Corpus-based tasks for learning Chinese: a data-driven approach

Smith, Simon

Original citation:
Smith, S. (2011) Corpus-based tasks for learning Chinese: a data-driven approach. The Asian Conference on Technology in the Classroom Official Conference Proceedings 2011. 48-59. ISSN: 2186-4705

Proceedings URL: http://iafor.org/actc_proceedings.html

Corpus-based tasks for learning Chinese: a data-driven approach

Simon Smith

(pp. 48-59)

The Asian Conference on Technology in the Classroom

Official Conference Proceedings 2011

ISSN: 2186-4705

Proceedings URL: http://iafor.org/actc_proceedings.html

iafor

The International Academic Forum

www.iafor.org

Corpus-based tasks for learning Chinese: a data-driven approach

Abstract

Over the last couple of decades, there have been many studies on the utility of data-driven learning (DDL) approaches to the acquisition of English and other Indo-European languages. Very little research has touched on DDL for Chinese, or indeed any corpus-based approaches to learning it. This is surprising, given the otherwise large choice of IT applications, including flashcards, online dictionaries, and stroke order practice software: certainly, it seems, people do wish to learn Chinese with computers.

Certain features of the Chinese language make it especially suited to a data-driven approach. In DDL, learners typically explore collocational and colligational patterns among words, but would not expect to be able to look at the internal structure of words using a corpus tool. The logographic Chinese writing system, however, allows the learner to investigate the ways that characters/morphemes pattern to form words.

We offered several corpus-based tasks to intermediate-level Mandarin learners, alongside traditional-communicative conversation classes. We describe these tasks, as well as some of the corpora and corpus interfaces used in our approach.

Introduction

The use of corpora and data-driven learning has been widespread in language teaching and learning for many years now. The importance of deriving language dictionaries, grammars and teaching materials from authentic sources is widely accepted. Since Johns (1991), the value of autonomous, student-centred language learning has been much discussed, with most scholars agreeing that the use of corpora is a mainstay of learner **autonomy** and task **authenticity**. The use of linguistic corpora in language learning often takes the form of concordance analysis by students, or data driven learning (DDL). Johns (1991) likens the language learner (on the DDL model) to a researcher, analyzing target language data and becoming familiar with the language through the regularities and consistencies encountered. Early users of DDL include Aston (1995), who assembled small corpora from CD-ROM collections of texts (on murder stories and hepatitis, among other topics), and assigned exercises on collocation and grammatical patterns on these topics. Tribble (1997) demonstrated so-called "quick and dirty" ways to assemble 30-40 thousand word themed corpora, using the Microsoft Encarta software. According to research (e.g. Bernardini, 1997; Cheng, Warren, & Xu, 2003), DDL can lend a strong sense of achievement to the serendipitous linguistic discovery experienced by some learners.

Quite a lot of research has been done on DDL, in terms of both learner evaluation and to a lesser extent learning outcomes; of 67 empirical studies of corpus use in the classroom located by Boulton (2009), "the vast majority of results are encouraging". A large proportion of DDL research and teaching practice, however, has been on English, given that language's position as international lingua franca. Chinese is also an emerging and widely studied world language, and many corpora of various sizes and purposes are available on the web and elsewhere (Chen and Huang, 2000; McEnery and Xiao, 2004; Sharoff, 2006). These resources have been used in dictionary production (Xiao,

Rayson, and McEnery, 2008) and grammatical exposition (Xiao and McEnery, 2004).

There have been some pilot studies on the use of authentic resources in Chinese, but corpora have not been used extensively for language teaching and learning. In one example, Wang (2001) developed a Chinese-English parallel corpus (a corpus where all the documents from one language are translated into the other). Wang uses extracts from the corpus to demonstrate differences between the English "now" and Chinese 现在. Students find that not only are there differences in shades of meaning, but the grammatical contexts in which they are used vary too.

Tao (2005), as part of the CALPER (Center for Advanced Language Proficiency Education and Research) project at Penn State, developed a 300000-word corpus of authentic examples of contemporary spoken Chinese. These materials were used to teach the features of natural conversation to advanced students, and to emphasize aspects of grammar such as the correct use of the particle 了.

Despite the efforts of the above-mentioned Chinese corpus researchers, take-up of corpus approaches and DDL in Chinese teaching has been limited. One reason for this is that many Chinese teaching institutions are constrained to a highly traditional teaching model. Often as a result of institutional policy on curriculum and materials, and because of the content of available textbooks, teachers of Chinese may adopt what in the EFL world would be seen as rather outdated methods, including pattern drilling, repetition and memorization. Another difficulty is that longer serving teachers may have become set in their ways, preferring known and trusted techniques, and taking up new approaches only reluctantly. Xi'an Jiaotong Liverpool University (XJTLU), Suzhou, where we are currently piloting materials, does not suffer from these constraints. As part of a western style, research-led university, we are free to adopt the materials and approaches that teachers and team leaders deem appropriate, including task-based, corpus-based and other novel approaches to learning, taking into account the genuine needs of students. The same is true of Xi'an Jiaotong University, Xi'an, and the Liverpool Confucius Institute, where we are planning to conduct more far-reaching pilot studies with much larger groups of students than are available at XJTLU.

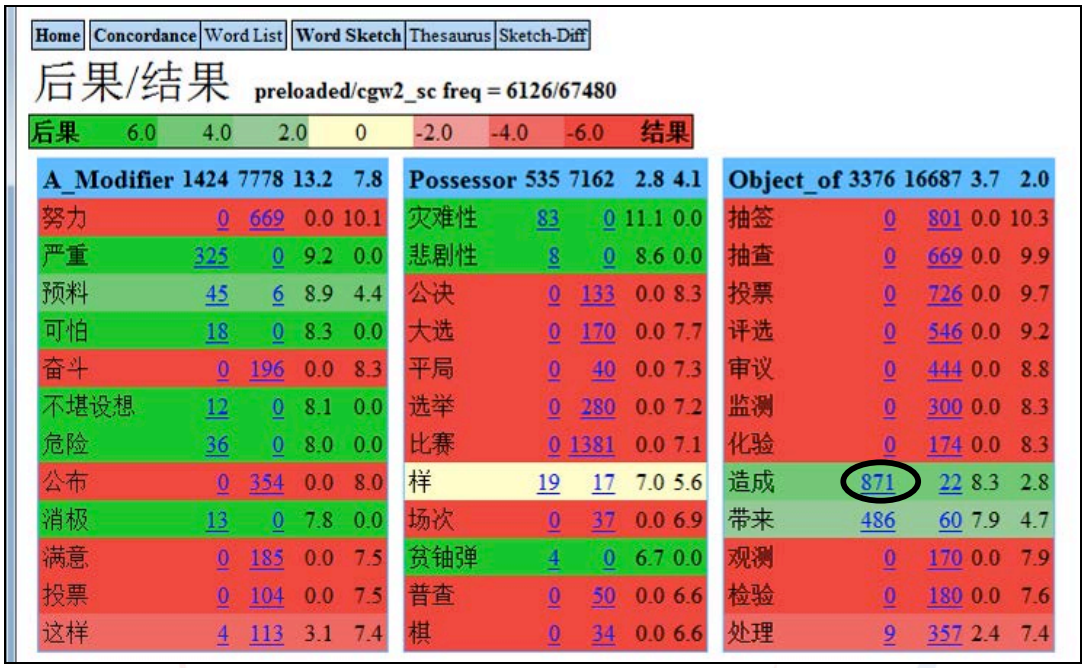
A number of existing web platforms provide various corpus use functions for learners of English and other alphabetic languages; one of the best known is Tom Cobb's Compleat Lexical Tutor, at www.lextutor.ca. We are not aware of any similar platforms for Chinese, but the Sketch Engine corpus query tool (SkE; Kilgarriff et al, 2004) is a useful DDL solution, having been successfully used in English classrooms by Smith (forthcoming) and Thomas (2008). SkE has some functions which can make corpus data more accessible to language learners than simple concordancing.

In this paper, we describe a number of corpus-based tasks, which make use of two of the special SkE functions, Sketch Differences and Word Sketches, as well as some modifications to traditional concordancing. These tasks can be used as supplements to traditional textbook themed units. Tasks corresponding to Wang & Shen (2008) units 17 (weather), 18 (health) and 20 (transport) were prepared, as well as Qiu et al (2008) unit 19, which is about hiking. The following sections describe the tasks in some detail.

Corpus task 1: *Please compare the use of* 结果 and 后果 [both meaning *consequence* or *result*]*, and find example sentences.*

The learner can complete this task by using two functions of Sketch Engine, Sketch Differences and the traditional concordance. The Sketch Differences output in Figure 1 shows the different

collocational properties of two words meaning *result* or *consequence* (see Xiao and McEnery, 2006). Those words shown in green are more likely to collocate with后果, which has a negative connotation, while the red backgrounded items are more likely to pair with the more neutrally oriented 结果. The learner will see immediately, for example, that 后果 tends to collocate with items such as 严重 (*serious*)，可怕 (*frightening*)，危险 (*dangerous*), and is frequently the object of the negatively oriented 造成 (*cause* [unpleasant consequence]).结果 is associated with 比赛 ([results of] a *match*), 投票 (*election*),满意 (*satisfactory*) and 公布 (*announce*).


Home | Concordance | Word List | Word Sketch | Thesaurus | Sketch-Diff

后果/结果 preloaded/cgw2_sc freq = 6126/67480

后果 6.0 4.0 2.0 0 -2.0 -4.0 -6.0 结果

A_Modifier 1424 7778 13.2 7.8					Possessor 535 7162 2.8 4.1					Object_of 3376 16687 3.7 2.0				
努力	0	669	0.0	10.1	灾难性	83	0	11.1	0.0	抽签	0	801	0.0	10.3
严重	325	0	9.2	0.0	悲剧性	8	0	8.6	0.0	抽查	0	669	0.0	9.9
预料	45	6	8.9	4.4	公决	0	133	0.0	8.3	投票	0	726	0.0	9.7
可怕	18	0	8.3	0.0	大选	0	170	0.0	7.7	评选	0	546	0.0	9.2
奋斗	0	196	0.0	8.3	平局	0	40	0.0	7.3	审议	0	444	0.0	8.8
不堪设想	12	0	8.1	0.0	选举	0	280	0.0	7.2	监测	0	300	0.0	8.3
危险	36	0	8.0	0.0	比赛	0	1381	0.0	7.1	化验	0	174	0.0	8.3
公布	0	354	0.0	8.0	样	19	17	7.0	5.6	造成	871	22	8.3	2.8
消极	13	0	7.8	0.0	场次	0	37	0.0	6.9	带来	486	60	7.9	4.7
满意	0	185	0.0	7.5	贫铀弹	4	0	6.7	0.0	观测	0	170	0.0	7.9
投票	0	104	0.0	7.5	普查	0	50	0.0	6.6	检验	0	180	0.0	7.6
这样	4	113	3.1	7.4	棋	0	34	0.0	6.6	处理	9	357	2.4	7.4

Figure 1 Sketch Differences for 结果 and 后果

Clicking on the blue link indicated by the ellipsis (871 is the frequency of the collocation in the corpus) takes the user to a concordance of sentences from the corpus in which 后果 is the object of 造成, as shown in Figure 2.

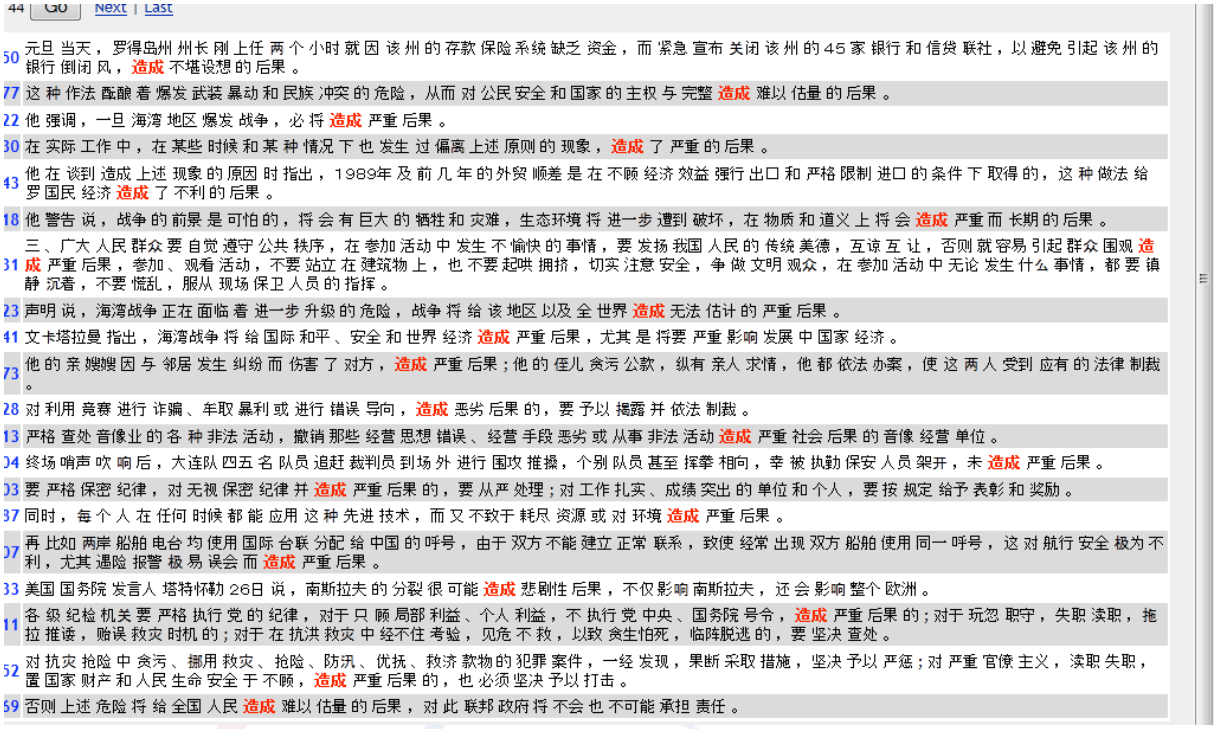


Figure 2 Concordance for后果 as the object of 造成

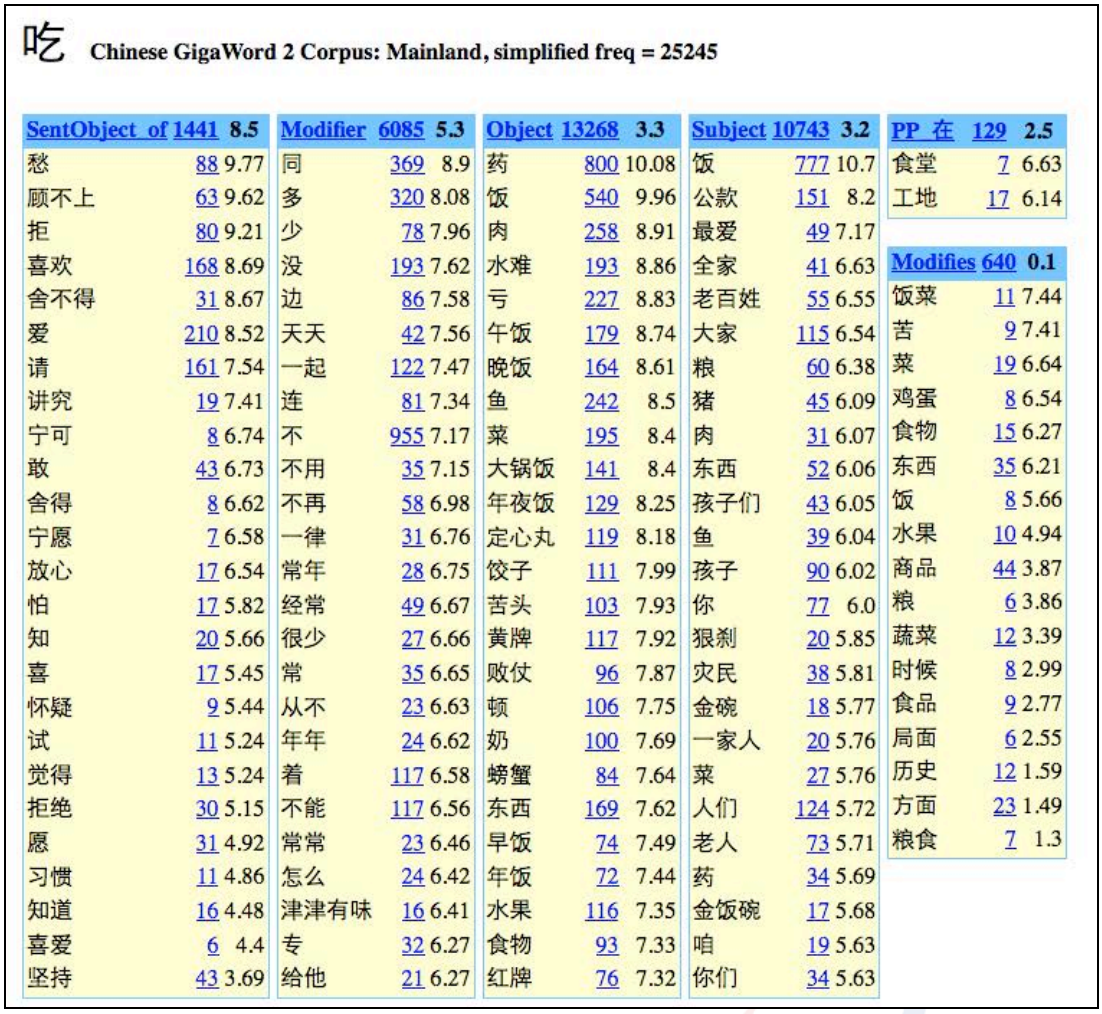
We do not, of course, expect intermediate learners to be able to read and understand every sentence in the concordance. Instead, we hope that they will look at the collocations and absorb some of the recurring patterns. We also set some general questions about the concordance, for example

1. In the first concordance line, where and when did the events take place? [New Year's Day; Rhode Island]

2. In the fourth from last line, starting 美国…, what is 国务院? How about 发言人? [State Department; spokesperson]

Learners can make an educated guess at the answers and confirm by looking them up on the web. Discovering this information as part of an attempt to understand an authentic (admittedly very short) text means that the new knowledge is likely to be retained, we believe. Also, it may inspire learners to ask themselves questions about particular words or structures they encounter; indeed, one activity we found effective was to ask learners to create questions based on concordance output for their peers.

Corpus task 2: *Please study the Word Sketch for* 吃 [*eat*]. *Classify the objects into several categories, and study example sentences in the concordance.*



吃 Chinese GigaWord 2 Corpus: Mainland, simplified freq = 25245

SentObject_of 1441 8.5		Modifier 6085 5.3		Object 13268 3.3		Subject 10743 3.2		PP_在 129 2.5	
愁	88 9.77	同	369 8.9	药	800 10.08	饭	777 10.7	食堂	7 6.63
顾不上	63 9.62	多	320 8.08	饭	540 9.96	公款	151 8.2	工地	17 6.14
拒	80 9.21	少	78 7.96	肉	258 8.91	最爱	49 7.17		
喜欢	168 8.69	没	193 7.62	水难	193 8.86	全家	41 6.63	Modifies 640 0.1	
舍不得	31 8.67	边	86 7.58	亏	227 8.83	老百姓	55 6.55	饭菜	11 7.44
爱	210 8.52	天天	42 7.56	午饭	179 8.74	大家	115 6.54	苦	9 7.41
请	161 7.54	一起	122 7.47	晚饭	164 8.61	粮	60 6.38	菜	19 6.64
讲究	19 7.41	连	81 7.34	鱼	242 8.5	猪	45 6.09	鸡蛋	8 6.54
宁可	8 6.74	不	955 7.17	菜	195 8.4	肉	31 6.07	食物	15 6.27
敢	43 6.73	不用	35 7.15	大锅饭	141 8.4	东西	52 6.06	东西	35 6.21
舍得	8 6.62	不再	58 6.98	年夜饭	129 8.25	孩子们	43 6.05	饭	8 5.66
宁愿	7 6.58	一律	31 6.76	定心丸	119 8.18	鱼	39 6.04	水果	10 4.94
放心	17 6.54	常年	28 6.75	饺子	111 7.99	孩子	90 6.02	商品	44 3.87
怕	17 5.82	经常	49 6.67	苦头	103 7.93	你	77 6.0	粮	6 3.86
知	20 5.66	很少	27 6.66	黄牌	117 7.92	狠剎	20 5.85	蔬菜	12 3.39
喜	17 5.45	常	35 6.65	败仗	96 7.87	灾民	38 5.81	时候	8 2.99
怀疑	9 5.44	从不	23 6.63	顿	106 7.75	金碗	18 5.77	食品	9 2.77
试	11 5.24	年年	24 6.62	奶	100 7.69	一家人	20 5.76	局面	6 2.55
觉得	13 5.24	着	117 6.58	螃蟹	84 7.64	菜	27 5.76	历史	12 1.59
拒绝	30 5.15	不能	117 6.56	东西	169 7.62	人们	124 5.72	方面	23 1.49
愿	31 4.92	常常	23 6.46	早饭	74 7.49	老人	73 5.71	粮食	7 1.3
习惯	11 4.86	怎么	24 6.42	年饭	72 7.44	药	34 5.69		
知道	16 4.48	津津有味	16 6.41	水果	116 7.35	金饭碗	17 5.68		
喜爱	6 4.4	专	32 6.27	食物	93 7.33	咱	19 5.63		
坚持	43 3.69	给他	21 6.27	红牌	76 7.32	你们	34 5.63		

Figure 3 Word Sketch output for 吃

Figure 3 shows the most salient and frequent collocations in which the verb meaning *eat* occurs in this particular corpus, presented by grammatical relationship with the keyword. It is of interest that the most salient collocation is 吃药 [*take medicine*]. The objects can be classified by the learner into items that are literally consumed (饭、肉 [*rice/food*; *meat*]), metaphors (亏、大锅饭), and items that are not genuine objects but have been interpreted as such by the software (顿 [measure word for a meal, normally followed by 饭; 吃水难 [a noun compound meaning *water shortage*]. Again, clicking on the links takes the user to the example sentences for each collocation.

Corpus task 3: *Find out the usual measure words for the following nouns occurring in Unit 19:* 石头、山、路 [*stone*, *mountain*, *road*].

Chinese nouns are usually preceded by measure words (量词, also known as classifiers) in the same way as rather marginal English cases such as *head* (of cattle) or *sheet* (of paper). The correct measure word varies from noun to noun, and therefore a nouns and the appropriate measure word have a strong collocational relationship. Measure words may also sometimes follow the noun, yielding a form similar to the plural in English.

Using a noun with the wrong measure word sounds unnatural, so learners are well motivated to learn the correct forms. It is likely that measure words learned through a process of research and discovery

are more likely to be retained. To answer the question, learners have to enter a corpus query language (CQL) command to request a concordance of all measure words in the corpus with the noun required. For example, they could enter [tag="q"] "石头" (q is the POS tag for measure word).

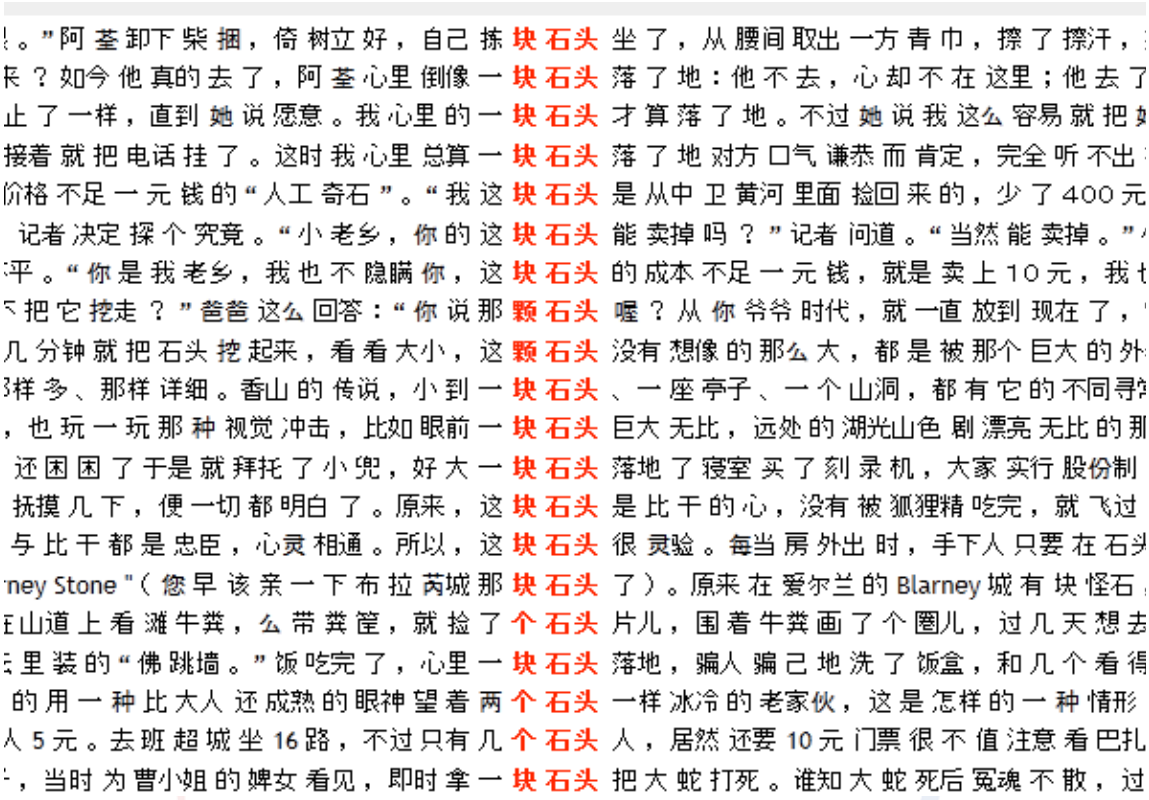


Figure 4 Concordance output for measure words followed by [*stone*]

An extract from the resulting concordance is shown in Figure 4 (in total it is 99 pages long). The learner will notice from the first page that 块 appears to be the most common classifier by far. The generic classifier 个, in the three instances where it occurs before the noun, refers to larger noun phrases that happen to include the following 石头 (for example in the third to last line in Figure 4, the reference is to “two people who looked as cold or expressionless as stone”. By inspecting this and later pages in the concordance, the learner may discover that the less common measure word 颗 is likely to refer to a smaller stone or pebble, often the kind that one might throw. To call up solid statistics on measure word usage in the corpus, the learner then requests a “Node forms” display, as shown in Figure 5.

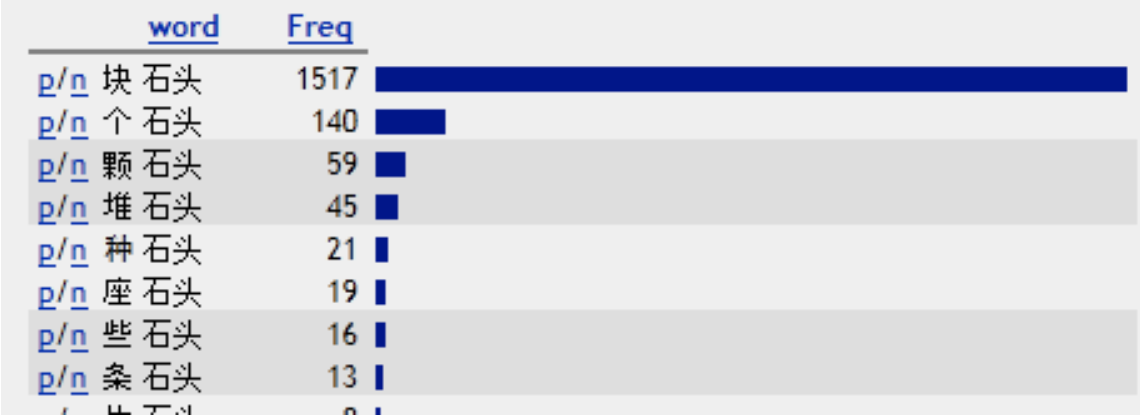


Figure 5 Node forms frequency display

The display shows that 块 is far and away the most usual measure word preceding *stone*. The learner may wish to reflect on why certain other measure words might appear (and can of course click on a link to a concordance for that collocation). The measure word 堆 would refer to a *pile* of stones, 种 to *types* or varieties of stones, and 些 to *several* stones, for example.

Corpus task 4: *Please find frequent words which include character X.*

Certain features of the Chinese language make it especially suited to a data-driven approach. In DDL, learners typically explore collocational and colligational patterns among words, but would not expect to be able to look at the internal structure of words using a corpus tool. The logographic Chinese writing system, however, allows the learner to investigate the ways that characters/morphemes pattern to form words. Most Chinese corpora (certainly those annotated with part of speech) are segmented into Chinese words (词) of one, two or more characters (字); but Chinese learners normally treat *characters* as the minimal unit to be learned, often memorizing the written form, and studying the words (often with related meanings) that the newly learned character participates in. It is as if an EFL learner were to learn new neoclassical compounds (such as *biology*, *biography*, *telegraphy*, *telescope*, *microscope*) by predicting and discovering meaning from compounds with the same prefix or suffix as others studied earlier.

Thus it is useful for learners using corpora to be able to call up a concordance for a particular character, and see which words come up, how they are used, and what their relative frequencies are. This function has been recently implemented in the SkE Chinese corpus interface.

做品位女人更好2005年9月7日星期三	如果说	性感魅力是女人外在的美丽，独立自信是女人内在的气质
在资料中夹了六个字：我们做朋友吧。	结果	，傻大姐将信据为己有，半个月之内向我狂送秋波和
起多年前秋日那次热吻，想再试一次，	结果	，松动的假牙使我们失去了一切兴致。80岁，坐在
，犹如一只扑火的蝴蝶，不顾一切，不计	后果	。我曾经告诉过你荆棘鸟的故事。那种鸟，找寻一生
无须...(P>(FONT c...我以前看过，	如果	是你...挺好玩的，很逗，(...你好痴情
。在这次高考，我有一个小小的心愿：	如果	分数合适的话，想考一个广州的大学。如果能够做到
分数合适的话，想考一个广州的大学。	如果	能够做到这一点就满足了。最后，真的是很舍不得大家
止些）其实我本来十分喜欢樱花大战的。	如果	把樱花大战算作恋爱养成类游戏的话，她应该是我唯一
翻字典而已。寒一下自己这几天以来的	成果	，内容为山东省的语文试题：角色露骨力能扛鼎量体裁衣
...2个大男人怎么好意思叫小名呢~16，	如果	他们没有做检事和律师，你觉得他们适合的职业是？
什么意思？(汗.....居然显示不出来.....	果然	是生僻字.....是(王录)这个字.....以下的?f均为
的很好.....回来的路上，我才发现煎饼	果子	都要2元了。盒饭成了4元了.....现在稍微好一
很大的经营利润空间，一瓶330毫升的	苹果	醋，经销商只需4元左右就可从厂家拿到，但它在
之大的利润空间，怎能没有诱惑力呢？	如果	您动心了，就赶快来精品的“生意眼”参与行动吧！
完全是两个概念。这里的醋饮可分为杂	果	醋饮、冰醋饮，橙、蓝枝、柠檬、醋饮等十几种
饮和家用烹饪的醋有很大的差别，保健	果	醋饮料可分别用橙、桔、猕猴桃、葡萄等水果汁经生物
醋饮料可分别用橙、桔、猕猴桃、葡萄等	水果	汁经生物工程技术制作出富含多种氨基酸、维生素、有机酸
酸、维生素、有机酸的新型保健营养浓缩	果	醋饮料。其特点：味道酸甜、具鲜果特有的芳香，不同
养浓缩果醋饮料。其特点：味道酸甜、具	鲜果	特有的芳香，不同水果果醋呈现不同色泽，冲饮方便。
点：味道酸甜、具鲜果特有的芳香，不同	水果	果醋呈现不同色泽，冲饮方便。明代医学家李时珍在《

Figure 6 Concordance of words including character 果

In Figure 6, the reader will notice the two words meaning *result* from Corpus Task 1. In fact 果 has

the core meaning of *fruit*, which (as with the English expression *bear fruit*) also carries the sense of *result*. An interesting exercise for the learner, here, would be to determine which of the corpus examples are of the edible sort, and which are abstract. The learner can also request the "node forms" display, as per Figure 7 (perhaps predicting, before so doing, which word containing 果 will turn out to be the most common—as you the reader may wish to do before glancing down).

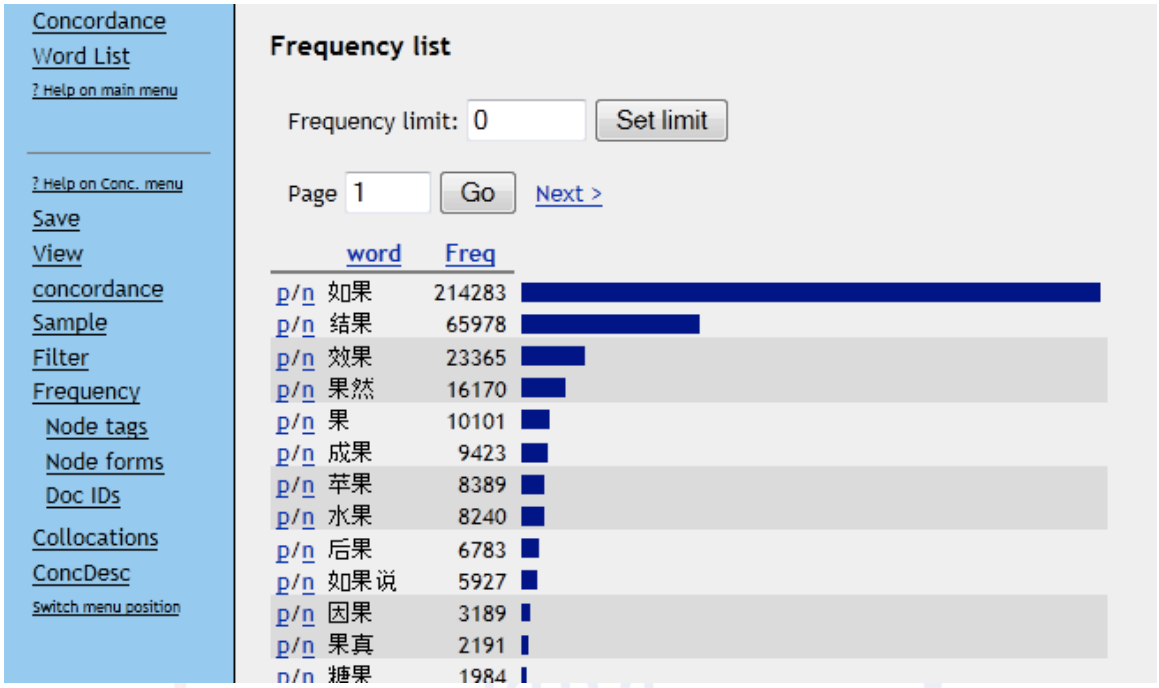


Figure 7 Node forms frequency display

By far the most frequently occurring word, then, is that meaning *if*, followed by the neutral result word 结果 (it can also be used as a conjunction meaning *with the result that*). As low as seventh in frequency is *apple*, followed by the standard word for *fruit*, with 后果 (the "negative consequence" of Corpus Task 1) in ninth place.

Corpus task 5: *Please identify the verb in these verb+object constructions.*

There is an important class of morphosyntactic structures in Chinese known as V+O compounds (离合词). In fact, some members of this class have already been seen in Corpus Task 2: 吃饭 means literally *eat rice*, but has come to mean eating a meal which may very well not include rice. The verb and object components of a V+O compound can be contiguous, in which case corpus segmentation algorithms treat the compound as a discrete word. The components may also be separated by the aspectual particles 过 or 了, yielding 吃过/了饭, as well as certain other types of material.

In the Academia Sinica Balanced Corpus (Chen and Huang, 2000), available only in traditional characters and on a web platform separate from that of SkE, the two components of V+O compounds are assigned a special tag [spo] or [spv]. It is possible to make a concordance of items with these tags, as shown in Figure 8.

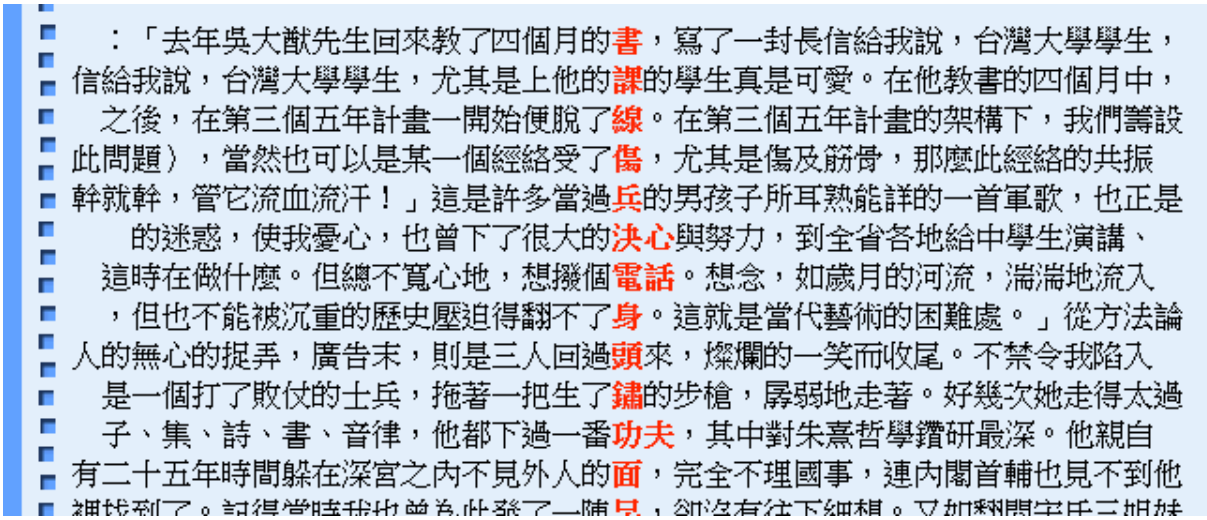


Figure 8 Sinica Balanced Corpus V+O compound concordance extract (Chen and Huang, 2000)

In the first line, the V+O compound is 教书 *teach* (literally *teach book*). In the concordance line, information about the time (4 months) spent teaching is given between the two components. The second line refers to 上课 (to go to class). This time, the inserted material is a pronoun indicating which teacher's class is being attended by students. The task is for students to find the verb component; in order to achieve that, the learners will have to understand what kinds of material are, and can be, inserted between the two components.

The five corpus-based tasks described above motivate students to learn through reflection and discovery. We followed Boulton (2009) in keeping the instructions clear, the tasks simple, and the focus on acquiring language rather than learning about corpus linguistics, as well as maintaining links with the textbook units being followed in the rest of the course. The questions are quite closely specified, and there are clear tasks to work on; however, there are many opportunities for motivated learners to go beyond the questions and discover the language for themselves.

Limitations

One limitation was that we did not have enough student participants to be able to conduct an effective pilot study. The tasks could not be a part of any credit-bearing study, so they were taken up by only a small number of keen volunteers. Most of these volunteers did, however, complete the tasks successfully, while reporting that they were both challenging and interesting.

It was pointed out above that Chinese study lends itself well to DDL because of aspects of its structure. One disadvantage, though, is that because of the challenging nature of the writing system, many students opt not to learn to read or write at all. This is regrettable, since it is clear that the lack of written input will impair the acquisition of speaking and listening skills, but it is a fact. There are some corpora available in Hanyu Pinyin transcription, such as the Lancaster Corpus of Mandarin Chinese (McEnery and Xiao, 2004) and the parallel corpora of Wang (2001), but for learners to derive any real benefit from corpus consultation, solid literacy skills are essential.

Future plans

We will provide training in the use of corpus interfaces to teachers and students in Xi'an and Liverpool. We will continue to produce a variety of tasks and exercises that will challenge and

interest students, motivating them to learn autonomously and inductively.

We will conduct a mixed methods study, using a combination of pre and post tests and feedback questionnaires, to establish the success or otherwise of our approach in a scientific way. **Research questions** will include:

- What domains of Chinese motivate students most (academic, business, cultural, general)?
- To what extent does the use of corpora help with learning in each domain?
- Does corpus use help with acquisition of grammar? Vocabulary?
- Does corpus use reinforce perception of collocations and patterns?
- Is learning through serendipitous discovery successful, or must exercises and tasks be explicitly provided for acquisition to take place?

Sketch Engine as a DDL platform

In addition to the functions reported above, Sketch Engine has a number of other features which could be turned to the advantage of the Chinese learner, and use will be made of these in future task development. There is, for example, a distributional thesaurus, which shows which words commonly occur in the same context as a user-supplied keyword, and are likely to be near synonyms of that keyword.

Concordances themselves are enhanced by the availability on SkE of a sentence mode, as well as the traditional KWIC mode, so that more may be gathered from the context. When accessing SkE's English corpora, concordance lines can also be ranked by quality using the GDEX ("good dictionary example") feature: a "good" example sentence is defined by Kilgarriff et al (2008) as one which is neither too short nor too long, which doesn't contain a lot of rare words or anaphors (which can sometimes only be resolved by looking outside the sentence), and is constrained by a few other parameters specified by the team. This feature is available for English corpora under SkE, but not currently for Chinese.

Although there are two Chinese corpora available on SkE, only the Gigaword newswire corpus offers the full functionality of SkE, with Word Sketches, Sketch Differences and the statistical thesaurus. The other corpus, Internet-ZH, has access to concordances only. In collaboration with the Sketch Engine team, we will make the additional functions available in due course.

Braun (2005) notes that corpus annotation schemes, for example for part of speech, are aimed at corpus specialists, and are often too complex for the needs of learners. Certainly the Gigaword and Internet-ZH do have a large number of tags, distinguishing for example many different types of nouns and verbs, and it is not especially convenient for the learner to have to type these in (as, for example, was necessary in our Corpus Task 3). In a development currently being implemented by the SkE team, it will be possible to request a concordance based on a keyword followed or preceded by an item belonging to a POS specified by the user, from a simplified list (noun, verb, measure word and one or two others) presented as a drop-down menu.

DDL and corpus methods have growing currency in English language teaching and learning, but are as yet virtually unknown for Chinese. In this paper, we have shown some examples of DDL exercises for Chinese, and have plans to extend and evaluate their use, expanding the repertoire of corpus-based teaching methods available.

References

Aston, G. (1995). Corpora in language pedagogy: matching theory and practice. In G. Cook, & B. Seidlhofer (Eds.), *Principle and practice in applied linguistics: Studies in honour of H. G. Widdowson* (pp. 257-270). Oxford, UK: Oxford University Press.

Bernardini, S. (1997). A 'trainee' translator's perspective on corpora. In *Proceedings, 1st International Conference on Corpus Use and Learning to Translate*, Bertinoro, Italy.

Boulton, A. (2009) Corpora for all? Learning styles and data-driven learning. In M. Mahlberg, V. González-Díaz & C. Smith (Eds.), Proceedings of 5th Corpus Linguistics Conference.

Braun, S. (2005). From pedagogically relevant corpora to authentic language learning contents. ReCALL, 17(1). 47-64.

Chen, K. and Huang, C. (2000). Sinica Corpus: Academia Sinica Balanced Corpus for Mandarin Chinese. Corpus available at http://www.sinica.edu.tw/SinicaCorpus (accessed 27 June, 2011)

Cheng W., Warren, M. & Xu, X. (2003). The language learner as language researcher: Putting corpus linguistics on the timetable. System 31(2), 173-186.

Johns, T.F. (1991). Should you be persuaded: Two examples of data-driven learning. In Johns, T.F. and King, P. (Eds.) *Classroom concordancing* (pp. 1-13), Birmingham: ELR.

Kilgarriff, A., Husak, M., McAdam, K., Rundell, M. and Rychlý, P. (2008). "GDEX: Automatically finding good dictionary examples in a corpus." *Proceedings of EURALEX*. Barcelona.

Kilgarriff, A., Rychlý, P., Smrž, P. & Tugwell, D. (2004). The Sketch Engine. Paper presented at EURALEX, Lorient, France, July 2004.

McEnery, T. and Xiao, R. (2004). The Lancaster Corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study, in M. Lino, M. Xavier, F. Ferreire, R. Costa, R. Silva (Eds.), *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC) 2004*, 1175-1178. Lisbon, 24-30 May 2004. Corpus available at http://www.lancs.ac.uk/fass/projects/corpus/LCMC/ (accessed 27 June, 2011)

Qiu, J., Peng, Z., Zhang, H. and Zhang, L. (Eds.), *Road to Success: Elementary (2)*. [成功之路．顺利编．第二册]. Beijing: Beijing Language and Culture University Press.

Sharoff, S. (2006). Creating general-purpose corpora using automated search engine queries. In M. Baroni and S. Bernardini (Eds.), *WaCky! Working papers on the Web as Corpus*. Bologna: Gedit. Corpora available at http://corpus.leeds.ac.uk/query-zh.html (accessed 27 June, 2011)

Smith, S. (forthcoming). Learner Construction of Corpora for General English in Taiwan. To appear in *Computer Assisted Language Learning*, September 2011.

Tao, H. (2005). The Gap Between Natural Speech and Spoken Chinese Teaching Material: Discourse Perspectives on Chinese Pedagogy. Journal of the Chinese Language Teachers Association, 40.2:1-24.

Thomas, J. (2008). Impatience is a virtue: Students and teachers interact with corpus data - now. In A. Frankenberg-Garcia (Ed.), Proceedings of the 8th Teaching and Language Corpora Conference (pp. 463 – 469). Lisbon: ISLA.

Tribble, C. (1997). Improvising corpora for ELT: Quick-and-dirty ways of developing corpora for language teaching. In J. Melia and B. Lewandowska-Tomaszczyk (Eds.), *PALC '97 Proceedings* (pp. 132-147). Lodz, Poland: Lodz University Press.

Wang, L. (2001). Exploring Parallel Concordancing in English and Chinese. *Language Learning and Technology*, Vol. 5, No. 3, Sep. 2001, 174-184.

Wang, Z. and Shen, W. (2008). *Chinese with me: an integrated coursebook (II)*. [跟我学汉语．综合课本（二）]. Beijing: Peking University Press.

Xiao, R. and McEnery, T. (2004). *Aspect in Mandarin Chinese: A corpus-based study*. Amsterdam: John Benjamins.

Xiao, R. and McEnery, T. (2006) Collocation, Semantic Prosody, and Near Synonymy: A Cross-Linguistic Perspective. *Applied Linguistics*, 27 (1). 103-129.

Xiao, R., Rayson, P. and McEnery, T. (2008). *A Frequency Dictionary of Mandarin Chinese: Core vocabulary for learners.* London: Routledge.

iafor