ICIC Express Letters
Part B: Applications
Volume **8**, Number **1**, January **2017**

 ISSN 2185-2766
pp. **185–192**

THE AUTOMATIC EXTRACTION OF COMMON ADJECTIVE BASED ON LARGE SCALE CORPUS

Zhimin Wang[1] and Lucy Zhao[2]

[1]College of Chinese Studies
Beijing Language and Culture University
No. 15, Xueyuan Road, Haidian Dist., Beijing 100083, P. R. China
wangzm000@qq.com

[2]School of East Asian Studies
The University of Sheffield
Western Bank, Sheffield, S10 2TN, UK
Lucy.zhao@sheffield.ac.uk

Received June 2016; accepted September 2016

Abstract. *The paper proposed a model for extracting common adjectives and built a common adjective statistic wordlist for second language teaching, by using the corpus of People Daily (PD) and broadcast television (BTV). In addition, a detailed analysis also led to the stylistic characteristics of various adjectives and their grades. Finally, we analyzed the strength of positivity vs. negativity of adjectives. Our research showed that the frequency of positive words is much higher than that of the negative words among the 16 pairs of adjectives. There are no adjectives staying from 40% to 60%, and the positive and negative words staying in polarization state.*
Keywords: Statistical time span, Distributing feature, Statistical wordlist, Frequency of word

1. **Introduction.** Adjectives, as one of three types of notion words in Chinese, are important for the development of Chinese textbooks and classroom teaching. To the best of our knowledge, there has not been much research done in terms of the criteria of distinguishing between frequently used adjectives. Experts in Chinese teaching and testing currently resort to personal experience and subjective judgement in selecting adjectives appropriate for various levels.

Various syllabus and reference books, depending on their aims include different number of adjectives. For instance, the Syllabus of Graded Words and Characters for Chinese Proficiency (SGWC) and Graded Chinese Syllables, Characters and Words (GCSC) [1], are both reference books for Chinese language teaching. They included 1173 and 1228 adjectives respectively. The Grammatical Knowledge base of Contemporary Chinese, as a knowledge base for computer processing [2], included 3156 adjectives. *Modern Chinese Dictionary* (Fifth Edition) included 5660 adjectives [3], many of which are rarely used in daily life. For teaching, it often requires common adjectives with high coverage, which still need to validate through the large-scale corpus.

Corpus-based studies have been conducted on qualitative adjective and state adjective [3,4]. These studies contribute to the syntactic distribution of adjectives. Nonetheless, to the best of our knowledge, there has not been any computational research on the frequency of common adjectives.

The rapid development of Chinese language teaching calls for reflection on many fundamental questions including the following questions: How many words should the first-year students master? What words should they master? In what sequence should these words be introduced in textbooks? At what rate should the newly introduced words recur? [5]

A possible way of solving these problems is to resort to large-scale corpus to provide an objective standard for selecting the frequently used words.

Thus, this paper suggests that Chinese common adjectives wordlist is to be set up and refine the extraction features of common adjectives. The rest of the paper is organized as follows. In Section 2, we describe the features of Chinese common adjectives and specify the corpus selection. In Section 3, we present model for the common adjectives extraction and show experimental results. In Section 4, we explore the stylistic features of common adjectives. In Section 5, we analyze the sentiment strength of common adjectives. Finally, we conclude with a summary and an outline of further research in Section 6.

2. Feature Extraction for Common Adjectives.

When we judge whether an adjective is frequently used, we not only need to consider its frequency and range, but also need to consider many factors like stability, corpus scale and statistic time span. The variation of a word in a historical period can be signified by its frequency at any given time.

National Language Resources Monitoring and Research Center publishes the statistics tables of vocabulary annually [6,7]. The tables use "year" as a unit of measurement and have a large amount of adjectives. However, we cannot judge directly the frequency of those adjectives only by "year", and those tables cannot be used for teaching. Through the statistic experiments using "day", "month" and "quarter" as the statistical time span, we found there are relatively a large number of adjectives appearing continuously at the node of "quarter". Therefore, this paper takes "quarter" as the feature extraction to create the segment.

The paper selects the corpora from People Daily (PD) and broadcast television (BTV) from 2005 to 2009. Both two kinds of corpora are representatives of written and spoken language embodying the changes of current Chinese language situation. Meanwhile, we specially made an experiment to verify the effect of using PD to extract Chinese word [8].

This paper takes "quarter" as an extraction features and divides corpora in PD or BTV in 5 years into 20 small files. Then it uses the Part-of-Speech Tagging System developed by Institute of Automation of Chinese Academy of Sciences. All adjectives in these 20 files are extracted and the distribution of those adjectives is counted.

The scale of corpora, in PD and BTV from 2005 to 2009, is 812,417,024 bytes and 787,218,432 bytes. After the word segment and tagging, this paper extracted all the data of adjectives at the nodes of "quarter" in two corpora, as shown in Table 1.

From 2005 to 2009, in 40 nodes of two corpora, the quantity of adjectives is between 2175 and 2844. However, if following the standard that the nodes of "quarter" cannot be zero, we found that the total common adjective in PD falls down to 1417, those of BTV

Table 1. The distribution of adjectives in corpora of PD and BTV

	Y	Q1	Q2	Q3	Q4	Qty Adj	Qty Adj
PD	2005	2285	2315	2253	2217	1417	1335
	2006	2227	2264	2302	2210		
	2007	2215	2300	2329	2175		
	2008	2222	2186	2242	2191		
	2009	2187	2256	2359	2355		
BTV	2005	2206	2332	2481	2375	1811	
	2006	2689	2749	2696	2733		
	2007	2783	2844	2794	2766		
	2008	2708	2750	2733	2714		
	2009	2761	2801	2823	2707		

down to 1811, and those occurring in two corpora at the same time down to 1335. By this way, we can effectively filter adjectives which do not usually appear in real corpus.

Meanwhile, according to the statistical result of adjectives in 40 nodes, we find that the total adjective among 20 nodes in BTV is larger than that of PD even though they have similar corpus scale. Adjectives at the nodes of "quarter" are no more than 2359, while that of PD reaches to 2844. This difference may be related to the style of two corpora. As a written media, the narrative style of PD is often rigorous, objective and fair, while most of the BTV programs are interviews and dialogue shows, paying more attention to people's mind and the emotional evaluation, thus reflecting the relaxation, liveliness, humor and exaggeration of the spoken language.

Any adjective appearing in the 20 nodes in this paper will be the entry of the statistical wordlist of common adjectives. The statistical wordlist of common adjectives makes all information at 20 nodes as the attribute field, and it can be drawn as a diachronic curve. Take the word "extraordinary (精彩 in Chinese)" as an example.

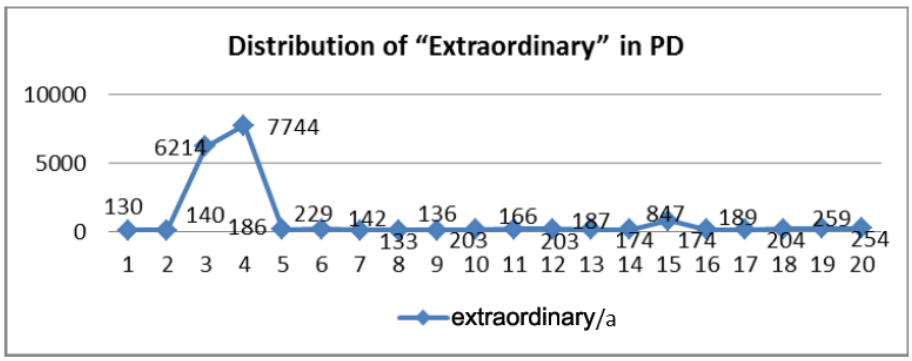

FIGURE 1. The frequency distribution of "extraordinary" in PD

The statistical wordlist records the information of "extraordinary" at the nodes of "quarter". We found that the "extraordinary" is not well-distributed. There are 100 to 300 in 18 quarters, while in the third and the fourth quarter in 2005, the total number reaches to 7744. At this point, both the frequency in the nodes of "quarters" and the average frequency cannot reflect the situation of its use.

3. **The Design of the Statistical Wordlist of Adjectives.** We proposed a model used for common adjective extraction by using the frequency and the stability.

$$U = \frac{\overline{f}}{stdev(f)} \tag{1}$$

In Formula (1), U is the usage degree of words. $\overline{f}$ is the average frequency of adjectives, and the calculation formula is shown as Formula (2); $stdev(f)$ is the standard deviation of adjectives, whose calculation formula is shown as Formula (3)

$$\overline{f} = \frac{f_1 + f_2 + \cdots + f_n}{n} = \frac{\sum f}{n} \tag{2}$$

$$stdev(f) = \sqrt{\frac{\sum (f - \overline{f})^2}{n - 1}} \tag{3}$$

In Formula (2) and Formula (3), n refers to the number of average frequency $\overline{f}$.

In this way, the statistical wordlist of common adjectives not only includes all information at 20 nodes, but adds attribute fields such as [sum], [aver], [stdev] and [U]. Such information is an important criterion to differentiate and analyze synonymous adjectives. For any group of adjectives, its extent of common usage can be obtained through the

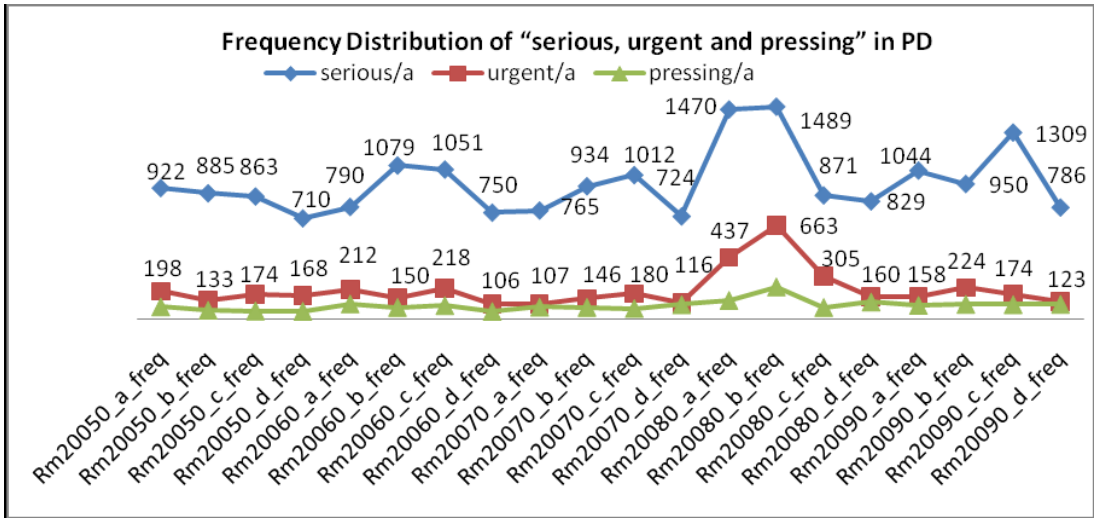

FIGURE 2. Frequency distribution of "serious, urgent and pressing" in PD

information at the nodes of "quarter" in the statistical wordlist. As we often say, when describing an event, "this is a very serious (严重 in Chinese) matter", "this is a very urgent (紧急 in Chinese) matter" or "this is a very pressing (紧迫 in Chinese) matter". And the difference between adjectives "serious, urgent, pressing" can be obtained from the curve charts.

In the above figure, the frequency of "serious" is much higher than that of "urgent", and "pressing", and its distribution is relatively uniform with medium fluctuation range and can be used as the teaching option.

For the synonyms with uneven distribution, frequency of "quarter" serves as a reference, so does other parameters like [U]. For example, in the wordlist, the frequency of the word "alone (孤独 in Chinese)" at the 18 nodes of "quarter" is higher than the word "lonely (孤单 in Chinese)", and it is very stable, as shown in Table 2.

TABLE 2. The quarter distribution of the synonymous adjectives "alone" and "lonely"

Q / W	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
Alone 孤独	25	40	22	27	22	49	30	27	36	42	28	20	19	31	35	16	18	35	26	31
Lonely 孤单	3	5	7424	7631	19	13	6	8	7	9	13	3	11	19	9	4	29	6	10	8

TABLE 3. The [U] analysis of "alone" and "lonely"

The [U] analysis of "Alone" and "Lonely"				
word	**sum**	**aver**	**stdev**	**U**
Alone 孤独	579	28.95	8.65402249033118	3.34526516800075
Lonely 孤单	15237	761.85	2314.05568497344	0.329227168104533

While the frequency of the word "lonely" at nodes of "quarter" is extremely non-uniform, surge from 5 times to over 7,000 times. When checking the original, we have found that there is a newly added column from the third quarter in the corpus which is about loneliness describing that when you feel lonely. This abnormal performance will directly lead to great changes in the [sum] and [aver] of the word "lonely". Therefore, if we judge only according to its [aver], we will get the result that "lonely" is more common than "alone", while if we judge by the [U] model, we will find that "alone" is more common than "lonely", as shown in Table 3. Thus, the [U] model, to some extent, truly reflects

the usage of words, and it does not change the sequence of word in the wordlist because their sudden changed frequency.

4. **The Discrimination of the Stylistic Features of Common Adjectives.** Stylistic features refer to the tendency and characteristics of the words reflected in their meaning. In general, it is divided into two categories: spoken language and written language. Compared with nouns and verbs, the stylistic features of adjectives are difficult to distinguish. The adjective statistical wordlist, based on PD and BTV can provide relevant data for distinguishing. For example,

A: He is very stingy (小气 in Chinese) with himself.

B: He is very mean (吝啬 in Chinese) with himself.

"Stingy" and "mean" are synonyms, and their differences show that "stingy" is oral language, while "mean" is written language. It is found that there is only "mean" in PD, which reflects that "mean" occurs continuously in the written language corpus. However, in BTV, there are both "stingy" and "mean", as shown in Table 4.

Table 4. Distribution of "stingy" and "mean" in BTV corpus

word	**aver**	**sum**	**stdev**	**U**
Stingy 小气	15.1	302	5.79382611515323	2.60622250303773
Mean 吝啬	16.05	321	7.04478156995808	2.27828213559438

Therefore, "stingy" can be basically identified as a spoken word. Most words that occur simultaneously in two corpora can be judged as the spoken words or the written words according to the data. Take the words "pretty (漂亮 in Chinese)" and "beautiful (美丽 in Chinese)" as an example.

Table 5. Comparison of the [U] of synonymous adjective in two corpora

Word of BTV	aver	sum	U	Word of PD	aver	sum	U
pretty 漂亮	764.35	15287	2.614980968	beautiful 美丽	194.6	3892	3.525696063
beautiful 美丽	536.15	10723	2.466663707	pretty 漂亮	84.7	1694	3.271706807
nice 好看	300.2	6004	2.404573588	nice 好看	30.9	618	2.88752527
beautiful 靓丽	42.85	857	1.41781951	beautiful 靓丽	7.85	157	2.717461088

From the wordlist based on BTV and PD, we can find that the sequences of the word "pretty (漂亮)" and the word "beautiful (美丽)" are different in BTV and PD. This can well prove that the spoken tendency of the word "pretty (漂亮)" is stronger, while the written tendency of the word "beautiful (美丽)" stronger.

5. **The Sentiment Strength Analysis of Common Adjectives.** As mentioned above, the statistical wordlist has 1811 and 1417 entries from two kinds of different corpora. We use features of words to try to rank all adjectives in the statistical wordlist. The higher the rank of a word is, the higher its frequency is, and the more important it is in corpora. Inspired by that, we design a proportion model as follows.

$$\text{Rate} = [\text{Sum}]/[\text{Total Frequency}]$$

[Sum] refers to the frequency sum of each adjective that occurs at the 20 nodes of "quarter". [Total Frequency] refers to the total frequency of all adjectives in the wordlist. It is a constant in the given sample.

Through the investigation, we found that the distribution of adjectives is pyramid. The higher proportion is, the smaller the number is. The statistical wordlists in two corpora have different amounts, so their rankings are also different in Table 6. The top ten words have 30% share, and it is called the first level. If the share reaches to over 50%, there

Table 6. The proportion distribution of adjectives in statistical wordlists

BTV			PD		
The level of W	Num	Freq Prop	The level of W	Number	Freq Pro
Level 1	10	30.63%	Level 1	10	30.78%
Level 1, 2	51	50.01%	Level 1, 2	39	50.33%
Level 1, 2, 3	552	90.02%	Level 1, 2, 3	422	90.02%
Level 1, 2, 3, 4	846	95.00%	Level 1, 2, 3, 4	665	95.01%
Level 1, 2, 3, 4, 5	1811	100%	Level 1, 2, 3, 4, 5	1417	100%

should be 51 words in BTV and 39 in PD, and among them, there includes the previous ten words so it is called the first and the second level. If the share reaches to over 90%, there should be 552 entries in BTV and 442 in PD. In this way, we got 5 levels according to proportion, and each level includes the words in the last level.

The first level has the least amount but the largest proportion. The first level adjectives in BTV are "big, good, manynew, small, high, important, same, old and long", while those in PD are "new, big, important, good, many, high, small, significant, harmonious and basic", as shown in Table 7.

Table 7. The first level adjectives in statistical wordlist of BTV and PD

BTV Wordlist			**PD Wordlist**		
W	**sum**	**rate**	**W**	**sum**	**rate**
big 大	518375	7.063713572223	new 新	195771	6.39656403311153
good 好	416764	5.67909625891285	big 大	183210	5.98614961616564
many 多	281765	3.83951242763909	important 重要	114933	3.75528701399905
new 新	243779	3.32189058292347	good 好	101851	3.32784959639805
small 小	240881	3.28240055749342	many 多	90533	2.95804859560245
high 高	163882	2.23316230073412	high 高	65172	2.12941074605506
important 重要	134129	1.82772864765604	small 小	58438	1.90938601205986
same 一样	88020	1.19941754256488	significant 重大	47917	1.56562595468484
old 老	81358	1.10863681467841	harmonious 和谐	47511	1.55236043018201
long 长	79157	1.07864456279037	basic 基本	36648	1.19742596546716

Most of the ten words are single syllable adjectives, and only 4 are two syllable form. They are basically positive vocabulary from the sentiment tendency. The proportion reflects the importance of the words in the wordlist. 10 words have 30% share of the total frequency, which reflects their frequent degree and importance, so they should be arranged in the primary stage of acquisition and can be written in the textbooks.

If the proportion reached to 50%, the statistical wordlist of BTV has 51 entries, and they are "big, good, many, new, small, high, important, same, old, long, different, real, serious, low, close, fast, few, strong, normal, hard, general, successful, great, not good, common, short, the highest, secure, basic, simple, clear, specific, happy, special, early, healthy, obvious, harmonious, giant, complete, black, the latest, effective, particular, nervous, difficult, easy, not bad, finefar and impossible".

The statistical wordlist of PD has 39 entries, and they are "new, big, important, good, many, high, small, significant, harmonious, basic, close, long, different, friendly, fine, strong, low, fast, advanced, outstanding, short, effective, grand, old, positive, serious, excellent, wonderful, giant, successful, specific, stable, complete, fundamental, the highest, wide, lonely, rich and difficult".

Adjectives in two corpora overlap each other. When studying Chinese, the foreign students can integrate the words form these two corpora. Chinese teachers can design their own contents for the adjective teaching according to their own needs, and can integrate

the words from 5 levels into teaching separately. Especially, the top 3 levels should be the focus of Chinese teaching.

The evaluation feature of words plays a very important role in discriminating the meaning of words. When judging two terms of meaning of "doubt ①" and"doubt ②", he found that the substantial clue to discriminate these two meanings in real corpora is the evaluation feature of the objectives (positive vs. negative) [9].

Emotional evaluation is one of the basic contents of adjective teaching, because in adjective teaching, there often involves the examples of positive and negative adjectives. For pairs of antonyms, it needs further study to figure out which are used more frequently, positive ones or negative ones, as it relates to which should be taught first. We found that the frequency of positive words is far higher than that of negative ones. Taking the BTV corpus as an example, we investigated 16 pairs of adjectives which are "big and small, good and bad, many and few, long and short, tall and short, new and old, true and false, far and close, fast and slow, strong and weak, happy and sad, safe and dangerous, clear and faint, simple and complex, exciting and boring as well as beautiful and ugly". Those phrases are mostly from the first and the second level designed by this paper.

For any pair of antonyms, we see two words as a whole which is equivalent to the positive and negative poles of power. The pointer will turn to the one that is used more commonly. We define the sentiment strength of the positive and negative adjectives as follows.

$$\text{Sentiment strength} = \text{sum}([\text{negative}|\text{positive}])/\text{sum}([\text{positive}]) + \text{sum}([\text{negative}])$$

In this way, we obtain the ratio of these 16 pairs of words, and find the interesting phenomena which is shown as Figure 3.

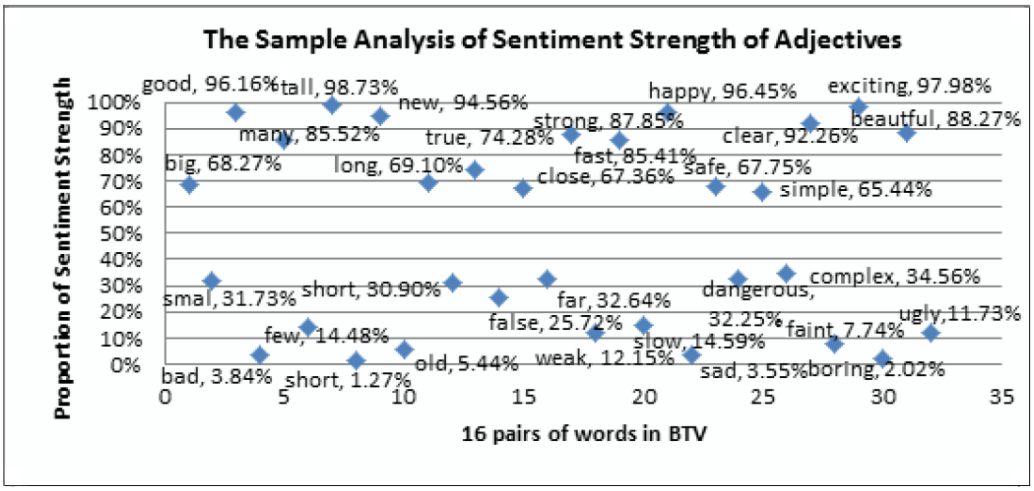

FIGURE 3. The sample analysis of sentiment strength of adjectives

From the above statistical results, we can see that the frequency of positive words is much higher than that of the negative words among the 16 pairs of adjectives. There are no adjectives staying from 40% to 60%, and the positive and negative words staying in polarization state. The frequency of positive words is significantly high, which indicates that people show strong preference to use positive adjectives, and the usage of adjective tends to be positive.

The trend of Chinese adjective development is that along with the improvement of people's cultural quality and linguistic attainment, they tend to use adjectives that can make others comfortable so as to communicate with them. Those kind and positive adjectives are very convenient and have a very good effect. People do not want to use negative and passive words. However, if speakers want to make jokes, show humor or

mock himself, they would use negative words. For example, "I am so embarrassed (狼狈 in Chinese)." or "I am very tired (疲惫 in Chinese)."

At the same time, we also found that in these 16 pairs of words, most of them express people's subjective evaluation, and positive words have very strong sentiment tendencies, having the highest strength ratio, and some of them reach to over 95%; while negative words have the lowest strength ratio, less than 5%, for example, good, bad, happy, sad, exciting and boring. The positive and negative adjectives which express the quality of objective things have a relatively weak sentimental difference, for example, big, small, long, short, true, false, close, far, simple and complex. If those words are distributed according to the average strength of positive and negative words, the average strength ratio of positive words/negative words is 8.3/1.7.

6. **Conclusions.** The paper proposed a model used for common adjective extraction and built a statistical wordlist for Chinese common adjectives. Meanwhile, through the sentiment strength analysis towards positive and negative adjectives, we find that the usage of adjectives tends to be positive. For the future work, we consider to examine the word extraction method for other part of speech in Chinese. Meanwhile, future studies will also expand the influence factors, so as to improve the results. The extraction method can offer some useful inspiration for named entity recognition and Chinese sentiment analysis.

Acknowledgment. The work was supported by the National Natural Science Foundation of China (No. 61170163); the Support Program of Young and Middle-aged Backbone Teachers for Beijing Language and Culture University; Funding Project of Education Ministry for Development of Liberal Arts and Social Sciences (16YJA740036); Wu Tong Innovation Platform of Beijing Language and Culture University (supported by "the Fundamental Research Funds for the Central Universities (16PT03, 14YJ160502)). The authors also gratefully acknowledge the helpful comments and suggestions of the reviewers, which have improved the presentation.

REFERENCES

[1] Y. Liu and J. Ma, The development of the graded Chinese syllables, characters and words: Exploring the new prospective of global Chinese education, *Chinese Teaching in the World*, no.1, 2010.
[2] S. Yu et al., *Introduction to Grammatical Knowledge Base of Contemporary Chinese*, Tsinghua University Press, 1998.
[3] P. Zou, *Study on Collocated Examples of Adjectives in the 5th Edition of Modern Chinese Dictionary*, Master Thesis, Sichuan International Studies University, 2012.
[4] G. Zhang, Typical feature of modern Chinese adjectives, *Zhongguo Yuwen*, no.5, 2000.
[5] J. Lu, Researches of applied linguistics in 21 century of China: Three key domains, *Zhongguo Yuwen*, no.6, 2000.
[6] National Language Resources Monitoring and Research Center, *Language Situation in China*, 2006.
[7] National Language Resources Monitoring and Research Center, *Language Situation in China*, 2010.
[8] Z. Wang and E. Yang, Chinese-teaching-oriented quantitative study of Chinese common verbs, *Journal of Language Teaching and Linguistic Studies*, no.1, 2012.
[9] Y. Yuan, On the sense extension mechanism and semantic construal strategy of the Chinese verb Huaiyi (怀疑), *Studies in Language and Linguistics*, no.7, 2014.
[10] W. Zhan, A review on word classification of Chinese from the perspective of sentence parsing by computer, *Zhongguo Yuwen*, no.2, 2013.
[11] M. Lewicka, J. Czapinski and G. Peeters, Positive-negative asymmetry or "When the heart needs a reason", *European Journal of Social Psychology*, vol.22, no.5, pp.425-434, 1992.
[12] J. Boucher and C. E. Osgood, The pollyanna hypothesis, *Journal of Verbal Behavior*, vol.8, pp.1-8, 1969.