A web-based vocabulary profiler for Chinese language teaching and research

Jun Da, *Middle Tennessee State University*

This paper describes the design and development of a web-based Chinese vocabulary profiler that is intended as a handy tool for Chinese language instruction and research. Given a piece of text, the profiler is capable of generating character and vocabulary profiles with frequency and availability information based on a comparison of the set of characters and words found in the text with a selection of other pre-compiled reference lists such as Da's Modern Chinese character and N-gram frequency lists, the HSK word list and customized vocabulary list provided by a user.

1. Introduction

Vocabulary profiling is a measure of the proportions of low and high frequency vocabulary used in a written text. In addition to frequency information, a profiler designed for lexical analysis of texts often provides other information such as the presence/absence of the set of words from the input text in other specialized word lists and their collocation with other words, etc. While it has been used in stylistic studies of written texts (e.g., Graves's (2004) profiling of Jane Austen's novels and letters), vocabulary profiling has also been suggested as a useful instrument in second language acquisition research and pedagogy (Laufer and Nation 1995). For example, Morris and Cobb (2004) examined the potential of using vocabulary profile as a predictor of TESL (Teaching English as a Second Language) trainees' academic performance and reported that vocabulary profile results correlate significantly with their course grades. Schuemann and Benz (2004) also find a vocabulary profiler to be a very valuable resource when they prepare materials and make decisions about how to address vocabulary in their ESL reading classes.

In terms of computing, the generation of vocabulary profiles is mostly straightforward: First, a text is segmented into individual words and/or phrases. Then, the list of words and phrases are compared with other pre-compiled vocabulary reference lists that contain frequency as well as other information. While manual profiling is both time consuming and prone to human errors, such a task can be accomplished easily by a computer program. For Western languages such as English and French, both standalone and web-based vocabulary profilers are readily available (c.f., for example, Nation's Range and Cobb's Online Vocabulary Profiler). However, a similar tool that is capable of processing (non-ASCII encoded) Chinese texts is still missing in the public domain.

In this paper, we describe the design and development of a web-based Chinese vocabulary profiler (c.f., http://lingua.mtsu.edu/chinese-computing/vp) that is intended as a handy tool for Chinese language instructors and researchers. Specifically, the online vocabulary profiler is designed to 1) generate distribution statistics about characters and words from Chinese texts provided by a user, and 2) provide both frequency and availability information about the characters and words from the text based on a comparison with other pre-compiled character and word (frequency) lists, which include, among others, Da's (2004) Modern Chinese character and N-gram frequency lists and the HSK (Hanyu Shuiping Kaoshi or 汉语水平考试) word list. In addition, the online profiler is designed to optionally construct vocabulary profile based on customized word list provided by a user. It is expected that the profiler will be useful for Chinese language instructors when they, for example, need to evaluate the suitability of reading texts for various levels of CFL (Chinese as a Foreign Language) learners. It can also be used by language researchers when they want to measure CFL learner's vocabulary knowledge and understand their lexical acquisition and performance.

2. A web-based vocabulary profiler

2.1. Definition

In inflectional languages such as English, individual words in running texts are delimited with blank space and punctuation marks. Further, lexical analysis often refers to concepts such as tokens (a count of every word in a text), types (unique words in a text), lemma (a headword and some of its inflected and reduced forms) and word families (a headword, its inflected forms and its closely related derived forms from affixation, etc.) (Nation 2001). In contrast, written Chinese texts contain strings of characters where no delimiters are used to mark the boundaries of individual words or phrases. Further, a word in Chinese could consist of one, two, three or (even) more characters. In terms of computing, it is often difficult, if at all possible, to decide if a particular sequence of two or more characters forms a single word, compound word or phrase. For example, we can argue that 看门 is a verb phrase containing the verb 看 and the noun 门. However, there is no surface clue for us to tell if 看门人 should be treated as a single word or compound word. Given the computational unpredictability associated with the identification of words or phrases in Chinese texts, we will use the term N-gram in the online vocabulary profiler to refer to text strings that contain more than one characters. We will also use the term vocabulary to refer to unique N-grams that form meaningful domains on their own, though no judgment is made on whether an individual meaningful domain falls into a single word, compound word or phrase. In this sense, our use of the term vocabulary roughly corresponds to the concept of type as defined by Nation (2001) for English.

2.2. Reference lists of characters and vocabulary

The web-based vocabulary profiler uses the following character and word lists as references for vocabulary profiling:

1. The HSK word list prescribed in 《汉语水平词汇与汉字等级大纲》, which is compiled by the China National Office for Teaching Chinese as a Foreign Language (国家对外汉语教学领导小组办公室). The list contains 8,882 characters, words and phrases that fall into four levels[1];
2. Da's (2004) character and N-gram frequency lists of Modern Chinese. The lists are compiled based on a 200 million character corpus containing a balanced selection of both imaginative and informative texts collected from the Internet[2];

[1] The official version of the HSK word list can be found in, for example, 《HSK 汉语 8000 词词典》. The online profiler uses an electronic version downloaded from http://www.chinese-forums.com/vocabulary/. The URL was last checked on 2006-04-16.

[2] C.f., http://lingua.mtsu.edu/chinese-computing.

3. Allanic's (2005) list of 1,440 Chinese characters for reading purposes (安雄《一级阅读字表》). According to Allanic (2005), his list of characters is compiled from several other Chinese character frequency lists as well as character lists prescribed in three CFL (Chinese as a Foreign Language) syllabi[3];
4. Zhang's (2005) list of 500 basic Chinese characters[4]. According to Zhang (2005), the set of characters is selected based on both their frequency and their representativeness of the structures and components of Chinese characters. A mastery of the basic set should help or enhance a learner's recognition or learning of other Chinese characters;
5. The List of Frequently Used Characters in Modern Chinese (《现代汉语常用字表》). The list contains 3,500 characters that fall into two levels: The first level contains 2,500 (more frequent) characters and the second 1,000 (less frequent) characters[5];
6. Lists of unique characters found in《三字经》,《千字文》and 《百家姓》[6]. The three character lists are compiled to provide some historical perspective for users when they build their own vocabulary profiles.

In addition to the above character and word lists, the online vocabulary profiler can also draw reference information from user-provided vocabulary list. That is, a user can compile a word list from his/her own sources and upload it to the system. The online profiler will in turn generate vocabulary profile by referencing information from the user's customized word list as well.

2.3. Character and vocabulary profiling

The vocabulary profiler is designed for open-access on the Internet through a web-based interface. It is developed on the FreeBSD platform using Apache as the front-end web server and MySQL the backend database. The PHP language is used for server-side scripting and JavaScript for some client-side functions.

The profiler is programmed with three modules to guide a user through the vocabulary profiling process: 1) the input processing module, 2) the N-gram segmentation module and 3) the profiling module. The input processing module handles both text input and customized word list from a user. The user can choose to upload a running text or a segmented text where individual words are delimited with blank space (to assist word segmentation). Both the input text and the optional customized word list can be replaced or updated during the (later) profiling process so that individual profiles can be built for each new text uploaded to the system during one working session. However, because of programming limitations, the input module can only process

[3] Allanic's reference lists include 1)《汉字频度表》, 2)《安子介汉字频率表》, 3)《汉字频率表》, 4)《现代汉语字频统计表》, 5) 《华夏文摘单字字频列表》, 6)《外国人基础 汉语用字表 草创》, 7)《汉字等级大纲·甲级字》, 8)《高等学校外国留学生 汉语言专业 教学大纲 》and 9)《法国中学汉语教学大纲》(c.f., http://www.fask.uni-mainz.de/inst/chinesisch/hanzirenzhi_papers_allanic.htm).

[4] Detailed list of the 500 characters is available in《汉语常用字字典》(秦旭平 and 程国富, 2005).

[5] C.f., http://www.moe.edu.cn/edoas/website18/info16840.htm. URL last checked on 2006-04-19.

[6] Original text of 《三字经》is taken from http://zh.wikipedia.org/wiki/%E4%B8%89%E5%AD%97%E7%BB%8F, 《千字文》 from http://www.guoxue.com/gxrm/gx_qzw.htm , and 《百家姓》 from http://zh.wikipedia.org/wiki/%E7%99%BE%E5%AE%B6%E5%A7%93. The three URLs were lasted checked on 2006-04-15.

Simplified Chinese texts encoded in the GB13000 standard and expect the customized word list to contain one word per line only. It is not able to reference other information (such as frequency) contained in a customized word list.

The N-gram segmentation module handles the identification of characters and N-grams from the input text. While character segmentation is done automatically, word segmentation can be accomplished either manually or automatically. If a user provides an already parsed text where individual words are delimited with blank space, the profiler will generate a list of words based on the user's input. However, if he/she chooses to let the system segment the input text automatically, the profiler will first generate lists of bigrams, trigrams and quadrigrams from the text and then require the user to inspect the N-gram lists to decide on the final set of N-grams to be used for profiling. To assist N-gram selection, the profiler will indicate if an individual N-gram is a possible word candidate or not based on information from a consolidated list of more than 140,000 Chinese words and phrases compiled by Da (2005)[7]. At the same time, a companion concordancer is also provided which is capable of constructing KWIC (key word in context) concordances for individual N-grams within the text.

With the text segmented into characters and words, the profiling module will generate both distribution statistics and vocabulary profile about the text. Distribution statistics include measures such as the total and unique number of characters; the total and unique number of words if the text is manually segmented by the user in advance or the total and unique number of N-grams where N represents text strings containing more than one but less than five characters if the text is automatically segmented. Vocabulary profile generated by the profiler includes both frequency and availability information about the set of characters and N-grams found in the text. For example, by comparing the list of characters with Da's (2004) Modern Chinese character frequency list and the HSK word list, the profiler will indicate the proportions of characters among the various frequency ranges and at the four different HSK levels. It will also indicate their presence in the customized word list which is optionally provided by a user. Similar frequency and availability information is also provided for N-grams or words from the text.

2.4. Possible applications

It can be foreseen that the vocabulary profiler has some useful applications in both Chinese language instruction and research. In the case of language instruction, for example, CFL instructors can use it to build character and vocabulary profiles to help them evaluate the suitability of reading texts for CFL learners at different levels and verify if the selected content meets the vocabulary instruction objectives set forth in a teaching syllabus. An instructor can also use it to build a new word list for a textbook through constructing vocabulary profiles for each individual text in the textbook.

In the case of language acquisition research, the vocabulary profiler can be used to measure a CFL learner's vocabulary knowledge. For example, writing samples of a CFL learner can be fed into the profiler so that information about his/her character and vocabulary knowledge (e.g., number of characters/words already learned and their relevant frequency) can be obtained,

[7] The list of 140,000 words and phrases were consolidated from six online sources in the public domain. For details of the six lists and the consolidated list, please visit http://lingua.mtsu.edu/chinese-computing/references/combined/index.php.

which in turn serves as indication of the depth and width of the learner's vocabulary acquisition and performance.

3. Concluding remarks

In this paper, we have described the design and development of a web-based vocabulary profiler for Chinese. While the author hopes that it is useful for Chinese instructors and researchers, the current version needs some further improvements. First, its capacity of Chinese text processing needs to be expanded to include both Unicode and Big5 encoded texts so that it will appeal to a wider array of users. Secondly, a batch processing capability needs to be added so that vocabulary profiling of a series of texts can be made more efficiently. The current version processes one piece of text at a time (though individual texts can be consolidated into one big piece for processing). A batch processing function would make it easier to generate individual profiles for each text so that comparisons can be made among those texts. Lastly, the profiler needs to include more reference lists of words or phrases with frequency information. The current version relies on reference lists that are currently available in the public domain. In comparison with the number of character lists, only a couple of vocabulary lists (notably the HSK word list) are used as reference lists in the current profiler. It would be more beneficial if we could include more specialized reference lists such as those prescribed in other teaching or language proficiency syllabi or those compiled from Chinese textbooks either for Chinese native speakers or CFL learners. While the profiler is programmed in such a way that reference lists can be added easily, it takes much more efforts to come up with those reference lists.

References

Allanic, Bernard. 2005. The "missing link" in the teaching of Chinese characters as a foreign language. Paper presented at the International Interdisciplinary Conference on Hànzì rènzhī - How Western Learners Discover the World of Written Chinese, Germersheim, Germany. (Online version at <http://www.fask.uni-mainz.de/inst/chinesisch/hanzirenzhi_papers_allanic.htm>. Last checked on 2006-04-15)

Cobb, Tom. Web VP V2.5. (Online version at <http://www.lextutor.ca/vp/>. Last checked on 2006-04-15)

Da, Jun. 2004. A corpus-based study of character and bigram frequencies in Chinese e-texts and its implications for Chinese language instruction. The studies on the theory and methodology of the digitized Chinese teaching to foreigners: Proceedings of the 4th International Conference on New Technologies in Teaching and Learning Chinese, ed. by Zhang, Pu, Tianwei Xie and Juan Xu, 501-511. Beijing: The Tsinghua University Press. (Online version at <http://lingua.mtsu.edu/academic/>. Last checked on 2006-04-15)

Da, Jun. 2005. Reading news for information: How much vocabulary a CFL learner should know. Paper presented at the International Interdisciplinary Conference on Hànzì rènzhī - How Western Learners Discover the World of Written Chinese, Germersheim, Germany. (Online version at <http://www.fask.uni-mainz.de/inst/chinesisch/hanzirenzhi_papers_da.pdf>. Last checked on 2006-04-15)

Graves, David Andrew. 2004. Vocabulary Profiles of letters and novels of Jane Austen and her contemporaries. Persuasion On-line. 26,1. (Online version at <http://www.jasna.org/persuasions/on-line/vol26no1/graves.htm>. Last checked on 2006-04-15)

Laufer, Batia and Paul Nation. 1995. Vocabulary size and use: Lexical richness in L2 written productions. Applied Linguistics. 16,307-322.

Morris, Lori and Tom Cobb. 2004.Vocabulary profiles as predictors of TESL student performance. System. 32,1,75-87. (Online version at <http://www.er.uqam.ca/nobel/r21270/cv/VP_LoriTom.htm>. Last checked on 2006-04-15)

Nation, Paul. Range. (Online version at <http://www.vuw.ac.nz/lals/staff/paul-nation/nation.aspx>. Last checked on 2006-04-15)

Nation, Paul. 2001. Learning vocabulary in another language. Cambridge, UK: Cambridge University Press.

Schuemann, Cynthia and Cheryl Benz. 2004. Corpus Linguistics Will Change the Way We Teach. Presentation at the 2004 TESOL Convention. California: Long Beach. (Online version at <http://www.instruction.greenriver.edu/avery/faculty/pres/TESOL04/CorpusArticle.htm>. Last checked on 2006-04-15)

Zhang, Xuetao. 2005. 学习汉字的新概念－为什么要学好５００个基本汉字. Paper presented at the International Interdisciplinary Conference on Hànzì rènzhī - How Western Learners Discover the World of Written Chinese, Germersheim, Germany. (Online version at <http://www.fask.uni-mainz.de/inst/chinesisch/hanzirenzhi_papers_zhangxuetao.htm>. Last checked on 2006-04-15)

国家语言文字工作委员会、(原)国家教育委员会. 1997. 现代汉语常用字表. 语言文字规范手册第三版. 北京: 语文出版社. (Online version at <http://www.moe.edu.cn/edoas/website18/info16840.htm>. Last checked 2006-04-15)

北京语言文化大学汉语水平考试中心. 2000. HSK 汉语 8000 词词典. 北京: 北京语言文化大学出版社.

秦旭平、程国富. 2005. 汉语常用字字典. 北京: 商务印书馆国际有限公司.