Chinese Corpus Resource Guide for Language Educators

by

Hongyin Tao
University of California-Los Angeles

Center for Advanced Language Proficiency Education and Research (CALPER)
The Pennsylvania State University
Copyright 2006. All rights reserved.
Fourth revised version, October 2006
What is a corpus?

A corpus (plural: corpora) is a principled collection of samples of natural language use, either written or spoken, which are usually stored as computer files. A written corpus can be gathered from a number of sources such as news media, literary works, or personal writings. A spoken corpus can be assembled from tape- or video-recorded narratives, interviews, conversations and the like, which would be transcribed into written texts. The size of a corpus can range from tens of millions of words to a few thousand. Larger corpora are usually required for big research projects such as writing dictionaries and major grammars, but so-called "mini corpora" consisting of several thousands of words can be extremely useful for language teachers. Once a corpus is built, we can use software tools to analyze it and produce word frequency lists, concordances and other useful types of output.

How can corpora be useful for Chinese language teaching?

There are many ways in which language teachers can benefit from language corpora. For example, from corpora you can find out how frequently words (or characters) are used in different discourse contexts by way of frequency lists. Table 1 below compares the frequency lists of a spoken corpus (conversation) and a written one (newspaper reports), and we can easily tell how they differ in character frequency. For one thing, while pronouns wo, ni, ta (我, 你, 他) are common in conversations, they do not appear at all in the first 20 most frequent characters in newspaper reports. Although this finding may seem obvious since news reports are normally impersonal, this type of information is not trivial and can even be very useful for textbook design.

<table>
<thead>
<tr>
<th></th>
<th>Conversation</th>
<th>Newspaper</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>是</td>
<td>的</td>
</tr>
<tr>
<td>2</td>
<td>的</td>
<td>一</td>
</tr>
<tr>
<td>3</td>
<td>等</td>
<td>国</td>
</tr>
<tr>
<td>4</td>
<td>个</td>
<td>年</td>
</tr>
<tr>
<td>5</td>
<td>就</td>
<td>存</td>
</tr>
<tr>
<td>6</td>
<td>不</td>
<td>了</td>
</tr>
<tr>
<td>7</td>
<td>一</td>
<td>如</td>
</tr>
<tr>
<td>8</td>
<td>我</td>
<td>人</td>
</tr>
<tr>
<td>9</td>
<td>这</td>
<td>中</td>
</tr>
<tr>
<td>10</td>
<td>了</td>
<td>工</td>
</tr>
<tr>
<td>11</td>
<td>你</td>
<td>有</td>
</tr>
<tr>
<td>12</td>
<td>有</td>
<td>大</td>
</tr>
<tr>
<td>13</td>
<td>的</td>
<td>会</td>
</tr>
<tr>
<td>14</td>
<td>么</td>
<td>上</td>
</tr>
<tr>
<td>15</td>
<td>她</td>
<td>是</td>
</tr>
<tr>
<td>16</td>
<td>对</td>
<td>家</td>
</tr>
<tr>
<td>17</td>
<td>说</td>
<td>这</td>
</tr>
<tr>
<td>18</td>
<td>在</td>
<td>民</td>
</tr>
<tr>
<td>19</td>
<td>后</td>
<td>为</td>
</tr>
<tr>
<td>20</td>
<td>好</td>
<td>的</td>
</tr>
</tbody>
</table>

Table 1: Character frequencies in natural language
If a textbook is computerized, we can run a frequency list and compare it with those of natural language corpora, as illustrated in Table 2. Here, if we judge only the occurrence of personal pronouns, we can be fairly sure that this particular textbook focuses more on speaking than writing. And indeed the two frequency lists look quite similar, apparently differing only in ranking. However, if we look closer, we can find notable differences regarding the use of certain particles and lexical items, which invite further investigation (e.g. the particle jiu 就 is highly frequent in natural conversation but does not appear in the textbook’s list; and the differential frequency of demonstratives zhe 这 and na 那 between the two corpora is also interesting.)

Table 2: Character frequencies in natural language and a textbook

<table>
<thead>
<tr>
<th>Conversation</th>
<th>Newspaper</th>
<th>Textbook</th>
</tr>
</thead>
<tbody>
<tr>
<td>是</td>
<td>的</td>
<td>我</td>
</tr>
<tr>
<td>的</td>
<td>一</td>
<td>你</td>
</tr>
<tr>
<td>那</td>
<td>国</td>
<td>是</td>
</tr>
<tr>
<td>个</td>
<td>年</td>
<td>们</td>
</tr>
<tr>
<td>就</td>
<td>在</td>
<td>的</td>
</tr>
<tr>
<td>不</td>
<td>了</td>
<td>好</td>
</tr>
<tr>
<td>一</td>
<td>和</td>
<td>这</td>
</tr>
<tr>
<td>我</td>
<td>人</td>
<td>有</td>
</tr>
<tr>
<td>这</td>
<td>中</td>
<td>去</td>
</tr>
<tr>
<td>了</td>
<td>工</td>
<td>不</td>
</tr>
<tr>
<td>你</td>
<td>有</td>
<td>一</td>
</tr>
<tr>
<td>有</td>
<td>太大</td>
<td>太</td>
</tr>
<tr>
<td>喝</td>
<td>吗</td>
<td>吗</td>
</tr>
<tr>
<td>么</td>
<td>上</td>
<td>他</td>
</tr>
<tr>
<td>他</td>
<td>是</td>
<td>学</td>
</tr>
<tr>
<td>他</td>
<td>对</td>
<td>家</td>
</tr>
<tr>
<td>说</td>
<td>这</td>
<td>中</td>
</tr>
<tr>
<td>在</td>
<td>民</td>
<td>在</td>
</tr>
<tr>
<td>后</td>
<td>为了</td>
<td>了</td>
</tr>
<tr>
<td>好</td>
<td>地</td>
<td>谢</td>
</tr>
</tbody>
</table>
A key instrument for further investigating the actual use of a particular linguistic structure or a lexical item in real language corpora is a concordance, which is sometimes referred to as KWIC (Key Word In Context). A concordance lists all the occurrences of a search word (key word) in a corpus. Typically the key word is centered on a line and the context is displayed around the keyword. The lines can be sorted according to different criteria (usually by the first word to the right of the keyword) to facilitate visualization; and the exact window of context can also be manipulated according to the user’s needs. In the following we use three examples to illustrate this useful tool.

Example 1: ba3 (把)

Many textbooks describe the ba (把) constructions as requiring a definite noun as its object (e.g., ta ba naige pingguo chi le 她把那个苹果吃了‘She ate the apple’), ignoring many instances of indefinite ba constructions. Although some researchers have rightly pointed out that indefinite objects may also be used with ba, their examples usually sound quite contrived. This is when a collection of real texts comes in handy. With a concordance of the keyword, we can easily find a number of good examples of the construction in question. The following is a sample concordance of the from a corpus of news wire texts, sorted by the first word to the right of the keyword, which evidences the common use of "ba + indefinite object."

Example 2: "把" (ba)

Many textbooks describe the ba (把) constructions as requiring a definite noun as its object (e.g., ta ba naige pingguo chi le 她把那个苹果吃了 ‘She ate the apple’), ignoring many instances of indefinite ba constructions. Although some researchers have rightly pointed out that indefinite objects may also be used with ba, their examples usually sound quite contrived. This is when a collection of real texts comes in handy. With a concordance of the keyword, we can easily find a number of good examples of the construction in question. The following is a sample concordance of the from a corpus of news wire texts, sorted by the first word to the right of the keyword, which evidences the common use of "ba + indefinite object."

Example 3: "把" (ba)

A key instrument for further investigating the actual use of a particular linguistic structure or a lexical item in real language corpora is a concordance, which is sometimes referred to as KWIC (Key Word In Context). A concordance lists all the occurrences of a search word (key word) in a corpus. Typically the key word is centered on a line and the context is displayed around the keyword. The lines can be sorted according to different criteria (usually by the first word to the right of the keyword) to facilitate visualization; and the exact window of context can also be manipulated according to the user’s needs. In the following we use three examples to illustrate this useful tool.

Example 1: ba3 (把)

Many textbooks describe the ba (把) constructions as requiring a definite noun as its object (e.g., ta ba naige pingguo chi le 她把那个苹果吃了 ‘She ate the apple’), ignoring many instances of indefinite ba constructions. Although some researchers have rightly pointed out that indefinite objects may also be used with ba, their examples usually sound quite contrived. This is when a collection of real texts comes in handy. With a concordance of the keyword, we can easily find a number of good examples of the construction in question. The following is a sample concordance of the from a corpus of news wire texts, sorted by the first word to the right of the keyword, which evidences the common use of "ba + indefinite object."

Example 2: "把" (ba)

Many textbooks describe the ba (把) constructions as requiring a definite noun as its object (e.g., ta ba naige pingguo chi le 她把那个苹果吃了 ‘She ate the apple’), ignoring many instances of indefinite ba constructions. Although some researchers have rightly pointed out that indefinite objects may also be used with ba, their examples usually sound quite contrived. This is when a collection of real texts comes in handy. With a concordance of the keyword, we can easily find a number of good examples of the construction in question. The following is a sample concordance of the from a corpus of news wire texts, sorted by the first word to the right of the keyword, which evidences the common use of "ba + indefinite object."

Example 3: "把" (ba)
Example 2: *qishi* (其实，‘actually’)

This next set of concordance lines shows that *qishi* (其实，‘actually’) is more often used as a discourse conjunction (located outside the main clause and linking large domains of discourse) than as a constituent conjunction (inside a clause and linking clause-internal elements). We can tell the difference between the two uses simply by looking at the display of the concordance lines. In the discourse conjunction use, for example, there are often punctuation marks before and after the key word, reflecting the independent status of the item in question. By contrast, the constituent conjunction is usually embedded in a clause as several of the following excerpts indicate.

真难为了，还真找来了。其实，当时咋处理的，根本没
他们还觉得什么全懂呢，其实，什么也不懂。组织老干部去搞
待业青年还挺欢迎他，其实，他连怎么写计划书，怎么运水
本，对，35 对，请客！其实，就是写给在我身边转来转去的
我们过去说是一星期工作六天，其实，加上星期三下午业务学习，
没有多少人的因素。其实，除了自由市场的小买卖
开始不好意思，很快就习惯了。其实，当美术模特儿非常累，有时
还不是走得影子也没有！其实，我早作了走的准备，告诉他
捕渔赚钱，很自由的。其实，我们回去，当局也知道，并不
舟山群岛的无人岛避风”。其实，真追究，没人相信，我们这船
富子女打他是“阶级报复”呢！其实，咱们的法是保护公民的合法权益
这是我们民族的习惯吧？其实，来了，我们也要调解，就算开庭
大叫了一场。他是让别人听。其实，我知道他是真爱上了我。
好像我们也是死尸似的。其实，尸体不可怕，人们是觉着我们这
全当我们是“死啦死啦地干活”。其实，我们才不等你说“默哀毕，散会
考试不及格，否则不会干这个。其实，我们有学问！有教养！
不安心工作的司机和电工多。其实，你不安心，去考研究生嘛，
他比妈妈瘦而且黑。其实，我真没什么要谈的。我的经历
现在的妇女们，都挺爱嫁我们；其实，嫁了还不是守活寡？她守活寡
觉得他好像完全是新的东西。其实，好多历史事件过去都学过。但
我不叫她当模特儿，这工作其实，也没多大劲。在时装表演队
女的，又有文化，就干了护士队。其实，也没那么正规，护士队也参加
你托托人，到时候就端上来了。其实，你不安心，去考研究生嘛，
卖这个，不行，利润看着大，其实，不大，做了三个月，我明白了
担把美军打到三八线……”，其实，不是那么回事，志愿军在最厉害的
这是以后的事…… 学生生活
文化大革命”…… 现在离休了，
许多人欢喜”石磨蓝”，
几乎所有的人都误会我是导游员，
错了，穿得也够花哨的；
这些工作全是老书记 - -
搞“斗、批、改” 时，号召，
派好几批人到各地去开交流会 - -
想了八百六十五回了。
其实 早已结束。名义上的结束，是一
其实 还能干事情，我和领导同志讲了
其实 那不是染的，是磨的，用“轻石”
其实 我一点也不知导游的技术，
其实 呢，就是本市的。女流氓。用不
其实 是强制性的，送我们下乡“练红心”
其实 是摸同行的底。实行改革了。
其实 她只要再多说几句，我可能给她

Example 3: kan-kan 看看
The third set of concordance lines reveals that the reduplication of kan (看, 'see, look') has a number of uses: directing attention (‘你们看看’), indicating a prolonged intensive action (‘拿到太阳底下再看看’), indicating a trivial action (‘看看表’), among others.

坐一会儿，拿上“大参考” 什么的
是想请文化界的人们下来听听
厅长调走了，新领导主张等等
电器维修》，你们上别处
电器也不是很便宜…… 你们
喘了口气， 拿到太阳底下再
你先准备、准备，等会儿我
大喇”，就坐着。老师进
推陈出新”呢! 摆好了，退几步
弄套光光鲜鲜的中山装， 叫他们
我把团委的介绍信给他看，他
否则会有麻烦。共方讲过：”
觉得现在胡耀邦、赵紫阳经常下来
这些孩子、鸡、猪，再回来
结婚以前，每年有假，来
孩子们说：不成? 老师您也不
看着编辑部的牌子，进来
编辑部的牌子，进来看看同志们，
呢！我当时就流泪了，他

看着 ， 十二点下班回家。不去可不行
看着 ， 反映反映我们在发展生产方面
看着 ， 不为最先， 不为最先, 不为最后
看着 ， 比我们搭得凶，《神拳》小人
看着 ， 还是他妈的“录取”! 这才乐起
看着 “老师走后，我问”大喇” 都准备
看着 ， 指挥我伸腿举胳膊什么的，说，
看着 “不行， 手高点儿”, 再看着，
看着 别看咱不代表国家, 咱还代表着
看着 ， “哥们儿! 好样的! 兄弟
看着 “带不回抛掉好了。” 他们很知道
看着 走走是大好事， 同时呢
看着 住房; 看看住住呢，再躲出去．
看着 ， 结了婚就不来了。不是四年一
看着 我们的孩子都多大了! 家里的
看着 同志们， 看看怎么编书……“这位
看着 怎么编书……” 这位老头儿，也是
看着 我， 以为我有重病, 伸手扶我
Because computer programs can search the corpus quickly, we can obtain a large number of examples of real language use in a very short period of time. This in turn saves valuable time for analyzing the language and preparing teaching materials. Furthermore, as language teachers, we can actually build a learner corpus from our students' language productions and use the various corpus-handling skills to uncover typical learner errors. In short, language corpora constitute a great data source for us to explore. And they benefit not only teachers and researchers but also motivate learners. In fact, corpora are increasingly being used as learning tools for students. Given that students nowadays tend to be well equipped with computer skills, they should be encouraged to make informed uses of corpus resources in conducting their own research and enhancing their learning. Some of the most illuminating examples can be found at Tim Johns' Data Driven Learning (DDL) website. Even though the examples are mostly English, similar methods can be easily applied to Chinese. (See Other Resources below for the URL to the DDL website.)

**Are there Chinese language corpora currently available to Chinese language teachers?**

Yes, there are quite a few Chinese corpora that are freely available on the internet. Here is a list of some of them.

**From Mainland China:**

► The Beijing Language and Culture University Institute of Language Information Processing has a searchable written Chinese corpus comprised of texts from the *People's Daily*, pre-modern and modern short stories and novels, encyclopedias, and a few other genres. There are two links to this corpus: one for word-based searches and the other for character-based searches.  

► The Peking University Modern Chinese Corpus is another source.  
  **URL:**  [http://ccl.pku.edu.cn/ccl%5Fcorpus/xiandaihanyu/](http://ccl.pku.edu.cn/ccl%5Fcorpus/xiandaihanyu/)
An online search system for the modern Chinese corpus developed by the Chinese National Commission on Language (国家语委) is available at:
URL: http://219.238.40.213:8080/

From Taiwan:

- The Academia Sinica has a Web-based Balanced Corpus of Modern Chinese (平衡语料库), consisting of texts mostly from Taiwanese newspapers. This corpus can be searched based on parts of speech information. It is also possible to search reduplicated forms.
  URL: http://www.sinica.edu.tw/ftms-bin/kiwi.sh

- The Academia Sinica also has a Digital Resource Center for Global Chinese Language Teaching and Learning (全球華語文數位教與學資源中心). It provides word frequency lists and a web-based collection of reading materials that can be searched with grammatical and semantic information.
  URL: http://elearning.ling.sinica.edu.tw/

From other parts of the world:

- The Lancaster Corpus of Mandarin Chinese (LCMC) was constructed by Tony McEnery and Richard Xiao, Lancaster University, UK. LCMC is a balanced corpus of Modern written Chinese, consisting of texts from mainland China. The corpus, including genres such as press reportage, press editorials, religious passages, skills texts, trade and hobbies passages, popular lore, biographies and essays, fictional literature, and so forth, is designed as a Chinese match of the Freiburg-LOB Corpus of British English (FLOB). It has a Web-based search interface with parts of speech information.
  URL: http://bowland-files.lancs.ac.uk/corplang/cgi-bin/conc.pl

- Serge Sharoff of Leeds University (UK) provides a web interface to search for portions of two news wire corpora (Xihua of mainland China and the Central News Agency in Taiwan). This interface also provides statistical information on search items.
  URL: http://corpus.leeds.ac.uk/query-zh.html

Consolidated Corpus from Multiple Chinese Speaking Regions:

- The LIVAC Corpus, or Linguistic Variation in Chinese Speech Communities synchronous corpus, contains texts from representative Chinese newspapers and electronic media of Hong Kong, Taiwan, Beijing, Shanghai, Macau and Singapore. It also provides concordance and frequency analyses. Because this corpus is constantly updated, it is possible to trace the use of expressions over time (within the time span of the corpus itself).
Multilingual Corpora Involving Chinese and Other Languages:

► The Virtual Language Centre in Hong Kong has a searchable online database which contains parallel texts of Chinese, English, Japanese, and French. This can be a useful resource for translation studies and for comparative analysis.

► The Babel English-Chinese Parallel Corpus consists of 327 English articles and their translations in Mandarin Chinese. The corpus contains a total of 544,095 words (253,633 English words and 287,462 Chinese tokens). It is online at:
   URL: http://bowland-files.lancs.ac.uk/corplang/babel/babel.htm

What Corpus Tools are Available as Freeware?

While there are many commercial software programs that can be used to prepare and/or analyze Chinese corpora, there are also a few programs which are available on the Web or as free downloads. A few of them are particularly valuable.

► DimSum Chinese Language Tool, by Erik Peterson, is a very useful Java-based program that can do word segmentation, English annotation, word lists, and Hanzi to Pinyin conversion, among other features. It runs on Windows, MacOS, and Linux systems.
   URL: http://www.mandarintools.com/dimsum.html

► ConcApp, by Chris Greavies, is a Windows-based free software program that can perform concordance (key word in context), collocation, and word frequency analyses.
   URL: http://www.edict.com.hk/PUB/concapp/

► AntConc, by Laurence Anthony, is a free program for Windows and Linux systems that can provide concordance, collocation, N-Gram and key word analyses. It works with multilingual texts.
   URL: http://www.antlab.sci.waseda.ac.jp/software.html

► A Corpus Worker’s Toolkit, by Hongyin Tao, is a collection of free software tools that can process Chinese texts, provide annotation, and perform a variety of corpus analysis tasks.
   URL: http://www.humnet.ucla.edu/alc/chinese/ACWT/ACWT.htm

► Conc is a concordance program for the Apple Macintosh. It was developed by SIL International and can be downloaded from their site.
   URL: http://www.sil.org/computing/conc/
Which Other Resources Are Available?

There are numerous websites, books, and articles on “corpus linguistics”, “language corpora”, and “Chinese language and linguistics” available. Here is just a small selection:

Websites:

► Corpus4U.Org is a Web-based discussion forum for Chinese and English corpus linguistics and applications. It is based in mainland China and has over 2500 registered users as of May 2006.
  URL: http://www.corpus4u.org/

► Marjorie K.M. Chan's ChinaLinks has a wealth of information about Chinese language and linguistics.
  URL: http://chinalinks.osu.edu

► Hongyin Tao’s Corpus Linguistics Course Web Page gives a brief introduction to East Asian language-based corpus linguistics.
  URL: http://www.bol.ucla.edu/~ht37/teach/222/222_info.html

► Tianwei Xie’s Learning Chinese On-line web page provides a variety of links to Web sites that are related to Chinese learning and teaching.
  URL: http://www.csulb.edu/~txie/on-line.htm

► Tim Johns' Virtual DDL (Data Driven Learning) Library has inspiring (non-Chinese) examples.
  URL: http://web.bham.ac.uk/johnstf/ddl_lib.htm

Books:

► Concordances in the Classroom: A Resource Book for Teachers by Chris Tribble and Glyn Jones (Houston: Athelstan, 1997) has many ideas for teachers with an interest in using electronic texts in the language classroom, even though it is English based.


► Exploring Spoken English by Ronald Carter and Michael McCarthy (Cambridge: CUP, 1997) is a practical guide to natural spoken English drawn from the CANCODE corpus. Although providing examples in English, it gives useful insights into using corpus data in teaching.

► Yuliaoku Yuyanxue (语料库语言学 Corpus Linguistics) by Huang Changning and Li Juanzi (Beijing: Commercial Press, 2002) is another introductory text to corpus linguistics.
Articles:


**Corpus Tutorial:**
The Center for Advanced Language Proficiency Education and Research developed a "Corpus Tutorial". This online tutorial is designed for language teachers and is aimed at enabling language educators to work with their own and other available corpora. The program consists of ten units of self-study devoted to corpus construction, and basic corpus-analytical techniques and applications. A beta-version is now available on CALPER's website at http://calper.la.psu.edu/corpus.php.
Contact CALPER Chinese Project Director:
Professor Hongyin Tao
Department of Asian Languages and Cultures
University of California-Los Angeles
290 Royce Hall
Los Angeles, CA 90095-1540
Email: ht37@ucla.edu

Contact CALPER:
James P. Lantolf and Karen E. Johnson, Co-directors
Center for Advanced Language Proficiency Education and Research
The Pennsylvania State University
5 Sparks Building
University Park, PA 16802-5203
Phone: (814) 863-1212  -  Fax: (814) 865-1316
E-Mail: calper@psu.edu

For more information visit our website at:
calper.la.psu.edu

This information was compiled and prepared for publication with funding from the U.S. Department of Education (CFDA 84.229, P229A060003). However, the contents do not necessarily represent the policy of the Department of Education, and one should not assume endorsement by the Federal Government.