The distribution of four-character idioms in Chinese news texts and its implications for CFL learning and instruction

Jun Da
Middle Tennessee State University
U.S.A.

1. Introduction

As in other languages, chengyu or idioms are a significant feature of the Chinese language. They are a means through which native speakers relate to their culture, history and tradition and create metaphysical or philosophical discourse beyond surface meaning (Zein 2004). Given their wide-spread occurrence in both spoken and written Chinese communication, idioms are also an indispensable part of CFL (Chinese as a Foreign Language) learning and instruction. The learning of idioms, however, plays out differently for CFL learners at various stages of their learning process. In beginning level CFL classrooms where learners have acquired very limited vocabulary and grammatical competence, idioms are used from time to time to assist learners' acquisition of basic language skills. An example that illustrates this practice is Chinese Link (Wu, etc. 2006), which offers a selective set of idioms (some of which also contain companion texts) as supplements in both its elementary and intermediate editions. The inclusion of a very limited number of idioms in such CFL textbooks is intended, among others, to provide cultural knowledge and enhance learner motivation. At the same time, simplified texts that illustrate the meaning of those idioms are provided to facilitate learners' acquisition of basic reading skills or serve as prompts for speaking activities. In terms of selecting idioms for beginning level CFL learning and instruction, our informal observation of common practices in several U.S. universities indicates that their choice is more likely to be influenced by instructors' personal preferences, students' interests and other pedagogical needs and constraints such as textbook design and the local community where CFL is learned.

However, idioms tend to play a more central role at the intermediate and/or advanced level of CFL learning and instruction. At this stage, CFL learners are more likely to have developed interests in acquiring competence in reading authentic written texts, where idioms themselves become part of the learning objectives. The change from an assisting role to the focus of learning is partly due to the fact that idioms are one of the language elements that make written Chinese texts (more) elegant (典雅) (冯胜利 2006) and hence are more difficult for CFL learners to comprehend because of (very) limited prior exposure[1]. Given the fact that Chinese idioms come in tens of thousands (e.g.,《中国成语大辞典》(Wang, etc. 1985) contains more than 18,000 entries.), it is extremely difficult, if pedagogically possible at all, to cover all or the majority of Chinese idioms in any CFL curriculum. Such a pedagogical constraint thus poses an immediate question for both CFL instructors and learners: Which idioms should be targeted and how much treatment should each of those idioms receive in, for example, a classroom-based curriculum? While individual judgment based on prior knowledge and experiences still plays a role in the selection of idioms for intermediate and advanced level CFL instruction, we believe that the selection of idioms for, for example,

[1] Our informal observation of college CFL textbooks such as Integrated Chinese (Yao, etc. 2005) and Chinese Link (Wu, etc. 2005), etc. used in the U.S. indicates that CFL learners are more likely to be exposed to spoken language at the beginning level. Written language becomes more dominant at the intermediate to advanced level.

developing competence in reading authentic news texts, can be better guided by a data-driven empirical approach.

In this paper, we present the findings of a corpus-based study on Chinese idioms whose purpose is to identify those idiomatic expressions that are most frequently used in authentic Chinese news texts. The objective of the study is to generate a reference list of four-character idioms that can be used for preparing reading and vocabulary learning materials for those who are interested in gaining competence in reading Chinese news texts. We choose four-character idioms as the focus of our study based on the assumption that they form the majority of idioms in Chinese. We believe that a measure of the frequency distribution of four-character idioms will provide us with representative statistics on the distribution of idioms in general in Chinese news texts and thus can serve as a good indicator of which idioms should be targeted for CFL learners.

2. This project

2.1 Chinese news text corpus

The current project builds on the author's previous study (Da 2005)[2] on identifying words and phrases used in Chinese news texts and uses the same corpus to obtain the frequency measure of four-character idioms.

The Chinese news texts used in the study were collected, using both manual and automatic methods, between the middle of 2003 and the end of 2004 from the news section of the World Forum website[3] (世界论坛网). The corpus contains a total of 27,965 news articles with more than 20 million Chinese characters. Those news articles fall into 15 different categories which include both international and regional news covering topics such as politics, science and technology, culture, education and sports, etc. A random sampling of those news articles indicates that the majority of them originated from different news agencies or media sources in Mainland China, Taiwan, Hong Kong, Macao, other Asian countries or regions, North America, and Europe that include, but not limited to 新华社, 凤凰卫视, 中央社, 路透社, and 美联社, etc. The corpus also includes a small number of news reports and commentaries written by individuals and published on the Internet. We chose news collections from the World Forum website based on the belief that while the selection of those news articles may have been influenced by the website's editorial polices as well as personal preferences of its news editor(s), the diversity in their sources, regions and subject areas nevertheless makes those news articles a good representation of journalistic Chinese that are currently used world wide.

2.2 Identification of four-character idioms

In order to help us identify four-character idioms in the news corpus, we first collected more than 28,000 Chinese idioms from various sources on the Internet[4], which include such websites as 国学网络, 中教网 and 汉典[5]. While the accuracy of our collection of idioms remains to be checked, we believe that the large number of idioms in the collection and their

[2] c.f., http://lingua.mtsu.edu/chinese-computing/newscorpus/.
[3] c.f., http://www.wforum.com/gbindex.html.
[4] For details of those 28,000 idioms, please refer to http://lingua.mtsu.edu/chinese-computing/chengyu.
[5] c.f., http://www.guoxue.com, http://www.teachercn.com, and http://www.zdic.net.

sources make them a good representation of all the idioms currently available in Modern Chinese[6].

With reference to the pre-compiled list of more than 28,000 idioms, we then used a computer script to automatically identify four-character idioms in our news corpus through a two-step process. First, news texts were segmented into continuous four-character strings. Secondly, each string was checked against the pre-compiled list of 28,000 idioms. For the sake of comparison, we also checked four-character idioms from the HSK vocabulary list[7] and obtained frequency information about those idioms in our news corpus.

2.3 Results and discussions

We found 4,900 four-character idioms from the news corpus whose frequency distribution is shown in Table 1. (More detailed information can be found at http://lingua.mtsu.edu/chinese-computing/newscorpus/.)

Table 1 Frequency distribution of four-character idioms in the Chinese news corpus

Frequency	695≥X>200	200≥X>100	100≥X>50	50≥X>10	10≥X>5	5≥X≥1	**Total**
Counts	5	22	107	1,136	813	2,817	**4,900**
Percentage	0.10%	0.45%	2.18%	23.18%	16.59%	57.49%	**100%**

As can be seen in Table 1, there are only a very limited number of idioms (a total of 134) at the high to intermediate frequency level (X>50), which makes up less than 3% of the 4,900 idioms found in the news corpus. In contrast, more than half of the idioms (2,817 or 57.49%) occur at very low frequency level (x≤5). Further, nearly 40% of the idioms occur at intermediate-low to low frequency level (50≥X>5).

To put the above statistics in perspective, we also calculated frequency statistics of all four-character idioms in the HSK vocabulary list. There are a total of 145 four-character idiomatic expressions in the HSK list, out of which 125 are also present in our news corpus. Their frequency distribution is shown in Table 2. Similar to those 4,900 idioms above, we find that those 125 idioms also span across a wide range of frequencies where the majority (more than 78%) occurs at intermediate-low to very low frequency level (X≤50).

Table 2 Frequency distribution of idioms from the HSK list in the news corpus

Frequency	323≥X>200	200≥X>100	100≥X>50	50≥X>10	10≥X>5	5≥X≥1	**Total**
Counts	3	9	15	52	18	28	**125**
Percentage	2.40%	7.20%	12.00%	41.60%	14.40%	22.40%	**100%**

The above statistics have some pedagogical implications for the teaching and learning Chinese as a foreign language. First, the large number of idioms we found in the news corpus suggests that if a CFL learner wants to obtain competence in reading authentic Chinese news texts, he/she needs to know or at least be exposed to a lot of idioms. If we assume that 30,000

[6] To the best of the author's knowledge, no empirical study has ever been conducted on estimating the upper limit on the number of idioms in Chinese.

[7] The HSK (Hanyu Shuiping Kaoshi, or Chinese Proficiency Test) vocabulary list is compiled by the China National Office for Teaching Chinese as a Foreign Languages (国家对外汉语教学领导小组办公室《汉语水平词汇与汉字等级大纲》) and contains 8,822 characters, words and phrases.

is the upper limit on the number of idioms available in Modern Chinese, then nearly one out of six will occur in news texts; Secondly, given the very skewed distribution of a small number of high-frequency idioms and a vast number of less frequently used idioms, we believe that the teaching and learning of idioms used in news texts can be made effective through a two-stage process: A CFL learner should first seek or be given the opportunity to master those most often used idioms so that vocabulary difficulty can be minimized at the onset of a learning program. Mastering of those idioms, for example, can be achieved through a concentrated study of those idioms with specially designed learning materials and the assistance of instructors. Once the set of high-frequency idioms are mastered, a CFL learner can then move on to acquire the large number of less frequently used idioms through contextualized extensive reading, with the help, for example, dictionaries and (occasional) assistance of an instructor or native speaker. Finally, given the fact that the HSK list contains a very small number of idioms that also distribute across a wide range of frequencies, we believe that the list alone is not sufficient to be used as reference for preparing learning materials designed for assisting CFL learners' acquisition of idioms in Chinese. Additional lists such as the one obtained in this study should be consulted so that we can obtain a better picture of idiom use in Chinese texts.

3. Concluding remarks

Chengyu or idiomatic expressions in Chinese are important aspect of the language that needs to be acquired by CFL learners if they want to develop competence in reading news texts. In this project, we identified 4,900 idioms from authentic Modern Chinese news text and found that a small number of them occur (very) frequently and the majority sparsely. Based on this observed frequency distribution pattern, we suggest that if we want to assist CFL learners to develop vocabulary competence for adequate reading comprehension of authentic Chinese news texts, an effective approach is to first offer them with a focused study of high-frequency idioms and then let them pick up the vast majority of idioms through extensive reading and with the help of dictionaries.

It should be noted that the idiom list we came up with in this study is based on news texts only and obtained through automatic means and with reference to a pre-compiled idioms list. Because of limited time and human resources, we did not manually check the accuracy of the list. Future research is needed to improve the quality of the list so that it will be more reliable and hopefully more useful.

References

Da, Jun. 2005. Reading news for information: How much vocabulary a CFL learner should know. International Interdisciplinary Conference on Hànzì rènzhī - How Western Learners Discover the World of Written Chinese. Germersheim, Germany.

Wu, Sue-mei, Weizhong Tian, Yanhui Zhang and Yueming Yu. 2005. Chinese Link: Zhongwen Tiandi. Upper Saddle River: Prentice Hall.

Yao, Tao-chung, Yuehua Liu, Liangyan Ge, Yea-fen Chen, Nyan-ping Bi, Xiaojun Wang and Yaohua Shi. 2005. Integrated Chinese (2nd Edition). Boston: Cheng and Tsui

Zein, Patrick Hassel. 2004. Idiomatic expressions in Chinese. [http://www.zein.se/patrick/chengyu.html]

冯胜利. 2006. 汉语书面用语初编. 北京: 北京语言大学出版社.

王涛等编著. 1985. 《中国成语大辞典》上海: 上海辞书出版社.

国家汉语水平考试委员会办公室考试中心. 2001.汉语水平词汇与汉字等级大纲(修订本). 北京: 经济科学出版社

Author's Biography

Dr. Jun Da is Associate Professor of Linguistics and Chinese in the Department of Foreign Languages and Literatures, Middle Tennessee State University, U.S.A.

Abstract

While the selection of idioms for beginning level CFL learning and instruction is mostly determined by individual preferences and pedagogical considerations, we propose that idiom selection targeting intermediate and advanced level CFL learners is better guided by a data-driven approach. Based on 20-million character corpus of Chinese news texts, we generated a reference frequency list of four-character idioms where only a small portion of the idioms occurs at high-frequency level and the majority at low frequency level.