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

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**.

	Conversation	Newspaper
1	是	的
2	的	一
3	那	国
4	个	年
5	就	在
6	不	了
7	一	和
8	我	人
9	这	中
10	了	工
11	你	有
12	有	大
13	啊	会
14	么	上
15	他	是
16	对	家
17	说	这
18	在	民
19	后	为
20	好	地

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.)

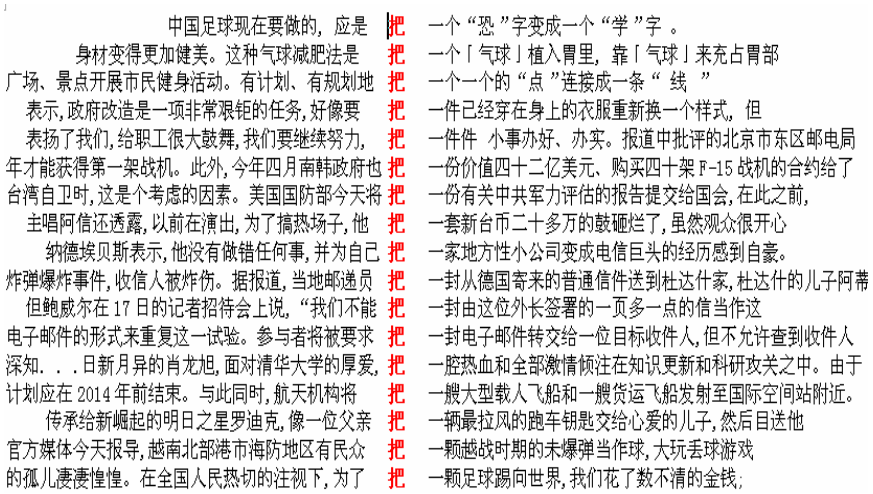
Table 2: Character frequencies in natural language and a textbook

Conversation	Newspaper	Textbook
是	的	我
的	一	你
那	国	是
个	年	们
就	在	的
不	了	好
一	和	这
我	人	有
这	中	去
了	工	不
你	有	一
有	大	太
啊	会	吗
么	上	他
他	是	学
对	家	儿
说	这	中
在	民	在
后	为	了
好	地	谢

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 把 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."



中国足球现在要做的，应是	把	一个“恐”字变成一个“学”字。
身材变得更加健美。这种气球减肥法是	把	一个「气球」植入胃里，靠「气球」来充占胃部
广场、景点开展市民健身活动。有计划、有规划地	把	一个一个的“点”连接成一条“线”
表示，政府改造是一项非常艰钜的任务，好像要	把	一件已经穿在身上的衣服重新换一个样式，但
表扬了我们，给职工很大鼓舞，我们要继续努力，	把	一件件小事办好、办实。报道中批评的北京市东区邮电局
年才能获得第一架战机。此外，今年四月南韩政府也	把	一份价值四十二亿美元、购买四十架F-15战机的合约给了
台湾自卫时，这是个考虑的因素。美国国防部今天将	把	一份有关中共军力评估的报告提交给国会，在此之前，
主唱阿信还透露，以前在演出，为了搞热场子，他	把	一套新台币二十多万的鼓砸烂了，虽然观众很开心
纳德埃贝斯表示，他没有做错任何事，并为自己	把	一家地方性小公司变成电信巨头的经历感到自豪。
炸弹爆炸事件，收信人被炸伤。据报道，当地邮递员	把	一封从德国寄来的普通信件送到杜达什家，杜达什的儿子阿蒂
但鲍威尔在17日的记者招待会上说，“我们不能	把	一封由这位外长签署的一页多一点的信当作这
电子邮件的形式来重复这一试验。参与者将被要求	把	一封电子邮件转交给一位目标收件人，但不允许查到收件人
深知. . .日新月异的肖龙旭，面对清华大学的厚爱，	把	一腔热血和全部激情倾注在知识更新和科研攻关之中。由于
计划应在2014年前结束。与此同时，航天机构将	把	一艘大型载人飞船和一艘货运飞船发射至国际空间站附近。
传承给新崛起的明日之星罗迪克，像一位父亲	把	一辆最拉风的跑车钥匙交给心爱的儿子，然后目送他
官方媒体今天报导，越南北部港市海防地区有民众	把	一颗越战时期的未爆弹当作球，大玩丢球游戏
的孤儿凄凄惶惶。在全国人民热切的注视下，为了	把	一颗足球踢向世界，我们花了数不清的金钱；

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
他们还觉着什么全懂呢，	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	，他连怎么造计划书，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,
没有多少人性的因素。	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" 嘛!	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	，我知道他是真爱上我了。
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	也是平常菜、传统菜，饭馆
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，志愿军在最历害的

这是以后的事…… 学生生活	其实	早已结束。名义上的结束， 是一
文化大革命" … … 现在离休了，	其实	还能干事情， 我和领导同志讲了
许多人欢喜" 石磨蓝" ，	其实	那不是染的，是磨的，用“轻石”
几乎所有的人都误会我是导游员，	其实	我一点也不知导游的技术,
错，穿得也够花哨的;	其实	呢，就是本市的。女流氓。用不
这些工作全是老书记 - -	其实	和他同岁搞的， 请他介绍吧。
搞”斗、批、改" 时，号召，	其实	是强制性的，送我们下乡“练红心”
派好几批人到各地去开交流会 - -	其实	是摸同行的底。实行改革了。
想了八百六十五回了。	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	看看	，十二点下班回家。不去可不行
是想请文化界的人们下来听听	看看	， 反映反映我们在发展生产方面
厅长调走了， 新领导主张等等	看看	， 不为最先， 不为最先, 不为最后
电器维修》。你们上别处	看看	， 比我们搭得凶， 《神拳》小人
电器也不是很便宜… … 你们	看看	， 又是一起车祸， 是香港车
喘了口气， 拿到太阳底下再	看看	， 还是他妈的" 录取"! 这才乐起
你先准备、准备， 等会儿我	看看	。" 老师走后， 我问"大喇” 都准备
大喇" ， 就坐着。老师进	看看	， 指挥我伸腿举胳膊什么的, 说,
推陈出新" 呢! 摆好了，退几步	看看	，" 不行， 手高点儿", 再看看,
弄套光光鲜鲜的中山装， 叫他们	看看	， 别看咱不代表国家, 咱还代表着
我把团委的介绍信给他看， 他	看看	， 说：“ 哥们儿! 好样的! 兄弟
否则会有麻烦。共方讲过： "	看看	， 带不回抛掉好了。" 他们很知道
觉得现在胡耀邦、赵紫阳经常下来	看看	、走走是大好事， 同时呢
这些孩子、鸡、猪， 再回来	看看	， 住住; 看看住住呢，再躲出去.
结婚以前， 每年有假， 来	看看	， 结了婚就不来了。不是四年一
孩子们说：不成? 老师您也不	看看	， 我们的孩子都多大了! 家里的
着编辑部的牌子， 进来	看看	同志们， 看看怎么编书…… “这位
编辑部的牌子，进来看看同志们，	看看	怎么编书… …" 这位老头儿, 也是
呢！我当时就流泪了， 他	看看	我， 以为我有重病, 伸手扶我

到天安门城楼子底下，站在那儿	看看	车再往回走。也是过两回马路
乱开心，这时也没话了，他	看看	四周围，把捆篾条的绳子都弄下
下了班，打打毛线，打打网球，	看看	戏和电视，就过了二十七天
现在也是这样，你们自己去摊上	看看	就明白了，要的价钱全一样。
上海人胆量大，干劲足。	看看	高第街的市场，人家一个个和上战
旁看观众。真的，你们	看看	他们的神态，看看他们的眼睛……

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.
URL: http://202.112.195.8:8089/ccir_login?input=*

► The *Peking University Modern Chinese Corpus* is another source.
URL: 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 (平衡語料庫), 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).

URL: http://www.rcl.cityu.edu.hk/english/livac/

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.
URL: http://www.edict.com.hk/concordance/default.htm

►*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.

►*Corpus Linguistics* by Douglas Biber, Susan Conrad, and Randi Reppen (Cambridge: CUP Press, 1998) is an introductory text to corpus linguistics.

►*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:

Carter, Ronald and Michael McCarthy (1995). Grammar and the Spoken Language. Applied Linguistics, *16* (2), 141-158.

Chan, Marjorie K.M. (2002). Concordancers and concordances: Tools for Chinese language teaching and research. *Journal of the Chinese Language Teachers Association*, *37* (2), pp. 1-58.

Chen, Jing and Hongyin Tao (2004). A usage-based study of preposed verbal quantification structures in Chinese. *Journal of Chinese Language and Computing*, *14* (2), 125-137, 2004. [Special Issue: Corpora, Language Use, and Grammar. Edited by Hongyin Tao.]

McCarthy, Michael and Ronald Carter (2001). "Size isn't everything: Spoken English, corpus and the classroom." *TESOL Quarterly* , *35*, 337-340.

McCarthy, Michael and A. O'Keeffe (2004). Research in the teaching of speaking. *Annual Review of Applied Linguistics*, *24*, 26-43.

McEnery, A., Z. Xiao & Y. Tono (2005). *Corpus-based Language Studies : An advanced resource book* . London : Routledge.

Ming, Tao & Hongyin Tao (forthcoming). Developing a Chinese Heritage Language Corpus: Issues and a Preliminary Report. University of California, Los Angeles, Asian Languages and Cultures Department.

Sun, Maosong (孙茂松) (1998). “取决”与“来源”小议 (Notes on *qujue* and *laiyuan*). 中国语文 (*Chinese Language*), *6*.

Tao, Hongyin (2000). Adverbs of absolute time and assertiveness in vernacular Chinese: A corpus based study. *Journal of the Chinese Language Teachers Association*, *3*, 53-73.

Tao, Hongyin (2001). Emergent grammar and verbs of appearing. *Contemporary Research in Modern Chinese*, (Japan), *2*, 89-100.

Tao, Hongyin (2002). The semantics and pragmatics of relative clause constructions in Mandarin narrative discourse. *Contemporary Research in Modern Chinese*, (Japan), *4*, 47-57.

Tao, Hongyin (2004). Fundamentals in spoken discourse analysis. *Yuyan Kexue* (Linguistic Sciences), *3,* 50-67.

Tao, Hongyin (2005). The gap between natural speech and spoken Chinese teaching material: Discourse perspectives on Chinese pedagogy. *Journal of the Chinese Language Teachers Association, 40*, 1-24.

Xiao, Zhonghua & Anthony McEnery (2004). *Aspect in Mandarin Chinese: A corpus-based study* . Amsterdam : John Benjamins.

Xiao, Zhonghua & Anthony McEnery (2006). Collocation, semantic prosody and near synonymy: A cross-linguistic perspective. *Applied Linguistics,* 27(1), 103-129.

Wang, Lixun (2001). Exploring parallel concordancing in English and Chinese, *Language Learning and Technology*, 5, 174-184.

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.