R. Liu, L. Wang

MONITORING THE USE OF ENGLISH LANGUAGE IN CHINA[1]

Abstract. Language monitor has been a research focus both in China and around the world. The present study utilizes a corpus-driven approach to identify new words in China Daily Corpus. To monitor the use of English language in China, a dynamic corpus of China Daily is built and statistical investigations are carried out. The present study identified a total number of 362 new words in China Daily during the year of 2011. Arising from multiple walks of life, these new words documented and reflected changes happened in China and abroad.

Key words. Language monitor, statistical investigation, new words

1. Introduction

The present English language monitor project is developed as a part of Chinese national language resource monitoring and research project to provide detailed up-to-date information on the linguistic situation and language development in English media published in China.

Language monitoring has been a recent research focus. Based on Linguistic theories and information processing technology, language monitoring is a multidisciplinary, large-scale, social language engineering project (He, et. al 2009). So far, agencies specialized in language monitoring include the Global Language Monitor, with a particular emphasis upon the usage of English language worldwide, and China's National Language Resource Monitoring and Research Center, with diverse projects focused on the usasge of Chinese languge in audio media (broadcast), print media (newspapers,

[1] Research project supported by the Ministry of Education (11YJC740067), Shanxi Scholarship Council of China (2012-041), Taiyuan University of Technology (The discipline construction and research of language information processing).

textbooks), TV as well as Chinese network. Each year these two organizations publish their reseach findings and realease new words and phrases. For instance, the 2010 Top Words released by global language Monitoring Network are spillcam, vuvuzela, the narrative, refudiate, Guido and Guidette, deficit, Snowmagedden 3-D, shellacking, simplexity. Chinese National Language Resource Monitoring and Research Center released the 2010 China top ten media new words as well. They are *Earthquake, the Shanghai World Expo, the Asian Games in Guangzhou, China Railway High-speed (CRH) train, low-carbon, microblog, currency war, Chang E II, China's 12th Five-Year Plan, Geili)*. We can see that the two groups of new words reflected respectively the international and domestic society and people's livelihood in 2010. «The actual status of the Chinese language is what we monitor», said Wang Tiekun, the Deputy Director of the Ministry of Education. «In recent years, one drastic change in the discipline of linguistics in China is that the State Language Commission began to publish regularly the annual report of language situation in China, which is a useful attempt to carry out investigation and study on the actual status of Chinese language usage and an important step taken to disclose the language information to the public. The significance lies in that it focuses public attention on language usage, helps them grasp national language status, perceive and respond to multiple changes and ambiguities arise from daily life in an effort to build a harmonious language life». It is stated in the langugage situation in China that Chinese language is developing well and language, as cultural asset and «soft power» of a nation, deserves nationwide attention. It also emphasizes the importance of constructing harmonious language life, from the perspective of which, languages and dialects are valuable national resources and efforts are needed to pursue their coexistence. Extensive research on Chinese language usage has been conducted, however, the use of the English language in China, mainly English usage in mainstream printing press, still awaits investigation. Therefore, monitoring English used in the mainstream printing press complements Chinese national language monitor project.

Investigations in English used in Chinese mainstream printing press can provide information on the state of cultural communication between China and the rest of the world. With the process of globalization, English is becoming ubiquitous. As a universal language, China gets to know western culture encoded in English language. Likewise, China diffuses its cultural heritage and the situation of China to the world through English language. In addition, the research findings can benefit English learners who are lack of the vocabulary to express their native culture in English. As an authority in the use of language, mainstream printing press may serve the purpose.

With advances in information technology, the recent years has witnessed an outburst of new words both in China and in western countries. The summary of these new words will yield a revealing insight into the state of the world and make people better grasp of the direction of social development. Li Yuming (2007) pointed out: «the purpose of releasing new words is to share information with the public». Language monitor technologies can not only track real-time language changes and developments, but also provide up-to-date language resources for information processing, linguistics and applied linguistics, lexicography, public opinion analysis and early warning, social and political services. Therefore, Language monitor project is of great significance. The purpose of this study is to monitor and analyze the usage of English language in Chinese mainstream printing press. By extracting new words and monitoring emergence and development of new words in the monitor corpus, a database of new English words unique to Chinese press is constructed, which may be used for the comparison with those released by the Global language Monitor and serve the field of second language acquisition and dictionary compilation.

A great number of new words concerning the social and livelihood developments emerged in the texts of Chinese mainstream printing press, which makes them the best candidates for language monitor. As China's official English-language newspaper and a reference to the foreign media, China Daily, plays a dominant role in

promoting Chinese philosophy and culture. Due to its authority in English usage, China Daily is chosen as the monitoring subject.

This study aims to build a large-scale corpus of Chinese mainstream printing press and conduct extraction and study of new words. The project will build a corpus for further study and at the same time monitor the usage of English language in China.

2. Trends in Corpus Linguistics

Since the construction of the first machine-readable corpus – Brown Corpus in 1960s, corpus linguistics has been developing rapidly. The recent trend is the third generation corpus (Leech 1991). The most notable features of the third generation corpus are incredible richness of data, with the number reaching hundreds of millions of words and dynamic updates which renew the corpus continuously or as required. «There is no limit to the magnitude and time span. The corpus is evolving the way language does» (Zhang 2001). The renewal of data provides a live circumstance for language.

Corpus construction is also a research focus in China with emphasis on Chinese corpus, English Corpus, parallel corpus and corpus of other foreign languages. In the field of English linguistics, currently built corpus are mainly English learners' corpus with the scale of one million words, i.e. Shanghai Jiaotong Daxue English of Science and Technology (JDEST) Corpus, Parallel Corpus of Chinese EFL Learners by Beijing Foreign languages University, Chinese Learner English Corpus jointly built by Guangdong University of Foreign studies and Shanghai Jiaotong University. These Corpus are common in that: 1) English learners' corpus are overwhelming; 2) the scale is relatively small, usually numbered in one million; 3) they are static and can not be updated.

3. Researches

3.1 Definition

New words are generally defined in two dimensions: time and form. Newmark (1988) defined new words as newly created words or

existing words with new meaning. Li Yuming (2007) pointed out: «words and phrases with new meanings are categorized as new words». Zou Gang (2006) classified new words as words out of vocabulary. As to whether words with digit constituent (MP3, Lhasa «3.14») fall into the category of new words, there is no definitive answer. In this project, new words are defined from a broad perspective. Words that emerged recently, existing words with new meanings, words with numbers are all within the scope of the new words. New Words and phrase in the 21st century take on some distinctive features. There is an increasing number of emotional words, affix word formation, and initiative or interactive word. Thus, extraction and tracking of new words can only be done on the basis of a clear definition.

3.2. New words extraction and identification technology

The difficulty of the research lies in the determination of boundaries while extracting new phrases, the identification of low frequency new words with fresh meanings. There is no relevant research in English new words in China so far. New words are to be extracted by means of statistics and linguistic rules.

4. Research Design:

4.1. The establishment of a dynamic corpus:

Sinclair's idea of monitor Corpus is realized in China through the establishment of Dynamic Circulating Corpus (DCC) by Professor Zhang Pu from Beijing Language and Culture University. The corpus is established on the relative time concept proposed by Professor Zhang Pu (2001) Compared with the previous corpus, the distinctive feature of DCC is diachronic approach, i.e. data are updated continuously, which provides the means of tracking the diachronic development of a word or phrase. However, DCC monitors the Chinese language. As a complementary part of language monitor in China, this research plans to set up a dynamic corpus of plain press to monitor the emergence, development of language phenomenon. Based

on the statistical analysis of this corpus, a bank of new words and phrases are built for further study.

Under the guidelines of Sinclair's monitor corpus and Professor Zhang Pu's practical approach to the establishment of DCC, our corpus made of plain press named China Daily is built. All the articles from China Daily are classified and stored in accordance with topics and constantly updated. The specific steps are as follows: 1) With the aid of open source software wget and httrack, html files are downloaded everyday and labeled in the format of «Year / Month / Day». An index of headlines, file names and hyperlinks is created daily to double-check the downloaded webpages and ensure files are complete and correct. 2) Convert the html files into files in the form of xml format and txt format. Programming is in command to remove advertisements, navigation bars as well as control characters generated by the conversion.

4.2. New words extraction

New words are extracted using statistical measures like frequency, mutual information and entropy. Semantic properties of the words extracted are checked by the introduction of the semantic dictionary «wordnet». Mutual information (MI) measures the mutual dependence of the two random words. It is a very effective measure widely used in the statistical language model for automatic extraction of words in measuring their collocation strength. Entropy helps to determine the boundary of a phrase.

The process of the new words extraction is illustrated as follows (Fig. 1).

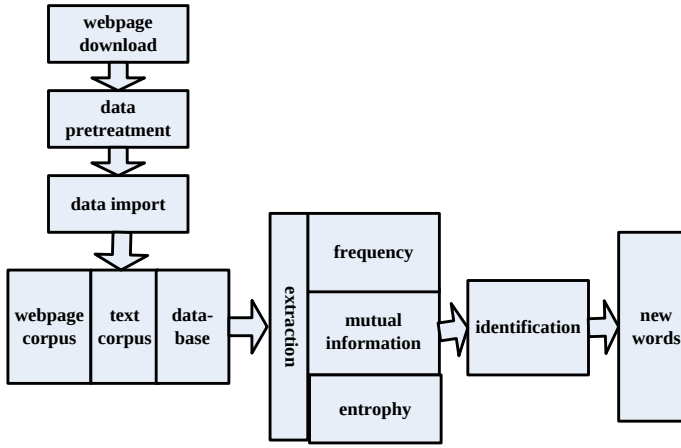

Fig. 1. The process of new word extraction

The overall research procedures are elaborated as follows:

1) Download original Webpage (authorized) from the website of China Daily to build a raw corpus.

2) Pre-treatment of the raw data includes classification and storage of the downloaded files in accordance with topic, format conversion, preliminary label and tagging.

3) Import data into corpus. We build two corpora, namely webpage corpus and text corpus. And database is built for further evaluation. Words and their statistical information (frequency, mutual information, entropy) are saved in the database for the convenience of extracting new words and phrases.

4) Extract automatically the new word candidates by means of frequency, spread curves, mutual information, etc.

5) Evaluate and identify the new words.

6) Identification the new words.

5. Experiment and results

The primary focus of the present study is the statistical investigation of new word extraction. The statistics, namely frequency, mutual information and entropy respectively play a

distinctive role in word extraction. Therefore, the determination of statistics is paramount. After many attempts, the most appropriate statistical boundaries are set. Table 1 lists the statistics for 2-, 3-, 4-, 5-grams extraction respectively.

Table 1. Statistics for 2-, 3-, 4-, 5-grams extraction.

Statistics	Frequency	Mutual information	Left Entropy	Right entropy
2-gram	>10	>0	>2	>2
3-gram	>9	>0	>2	>2
4-gram	>6	>0	>2	>2
5-gram	>5	>0	>2	>2

After statistical computation and artificial identification, a total number of 362 new words and phrases are identified from the 2011 China Daily Corpus.

In a close examination of the words and phrases, we find that they come from a variety of sources. They arise with the emergence of advanced technology, social problems, natural disasters, conflicts and other developments in all walks of life.

18 out of 24 words concerning disasters come from 2011 Japan earthquake, a magnitude 9.0 earthquake and the 5th biggest in World history and the subsequent tsunami, 40.5 meters (133 feet) at Miyako. The earthquake and tsunami damaged the Fukushima Nuclear plant and lead to a nuclear disaster in Japan and its neighboring countries. Other major international incidents are reported in Chinese press. Among others, the «Arab Spring» stands out. The Arab Spring is a volutionary wave of demonstrations, protests, and civil wars in the Arab world that began on 18 December 2010. To date, rulers have been forced from power in Tunisia, Egypt, Libya, and Yemen; civil uprisings have erupted in Bahrain **(http://en.wikipedia.org/wiki/Arab_Spring - cite_note-reutbahdor-5)** and Syria; major protests have broken out in Algeria, Iraq, Jordan, Kuwait, Morocco, and Sudan; and minor protests have occurred in Mauritania, Oman, Saudi Arabia, Djibouti, and Western Sahara. Related words and phrases

include *colonel Gadhafi, Libyan rebel, the Southern Sudanese, the Syrian crisis, National transitional council, the Libyan crisis, the Military council, Gadhafi's death, the muslim brotherhood, Egypt and Tunisia, embassy in Cairo, Cairo's Tahrir square, the crisis in Libya, Libyan leader Muammar Gadhafi, the National transitional council, the unrest in Syria, North and South Sudan, Saleh to step down, a no-fly zone over Libya, the «Arab Spring», unrest in the Middle East* etc.

As a communicating media, China Daily also reports on China's domestic affairs. On Oct 5th, Naw Kham and his gang members masterminded and colluded with nine Thai soldiers in an attack on two Chinese cargo ships, the Hua Ping and Yu Xing 8, on the Mekong River, and 13 Chinese fishermen were murdered. In April 2011, under Naw Kham's instructions, several of his gang members also kidnapped Chinese sailors and hijacked cargo ships in exchange for ransom. In Nov 2011, China, Laos, Myanmar and Thailand established a security mechanism and conducted joint law enforcement operations to safeguard the river. The crime ring was busted in a joint operation by police from China, Laos, Myanmar and Thailand and brought to China for trials. Words reflecting this incident, namely joint law enforcement, 13 Chinese sailors, are extracted. Besides that, wording reflecting social problems like school-selection fees, the guo meimei incident, begging in vehicle lanes, buildings with glass curtain walls, fake seeds and gutter oil account for a large part of the new words and phrases identified. These words show pubic concern for education, the widening gulf between the rich and the poor, and security issues in architecture, agriculture and food. To combat social problems, the government is taking measures, which is seen from words and phrases like the basic healthcare, the low- and middle-income groups. Wen Jiabao, the premier of China, said «We will vigorously adjust income distribution, increase the incomes of low- and middle-income groups and enhance people's ability to consume». China introduced in January 2011 purchase restrictions on housing and its tight credit and tax policy on multi-home buyers. The rules limit every local household to buy two housing units and confine every non-local family to just one are maintained ever since to curb speculative and

investment demand in the housing market. «The government will not loose restrictions on home purchases and mortgages, which are the most effective controlling measures against speculation». Said the Ministry of Housing and Urban-Rural Development. As a new policy, the use of words are a little bit confusing at first. Borrowing limit, purchase restrictions, home purchase restrictions all refer to the same thing. As time goes by, borrowing limit has been eliminated which reflects the development of a new word.

In addition, advances in various fields are reported. Weibo users, micro blog accounts, a micro blog post, use of micro blogs become popular due to the ubiquity of micro log. China's first aircraft carrier has been long expected and boosted China's morality. With the promotion of traditional Chinese medicine, tcm therapy is popularized.

To sum up, these new words reflect the changes occurred in various fields both in China and around the world. Monitoring the use of English language in Chinese printing press reflects the China's concern for the world and also displays a whole view of China to the world. As a part of Language monitor project in China, the China Daily corpus and the new word extraction experiment provides a bird view of what's going on in China and the world.

References

[1] *Leech G.* 1991. The state of the art in corpus linguistics. In K. Aijmer and B. Altenberg, eds., English Corpus Linguistics: Studies in Honour of Jan Svartvik. London: Longman.

[2] *Li Y. M.* 2007. Thoughts on the issue of new words. http://www.gmw.cn /01gmrb/ 2007-08/24/ content_660188.htm

[3] *Newmark P.* 1988. A Textbook of Translation. Shanghai: Shanghai Foreign Language Education Publishing.

[4] *Sinclair J.* 1991. Corpus, Concordance, Collocation. Oxford: Oxford University Press.

[5] *Zhang Pu*. 2001. «On Cybernetics and Dynamic Updating of Language Knowledge». Applied Linguistics, 2(1): 71–76.

[6] *Zou Gang*. 2006. Automatic new word extraction on Internet.

[7] **http://www.languagemonitor.com/top-words/top-words-of-2010/**

[8] **http://www.moe.edu.cn/publicfiles/business/htmlfiles/moe/moe_1485/201012/113648. html**