Sketch Engine: a sense discrimination engine for English, Chinese and other languages

Adam Kilgarriff[1], Pavel Rychlý[2], Simon Smith[3], Chu-Ren Huang[4], Yiching Wu[5], Cecilia Lin[3]

ssmith@mcu.edu.tw

1. Background: corpora and concordances

Analysis of text and spoken language, for the purposes of second language teaching, dictionary making and other linguistic applications, used to be based on the intuitions of linguists and lexicographers. The compilation of dictionaries and thesauri, for example, required that the compiler read very widely, and record the results of his efforts – the definitions and different senses of words – on thousands, or millions of index cards.

Today's approach to linguistic analysis generally involves the use of linguistic corpora: large databases of spoken or written language samples, defined by Crystal (1991) as "A collection of linguistic data, either written texts or a transcription of recorded speech, which can be used as a starting-point of linguistic description or as a means of verifying hypotheses about a language". Numerous large corpora have been assembled for English, including the British National Corpus (BNC) and the Bank of English. Dictionaries published by the Longman Group are based on the 100 million word BNC, and corpora are routinely used by computational linguists in tasks such as machine translation and speech recognition.

The BNC is an example of a **balanced corpus**, in that it attempts to represent a broad cross-section of genres and styles, including fiction and non-fiction, books, periodicals and newspapers, and even essays by students. Transcriptions of spoken data are also included; and this is the corpus that is used with the English Sketch Engine. Academia Sinica and Lancaster University, respectively, offer balanced text corpora of Taiwan and mainland China Chinese. Although very useful for many text linguistics and lexicographical applications, both of these corpora are too small to provide sufficient training data for sense discrimination.

The corpus chosen for Chinese Sketch Engine is the Linguistic Data Consortium's Chinese Gigaword. It is by any standards large, at around 1.12 billion Chinese characters: it contains 286 newswire stories, taken from CNA Taiwan (735 million traditional characters) and the mainland's Xinhua news agency (380 million

[1] Lexical Computing, UK
[2] Masaryk University, Czech Republic
[3] English Language Center, MCU
[4] Institute of Linguistics, Academia Sinica
[5] National Tsinghua University

simplified characters). Unlike the other two Chinese corpora mentioned above, Gigaword has no part-of-speech tagging, nor any other linguistically meaningful mark-up.

Central to corpus analysis is the context in which a word occurs: J R Firth pointed out that information about meaning can be derived from surrounding words and sentence patterns: "You shall know a word by the company it keeps", as he famously stated in 1957. A convenient and straightforward tool for inspecting the context of a given word in a corpus is the **KWIC** (keyword in context) **concordance**, where all lines in the corpus containing the desired keyword are listed, with the keyword at the centre. Figure 1 shows such a concordance.

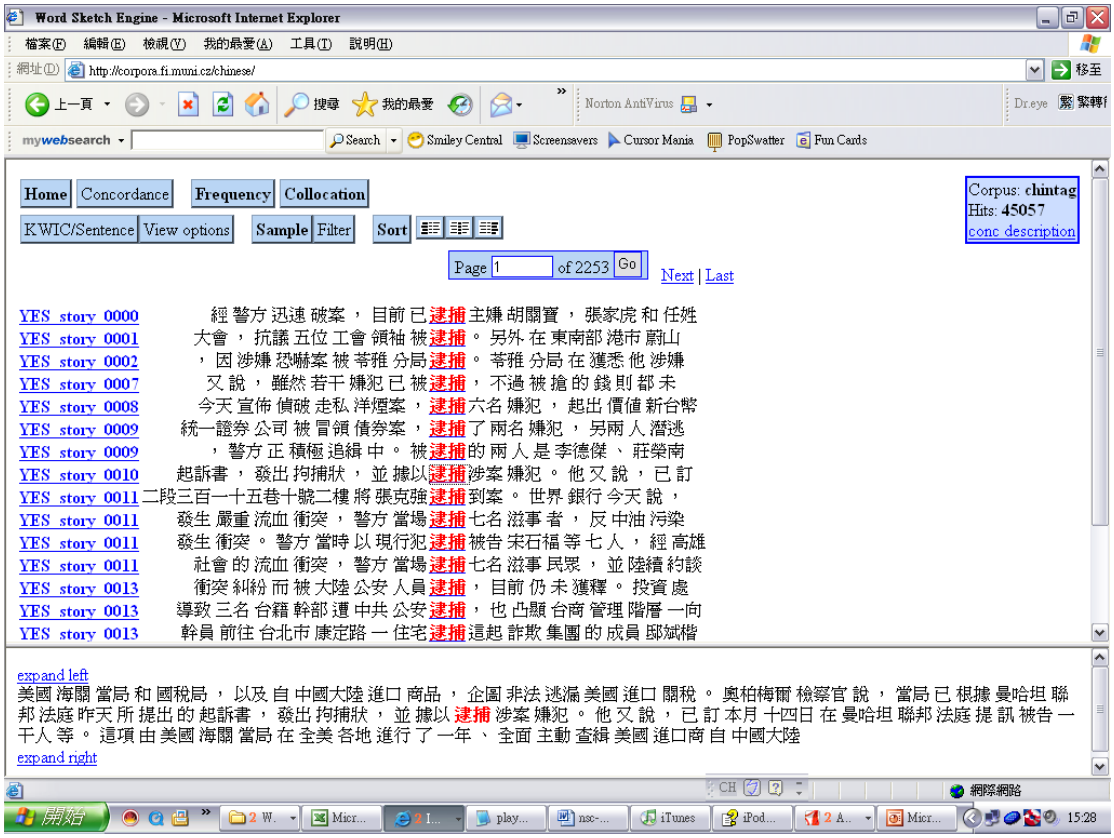

Figure 1 Excerpt from KWIC concordance of the word 逮捕 from the Gigaword corpus, generated by Chinese Sketch Engine

That Figure 1 includes only an excerpt from the full concordance is probably fairly obvious: in a large corpus like Gigaword, a word such as 逮捕 would occur dozens if not hundreds of times. So while KWIC might help a lexicographer or a student of Chinese to see, for example, that the word in question often takes a particular kind of subject (such as 警方) any comprehensive analysis taking into account all the occurrences of the keyword would not be practicable.

2. Corpus query tools

Various tools are available for exploring word context in corpora, determining by a statistical analysis which words are likely to appear in collocation with which others. Often, the statistic involved is **mutual information** (MI), first suggested in the linguistic context by Church and Hanks (1989).

Oakes (1998:63) reported that co-occurrence statistics such as MI "are slowly taking a central position in corpus linguistics". MI provides a measure of the degree of association of a given segment with others. Pointwise MI, calculated by Equation 1, is what is used in lexical processing to return the degree of association of two words *x* and *y* (a **collocation)**.

(1) $$I(x; y) = \log \frac{P(x \mid y)}{P(x)}$$

Where one constituent of a collocation could scarcely occur other than in the company of the other (as with "Hong" and "Kong", perhaps), MI will be positive and relatively high. Zero MI indicates, in principle, that two items are contiguous by chance, and that they are independent of each other (although it is quite difficult to make out a case for independence when word order is clearly constrained by rules of syntax). A negative MI suggests that the items are relatively common, but in complementary distribution: ungrammatical sequences such as "the and" would come into this category.

The SARA tool, widely used with the BNC, and the Sinica Corpus user interface both offer an MI analysis of the corpus contents. Such tools, however, suffer from two important constraints: first, when considering the context of a word, an arbitrary number of adjacent words to the left or right is taken into account, ignoring **discontinuous collocations**, which occur when other words (in particular **function words** like *the* and *of*) are found between the collocation components. To illustrate the problem, imagine that we wish to determine which of two senses of the English word *bank* ("the bank of a river", or "financial institution") is more common. If the strings *river bank* and, say, *investment bank* are frequent, there might be enough evidence on which to make a judgment. But such an analysis would ignore *Bank of Taiwan* and *bank of the river*, where the important collocates are not adjacent to the keyword, even though *Taiwan* and *river* stand in the same grammatical relationship to the keyword as *investment* and *river* in the other example.

The second constraint is that a list of collocates of some keyword could include, undistinguished, items of any **part of speech** (POS: noun, verb and so on) and of any syntactic role (such as subject or object). This sort of grammatical information can provide useful clues for sense discrimination, which standard corpus analyses are unable to take advantage of. Consider again the word *bank*, which has at least two verbal senses, illustrated by *The plane banked sharply* and *John banked the money.* The first of these is an intransitive verb – it cannot take an object. Thus, if an object is observed in the sentence featuring the keyword, the chances are that forms of the verb

bank properly belong to the second sense.

3. Sketch Engine features

One corpus query tool which overcomes these limitations is the Sketch Engine, developed by Adam Kilgarriff and Pavel Rychly, and described by Kilgarriff, Rychly, Smrz &Tugwell (2004). The description of the Sketch Engine which follows draws from that source, and from the Sketch Engine website www.sketchengine.co.uk.

The Sketch Engine is embedded in a corpus query tool called Manatee, and offers a number of modules. There is a standard concordance tool, whose output is shown at Figure 1. It allows the user to select, as a keyword, either a lemma (in which case the keyword *bank* would yield results for all of *bank, banks* and *banking* for example), or a simple word-form match. The user may also specify the size of the window (the numbers of words to the left and right of the keyword) that he wishes to view. Word frequency counts are also available, and the user may define a subcorpus (in the case of the BNC, on which the English version of Sketch Engine is based, one can choose different parts of the corpus such as fiction or non-fiction).

The Sketch engine can produce thesaurus lists, for an adjective, a noun or a verb, the other words most similar to it in their use in the language [10]. For instance, the top five candidates similar to the verb *kill* are *shoot* (0.249), *murder* (0.23), *injure* (0.229), *attack* (0.223), and *die* (0.212). It also provides direct links to the Sketch Differences which lists the similar and different patterns between a keyword and its similar word. For example, both *kill* and *murder* can occur with objects such as *people* and *wife*, but *murder* usually occurs with personal proper names and seldom selects animal nouns as complement whereas *kill* can take *fox, whale, dolphin*, and *guerrilla*, etc. as its object.

The novelty of the Sketch Engine lies in its ability to produce **word sketches**. The word sketch for the verb 逮捕 is shown at Figure 2. It will be seen that occurrences of 逮捕 in the corpus are presented according to the grammatical context in which they occur, along with a frequency count and a salience count (this statistic is based on mutual information). Thus the most salient collocate of 逮捕 to act as object is 嫌犯, while the most salient subject collocate is 警方. The most frequent (and second most salient) modifier is 當場. This is a rather satisfying finding: one would certainly expect, in the real world, to learn that the principal agents of arrest are the police, that those arrested are suspects, and that characteristically arrests occur on the spot, or at the scene of the crime!

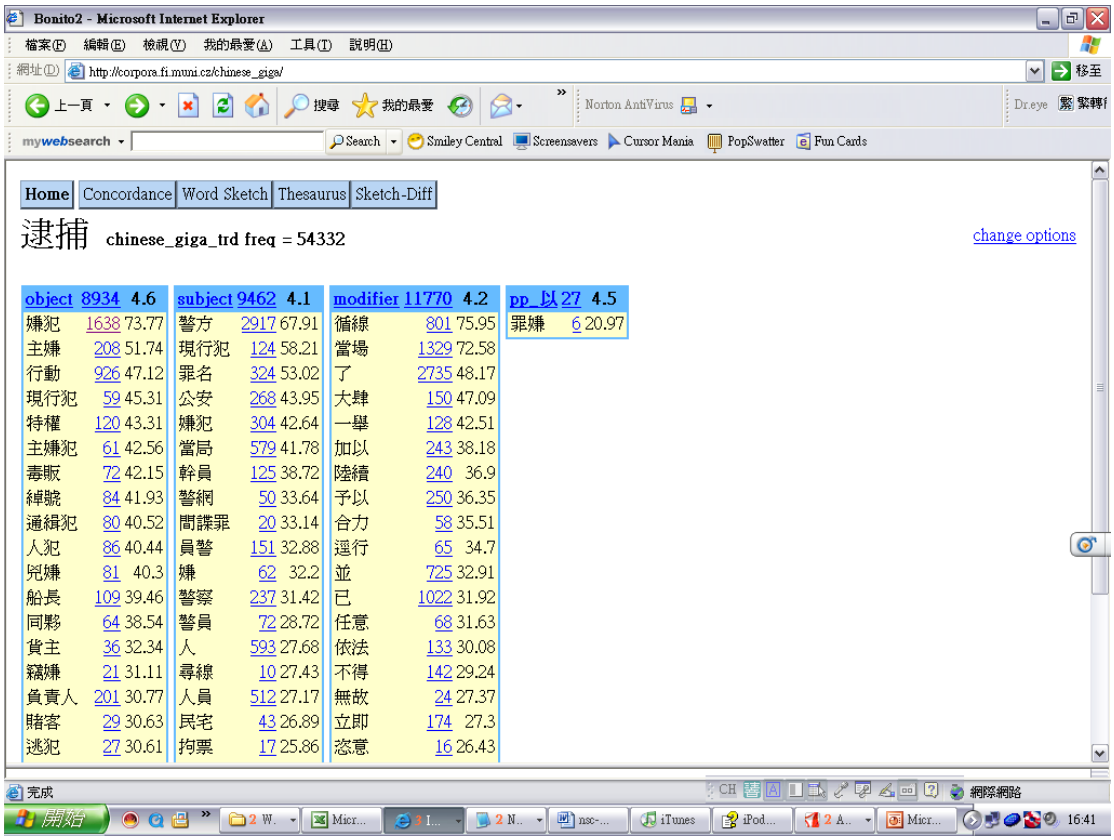

Figure 2 Word sketch for the lemma 逮捕

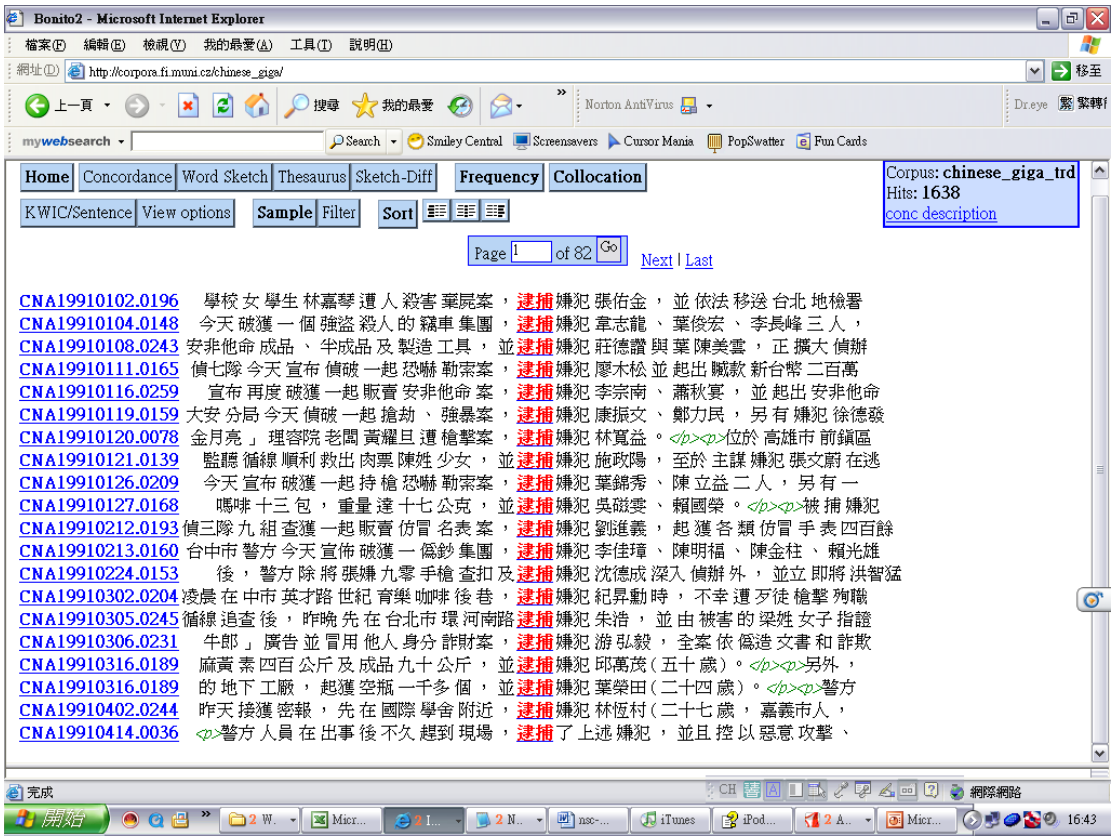

Figure 3 Sketch engine concordance for 逮捕...嫌犯

Clicking on the frequency count for 嫌犯 *suspect* yields the concordance shown at Figure 3.

The English version of Sketch Engine handles discontinuous collocations of verb and object such as *express our deep concern* as well as the canonical *expressed concern* where verb and object are adjacent; from Figure 3, it will be noticed that 逮捕嫌犯 and 逮捕了上述嫌犯 are both found, and incorporated in the statistical analysis of 逮捕嫌犯. Other patterns, including *the concern expressed* and *expressed by the infinitive* (a passive form) are appropriately dealt with in the English system, while the Chinese version is thus far less powerful in such respects.

At last year's conference, we presented a paper (Wu, Smith & Huang, 2005) which analyzed and compared the use in English and Chinese respectively of the apparently equivalent verbs *express* and 表示, finding that the two forms exhibit considerable differences. The Sketch Engine was used for the English part of the analysis, and revealed some interesting findings: for example, the PI, a native speaker of English, was not aware that the sentence *The justification for this was <expressed> to be that it would thus be open to a court at a later date to review the matter of sex determination* was a possible sentence of English.

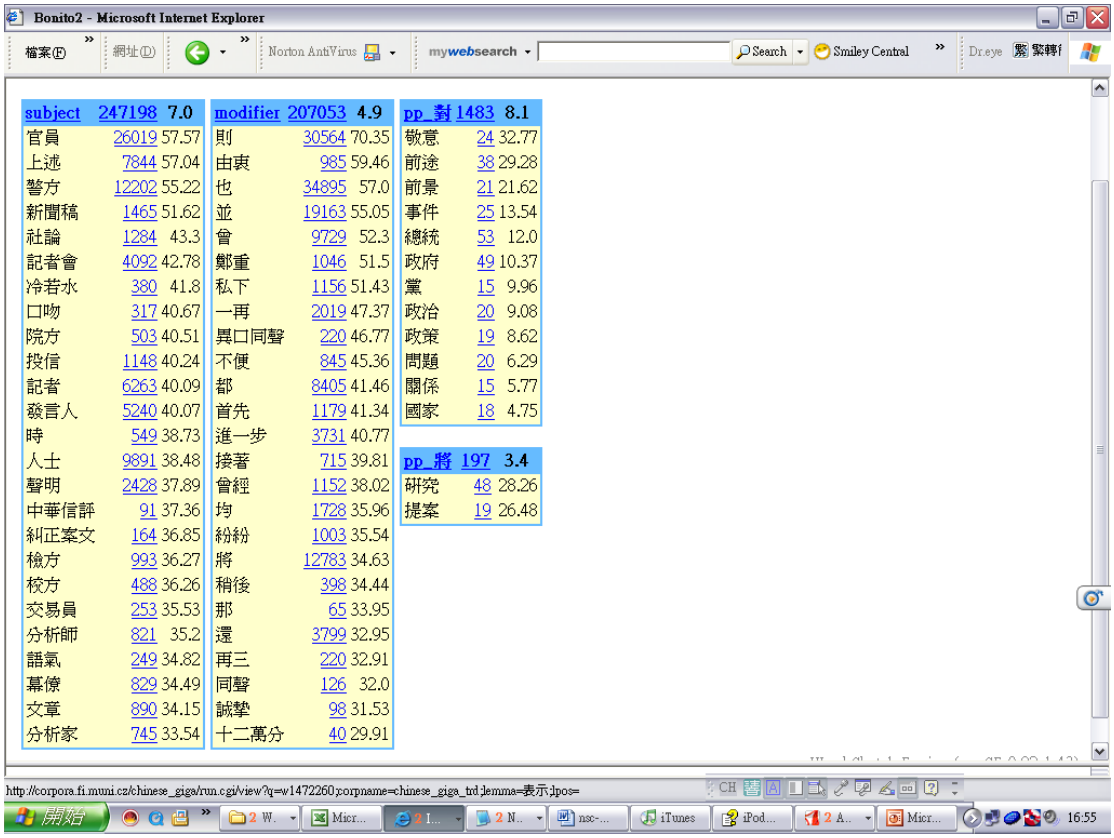

Figure 4 Word sketch for 表示

Figure 4 shows that 表示 does not occur with an object at all in the corpus; it is found with sentential complements such as 警方表示，今天清晨零時許，有三名男

子搭乘計程車到金月亮理容院. This mirrors the English legalistic usage, noted above, very closely. Another word choice, 表達, maps more closely to the standard English use of *express*, as seen in Figure 5 (as was surmised by a questioner at last year's conference!). Here are found objects such as 意見 and 立場.

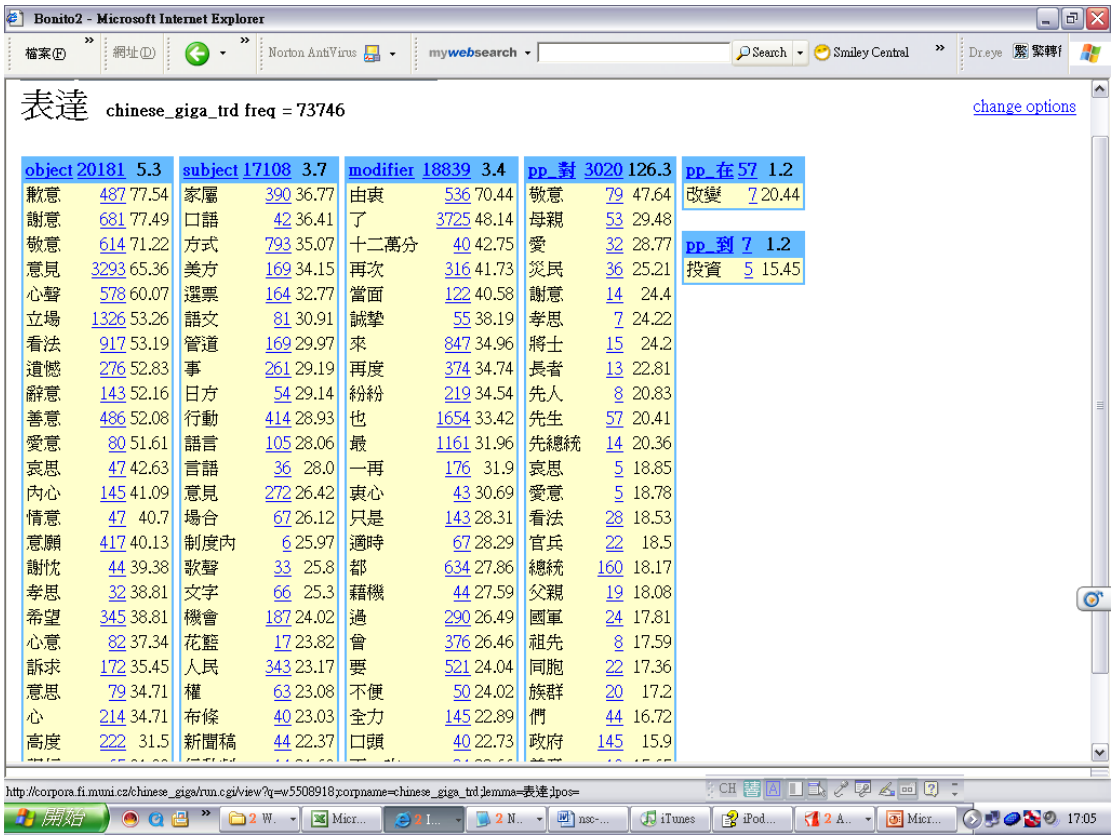

Figure 5 Word sketch for 表達

4. Grammatical relations

Altogether, the Sketch Engine defines 27 **grammatical relations** for English. As well as the subject and object relations, adverbial modifier, and/or, and prepositional relations are available. The grammatical relations are defined using regular expressions over part-of-speech tags; thus, (2) is used to retrieve the verb-object relation in English.

(2) `1:"V" "(DET|NUM|ADJ|ADV|N)"* 2:"N"`

In (2), the 1: and 2: identify the two collocate components. Between the components, zero or more (denoted by the *) words may appear. If any do appear, they may be determiners (*the* or *a*), numbers, adjectives, adverbs or nouns. Other rules are also required for the verb-object grammatical relation (for example to capture the passive form mentioned above). So far nine grammatical relations are available in the Chinese Sketch Engine. These include subject and object relations, possession (with 的 for example), nominal, adjectival and stative verb modification of head nouns,

and/or (using conjunctions and 逗號), and the characteristic association of nouns with classifiers (量詞), demonstrated in Figure 6. The formalisms for three of the relations are given below.

(3) *and/or*

```
1:any_noun listcomma 2:any_noun & 1.tag = 2.tag
1:any_noun conj 2:any_noun & 1.tag = 2.tag
```

(4) *object/object_of*

```
1:trans_verb particle? long_np
```

(5) *subject/subject_of*

```
2:common_noun 1:any_verb
```

By (3), either a conjunction or a 頓號 may occur between the two nominal components. Both must share a tag (for example, they should both be either common nouns or proper nouns). In (4), the particle between the transitive verb and the NP is optional, as shown by the question mark. (5) is self-explanatory.

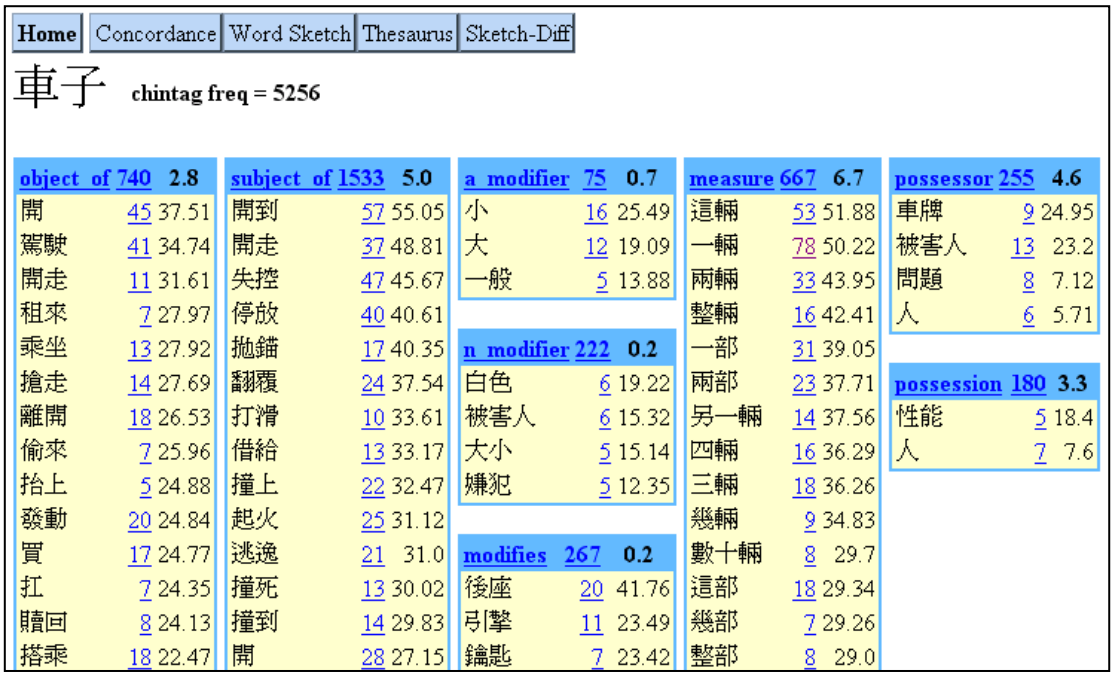

Figure 6 Word sketch for the lemma 車子

Another important task is to refine the set of grammatical relations (essentially grammar rules) for Chinese. The number of grammatical relations should be expanded from the current 9 to a figure similar to the 27 required for English, or the 23 used for Czech. The procedure will be to study large sections of Gigaword and other corpora, as well as other texts, and determine what characteristic patterns emerge. Then each grammatical relation will need to be encoded in the form of regular expression accepted by the Sketch Engine. Each relation will probably require several rule formalisms, and will differ from similar English relations in key respects.

Recall Figure 2 above: Chinese Word Sketch correctly shows that the most common and salient object of 逮捕 is 嫌犯, the most common subject 警方 'police'; and the most common modifier 當場. The reader may have noticed, though, that all does not seem quite as it should be with regard to 逮捕: 嫌犯 occurs both as the subject and object of the verb! A typical Gigaword context is 陸續將其餘十二名嫌犯逮捕, where the object has been fronted by 將. The verb-object relation needs to be equipped to handle fronted objects preceded by 把 and 將. The sample rule at (6) covers such situations.

(6) (把|將) ("N"的) "(DET|NUM|ADJ|ADV|N)"* 2:"N" 1:"V"

In (6), most of the elements are optional (in brackets, or with an asterisk wildcard). It validates verb-object sentence fragments with and without 把: 把功課交 (給老師, for example), 把你的功課交…, 功課交… would all be identified as verb-object relations by this rule.

Another kind of case role error can occur in sentences like 功課還沒有寫完, where the object 功課 precedes the verb. One solution we plan to try is to test for an object **after** the verb, and if there is none assume that the object has been fronted.

5. Future work, and concluding remarks

Initially, Word sketch and other Sketch Engine modules were implemented on a corpus-wide basis only. Now, though, colleagues at Academia Sinica have created two sub-corpora, one of Taiwanese and the other of mainland China data. This means that it will in the future be possible to conduct comparative studies of the two writing styles, using the Sketch Engine. There are also wide applications in the localization adaptations of language related applications.

Before a corpus can be used by the Sketch Engine, it must be segmented into word tokens (in the case of a language like Chinese which does not indicate word boundaries by white space) and then tagged for part of speech. Thus, the Chinese Gigaword corpus was automatically segmented and tagged as a pre-processing step. There remain, however, tagging and segmentation errors which need to be tackled: because 吃 and 吃飯, for example, are listed separately as lemmas, it is not possible for the Sketch Engine to take account of 飯 as a potential collocate of 吃, except where the object component of the compound is separated from the verb component by other material (as in 吃過飯).

In a lexicography application, the decision on which senses to allocate to a word probably should not be completely decoupled from the decision on what group of morphemes (ie characters) constitutes a word. But at the point when the lexicographer consults Sketch Engine, the word segmentation is already done. It would be no trivial matter to implement a "fuzzy" function to allow searches for non-canonical lemmas (i.e. lemmas that are segmented differently from the standard corpus); a potentially useful option to ignore segmentation and treat each Chinese character as a lemma could be put in place more straightforwardly.

The Academia Sinica Balanced Corpus tagset adapts to the fact that Chinese has a freer word order than English by incorporating semantic information with the grammatical category. For instance, locational and temporal nouns, proper nouns, and common nouns each are assigned a different tag. Verbs are sub-categorized according to activity and transitivity. Such information is not available in the BNC tagset and hence not used in the original Sketch Engine design. It is planned to enrich the collocational patterns with the annotated linguistic information from the Sinica Corpus tagset.

An online resource will in the first instance be made available to the academic community, enabling researchers to create their own word sketches dynamically. Those working on word sense discrimination should be able to start using it immediately, and those involved in pedagogical research, or the teaching of Chinese to non-native speakers, should find it useful too.

Ultimately, it is hoped that the Sketch Engine could form part of a Chinese CALL (Computer aided language learning) platform, for the benefit of foreign learners. It could also be adapted for native Chinese elementary school students, who are beginning to learn writing skills.

Bibliography

Church, K. W. and Hanks, P. (1989) Word association norms, mutual information and lexicography. In *Proc. 27th Annual Meeting of ACL*, Vancouver. 1989: 76-83

Crystal, D (1991) *A Dictionary of Linguistics and Phonetics.* Oxford: Blackwell.

Firth, J.R. (1957) A synopsis of linguistic theory, 1930-1955. In Palmer, F.R. (ed) (1968) *Selected papers of J.R. Firth 1952-9.* Harlow: Longman

Kilgarriff, A, Rychly, P, Smrz, P & Tugwell, D (2004). The Sketch Engine, in *Proceedings of EURALEX*, Lorient, France, July 2004

Oakes, M (1998) *Statistics for Corpus Linguistics*. Edinburgh University Press

Wu, Yiching, Simon Smith and Chu-ren Huang, 2005. "How to express 'express' -Pedagogical reflections from web resources", 2005 Conference and Workshop on TEFL and Applied Linguistics, Taoyuan, pp 433-440