Vocabulary Entry of Neologism
A Lexicographical Project aided with NLP Application

Catherine Dahlberg
School of Foreign Studies
Nanjing University, Nanjing, China
22 Hankou Road, Nanjing, China 210093
nanwai@hotmail.com

Tracy Qian Liu
School of Foreign Studies
Nanjing University, Nanjing, China
22 Hankou Road, Nanjing, China 210093

Carolyn Fangya Chen
School of Foreign Studies
Nanjing University, Nanjing, China
22 Hankou Road, Nanjing, China 210093

Summary

This paper describes a lexicographical project utilizing natural language processing tools, where the entries of a glossary are identified by Wordsmith and HC2009YLCL v. 3.0 from a corpus segmented by ICTCLAS 3.0.
The researchers of this project identified two important issues: 1) the entry test as canonized by Chinese lexicographers and 2) the position of neologism in language development. The researchers made a detailed analysis of the nature of neologism and offered a new taxonomy. In addition, a brief evaluation of technology application is included.

Key words: creative language use, neologism, NLP processing, novelty detection, lexicography

Introduction

Internet has greatly promoted language change. A neologism used on the Internet is spread almost instantly to readers who are miles away from the physical location of the creation of this word. In today's wired world, neologism appearing on the web also enters verbal communications in people's real life, a phenomenon itself suggesting the driving urge of man to copy creative expres-

sions. In China1, Internet, since its introduction in 1995, has provided rapid sharing of information to 40 million registered users. It has greatly amplified the influence of neologism, by giving web browsers a chance to quote a neologist unlimited times to an infinite number of audience who might adapt intuitively this very expression at various degrees to suit certain contexts. In this process, a neologism gradually transforms into a real-life linguistic being.

Literature review

Lexicographers have come to understand the difference between evidential and general dictionary. But in China, where the concept of evidential dictionaries is under appreciated, linguists perceive neologism as an uncomfortable variant threatening the stability and comfort of *their* language system. Yang blames the calque expressions in Chinese as having caused confusion in use and deformation of Chinese language (2007: 121–122). He criticizes phrases that frequent TV and cover stories for their violating the structural order of Chinese and challenging established norms (ibid.).

To minimize the impact of variants to their linguistic comfort, scholars established canons in lexicography, which will be consistently referred to in this paper as the entry test. The entry test examines three things: 1) the frequency of usage, 2) the scope of users, and 3) the resistance to change of a phrase. Canonism, *prima facie*, offers a nice operational guidance on lexicography … until one asks specific questions. For example, how do people know if a word, say W, is frequently used? Su and Huang explain that the words are frequent because a number of dictionaries have listed them (2006). The researchers feel sorry for such "frequency" view and wish to provide some background information about Chinese lexicography before moving on to topics on corpus linguistics. The pre-corpora lexicographers had no access to statistical evidence of a particular word except by relying on their own assumption. Now statistical evidence is available, but lexicographers still cling to the old practice. The dictionary data from Su and Huang offer no creditability at all, for they could have been either a perception or coincidence. Worse is the case if some lexicographers simply trusted other scholars' judgement by copying them.

Corpus linguistics offers a better understanding of usage frequency. Meyer explains that computational linguists base their analysis of language on "real data", that is, on evidence provided by corpus and retrieved records of language usage (2002: xiii). The question for Chinese linguists is what kind of corpus to use? Meyer explains why balanced large corpora are the solution (2002: 15). His idea is followed by the Centre for Chinese Linguistics of Peking University

[1] People's Daily (story 2008/11/3) claims 1990 as the year of China's initial Internet connection, downloaded 2009/7/2 from http://english.people.com.cn/90002/95607/6526583.html. Note that the 1990-er Internet was in fact privileged technology designated for academic and research purposes. Access for general China was made available in 1995 by China Telecom.

in its 477-million-character CCL corpus. Unfortunately, pertinent to this project, sorting neologism records from among millions of CCL data is a problem. Other linguists have built smaller corpora for specialized purposes. Su and Huang mention annual dictionaries made available by synchronized corpora (2007). Meyer cites linguists' idea of a monitor corpus (2002: 15).

The researchers in this project also found literature discussions on the legitimacy of language change. Aitchison, who examines language from the socio-historical perspective, emphasizes the importance of careful study of language development, particularly the descriptive approach that examines the "frayed edges" of language and gives enough credit to expressions which conflict with grammar (1997: 104). She concludes that new expressions are essentially a reflection of the changing face of the world (ibid.).

This digital glossary aims to detect neologism actively used in journalism, and the first question is how does it get into media? Yuan summarizes channels of neologism in Chinese as 1) loan words, 2) borrowing from dialects, 3) new meanings injected into an existing word, and 4) coinage (2005). Dai asserts that the adoption of neologism in media is motivated by the need to communicate with the users of neologism (2004: 69). The researchers are not quite content with their explanation. It is necessary to further develop the ontology of neologism, which is a question that may be properly answered by discussions on: 1) the taxonomy of neologism and 2) how people respond to neologism.

Project methodology

Raw data of entertainment news dated between 2005 and 2009 are collected and this corpus is then segmented by ICTCLAS 3.0, so that a Wordsmith key word list is created, lemmas of which then serve as key words for HC2009YLCL v. 3.0 concordancing to identify neologism entries for the glossary.

There are three major steps in this project: data collection, entry list building, and glossary compiling.

Step 1: Data collection

1.1. Internet portals

The researchers identified a number of portals which pool various sources of news service as the data source. They are http://www.yahoo.com.cn, http://www.longhoo.net, http://www.sina.com, http://www.sohu.com, and http://www.tom.com.

1.2. Data extraction

Each researcher is assigned with the collection of 200 sentences, and a total of 573 sentences from 60 articles are actually extracted. A sentence is defined as a body of text within two consecutive periods. The concern about copy right issues is addressed by ratio control and random sequence. Here is the operational routine established in this project: first, the researcher selects a news story

which contains at least 30 sentences; second, by clicking the “page down” button, the researcher positions the cursor to the last period in the article; third, going backward, the researcher extracts a running sentence into the corpus; fourth, the researcher places the cursor at where she had stopped in step three and randomly selects a period that is above it. By repeating step three, the researcher collects 10 sentences from an article without actually copying the story, thus gaining some legitimacy of corpus for research purposes.

Using a period as the sentence boundary is thought to simplify the process. However, the 32,000 characters in the raw data still came as a surprise, which prompted the researchers to examine the style of entertainment news. In the section titled “Findings”, the researchers will report related details.

data treatment

For easy analysis, the researchers regard each character as a lemma in this project. Theoretically, a neologism is a (multi-syllable) word, but the researchers encountered the problem of physical spacing of words in Chinese text, whose iconic layout lacks a visible boundary. To make things worse, Chinese is full of parsing ambiguity. Yu and Zhu provided many examples of ambiguous sentences which are pre-NLP formatted (2006).

Previous research by Chinese computer scientists focused on word segmentation. To illustrate the features of Chinese text, a pair of raw and treated sentences is concordanced in Wordsmith for a lemma. Table 1 is the result.

Table 1. Concord results of “chao”

sentence	treat	Concord
超女张靓颖已在歌坛摸爬滚打一年有余，但这次她与周笔畅、刘亦菲、弦子、薛之谦入围的 却是最佳新人奖。	no	0
超/v 女/nz 张靓颖/nr 已/d 在/p 歌坛/n 摸爬滚打/vl 一/m	seg	超/v 女/nz 张靓颖/nr

ICTCLAS is selected as the segmenter in this project. Before turning to it, the researchers had tried out a number of tools including the MS Word and HC2009YLCL, but dismissed their application for various reasons.

ICTCLAS provides a combined segmenting and tagging treatment, where the bonus POS tagging does not interfere with the project. Proofreading for any segmenting errors is supplemented, and in the case of ambiguity, manual reseg was done according to linguistic sense. In this way, the researchers collected a processed corpus of journalism.

Step 2: Entry list building

2.1. Wordsmith Wordlist

Wordsmith is used to generate a list of words from the segmented corpus. In pilot studies, the researchers identified that the ideal parameters for neologism are length between 2 and 4, and frequency under 3. A two-syllable list was

made by following the low frequency parameter (between 2 and 3), and a few other lists were made at varying parameter values.

2.2. Wordsmith Key Word

The two-syllable list was used as reference and 0.5% setting was used to make a Key Word list which contains a modest amount of neologism.

2.3. Supplement list

Software novelty detection in the previous two steps did not reach the full potential of the corpus. Fortunately the research team consists of experts in language, who possess excellent memory of and impressive vocabulary in neologism. So human knowledge is tapped into to make a quality entry list.

Earlier in step 1.3, ICTCLAS made a noticeable amount of mistakes. The researchers collected information on a discrepancy between the claimed 98.45% accuracy rate and the actual 95% accuracy in lemma processing. This discrepancy suggests an incompatibility between ICTCLAS and neologism. The researchers initially tried to use NLP errors as a sign of neologism. But further research suggested that this thought was over simplistic. Details on this part will be provided in the section titled "NLP evaluation".

HC2009YLCL concordancing replaced corpus proofreading. The short list made available in step 2.2 provided an initial amount of seed lemmas. The flexibility of Chinese is taken advantage of, so that a seed is concordanced to identify additional neologism. Before going too far, the researchers wish to offer an example to explain the flexibility of the collocation power in Chinese. In a two-syllable word, *daxue*, *-xue-* is a free lemma that can be combined with another lemma to form a new word *xuexiao*. The researchers thought that a list of seeds which possess good neologism productivity can help identify enough glossary entries if queried in HC2009YLCL.

HC2009YLCL proves an efficient method of novelty detection as it returns more and better results than Wordsmith. It concordances sentences[2], and a maximum of eleven concords can be recalled in a search.

Step 3: Glossary compiling

The digital glossary is saved as an Excel file. There are a few sections in this glossary, and the major sections are the index page and the entry section.

3.1. Index page

The index page is designed to provide glossary users with access to quick lookup. Lemmas, as listed on the Key Word list and the supplement list, are called the seeds on the index page, since they link a group of new words. A few

[2] Two periods are set as the boundary of a sentence.

miscellaneous records of neologism where no lemma can be found as their common index is assigned to a number.

3.2. Entry section

There are over 220 entries. For easier use, the entry section is broken down into three alphabetically listed entry sheets. Next to each entry is a horizontal array of its definition, grammar explanation and example phrases. The researchers gave each entry a definition based on their linguistic knowledge and the understanding gained through corpus lookup.

Findings

Low frequency of neologism

Wordsmith Key Word indicates a tendency of low frequency of neologism distribution in the corpus. The researchers followed a frequency curve from 7 to 3 before capturing enough records from the corpus.

Interestingly, HC2009YLCL also shows a tendency of low frequency. For example, among a total of 43 concordances of men, only 11 are neologism. In most cases, there was only a single record of a neologism.

Relating findings in NLP application, the researchers conclude that low frequency words should be used in neologism search in a corpus.

Concentration in seeds

Despite the above finding of low frequency, a small group of lemmas show a tendency of high productivity of neologism, as is noticed by the researchers. Table 2 lists a few productive phrases in the corpus.

Table 2. Lemmas with high concentration of neologism

	fensi	*pinpai*	*qijian*	*quan*	*shanzhai*	*shijian*	*yiren*	*zu*
C_{all}	12	9	9	17	23	15	19	13
C_n	12	9	7	17	17	15	19	11

Note: C_{all}=all results, C_n=result of neologism concordancing

NLP evaluation

Despite its consistency, ICTCLAS is not suitable for neologism detection. The researchers suspect that the conflict between neologism and grammar (rules on which the ICTCLAS is built) is probably caused by dated linguistic information on the developers' part. In this project, POS and segmenting of inicinym[3] is a big challenge to technology. Below are a few examples of mistakes.

[3] This is a term given to a specific type of neologism. Discussion on inicinym is provided in the section titled "On neologism".

NLP tagging error:	yule/v	quan/n	1 (12)
	yule/n	quan/v	2 (12)
NLP segmenting error:	fei/b	lin/ng	2 (2)

As mentioned in the methodology, the researchers had hoped that all NLP mistakes can be contributed to neologism. But the truth is, NLP developed with a dated grammar system simply fails to recognize unfamiliar data, whether they are neologism or not. Among the NLP mistakes, a majority is not caused by neologism. In general, ICTCLAS had a difficult time processing person names (PN), film titles, and some proper names. The following example shows a mistake in segmenting a PN and a POS error of a conventional word.

Fengxiao/nr gang/d niandu/n da/d zuo/v 《/wkz ye/tg yan/vg 》/wky
Tr. Feng Xiaogang's masterpiece of the year "*Yeyan*"

In this phrase, Fengxiaogang (a PN) is segmented into two parts, where Fengxiao is considered as a PN, and gang is tagged as an adverb. Two words after this, dazuo (segmented into two parts) is tagged as adverb+verb, but the correct tagging is adj.+noun. To give some credit to technology, the confusion is caused by the ambiguity of lemmas of gang, da, and zuo. But mistakes like this show that NLP technology still needs improvement.

The film title in this example (yeyan) shows another issue of ambiguity. The machine's tagging as [temporal]+verb is correct, but this word has a dual tagging as [temporal]+noun, in which case human intelligence has a better application than NLP technology to offer a sensible solution.

NLP processing is one thing, and the novelty detection needed in this project is another. Inicinyms obviously challenged ICTCLAS. For example, xiuchang is correctly segmented, but NLP is not able to distinguish a neologism of an isonym from a conventional use. Another type of neologism, the structural frame, can be a challenge, too. In the section titled "On neologism", the researchers will explain why.

When comparing ICTCLAS with HC2009YLCL, the researchers are impressed by the latter for its smart processing of PN and neologism. A possible explanation for that is the correlation between the make and year of a technology and the language development in vocabulary and grammar.

Related discussions

On entry test

The entry test is meant to reject neologism, but it no longer fits lexicography. Here is an example to illustrate why the frequency canon fails to address the reality. *Xiu* has over 46,000 records in the CCL corpus, but only 61 of them are in the neologism sense. If the canon of frequency is followed, obviously *xiu* as a neologism does not deserve academic attention. Or does it?

Generally speaking, there is a correlation between time and frequency. Old words have high frequency as time rewards them with an accumulated occurrence. Neologism, on the contrary, naturally lacks enough time to get established. As in the case of xiu, the time factor guarantees a much higher frequency of the conventional isonym than neologism. The researchers object using frequency as an entry test. Instead, they propose giving neologism a separate status and making entry test to focus on the features of neologism.

As for the canons of general audience and stability, the researchers think they are impractical. A Neologism, by definition, is a new word. Internet may have facilitated the encountering of neologism, but the scope of its usage is probably still small. Again, neologism should be studies as a separate subject from conventional words. The scope of a conventional word belongs to the past, but that of a neologism to the future. It will be too early to assess the scope of usage in the case of a neologism if only the current (or to be exact, the initial) popularity is examined. Besides, a careful lexicographer should conduct a full range check of the scope of all the words for his or her dictionary, instead of applying the entry test only to neologism as currently practised.

Then what is the perspective that may suit the reality of a neologism? Suppose A is a new word and B old. Stability of B can be easily measured by looking into its past. But no one can predict the future of A, so there is no way to compare the stability of A with that of B. There is, instead, a possibility to monitor the user profile of A. If A initially appeared as a blog expression and soon gained momentum through repeated online quoting, and finally if it successfully incarnates in some kind of traditional fields, for example, in prints, on the radio, and in speeches, it is safe to assume that the scope of usage of this word has undertaken an expansion. The researchers of this project are interested in the process in which a neologism duplicates from its initial virtual identity into actual usages in business, in news reports, and in daily expressions.

Unfortunately, some people are simply determined to neglect neologism, especially in the case of an established one. Xiu, if looked from the temporal perspective, should no longer be considered as new. Many years ago, Lou described expressions of xiu which had been actively used in Taiwan (1993). In the next decade, established knowledge of xiu never led to dictionary entry. In this glossary, the researchers have identified a number of xiu-group entries, such as gerenxiu, xiuziji, and xiuchang. Since 1993, this word has expanded its scope in meaning. The researchers can identify at least two types of neologism: a) a transliteration of “show” (as explained by Lou (1993)), and b) a fixed expression meaning new performers in sports or entertainment industry.

Is it possible to acknowledge the historical position of neologism? First the researchers would like to establish this argument: newspaper is a public institution and it does not favour the usage of neologism. The media presence of neologism simply reflects the updated face of language at the time of being. If news reporters choose to use a neologism, it means there is a perception that this

word is understood by a large audience. Thus being used in journalism reflects an established status of a neologism in the society. The researchers argue that access to journalism should be considered as some kind of entry test, since media is not the source, but the channel of the circulation of a neologism. The researchers further argue that stability should be replaced by usability. Many words in a dictionary are simply "dead". Why not clean up some dead entries to make room for neologism?

On neologism

There are two types of neologism, one is a coinage, the other is a new use of an existing phrase when the speaker has created a new meaning for it. The researchers would like to describe the latter as a meaning injection, since it seems that the A-sense meaning is injected from outside. For easy reference, the researchers name a neologism created this way an inicinym.

What exactly is a new word? There is a paradox of neologism. One has to possess a repertoire of B to ground the comparison where A can be recognized as new. Now this is exactly how the paradox exists: a new word is no longer new once it is known. In the case of neologism, while bringing the perception of the creativity in a word, the encounter of A takes away its newness.

The complexity of this paradox can be further examined from various angles. The researchers will take you to a speculative discussion on the route of the circulation of an inicinym First, the researchers put the factor of individuality aside and only examine the time factor. Suppose a homogeneous group p encounters A. At a given time t, its subgroup, p1 completes the encountering ahead of the rest, p2. As a result, p1 recognizes A, but p2 does not.

Second, the time factor is replaced by the growth factor. A reaches p which is a mixture of grown-ups g and children c. The repertoire of B is available to g but not to c. As a result, despite simultaneous encounter, g sees A as a neologism, but c acquires this word as a conventional expression.

And finally, personal preference is introduced. Still in a t-instant model among every member of a group e. Subgroup e1 welcomes A, subgroup e2 lacks sensitivity to vocabulary, and subgroup e3 rejects A. When the newspaper starts to be filled with A phrases, e2 may be gradually assimilated into e1, but e3 will not. Then how does A propagate among its users? This is a complicated issue so difficult yet so interesting that we have to research into.

Here are additional thoughts on neologism. Besides asserting that low frequency is the nature of new expressions, the researchers argue that neologism is designed to enter the body of language. Neologism suits the need of people in a way that no other words can. The researchers think that the use of neologism is caused by the urge for creativity which is a unique function of intelligence. A neologist enjoys making new words, and such a manipulation of language is an integral aspect of intelligence. The researchers think the fact that a neologism

gains popularity suggests that people in general appreciate creativity and new perspectives. It is part of human nature to adopt a new word once it is created. After all, language is man's tool to index and explore the world. A neologism means a new outlook of the world.

Neologism calls for better solutions in novelty detection. Ambiguity in the case of isonyms makes it very hard for machine to recognize neologism. Shanzhai is either 1) a conventional word (whereas it is a compound) or 2) an inicinym (a single word). NLP correctly segments it in a running text, but it makes no attempt to POS tag its structure accordingly. Inicinyms in this project were hand picked to make up for the missing NLP available for neologism detection.

Another challenge to NLP is pattern extraction. The researchers call neologism identified in this way structural frame. Though not mentioned in the literature, structural frame exists as a third form of neologism, which is observed in the media. What is it? Simply put, it is a pattern that has been extracted by human intelligence to be used for word formation. For example, hen+[adj.]+hen+[adj.2]. A hen+[adj.] is a conventional phrase, and no grammar ever prohibits putting two of them together, but doubling a hen- structure is a neologism. In this word, the second adjective is always in a two-syllable format. Users of neologism believe that it was created by an actress whose name is Zhong Xintong (Ajiao)4. The latest use of it is the commercial of Xtep "The athlete is henkuhenqiangda, his rival is henruohenkelian".

Conclusion

Lexicographers should give credit to web language and new words used in the media. The entry test contradicts with the nature of neologism. A more reasonable approach to a specialized dictionary should be a corpus-based methodology that is developed with evidential references. NLP application can be utilized to improve the efficiency in a project.

Neologism is a complicated phenomenon of intelligent creativity including coinage, inicinym, and structural frame. The flexibility in neologism is a challenge to NLP technology. Current programs developed for Chinese NLP are adequate in conventional tasks, but they lack competence in novelty detection. Human detection of neologism relies on encyclopaedic knowledge of lexemes, sensitivity to creative collocation, abstraction of "rules" of expression construal, and calibration of semantic deviant. NLP application should take human-specific factors into consideration to enhance the utility of neologism detection.

Academic attention should be paid to the spontaneous linguistic changes caused by free online publishing and easy content sharing. Though it conflicts with language norms, neologism deserves detailed study by linguists.

[4] Her speech of "*henshahentianzhen*" (很傻很天真 tr.: I made a fool of myself. I was too naïve.) soon gained popularity in 2008.

This glossary is compiled with such philosophy in mind, that novelty in language usage claims its position in the vast spectrum of human history. In this sense, this glossary is designed not to serve the general public but to describe current language development. This project hopes to capture the nature of neologism. The researchers call for serious academic considerations of neologism and the changing face of language.

Additional information

1. ICTCLAS is developed by the Institute of Computing Technology of the Chinese Academy of Sciences. The researchers used this technology to segment raw data in a corpus. ICTCLAS is based on HHMM (Hierarchical Hidden Markov Model). It is able to identify the boundary of a chunkable item in a near-human-perception fashion. It can process 1 million characters at the same time.
2. HC2009YLCL is developed by Nan Yanfei (Cheng Nanchang), of the College of Literature of Guangxi University for Nationalities. The researchers used this technology to concordance a corpus. Frequency count is a useful utility of this program, where the boundary of a phrase is automatically identified. It can process 60,000 characters at the same time.

References

Aitchison, Jean. Language Change. Tr. by Xu Jiazhen. Beijing: Yuwen Press. 1997.

Catoni, Bruno. Lexical Resources for Automatic Translation of Constructed Neologisms. //LREC, Marrakech, Morocco. 2008. Retrieved June 12, 2009 from http://www.lrec-conf.org/proceedings/lrec2008/pdf/247_paper.pdf

Cui, Shiqi; Liu, Qun; Meng, Yao; Yu, Hao; Fumihito, Nishino. New Word Detection based on Large Scale Corpus. // Journal of Computer Research and Development. 43 (5), (2006). p. 927–32.

Dai, Qingxia (Ed.). Introduction to Sociolinguistics. Beijing: Commerce Press. 2004.

Janssen, Maarten. Lexical vs. Dictionary Database. // COMPLEX, Budapest, Hungary. 2005. Retrived June 10, 2009 from http://maarten.janssenweb.net/Papers/COMPLEX2005-mjanssen.pdf

Janssen, Maarten. NeoTrack: semiautomatic neologism detection. // APL, Lisboa, Portugal. 2005. Retrieved June 14, 2009 from http://maarten.janssenweb.net/Papers.

Kilgarriff, Adam. What computers can and cannot do for lexicography. Retrieved June 12, 2009 from http://www.kilgarriff.co.uk/Publications/2003-K-AsialexKeynote.doc

Lee, Min-Jeh; Huang, Chien-Kang; Chien, Lee-Feng. Automatic Construction of a Bilingual Dictionary for Spoken Language Processing Applications. // Oriental COCOSDA99, Taipei:Academia Sinica. p. 37–40.

Lou, Chengzhao. On *Xiu* and more. // Chinese Translators Journal. 5(3), 1993. p. 45–6.

Meyer, Charles. English Corpus Linguistics. Port Chester, NY, USA: Cambridge University Press. 2002.

Su, Xinchun; Huang, Qiqing. Development of Neologism and the Canons of Prescriptive Dictionary Compiling. // Lexicographical Studies. (3), 2003. p. 106–13, 15.

Sun, Maosong; Huang, Changning; Fang, Jie. Quantified Research in the Collocation of Chinese. //Zhongguo Yuwen [Chinese]. 256 (1),1997, p. 29–38

Sun, Honglin. Features of Collocation in Discourse. //1998 Zhongwen Xinxichuli Guoji Huiyi Lunwenji [Proceedings of the International Symposium on the Digitalized Processing of Chinese 1998]. Huang, Changning (Ed.). Beijing: Tsinghua University Press. p. 230–6.

Wen, Duanzheng. On Dialect and Popular Sayings. Originally published in Linguistic Research, 30 (1), (1989). Selected Papers on Linguistics by Wen Duanzheng. Shanghai: Shanghai Lexicographical Publishing House. 2003. p. 339–50.

Yang, Xipeng. Research on Foreign-Origin Vocabularies in Chinese. Shanghai: People’s Press of Shanghai. 2007.

Yu, Shiwen; Zhu, Xuefeng; Li, Feng. Design of a Lexicon Database of Contemporary Chinese and its Application. // Chinese Teaching in the World. (2), 1999. p. 39–46.

Yu, Shiwen; Zhu, Xuefeng. Yuwen Xiandaihua yu Hanyu Xinxichulijishu [Modern Technology applied to Chinese and Chinese Digitalization]. // Yuwen Xiandaihua Luncong [Papers on the Modernization of Chinese], vol. 6. Su Peicheng (ed.). Beijing: Yuwen Press. 2006. p. 176–89.

Yuan, Jiheng. Mechanism of Chinese New Words. // Journal of Kaifeng Institute of Education. 25 (1), 2005, p. 32–3.