质-无生物-人工物-钱财
衣服	√	×	具体事物-非生物-人工物-衣物	实体-万物-物质-无生物-人工物-钱财-衣物
羽毛	√	×	具体事物-构件-身体构件	实体-部分-部件

嗓子	√	×	具体事物-构件-身体构件	实体-部分-部件

Table 10 shows the semantic classes of words that can be combined with "爱惜(aixi)".

TABLE X. SEMANTIC CATEGORIES OF WORDS THAT CAN BE COMBINED WITH "爱惜(AIXI)"

词语类别	可以与“爱惜”搭配	实例
动物、植物类的词	√	爱惜千里马、爱惜小树苗
描述人类的词	×	爱惜同学、爱惜孩子
身体构件类的词	√	爱惜眼睛、爱惜嗓子
具体物品类的词	√	爱惜衣服、爱惜公物
时间类的词	√	爱惜时间
情感类的词	×	*爱惜亲情、*爱惜友情

Through the data shown in Table 9 and Table 10, it can be found that:

1) "爱惜(aixi)" can be collocated with the words of animals and plants, such as "爱惜千里马" and "爱惜小树苗", but "珍惜(zhenxi)" can not be collocated with the words of animals and plants;

2) "爱惜(aixi)" can be collocated with the words of body components, such as "爱惜眼睛" and "爱惜嗓子", while *" 珍惜(zhenxi) "* cannot be collocated with the words of body components;

3) The collocation object of *"爱惜(aixi)"* tends to be the actual specific items, such as "things, public property, clothes" and so on;

C. Cases of words that can only be combined with "珍惜(zhenxi)"

Table 11 is a list of words that can only be combined with "treasure". Table 12 shows the semantic classes of words that can be combined with "珍惜(zhenxi)".

TABLE XI. A SEMANTIC CLASS OF WORDS THAT CAN ONLY BE PAIRED WITH "珍惜(ZHENXI)"

宾语	跟爱惜搭配	跟珍惜搭配	北大语义分类体系	HowNet 语义分类体系
良机	×	√	/	/
友谊	×	√	抽象事物-心理特征-情感	实体-万物-精神-情感
感情	×	√	抽象事物-心理特征-情感	实体-万物-精神-情感
机遇	×	√	/	/
年华	×	√	时间?	实体-时间?
和平	×	√	/	/

TABLE XII. A SEMANTIC CLASS OF WORDS THAT CAN ONLY BE PAIRED WITH "珍惜(ZHENXI)"

词语类别	可以与“珍惜”搭配	实例
动物、植物类的词	×	*珍惜千里马、*珍惜小树苗
描述人类的词	×	*珍惜同学、*珍惜孩子
身体构件类的词	×	*珍惜眼睛、*珍惜嗓子
具体物品类的词	×	*珍惜衣服、*珍惜公物
情感类的词	√	珍惜亲情、珍惜友情
时间类的词	√	珍惜时间、珍惜时光

Through the data shown in Table 9 and Table 10, it can be found that:

1) "珍惜(zhenxi)" can be collocated with emotional words, such as "珍惜亲情" and "珍惜友情", "爱惜(aixi)" can not be collocated with emotional words;

2) The object of *"珍惜(zhenxi)"* tends to be spiritual and other abstract words, such as "friendship, affection, opportunity, peace" and so on;

3) The objects dominated by *" 珍惜 (zhenxi)"* are generally things that are easy to "disappear", such as "good opportunities", "friendship" and "opportunities".

V. VISUALIZATION

After obtaining the collocation information of near-synonyms, we can use the visualization technology to show the collocation relationship of near-synonyms in the vocabulary network, and help learners to summarize the usage rules of near-synonyms.

According to the different categories of collocation words, we can present collocation words from two perspectives:

(1) Present collocation words according to the categories of syntactic distribution;

(2) Collocation words are presented according to the semantic category to which the words belong.

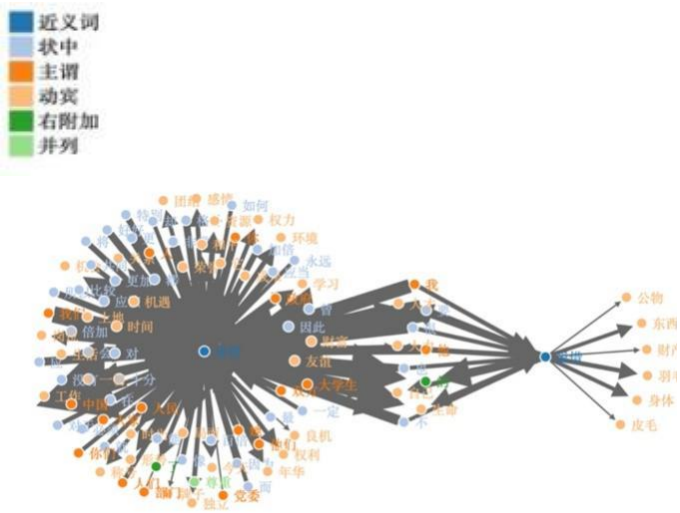

Fig. 2. The top 100 cases of collocation entries about '爱惜(aixi)' AND '珍惜(zhenxi)' currently included in the database

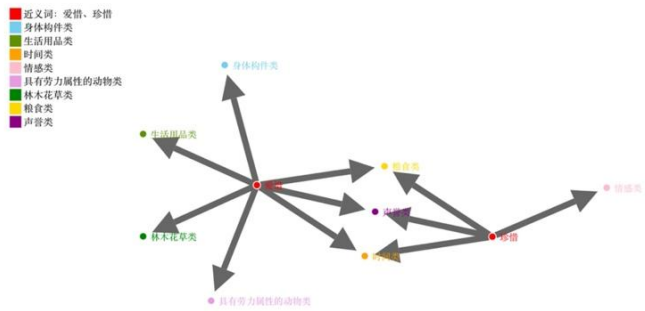

Fig. 3. Network of '爱惜(aixi)' AND '珍惜(zhenxi)' according to semantic categories

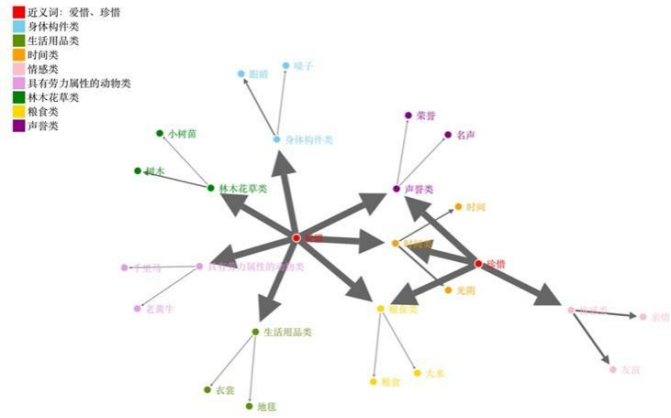

Fig. 4. Network of '爱惜(aixi) AND '珍惜(zhenxi)' according to semantic categories

From these figures, we can observe that:

(1)The objects collocated by '珍惜(zhenxi)' have the characteristics of "disappearing" easily, such as "feelings" will "disappear";

(2) The object collocated by '爱惜(aixi) is characterized by its susceptibility to "damage", such as "small saplings", due to its fragile nature. Therefore, it is essential to cherish it.

VI. Framework of Exploring the Differentiation of Near-Synonyms

In this paper, the analysis mode of 'Nuance of Dimensionality + Hierarchy of Words' is extended to form an automatic process of obtaining resources for discrimination of near-synonyms:

(1) Word Selection: Selecting near-synonyms that are likely to be confused by Chinese second language learners;

(2) The dependency parser is employed to extract the primary collocations of near-synonyms through a collaborative approach between humans and machines;

(3) Highlight the nuance among near-synonyms by investigating the semantic category distinctions of their collocations;

(4) Visualizing the differences in syntactic and semantic usage of near-synonyms.

VII. Conclusions

In order to help Chinese second language learners grasp the nuance of near-synonyms and construct lexical network knowledge related to near-synonyms more conveniently and clearly, this paper proposes a framework for obtaining near-synonym discrimination resources based on the technologies of automatic collocation extraction and semantic clustering, we also use visualization technology to associate the differences in syntactic and semantic usage of near-synonyms, so as to realize visual representation of knowledge of near-synonyms discrimination, which is more friendly to Chinese second language learners.

Acknowledgment

I would like to express my sincere gratitude to all those who have contributed to this paper. I am particularly indebted to my postdoctoral supervisor, Professor Yanbin Diao, for his invaluable feedback and meticulous revisions to the manuscript. I am also deeply grateful to my doctoral supervisor, Professor Weidong Zhan, for his continuous guidance and support throughout the course of this study.

References

[1] Chu Zexiang, Liu Qi. On the Practicality of Synonym Differentiation. Applied Linguistics, 2021(04): 82-92.

[2] Fu Huaiqing. Some Issues in Synonym Research. Chinese Language,2000(03):221-227+286.

[3] Fu Huaiqing. "Word Meaning Component-Pattern" Analysis (Adjective Words). Chinese Language Learning,1997(03):31-35.

[4] Fu Huaiqing. "Word Meaning Component-Pattern" Analysis (Noun Words). Chinese Language Learning,1997(01):24-28.

[5] Fu Huaiqing. "Word Meaning Component-Pattern" Analysis (Verb Words). Chinese Language Learning,1996(05):3-9.

[6] Fu Huaiqing. Analysis and Description of Word Meanings. Beijing: Foreign Language Teaching and Research Press,2006

[7] Li Qiang, Yuan Yulin. A Study on Synonym Noun Differentiation Method Based on Object Role. World Chinese Teaching, 2014,28(04):519-531.

[8] Xing Hongbing. Study on Collocation Knowledge and Second Language Vocabulary Acquisition. Application of Language and Text, 2013(04): 117-126.

[9] Xing Hongbing. Extraction of Lexical Knowledge Based on Corpus and The Compilation of Extroverted Dictionary. Lexicography Research, 2013(03): 36-41+94.

[10] Xing Hongbing. A Frequency Dictionary of Verb Usage for Teaching Chinese As A Foreign Language, Proceedings of The International Symposium on Lexicology for Learning Chinese as A Foreign Language.,2005:84-100.

[11] Zhu Dexi. A Grammar Handbook. Beijing: The Commercial Press,1982/2016.