

Learning Chinese with the Sketch Engine

Adam Kilgarriff, Lexical Computing Ltd., UK

Nicole Keng, Xi'an Jiaotong Liverpool University, China

Simon Smith, Coventry University, UK

Wei Bo, Foreign Language Teaching and Research Press, Beijing

Introduction

For the last twenty years, the world of English language teaching has come to realize the value of corpora and has worked out how to use them in the preparation of teaching materials, dictionary-making, syllabus design, and in the classroom. In the last few years, the teaching of Chinese has grown rapidly. Thus it is desirable to transfer the expertise across, so Chinese teaching benefits from the experience of corpus use in English teaching.

Two central components of corpus use are: (1) the corpus itself; (2) corpus analysis software. The Sketch Engine (Kilgarriff et al 2004) provides both. Chinese Gigaword is a very large corpus of Chinese journalism (and there is also zhTenTen, a large Chinese web corpus, available). The software offers a wide range of functions for corpus exploration. In this paper we introduce the Sketch Engine and its core functions, with examples drawn from Chinese.

The Sketch Engine

The Sketch Engine is designed for anyone wanting to research how words behave. It is a Corpus Query System incorporating word sketches. In this paper, the key features of Sketch Engine will be introduced, including some useful tasks for teachers and learners of Chinese.

Basic concordance

A concordance is a list of every occurrence of a given word in the corpus. The concordance shows the word itself in a different colour, then the parts of the sentence right before and after the keyword.

Figure 1 shows that after logging in on Sketch Engine website, it is easy to find various corpora (with more than 42 different languages covered).

Corpus name	Language	Tokens	Words
Chinese GigaWord 2 Corpus: Mainland, simplified	Chinese, Simplified	250,124,230	205,031,375
Internet-ZH	Chinese, Simplified	277,931,664	198,205,344
zhTenTen (10M)	Chinese, Simplified	11,028,308	9,012,125
British Academic Spoken English Corpus (BASE)	English	1,252,256	1,186,290
enTenTen08	English	3,268,798,627	2,759,340,513
enTenTen12	English	12,968,375,937	11,191,860,036
FinnishWaC	Finnish	144,972,820	112,389,123
esTenTen11 (Eu+Am, Freeling)	Spanish	9,797,406,054	8,380,202,011

Corpus name	Language	Tokens	Words
Arabic web corpus	Arabic	174,239,600	407,005
arTenTen	Arabic	6,637,387,738	5,794,161,583
CHILDES Afrikaans Corpus	Arabic	33,134	26,020
BasqueWaC	Basque	123,856,183	99,719,584
BengaliWaC	Bengali	13,719,158	11,761,881
BulgarianNC	Bulgarian	26,518,884	20,974,953
BulgarianNC2 nonweb	Bulgarian	27,721,533	22,398,403
BulgarianNC2 web	Bulgarian	545,637,740	419,509,472
BulgarianNC2 web 10M	Bulgarian	9,478,549	7,693,925
CHILDES Catalan Corpus	Catalan	277,816	209,525
Internet-ZH (10M)	Chinese, Simplified	9,431,058	6,229,745
zhTenTen	Chinese, Simplified	2,106,661,021	1,729,867,455
Chinese GigaWord 2 Corpus: Taiwan, traditional	Chinese, Traditional	455,526,209	382,600,557
ChineseTaiwanWaC	Chinese, Traditional	349,198,060	259,156,002
ChineseTaiwanWaC (Universal Sketch Grammar)	Chinese, Traditional	349,198,060	259,156,002
CHILDES Croatian Corpus	Croatian	389,674	286,765

Figure 1: Start page

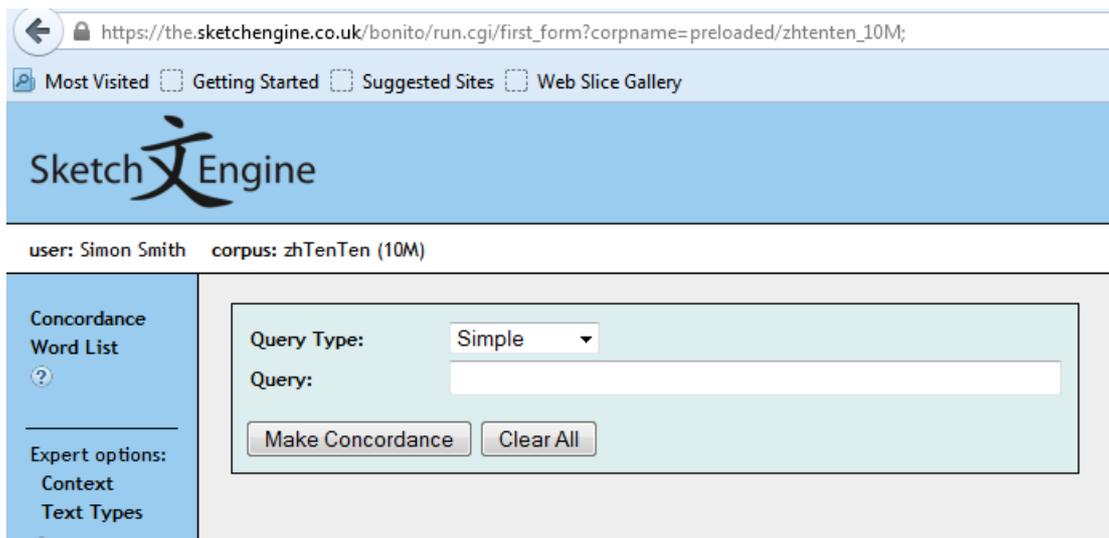


Figure 2: Concordance form

Any keyword (single or multiple character) keyword can be input at this interface. Figure 3 shows that the first of pages of occurrences of the keyword 公车 in a general web corpus of 10 million words. 公车 has three distinct meanings: in standard Mandarin, it refers to an *official or government vehicle*, while informally; especially in Taiwan (and apparently in Xi'an) it is used to mean simply *bus*, and there is a historical meaning, “selecting politicians to be officials”, as in the second instance below.

又去过苏州、无锡，那个时候仅管是一种穷开心，挤**公车**，住“浑堂”（旅馆，招待所要介绍信），还是有一代表的资产阶级维新派。1895年，康有为发动了著名的“**公车**上书”，成为维新运动的起点。维新派主张开国会，立宪法青年旅馆去吧，坐一下，拍点相片，3点半我就要坐**公车**去火车站了。一路走向旅馆，在路边小店我阿Mae热心的向她介绍我在武汉一星期吃的美食，还告诉她坐**公车**到哪里吃什么，在学校的什么地方又吃什么。有人聊天不过干净就好啦，160元/标间/天。安顿好后，坐**公车**回到火车站（西安坐公交车简直太方便简单了，有景点可元（外面大街上羊肉泡馍普通的7元）。后来在西安坐**公车**时发现了真的老孙家，在端履门！我们想，一般的都她家在东河区，也就是旧城区，我们从新城区乘5路**公车**过去。包头旧城区和新城区相比人口密度高很多，因此通过财基处由省直机关政府采购中心统一采购。院投资企业**公车**配备按照本企业的有关规定执行。</p><p>第三章车辆使用设专职驾驶员岗位，不得用于经营活动。第十三条禁止使用**公车**参加婚丧、钓鱼、游玩等非公务活动，杜绝**公车**私用等禁止使用**公车**参加婚丧、钓鱼、游玩等非公务活动，杜绝**公车**私用等不正之风。因特殊情况私事用车时，经批准后按。第二十条驾驶员不得私自将车辆交给他人驾驶，严禁将**公车**当驾驶练习车。第二十一条专职驾驶员年龄原则上不超过55安排到龄专职驾驶员转到其他合适的岗位。</p><p>第六章**公车**自驾第二十二条**公车**是指因**公车**需要由我院非专职**公车**需要由我院非专职的持有驾驶证的工作人员驾驶的单位**公车**、私车公用车辆和社会租用车辆。考虑到省级机关尚未私车公用车辆和社会租用车辆。考虑到省级机关尚未实施**公车**改革，暂不鼓励私车公用；为了充分利用现有**公车**资源，未实施**公车**改革，暂不鼓励私车公用；为了充分利用现有**公车**资源，不提倡租用社会车辆；有条件的单位可以实行**公车**资源，不提倡租用社会车辆；有条件的单位可以实行**公车**自驾。第二十三条**公车**自驾车辆在使用过程中产生的养护费人提出批评直至纪律处分。第二十八条若发生长期占用自驾**公车**现象，对单位负责人和直接责任人提出批评直至纪律处分对单位负责人和直接责任人提出批评直至纪律处分；若发生**公车**私用，一经查实，所发生的直接和间接费用均由用车，所发生的修理费和赔偿费等全部由驾车人员个人承担；**公车**私借发生的交通事故，要追究该**公车**驾驶员的责任，所驾车人员个人承担；**公车**私借发生的交通事故，要追究该**公车**驾驶员的责任，所造成的经济损失也由该**公车**驾驶员个人

Figure 3: Concordance for 公车

A glance at the KWIC (KeyWord In Context) concordance output of Figure 3 brings out the two meanings of the word quite clearly, and even gives us a rough idea of the distribution. In many cases, though, the difference between usages is not so clear cut: a term, especially an abstract term, may have several senses, difficult to pick out by scanning concordance lines. There may be many thousands of pages of concordance output, rather than a mere four, and it is the higher frequency words which typically have many nuances of meaning. To capture these nuances, we need to explore the collocational behaviour of keywords and their distribution across the corpus; the Sketch Engine provides a number of tools to allow just that.

Character search

Students of Chinese need to acquire vocabulary *and* learn the characters that make up the new words: this is one the features of the language that makes it tough to learn. An effective and popular strategy for internalizing newly-learned characters is to look up other words which share one of the characters of a recently learned item of lexis. In Sketch Engine, it is possible to get a concordance and frequency lists for all words which incorporate a particular character, such as the 果 in 结果 (*result*). A student learning 结果 for the first time would very likely be interested to know what other words incorporate the character 果, and he or she can use the Sketch Engine to find out which are the most common in a given corpus. As usual with Sketch Engine, clicking on a hyperlink calls up a concordance of examples of that word in use.

Page [Next >](#)

word	Freq	
p/n 如果	214283	
p/n 结果	65978	
p/n 效果	23365	
p/n 果然	16170	
p/n 果	10101	
p/n 成果	9423	
p/n 苹果	8389	
p/n 水果	8240	
p/n 后果	6783	
p/n 如果说	5927	
p/n 因果	3189	
p/n 果真	2191	
p/n 糖果	1984	
n/n 果子	1914	

Figure 4: Frequency list of all words including the character 果

We can see from Figure 4 that 如果 (*if*) is far and away the most common word that includes 果. Following that is 结果 itself, and after that 效果 (*effect*, similar in meaning to *result*). There are several other “result” related words in the list, including 成果, used to refer to a positive outcome, and 后果, which signals a negative consequence of some action. The student will be interested (and even entertained, perhaps) to learn that the words for *fruit* (水果) and *apple* (苹果) are also related.

Measure words

Chinese measure words (量词, also known as classifiers) are used alongside nouns when the latter are qualified by numerals or determiners. The most common measure word, used with the majority of nouns is 个 (as in 一个人, *a person*). However, a substantial minority of nouns are associated with a different measure word, which in the case of abstract nouns often reflects the object’s shape. But there are no easy rules, and measure words have to be learned and memorized. The Sketch Engine allows the student to see quickly which measure word collocates with a given noun, as shown in Figure 5, and as usual click on the links to see concordance examples of the measure word + noun collocations. The measure word associated with 飞机 (*plane*) is indeed 架. The motivated student may wish to investigate what has prompted use of the other measure words in the display. The second-ranking measure word 班, for example, is used in combination with 飞机 (*plane*) to mean *flight*, as in *the next flight to London*.

<u>word</u>	<u>Freq</u>
p/n 架飞机	891
p/n 班飞机	93
p/n 次飞机	62
p/n 种飞机	54
p/n 个飞机	54
p/n 条飞机	23
p/n 趟飞机	18
p/n 批飞机	16
p/n 分飞机	12
p/n 点飞机	11
p/n 日飞机	11

Figure 5: Frequency of measure words associated with 飞机

Word Sketch

The word sketch is a distinctive feature of the Sketch Engine. It shows, in a convenient one-page summary, a list of words that commonly collocate with the keyword.

Figure 6 shows the most frequent collocations in which 说 (*speak/say*) occurs, in Chinese Gigaword, a 200 million word newswire corpus. It also presents the grammatical relationship with the keyword. Not unexpectedly, the most frequent subject collocating with 说 is 他 (*he*). The next is 她 (*she*). There are also subjects like 记者 (*journalist*), 申明 (*declamation*), 报告 (*report*) or even 发言人 (*spokesperson*), which reflects the language of news articles. The most frequent object collocation is 说话 which means *say a few words* in most of the example sentences. The second most frequent collocation is 说 *to be honest; truth to tell*. Other objects include languages such as 普通话 (*Mandarin*), along with lower-ranking near-synonyms 汉语 and 中国话, as well as 藏语 (Tibetan) and 英语 (*English*). The object can also convey a genre or type of speech, as in 脏话 (*bad language*), 笑话 (*joke*) or 相声 (*xiangsheng*, a traditional type of Chinese humorous dialogue).

The most common modifier used with 说 is 来 which is not really related to its literal translation *come*. It is used in the sentence to represent “For”.... For example, 对伊拉克来说 (For Iraq) or 就全国来说 (For the whole country). The second word 还 is used the similar sense as 来 to represent “also”. The other common words are adverbs like 激动 (*thrilled*), 高兴 (*happy*), 感慨 (*rueful*), 动情 (*moved*), 兴奋 (*excited*), and 自豪 (*proud*).

说

Chinese GigaWord 2 Corpus: Mainland, simplified freq = 736823 (2945.8 per million)

Subject	609585	8.4	Modifier	82540	3.3	Object	5366	0.1	Modifies	719	0.0
他	139180	11.86	来	18717	11.16	实话	174	9.98	心里话	13	8.21
她	13831	9.3	还	13168	10.39	空话	88	8.98	话	220	8.17
记者	19637	8.92	可以	4416	9.99	话	425	8.87	PP_向 667 0.8		
声明	8983	8.8	激动	1938	9.52	心里话	73	8.62	报界	38	9.46
报告	9488	8.76	高兴	2212	9.39	普通话	64	7.8	新闻界	95	8.28
报导	7330	8.5	感慨	1463	9.11	句	82	7.64	群	11	5.88
发言人	7094	8.44	着	1726	8.92	真话	33	7.61	报告	69	4.37
官员	6627	8.29	应该	1119	8.53	新闻界	80	7.48	媒体	12	3.57
公报	6089	8.28	此间	1064	8.5	报界	29	7.2	记者	151	2.79
李鹏	6192	8.25	动情	921	8.48	老实话	24	7.19	PP_把 133 0.3		
江泽民	6047	8.16	兴奋	876	8.4	遍	29	7.14	道理	9	6.63
负责人	4996	7.87	自豪	779	8.24	英语	54	7.13	PP_在 85 0.1		
消息	4764	7.83	都	1736	8.09	假话	21	6.98	嘴	10	6.68
人士	4605	7.66	却	859	8.06	实在话	17	6.69			
文章	3613	7.55	常	712	8.05	汉语	33	6.61			
代表	4819	7.43	所	1150	8.02	脏话	14	6.38			
钱其琛	2910	7.23	接着	665	8.01	话题	23	6.26			
他们	3745	7.14	再	969	7.88	坏话	10	5.92			
谈话	2722	7.14	又	1205	7.82	滋味	12	5.9			
专家	3003	6.99	曾	822	7.76	中国话	10	5.88			
人	4258	6.94	则	661	7.61	笑话	10	5.84			
评论	2261	6.9	没	581	7.45	相声	11	5.81			
社论	2198	6.87	不能	569	7.32	声	18	5.8			
我	2129	6.55	别	401	7.22	藏语	10	5.7			
教授	1869	6.53	连声	339	7.06	事	78	5.62			

Figure 6: Word Sketch output for 说

After clicking on the number next to the word 心里话, (the collocation frequency, 73 in this example) we see a concordance for the collocation (see Figure 7). The HTML-like formatting here shows paragraph boundaries.

监督、面对面的监督。一些群众也反映：人民代表为群众说心里话，威信更高了。我国完成服装号型标准修订工作了，报上登了文章，村里才知道你干的是县委书记。说心里话，瞧你那打扮，一件油渍麻花的青棉衣，布棉鞋了理满头的银发对记者说：“这次会议大家畅所欲言，说了心里话，只要心合在一块，一切都好办”。说着，了几年水果摊的孙大娘说：“晓平办事又公道又热心，说句心里话，不看国家就看他这个人，我们也要好好了戏。这些剧目大都取材于农村生活，演农民身边事，说农民心里话，生活气息浓郁，形式多样，深受广大农民喜爱电影太少了……”许多在学校、在家里不敢说或顾不上说的心里话，在孩子们的节日里统统倾吐出来。科普工作的科技人员，有机会向省委、省政府领导同志说心里话，反映基层实际情况创造了良好的环境。了一些不错的画，他想念北京的朋友。他说，人们开始说心里话时，这世界会十分美丽。四川解开“债务链”拉起了家常。任久荣老大娘只想和真办实事的好市长多说点心里话，可又不知说什么好，只是念叨着：“我来农家走走。走多了，农民都认他做朋友，也跟他说心里话。转眼暮霭降临，老农林源明跑来，请老朋友，现在他们的儿子却有幸成为一名为同胞救死扶伤的医生，说句心里话，还是共产党好啊！”求医问药新华社北京就把他拉到一边悄声问：“你我都是四川人，你对我说心里话，上三峡工程到底是好还是不好？”唐登清笑着完全能走向新生”。华日厂厂长陈励君言词诚恳：“说心里话，看到西冷厂这几个月的飞速发展，我真替它其他七项全能运动员正紧追不舍，这也是一种激励。不过说心里话，只有我能夺冠军，因为我想这么做。”通过和这些筑路兵相处，才真正认识到了军人的伟大。说句心里话，他们图的什么，还不是各族人民的安宁和国家。“风险一条船，才能齐心划桨”。有的农民对记者说了心里话，“现在各种各样的服务，除了从不收钱到相继有5000多人走进了他们的直播室，既侃天下事，也说心里话，再热的话题，也不惧直面现实。随着写照。战士们最大的乐趣就是伴着着吉他同唱一曲“说句心里话，我也想家……”但他们耐住了寂寞。经济、精外语的未来夫人到芝英共图大业。“说心里话，娶女大学生为妻，绝不是为了赶时髦，给自己这一活动内容包括：我为母校献良策，提合理化建议；说句心里话，为母校留言；植一片绿茵，营造毕业生林；

Figure 7: Concordance for 说...心里话

Thesaurus

The thesaurus is another function which shows more relevant words (see Figure 8). It represents the words that share most collocates with the keyword and looks at distributions to show the words occur in the same contexts with the keyword. Usually these are words which are close to the keyword in meaning.

The word 表示 (*indicate; state*) in Figure 8 is the most common word sharing a meaning with 说 in the journalism corpus. 认为 (*think; consider*) is another formal word with the same meaning as 说.

Lemma	Score	Freq
表示	0.562	222142
指出	0.547	117059
强调	0.494	80782
认为	0.488	178028
宣布	0.405	90484
发表	0.392	88600
谈	0.391	14844
介绍	0.386	124870
接受	0.371	67910

Figure 8: Thesaurus output for 说

When clicking the first word “表示”, which is a more formal term for “say”, it shows the different use of these two words (see Figure 9). It shows that these two words share many subjects collocates, for example, 他们 (*they*), 负责人 (*person in charge*), 官员 (*official*) and 发言人 (*spokesperson*). The difference is that 表示, being more formal, is usually used in more formal contexts, for example with 遗憾 (*regret*), 敬意 (*respect*), 谢意 (*thanks*) and 忧虑 (*worry*). This is the function in Sketch Engine called “Sketch Diff”, which displays the similarities and differences between similar words. The next section will introduce more details about this function.



Figure 9: Sketch Diff for 说 and 表示

Sketch Diff

In Chinese, many words share the same meaning. The use of the word depends strongly on its context. It is usually difficult for Chinese learners to decide which is the most appropriate in a certain context. Sketch Difference is a unique function in Sketch Engine to help learners understand how similar words differ in order to help with the choice of the right words.

Figure 10 shows the Sketch Diff for two similar words: 成立 and 建立 whose translation in English means *establish*. The words shown in green (in the lower part of each table) are more likely to collocate with 成立, while the red ones (the upper part of each table) are more likely to collocate with 建立. The white ones collocate equally

with both. From the examples in Figure 10, we can see that 成立 is usually used with words related to social organization based on people's decision making, like 大会 (meeting), 仪式 (ceremony), 小组 (team) or 委员会 (committee). 建立 is usually used in the context about establishing some intangible items, such as 关系 (relationship), 制度 (regulatory system), 责任制 (responsibility), 体系 (organizational system), or 基础 (foundation).

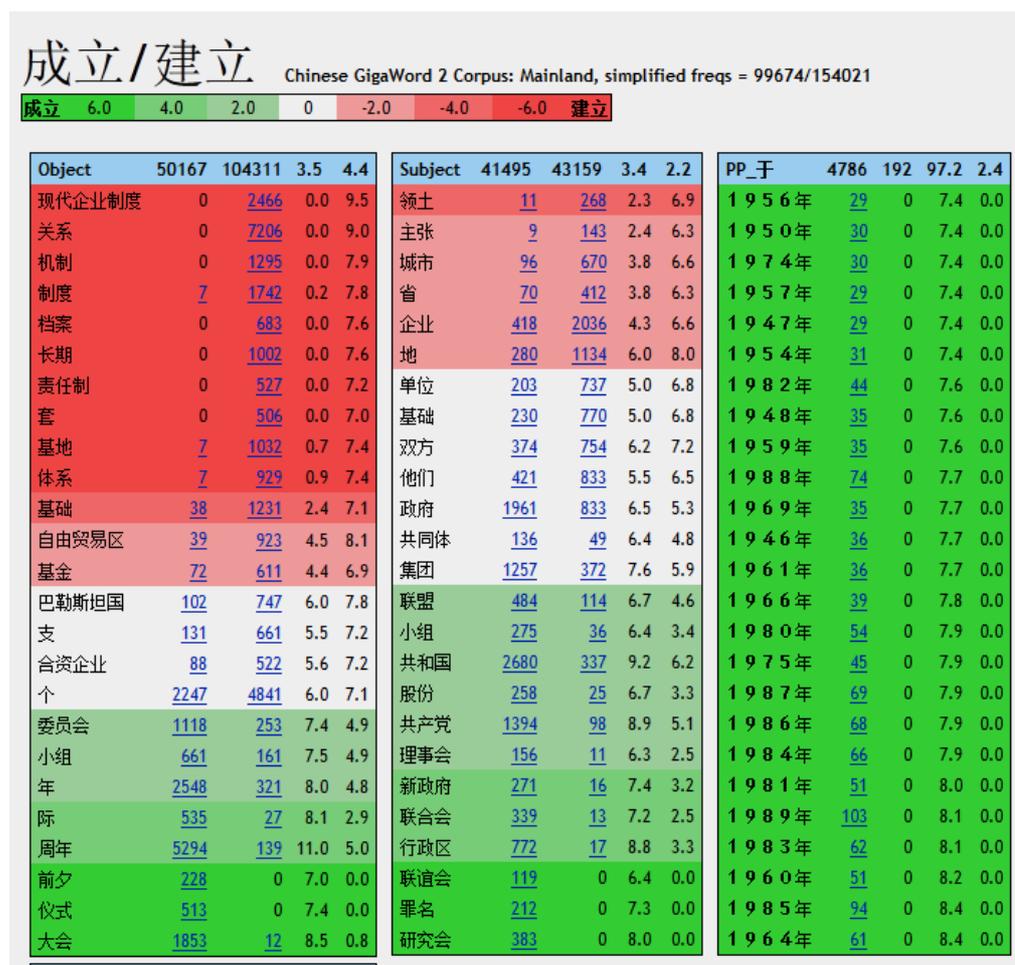


Figure 10: Sketch Difference for 成立 and 建立

When we click on the number next to the objects, we see the example sentences. Figure 11 shows the examples of 建立关系 (establish relationship). For instance, the first sentence shows 英国政府正谋求同中国政府成立更好的工作关系 (“The

British government is trying to establish a better work relationship with the Chinese government.”)

在剑桥举行的一个会上说，英国政府正谋求同中国政府建立更好的工作关系。他指出，这不仅因为英国政府本身有亲密的、富有成效的合作关系”，同捷克斯洛伐克和匈牙利建立伙伴关系的政策。他强调同西欧、特别是同法国建立友好匈牙利建立伙伴关系的政策。他强调同西欧、特别是同法国建立友好关系的重要性。他认为，同美国和加拿大保持亲密关系深圳，邀请港澳台新闻界人士、旅游业同行来深观光考察，建立合作关系；另一方面调整发展战略，集中财力、物力兴办有合作的信心。目前深圳赛格集团已和一批国际知名的厂商建立良好的合作关系，5年来引进外资1.1亿美元。

樱内指出，在当前的国际形势下，日中两国携起手来建立坚如磐石的友好关系比任何时候都显得重要。他希望通过认为，海部出访汉城的目的首先在于通过访问，同南朝鲜“建立面向未来的合作关系”。日本过去曾对朝鲜半岛实行了长达是解决悬案，促进双边贸易发展。1965年日本同南朝鲜建立外交关系后，双方一直没有解决旅日第三代以后南朝鲜人总统卢泰愚今天上午在汉城举行的第二轮会谈中，就双方建立伙伴关系的原则达成了协议。

会谈中，海部和项原则：日本和南朝鲜进行交流、合作，增进理解，以建立真正的伙伴关系；为亚太地区的和平、和解、繁荣、开放起死回生。合肥汇通商厦在工商联的帮助下，与26家企业建立了长期的业务关系，生意越做越红火。一些企业经常向的斯图加特大学等具有世界一流同类专业的院校都和他们建立了各种形式的合作关系。

南京太平天国历史博物馆建馆、棉织品服装的系列化加工出口，同90多个国家和地区建立了正式贸易关系，使一向没有多少竞争力量的河南服装走向以来，已与全国28个省(市)、130多个地、市建立了比较固定的贸易关系，一些省、市还向这里派出了进行了会谈。这两家银行与中国人民银行和中国银行都已建立了正常的业务关系。

代表团是21日到达新西兰中国驻马绍尔群岛共和国大使馆临时代办顾思聪，对中、马正式建立外交关系表示高兴。

卡布阿总统还对中国在两国2月2日电。罗马尼亚和斯威士兰王国两国政府2月1日决定建立大使级外交关系。据罗新社报导，两国是通过其驻伦敦1975年3月就任泰国总理并于同年7月1日促成了中泰两国建立正式外交关系。

快讯：王秀兰超一项世界纪录新华社，18个地方分会，与美、德、日、苏等11个国家建立了良好的交流合作关系。目前，全国已初步形成了适应较大幅度的增长。这家银行去年还与27家国外金融机构建立了代理行关系，将国外代理行增加到4000多家，出口收汇

Figure 11: Example sentences for 建立关系

Conclusion

After twenty years of corpus work to support English Language Teaching, there is a substantial body of knowledge on how corpora can be put to good use in language education. Much of this will transfer to the teaching and learning of other languages, for example Chinese. A crucial component is the corpus query tool. In this paper we have introduced one leading corpus tool, the Sketch Engine, which is available to everyone over the Internet, and we have shown how it can be used, together with a corpus of Chinese which is already loaded into it, by learners of Chinese. We have introduced its core functions: concordances, character search, word sketches, thesaurus and sketch differences. We believe the Sketch Engine is a useful tool for learners to explore the structure, grammar and collocations of Chinese words and phrases.

References

Adam Kilgarriff, Pavel Rychly, Pavel Smrz, David Tugwell. 2004. The Sketch Engine. Proc EURALEX, Lorient, France; Pp 105-116.