International Journal of Knowledge and Language Processing
Volume **2**, Number 3, July **2011**

www.ijklp.org
KLP International ©2011 ISSN 2191-2734
pp. 14-20

Recognition of the Metaphoric Neologisms

LI Hui [1], Feng Minxuan [2]

[1] School of Chinese Language and Literature
Tsinghua University
Beijing, 100084, China
lh9743@126.com

[2]School of Chinese Language and Literature
Nanjing Normal University
Nanjing, 210046, China
fennel_2006@163.com

Received June 2011; revised July 2011

ABSTRACT. *Metaphoric neologism refers to units of phonetic symbols or characters used to represent new meanings by metaphor. In recent years, neologisms appear in large numbers as a linguistic phenomenon and play a significant role in social media, which attracts our attention. The intention of this paper is to analyze the structure of the metaphoric neologisms and their semantic characteristics. Experiments on the corpus constructed from Sohu website are made in order to build rules for recognition of unknown words. Morphological Productivity and Mutual Information are used as tools of statistics, achieving an encouraging ratio of precision and recall in the experiments of neologism identification.*

Keywords: neologism, metaphor, recognition

1. Introduction. In 2011, "the top 10 neologisms of the year 2010" were published by Baidu website and Southern Weekly as follows — "甲型 H1N1 流感(Swine Flu)、躲猫猫(Hide and Seek)、钓鱼执法(entrapment)、临时性(temporary)、被时代(times of by)、蜗居族(Dwelling Narrowness)、压力差(Differential Pressure)、70 码(70 yards)、穷二代(the Poor Successors)、翻墙(scale the wall)". Presently, neologisms are increasingly common in online chats, forum talks, daily communications, etc. New words or phrases are innovated, borrowed from other dialects/ classical Chinese/ foreign languages/ jargons, or updated with novel meanings.

Previous studies laid emphasis on the relation between literal senses and metaphoric senses in the changes of lexical senses. Zhu (2008) confirms that "Metaphoric Projection from the source domain/specific domain to the target domain/abstract domain is an effective channel to the burning of new meaning." Owning to the metaphor, people take advantage of concrete objects to describe abstract ideas, which links a connection among different concepts. Despite recent progress reviewed in literatures, there is no generally definition concerning Metaphoric Neologism (MN). In our opinion, MN is a structure with two or three syllables, which not only illustrates the concise linguistic features, but also gives a vivid mirror of novelty chasing and jokes of social changes.

Our study of the MN is aimed at the identification of unknown words. According to the "Language Situation in China" (2006) issued by the Ministry of Education in China, there are 171 neologisms listed in the document, such as “奔奔族(rushing clan)、独二代(Alone II)、国际高考移民(NCEE Immigrants)etc. In the examples, the character “代” in 独二代 originally means a generation, while in the phrases such as “富二代/穷二代/官二代", it refers to the generations with negative connotation. This paper is making efforts to identify possible patterns in these neologisms and turn these patterns to rules which are then used to recognize them.

Our paper is divided into four main sections. The first section deals with corpus data and focuses on semantic analysis. Section 2 is the experiment of recognition of MN with combination of rules and statistics. Illustrations of results and problems are presented in the third section. The last section is about the conclusion and future steps.

2. **Corpus Analysis.** The following resources are constructed and used in the study:

Closed Corpus:

Sogou Cell Corpus, 2.3M. 86,076 items of words
Two years(2007-2009) popular neologisms.
Covering topics such as entertainment, sports, arts, medicine, natural science, humanities, social science and almost all aspects of daily life.

Training Corpus:

half-year (09.2009-02.2010) news from Sohu webpage, 433k

Language Tools:

ICTCLAS (Institute of Computing Technology, Chinese Lexical Analysis System) 1.0 version, which is functioned as a system of word segmentation and POS tagging.
Segmentation Speed per Computer: 996 KB/s
Accuracy of Segmentation: 98.45%

2.1. **Word Family.** The core words of the neologisms accelerate the propagation of new word family. Song (2007) also argues that “neologisms with two or more syllables often have a central character and accumulate into a family with parallel membership”. People use analogy as an important mechanism to expand their vocabularies. In other words, semantic features combined with word family could improve the strength of neologism. From the data collected, we found phrases listed in Figure 1, the examples of which are given below.

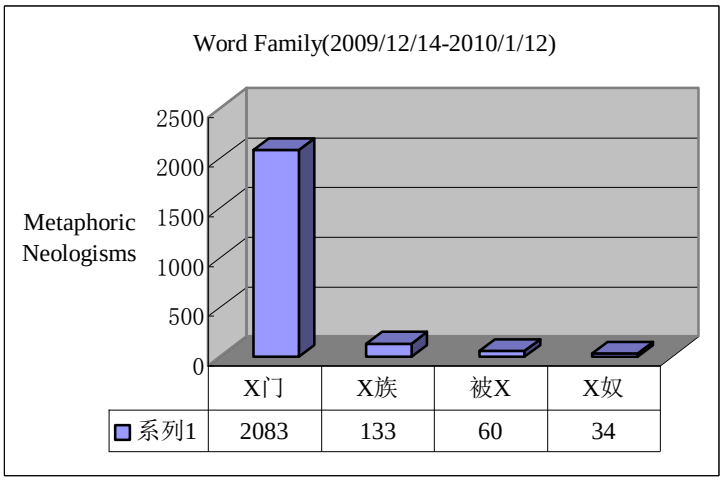

FIGURE 1. Word Family

2.1.1. *族 (Bird People): Originally the morpheme is used to refer to people with same origin or genes, now shifted to people with similar living attitude or particular personal needs,. Such as "啃老族" (Neet), which refers to some young people who do not work but live off their parents. "装嫩族" (Grups), refers to people who are in their 30s or 40s but act like they're in their 20s.*

2.1.2. *被(Particle or Preposition): Now contains more deprecating of reality in the phrase like "被就业(Being Employed)/被增长(Being Increased)", which clearly conveys the negative emotion in society.*

2.1.3. *奴(Slave): Now pointed to people thrive for living. Take "房奴"(Housing Slave) as an example, in the reality slave-owners should be human beings, however, in this phrase, the non-living property becomes the owner while people become the slaves and hard-earned money is filled into the hole of housing loaning. Although these people might have big apartments in urban areas, they can hardly achieve the matching living standard.*

2.2. **Euphemism.** The occurrences of Euphemism in the neologisms are concern us. People use euphemism to express their thoughts more elegantly and suggest their dissatisfaction indirectly. During the process of creation, people not only receive the acknowledgement of social participation, but also the great entertainment, which is one of the reasons why neologisms are widely accepted. Cases in point can be seen in the following examples 2.2.1. *剩女(Leftover Girls):Women who are highly educated with successful career* but remain single *after the best getting-married ages).*

2.2.1. *杯具(Tragedy): Usually used to refer to containers such as cup, glass, or bottle etc, but currently also used by young people to describe unhappiness, failure, or dissatisfaction in a strong flavor of banter due to the fact that it is a homophony of the word "悲剧".*

3. Identification. In the decades, Chinese information processing is still facing the difficult problem – recognition of unknown words. No clear advancement has so far been seen in the area of identification of fast-emerging MN. In this paper, we try to combine rules and statistics methods together to identify MNs.

3.1 Statistics.

3.1.1. Morphological Productivity. In MN, some Chinese characters always have specific positions in phrases, like "族" usually appears in the end, and "被" often in the prefix. Therefore we introduce the Morphological Productivity (MP) and the formula is quoted as follows:

$$MP=n_1/\mathrm{N} \quad (1)$$

Where N denotes the number of characters in a particular category in specific texts; n_1 is the amount of one-time appearance word forms of the set of characters. Qin (2004) pointed out that words with relatively high MP have a tendency to produce neologism. For instance, " X 虫" presents 200 times in the texts but the one-time occurrence is only 10, so the rate is :10/200=5%. In our test, n1 is extended to the quantity of form of one-time appearance of MN.

3.1.2. Mutual Information. Mutual Information (MI) is used to describe the interdependent relations of two words. When MI is greater than a threshold value, the words can be judged as MN.

$$MI(w_1,w_2)=lgP(w_1,w_2)/(P(w_1)*P(w_2))=lg\ \ (f(w_1,w_2)/(f(w_1)*f(w_2)) \quad (2)$$

$P(w_1,w_2)$is the probability of occurrence of two adjacent characters w_1 and w_2 in the text. P (w_1) is the probability of occurrences of w_1, and $p(w_2)$ is the probability of occurrences of w_2. The greater the value of MI, the stronger tightness; the smaller the value of MI, the weaker tightness.

3.2. Rules. A,B,C,D represent four random Chinese characters, the two-character word string can be expressed as AB, three as ABC. As MN mainly appears in triple string, so the test string is chosen as ABC, here are the algorithms:

Rule 3.2.1. X 族： while(C== '族') {
if ((tagnum(AB)==1) new(ABC)=2;
else {if (wordtype (A)=='N' && (wordtype(B)=='N' || wordtype(B)=='V' || wordtype(B)== 'A')) new (ABC)=1;
else if(wordtype(A)=='V' && (wordtype(B)=='N' || wordtype(B)=='V' || wordtype(B)=='A')) new (ABC)=1; }}

Indication 3.2.1. : If AB has only one tag then mark it for later artificial proofread; if A is a noun and B is a noun/verb/adjective, or A is a verb B is a noun/verb/adjective and then mark the phrases as a MN.

Rule 3.2.2.被 X：while（A==’被’）{
if ((tagnum(BC)==1) new(ABC)=2;
else if (wordtype （B）==‘V‘ && (wordtype(C) == ‘N’ || wordtype (C) == ‘V’ || wordtype (C) ==’A’)) new(ABC)=1;}

Indication 3.2.2.: If BC has only one tag then mark it for later artificial proofread; if B is a verb, C is a noun or verb or adjective, then mark the phrases as a MN.

Besides these available rules and calculation, filter is also indispensable for recognition of words such as “防盗门”, “走后门”, which conforms to the rules but not belongs to the neologism. We develop a list of words and phrases in the Modern Chinese Dictionary (5th edition), to filter nine MN families. If a phrase matches with the phrase in the list, then it is not a MN.

3.3 Results. Precision (P), recall(R), and F-Measure are adopted to evaluate the performance of experiments.

P= correct MN/all recognitions
R=correct MN/existed MN in the corpus
F= 2*P*R/ (P+R)

After labeling-out and word-extraction, the corpus consists of 728,472 words, ICTCLAS is applied into segmentation and part of speech tagging, and then filtering is implemented based on JAVA regular expression, rules related and artificial proofread, part of the results are shown in the table 1 below(*word is unrecognized new word, MP is 0.2, threshold of MI is 2.50).

TABLE 1. Part of the Results

X族	Freq	X门	Freq	被X	Freq	X奴	Freq
奔奔族	41	相声门	16	被残疾	3	房奴	3
快闪族*	9	电话门	14	被增长	1	车奴	1
啃老族	3	代言门	4	被代表*	1	卡奴	1
钟摆一族*	2	间谍门*	3	被就业	1	…	…
蚁族*	2	芯片门	2	被被捕*	1	…	…
80后一族*	1	牌坊门	1	被中产*	1	…	…
月光族	2	窃听门	1	被被了*	1	…	…
房奴一族*	1	…	…	…	…	…	…
游牧族*	1	…	...	…	…	…	…
暴走族*	1	…	…	…	…	…	…
布波族	1	…	…	…	…	…	…
sum	64		130		351		6

	X族	X门	被X	X奴
MP	0.078125	0.253846	0.931624	0.50
R	70.31%（45/64）	92.68%（38/41）	55.56%（5/9）	100%（5/5）

	甲流	杯具	洗具	剩女
MI	3.9945	4.2164	3.2463	3.1137

Seemly, the MP value is mostly higher than the threshold except X, and the value of MI shows the tightness of euphemism. Then it is the evaluation results:

TABLE 2. Results

Correct recognitions	**All recognitions**	**existed neologism in the corpus**	**P**	**R**	**F**
97	155	118	62.58%	82.20%	71.06%

The F value is only 71.06%, so we introduce the filter, here is a list of "x 门".

TABLE 3. x 门

X门	**X门**
车/院/寝室/…门	串门
冷/热门	师门
开/关门	出远门
国门	罗生门
石/铁/钢筋…门	走/开后门
东/西/南/北门	防盗门
掌门（人）	X部门
…	…

Eventually, R value is improved over 85%, which is quite satisfying.

TABLE 4. Results after Filter

Correct recognition	**all recognition**	**existed neologism in the corpus**	**P**	**R**	**F**
97	109	118	88.99%	82.20%	85.46%

4. Conclusion. As a fresh and humorpus figure of speech, WNs are very popular in our society and play a very important role in daily communication. In this work, we analyze the linguistic features of MN, with rules and calculation to make small-scale tests on the corpus. The results are in our expectation. But there are some problems which should not be ignored.. First, the MP value of X 族 is lower than expectation, but "奔奔族" appears 41 times, which accounts for 64.06% of "X 族". Evidently, all the occurrences of "x 族" belong to the MN, so in fact the productivity of "x 族" is not that low.

So we use Baidu to obtain search results for the searching string "x 族", and conduct a survey on the top 10 pages, The result proves that phrases like "恐归族/上班族/蚁族/淘乐族" occur two or three times, the rest phrases in the family appear only once. And the MP value achieves over 80%. Thus by enlarging the scope of corpus, news, the MP value can be significantly increased.

Moreover, a few MNs, like "X 一族" and "被被 X" were not recognized. More specific

rules are needed. Last but not least, some homophonic words cannot be dealt with in our experiments. In the next step, we should combine the semantics with phonetics. A database can be built with records of pronunciations of Chinese characters. If we want to query the synonymous terms "囧" and "窘", which also have the same pronunciation, the same phonetic transcription will be extracted and marked.

Acknowledgement. This work is supported by the funds of social science of Jiangsu Province（10YYB007）and Nation（10CYY021）.

REFERENCES

[1] Dai Shuaixiang, Zhou Changle, HuangXiaoxi, Yangyun, Wang Xuemei. Computational Model of Metaphor and its Application in Metaphorical Classification [A]. *Computer Science*. 2005, (32).

[2] Han Yan, Yao Jianmin, Zhu Qiaoming, Zhang Jing. Study on Chinese OOV Identification without Domain Restriction [A]. *Journal of Zhengzhou University(Nat. Sci. Ed.)*. 2008, (9).

[3] He Min, Gong Caichun, Zhang Huaping, et al. Method of New Word Identification based on large—scale Corpus. *Computer Engineering and Applications*, 2007, 43(21).

[4] Jiang Aoshuang. The Features and Patterns of New Metaphoric Words [A]. *Journal of Honghe University*. 2004, (12).

[5] Lin Ling. How Chinese New Words Prevailing in Internet Survive [A]. *Journal of Chengdu University(Social Science)*. 2008, (2).

[6] Liu XiaoMei. Study on Neologism of Mandarin of the time[D]. *Doctoral Dissertation of Xiamen University*. 2003, (6).

[7] Qin Haowei, Bu Fenglin. Research on a Feature of Chinese New Word Identification [A]. *Computer Engineering*. 2004, (12).

[8] Song Zuoyan. Character-formalization and the Extraction of Unlisted Words[A]. *Journal of Peking n University(Philosophy and Social Sciences)*. 2007, (3).

[9] Su Donghua. The Study on the Rhetoric Lexicalization of New Words [A]. *Dissertation of Master Degree of Jinan University.* 2006, (6).

[10] Zhang Yisheng , Xu Xinyuan. A Tentative Analysis of the Word Family X Ke: A New Exploration into the Relationship between Lexicalization and Grammaticalization[A]. *Applied Linguistics.* 2008, (11).

[11] Zheng Jiaheng , Li Wenhua. A Study on Automatic Identification for Internet New Words -- According to Word-Building Rule [A]. *Journal of Shanxi University (Nat. Sci. Ed.)*. 2002, (25).

[12] Zhu, Honglei. Metaphoric studies on Internet New Words[A]. J*ournal o f Zhengzhou Institute of Aeronautical Industry Management (Social Science Edition)* [A]. 2008, (6).

[13] Martin J H. A Computational Model of Metaphor Interpretation. *Boston. Academic Press.* 1990.

[14] Neshanic C L. Computation for Metaphors, Analogy, and Agent. *Springer*. 1999.