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).

![Corpora Table]

**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.

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.
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.

<table>
<thead>
<tr>
<th>word</th>
<th>Freq</th>
</tr>
</thead>
<tbody>
<tr>
<td>架 飞机</td>
<td>891</td>
</tr>
<tr>
<td>班 飞机</td>
<td>93</td>
</tr>
<tr>
<td>次 飞机</td>
<td>62</td>
</tr>
<tr>
<td>种 飞机</td>
<td>54</td>
</tr>
<tr>
<td>个 飞机</td>
<td>54</td>
</tr>
<tr>
<td>架 飞机</td>
<td>23</td>
</tr>
<tr>
<td>班 飞机</td>
<td>18</td>
</tr>
<tr>
<td>对 飞机</td>
<td>16</td>
</tr>
<tr>
<td>分 飞机</td>
<td>12</td>
</tr>
<tr>
<td>点 飞机</td>
<td>11</td>
</tr>
<tr>
<td>日 飞机</td>
<td>11</td>
</tr>
</tbody>
</table>
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 (declaration), 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).
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.
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. 但 (think; consider) is another formal word with the same meaning as 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.

**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**