# FROM CORPUS TO GRAMMAR: AUTOMATIC EXTRACTION OF GRAMMATICAL RELATIONS FROM ANNOTATED CORPUS

**Chu-Ren Huang**
*The Hong Kong Polytechnic University*

**Jia-Fei Hong***
*National Taiwan Normal University*

**Wei-Yun Ma**
*Academia Sinica, Taiwan*

**Petr Šimon**
*The Hong Kong Polytechnic University*

ABSTRACT

Automatic extraction of grammatical knowledge from corpora has been one of the ultimate goals and challenges of corpus linguistics. We present in this paper [1] one of the approaches to this challenge in Chinese corpus linguistics by introducing our recent work using the Sketch Engine (SkE, also known as Word Sketch Engine)[2] platform to automatically extract grammatical relations from PoS-annotated Chinese corpora. The SkE approach requires both giga-word size corpora and comprehensive lexico-grammatical information of the language in question. On the one hand, corpus size is crucial as the automatic extraction of grammatical relations requires enough instances of the relation pairs, which in turn require an exponential jump from the million-word size corpus for observation of single lexical items. On the other hand, lexico-grammatical information is crucial to the identification of potential relational pairs based on local context. The quality of such extraction is dependent on the quality of available lexico-grammatical knowledge. We show that a comprehensive lexical grammar, based on Information-based Case Grammar (Chen & Huang 1990) and covering over 40 thousand verbs greatly help the accuracy and recall of grammatical relation detection. The paper concludes by underlining the importance of

* Corresponding author: jiafeihong@ntnu.edu.tw

*Linguistic Corpus and Corpus Linguistics in the Chinese Context*
Journal of Chinese Linguistics Monograph Series 25 (2015): 192-221

2409-2878/2015/25-0010

integrating existing grammatical information to meet the challenge of automatic extraction of grammatical knowledge from large corpora.

KEYWORDS
**M**andarin Chinese **G**rammatical knowledge **A**utomatic extraction
**L**exical grammar **S**ketch engine

## 1. BACKGROUND

The original goal of corpus-based studies was to provide 'a body of evidence' for linguistic studies (Kucera & Francis 1967). By this design, corpus linguistics studies involve observation of both the linguistic data contained in the corpus as well as statistics based on word and part-of-speech distributions. Generalizations are then made based on these observations, in the tradition of 'computer-aided armchair linguistics' as described by Fillmore (1992). More recently, automatic acquisition of grammatical information has become one of the most important research topics in corpus linguistics, with improvements in electronic data acquisition and preparation, as well as advances in language technology. This research enables corpus linguistics to have more synergy with its neighboring disciplines, such as computational linguistics, computational lexicography, as well as theoretical linguistics. Previous works that made significant contribution to the study of automatic extraction of grammatical relation include Sinclair's (1987) work on KWIC, Church and Hanks' (1989) introduction of Mutual Information, and Lin's (1998) introduction of relevance measurement.

In Chinese corpus linguistics, Sinica Corpus is the first sharable modern day corpus for Mandarin Chinese (Huang & Chen 1992); even though computational studies of Chinese can be traced back to 1960's (e.g. Dougherty 1969 and Wang 1969). While Sinica Corpus immediately generated some research in 'computer-aided armchair linguistics', e.g. Huang (1994); it also supported research on automatic extraction of grammatical information from very early on. Two examples are Redington et al.'s (1995) work on automatic acquisition of parts-of-speech, and Huang et al.'s (1998) study of extraction of semantic classes based on classifier collocation.

A major and critical difference between earlier studies of automatic extraction of grammatical information (e.g. Church & Hanks 1989; Lin 1998; Huang et al. 1998) and recent approaches is the explicit exploitation of grammatical knowledge. Earlier works rely solely on stochastic information, hence focusing on the 'computer-aided' part. These studies prove to be quick and efficient but unable to reliably attain more sophisticated information, such as grammatical functions. Recent studies incorporate some available grammatical knowledge, such as predicate argument structure information, into the automatic extraction process. In a way, these efforts are trying to formalize the more mechanic part of the grammatical knowledge of armchair linguists in order to discover richer information from corpus and allow linguists to achieve even deeper insights based on automatically extracted patterns. It is important to note, however, the size of available corpora also plays a crucial role in the development in automatic extraction of grammatical knowledge. Automatic extraction of grammatical information relies on the statistic significance of certain linguistic patterns in the corpus, which in turn requires repetition of certain collocations. As number of linguistic collocations is restricted by the frequencies of the elements involved, many mid to low frequency collocations cannot be reliably attested unless the corpus size is over a billion words or bigger in size.

The current paper reports recent developments of such studies on Chinese. Given in the context of the second round-table conference on Chinese corpus linguistics, it should offer some insights on the progress of this field, in comparison to studies reported in T'sou et al. (1998), a volume collecting the first round-table conference on Chinese corpus linguistics.

In what follows, we will first introduce the recent construction and annotation of the Chinese Gigaword Corpus, followed by introduction of the Sketch Engine (SkE, Kilgarriff et al. 2004) approach towards automatic grammatical information extraction. We then introduce our implementation of the SkE for Chinese and discuss how such large-scale implementation can be evaluated. Lastly, we illustrate how the Chinese Word Sketch (CWS) can be applied in linguistics with a study on verbs of ingestion. The paper ends with a session of concluding remarks.

## 2. GARGANTUAN CORPUS AND ITS ANNOTATION: CHINESE GIGAWORD CORPUS

With growing interest in Chinese language processing, many

Chinese corpora of modern Chinese have been assembled and released with query tools in recent years. For example, the Sinica Corpus (CKIP 1995, 1998) developed by Academia Sinica in Taiwan contains 5.2 million words with part-of-speech (POS) tags while the Chinese corpus developed by the Center for Chinese Linguistics (CCL corpus) at Peking University contains 85 million Chinese characters. Both corpora offer the keyword-in-context (KWIC) function for inspecting the context of a given keyword through their web interfaces. However, there are two major restrictions to use both popular online corpora to obtain deeper and comparable Chinese grammatical information. One difference is that although the Sinica Corpus is segmented and POS-tagged, CCL is not segmented and tagged. Therefore it is unable to make deeper syntactic analysis via CCL and is also difficult to compare the syntactic behaviors of a given word between Taiwan and Mainland China. The other difficulty is that only utilizing KWIC concordance is not sufficient to capture and display complete and organized grammatical information of a given word or linguistic unit.

Several other existing linguistic annotated corpora of Chinese, e.g. Penn Chinese Tree Bank (Xia et al. 2000; Xue et al. 2002), Sinica Treebank (Chen et al. 2003), provide more elaborate annotations. The richness of the annotated information belies a different problem: they are all extremely labor-intensive to build and are typically much smaller in size than other corpora. Hence treebanks lack sufficient distributional information for automatic discovery of linguistic generalizations.

To overcome the difficulties described above, a much larger corpus is needed. It turns out that such a corpus was available since 2003, albeit without segmentation or POS-annotation. The Chinese Gigaword Corpus (CGW) was released by Linguistic Data Consortium (LDC) in 2003 (1st Edition, Graff & Chen 2003) and 2005 (2nd Edition, Graff et al. 2005). The first edition of CGW contains about 1.12 billion Chinese characters, including 735 million characters from Taiwan's Central News Agency (CNA) from 1991 to 2002, and 380 million characters from Mainland China's Xinhua News Agency (XIN) from 1990 to 2002. The second edition of CGW contains over 1.29 billion Chinese characters, with the crucial addition of Singapore's Lianhe Zaobao newspaper (ZBN) and additional data from both CNA and XIN. CNA uses the complex character

form and both XIN and ZBN use the simplified character form. CGW has three major advantages for the corpus-based Chinese linguistic research: (1) It is large enough to reflect the real written language usage in either Taiwan or Mainland China, and Singapore to a lesser extent. (2) All text data are presented in a SGML form, using a markup structure to provide each document with rich metadata for further inspecting. (3) CGW is appropriate for the comparison of the Chinese usage among Mainland China, Singapore and Taiwan because it provides the same newswire text type, and these news texts were almost published during the overlapping time period.

A challenging task is to segment and POS-tag such huge amount of corpus efficiently while ensuring high quality without manual checking. Given the corpus size, it is clearly not possible to adopt the semi-automatic approach of human-aided machine tagging to reach the task in the limited time. And given adoption of full-automatic tagging strategy, maintaining high annotation quality is still a major technological challenge. The challenge was met with the release of Tagged Chinese Gigaword in 2007, with an updated version released in 2009 (Huang 2007, 2009). We will describe both the CGW and the methodology adopted to for its automatic annotation in this section.

## 2.1 Content and File Format of CGW

Table 1 (from Huang 2009 and Graff et al. 2005) describes the content of Tagged CGW 2nd Edition in terms of source of the data, number of documents per source, number of characters per source, and number of words per source after tagging.

**Table 1** Basic Information of Chinese Gigaword Corpus 2nd Edition

| | **Source** | **Characters (x 1,000)** | **Words (x 1,000)** | **Documents (x 1,000)** |
|---|---|---|---|---|
| CWG Second Edition | CNA (Central. News Agency) | 792,195 | 501,456 | 1,769 |
| | XIN (Xinhua News Agency) | 471,110 | 311,660 | 992 |
| | ZBN (Lianhe Zaobao) | 28,066 | 18,632 | 41 |
| | TOTAL | 1,291371 | 831,748 | 2,803 |

An interesting fact from this overall description is that the average word length is 1.55 characters for the CGW, while the average word length for the sub-corpora is 1.58 (CNA), 1.51 (XIN), and 1.51 (ZBN) respectively. These are longer than the typical word length of around 1.3 obtained from balance corpora. As the automatic annotation should be biased towards shorter words (as more unknown words will be treated as single-character words), this result strongly suggests that the news report genre favors longer words.

In CGW, each file contains all documents for the given month from the given news source. Hence the metadata of each file contains information of year and month of publication. This information is crucial for comparative studies and for extraction of comparable corpora from different Chinese speaking communities. All text data in CGW are presented in a SGML form, using a very simple, minimal markup structure. The markup structure, common to all data files, can be illustrated by the following example:

(1) Example of a news document in CGW

```
<DOC id="CNA19910101.0003" type="story">
<HEADLINE>
捷運局對工程噪音採多項防治措施
</HEADLINE>
<DATELINE>
(中央社台北一日電)
</DATELINE>
<TEXT>
<P>
台北都會區捷運工程正處於積極趕工階段,…
</P>
<P>
淡水線工程進度百分之三十六點一九,落後百分之二點六七,…
</P>
</TEXT>
</DOC>
```

For every "opening" tag (DOC, HEADLINE, DATELINE, TEXT, P), there is a corresponding "closing" tag. The "id=" attribute of DOC consists of the 3-letter source abbreviation (in CAPS), an 8-digit date string representing the date of the story (YYYYMMDD), a period, and a 4-digit sequence number starting at "0001" for each date (e.g. "CNA19910101.0003"); in this way, every DOC in the corpus is uniquely identifiable by the id string.

### 2.2 Design of Automatic Annotator

The annotation of a Chinese raw corpus involves two main tasks: word segmentation and POS tagging. In order to speed up the process and to maintain high quality at the same time, our automatic annotator has the following characteristics: (1) The annotator takes advantage of the characteristics of CGW. (2) The annotator has the capability to process a large corpus efficiently, which means the program is robust, and hardware resources used by the program are carefully managed. (3) The annotation format meets the requirements of the intended corpus query tool (i.e. Chinese Word Sketch based on SkE). (4) The annotator generates some records of annotation process for speeding up human examination if human examination is still decided to be done in the future. For instance, several word types are more difficult to be correctly identified. The annotator records the list of these unreliable words. If human examination is undertaken in the future, human annotators will only need to examine these records and get much better overall quality in a limited time.

We adapt two attested reliable tools for both segmentation and pos-tagging since the size of CGW makes it infeasible to rely on manual checking as the main quality assurance method, and there is no gold standard available for automatic checking. Ma and Chen (2005) is adapted for word segmentation while HMM algorithm incorporating Tseng and Chen's (2002) morpheme-analysis-based method is adopted for POS tagging with special emphasis of tagging new words. Both tools have been used extensively over the Sinica Corpus, the Sinica Treebank, as well as other text databases by the Academia Sinica team. Hence we have knowledge of error analysis and extensive documentation on the kind of errors that these two tools are likely to generate. Hence we are able to

selectively check and enhance the quality without having to manually check the complete annotation output.

### 2.3 Focus on Out-of-Vocabulary Words for Annotation

Out-of-Vocabulary words (OOV, also referred to as unknown words or new words) are the biggest challenge to automatic annotation, especially when the document size is big and manual correction cannot be performed. The problem is aggravated in Chinese segmentation because most characters can stand alone as a word, hence there is no way to know precisely where the OOV words are and which string of characters is an OOV word. Hence in annotation of Chinese text, OOV requires both prediction and identification. OOV words are the single most significant cause to bring down the performance of word segmentation methods. The percentage of OOV words is especially high in news reports – on average 3% to 5% new words within a news document.

Most popular segmentation technologies (Chiang et al. 1992; Tseng et al. 2005) use corpus-based statistical methods for identifying OOV words with high frequencies and use morphological rules for those with low statistics. However, for these corpus-based statistical methods, they usually suffer a problem that phrases or partial phrases are easily incorrectly identified as words because of their statistical significance in a corpus. Even character strings with strong statistical associations but without any lexical relation are likely to be incorrectly identified as words by this method. On the other side, OOV words with high frequency within a document but typically low frequency in the whole corpus are difficult to identify. This situation is more serious while processing newswire text data. For newswire text data like CGW, a document usually focuses on the same event or subject, and the keywords of a text are often OOV words which are repeated frequently in a news document, but are likely to occur in other parts of the corpus.

Therefore, for statistical methods of our word segmentation, we mainly rely on the document-based statistical information instead of corpus-based statistical information so that the locality of the keywords in a newswire document is fully utilized. Because all text data of CGW are presented in a SGML form, it is convenient to separate CGW into

individual documents using a simple SGML parser. We proposed two strategies of word segmentation by pseudocodes shown in (2) and (3).

(2) Strategy A

For each newswire document-$d_i$
Begin
  Calculate statistical information-$s_i$ from $d_i$
  Extract out new words-$nw_i$ by referencing $s_i$ and
    (probabilistic) morphological rules
  Segment $d_i$ by referencing the basic lexicon and $nw_i$
  Release memory resources for $d_i$, $s_i$, $nw_i$
End

(3) Strategy B

For each newswire document-$d_i$
Begin
  Calculate statistical information-$s_i$ from $d_i$
  Extract out new words-$nw_i$ by referencing $s_i$ and
    (probabilistic) morphological rules
  Release memory resources for $d_i$, $s_i$, but keep the record
    of $nw_i$
End
For each new word in the collection of all *nw*, accumulate
  its frequency from the records of all *nw* and collect those
  new words which accumulated frequencies are higher
  than a threshold. The filtered collection is named as
  NewWordLexicon
For each newswire document-$d_i$
Begin
  Segment $d_i$ by referencing the basic lexicon, $nw_i$, and
    NewWordLexicon
  Release memory resources for $d_i$, $nw_i$
End

In Strategy A, while segmenting a given document, only the basic lexicon and extracted new words of the document are referenced. In Strategy B, while segmenting a given document, we also reference NewWordLexicon collected from other documents. But two things are worth noticing: One is that in NewWordLexicon only new words with high accumulated frequency are covered, which means these words have high reliability as real words. Another is that when referencing these statistics, the statistics of a given document should still play a more important role than NewWordLexicon for resolving segmentation ambiguity.

In addition to fully utilizing locality of newswire data text, Strategy A or B also has another advantage: the memory resource is always controlled within the range of a document, which also means the total processing time will be much shorter than corpus-based statistical methods because the searching space of document-based statistical information is much smaller than the searching space of corpus-based statistical information.

In order to exhibit substantial linguistic differences under consistent querying environment for CNA and XIN, it is necessary to use a unified basic lexicon and POS tag set for annotation. The basic lexicon we used consists of three sources: (1) Sinica lexicon with 80,000 word entries. (2) A 50,000-word set collected from Sinica Corpus 3.0. (3) Xinhua new-words lexicon, which collects 5,000 new words frequently used in Mainland China. We adopt Sinica Tagset as the uniform POS tagset for CNA and XIN.

### 2.4 Annotation Format

We utilize Sketch Engine as the corpus query tool. Besides traditional KWIC function, the engine would automatically generate a one-page, corpus-derived summary of a given word's grammatical and collocation behaviour, such as the distributions of its subjects, objects, preposition objects, and modifiers, by consulting grammatical relations for Chinese. The grammatical relations are defined using regular expressions over POS tags. The more elaborate grammar relations are, the more precise querying results will be obtained.

Therefore in order to facilitate the design of flexible and elaborate grammar relations of Chinese Word Sketch, we adopted mixing POS

tagging strategy: after segmentation and HMM-based tagging process, each word is annotated with the basic POS, such as "陳(Nb1)". And for most words, their basic POS's can be further converted into finer-grain POS's, such as"陳(Nbc1)" by consulting the basic lexicon. However, OOV words are annotated with basic POS's which are obtained by the prediction of the tagger since there is no lexical information available for them. The final annotation results of (1) above can be illustrated by (4). In this illustration, bold and shaded characters represent OOV words and their predicted basic POS's, the others represent words and their POS's from the lexicon or morphological rules.

(4) An annotation example

<DOC id="CNA19910101.0003" type="story">
<HEADLINE>
**捷運局(Nc)** 對(P31) 工程(Nac) 噪音(Nad) 採(VC2) 多(Neqa) 項(Nfa) 防治(VC2) 措施(Nac)
</HEADLINE>
<DATELINE>
((PARENTHESISCATEGORY) 中央社(Nca) 台北(Nca) 一日(Nd) 電(VC2))(PARENTHESISCATEGORY)
</DATELINE>
<TEXT>
<P>
台北(Nca) 都會區(Ncb) 捷運(Nad) 工程(Nac) 正(Dd) 處於(VJ3) 積極(VH11) 趕工(VA4) 階段(Nac) ，(COMMACATEGORY) …
</P>
<P>
**淡水線(Na)** 工程(Nac) 進度(Nad) 百分之三十六點一九(Neqa) ，(COMMACATEGORY) 落後(VJ1) 百分之二點六七(Neqa) ，(COMMACATEGORY)…
</P>
</TEXT>
</DOC>

2.5 Implementation and Evaluation

Strategy A was adopted for first version of Tagged Chinese Gigaword (Huang 2007) and Strategy B with specific heuristic rules to correct common errors found in the first version was adapted for Tagged Chinese Gigaword Version 2 (Huang 2009). An array of machine was used to process CGW, which took over 3 days to perform.

As mentioned, manual verification is not possible given the size of CGW, hence we opt for evaluation of randomly chosen documents. We randomly picked one document from CNA per season and one document from XIN per year. Then there are total 48 documents of CNA and 12 documents of XIN. They are regarded as testing data set for evaluation. These 60 documents are carefully checked by a linguist. The annotation performance is provided in Table 2.

**Table 2** Evaluation result

| | RefWord# | TestWord# | MatchWord# | Recall[a] | Precision[b] |
|---|---|---|---|---|---|
| CNA | 12500 | 12416 | 12186 | 0.97 | 0.98 |
| XIN | 4002 | 3945 | 3790 | 0.95 | 0.96 |

[a]Recall=MatchWord#/RefWord#

[b]Precision=MatchWord#/TestWord#)

| | MatchWord# | MatchPOS# | POS Precision[c] |
|---|---|---|---|
| CNA | 12186 | 12033 | 0.99 |
| XIN | 3790 | 3725 | 0.98 |

[c]POS Precision=MatchPOS#/MatchWord#

The evaluation result shows that our automatic annotator performs very well in either CNA or XIN. The segmentation performance of XIN is a bit lower than CNA probably because most of the words in our basic lexicon are collected from Taiwan sources. In other words, the proportion of new words of XIN is higher than CNA, and these new words caused more segmentation mistakes.

## 3. AUTOMATIC ACQUISITION OF GRAMMATICAL KNOWLEDGE: IMPLEMENTING CHINESE WORD SKETCH WITH SKE

### 3.1 Sketch Engine: A Platform for Automatic Acquisition of Grammatical Information

Kilgarriff and colleagues' work on Sketch Engine (SkE) took an important step forwards in automatic linguistic knowledge acquisition (Kilgarriff & Tudgell 2002; Kilgarriff et al. 2004). The main claim is that a 'gargantuan' corpus[3] contains enough distributional information about most grammatical dependencies in a language such that the set of simple collocational patterns will allow automatic extraction of grammatical relations and other grammatical information. Crucially, the validity of the extracted information does not rely on the preciseness of the rules or the perfect grammaticality of the data. Instead, SkE allows the presence of ungrammatical examples in the corpus and the possibility of collocational patterns to occasionally identify the wrong lexical pairs. SkE assumes that these anomalies will be statistically insignificant, especially when there are enough examples instantiating the intended grammatical information. In addition, SkE relies on Salience measurement to rank the significance of all attested relations. Salience is calculated by MI of a relation multiplied with the frequency of the relation, in order to correct MI's bias towards low frequency items. SkE follows Lin's (1998) formulation of MI of relations, where $||w_1, R, w_2||$ stands for the frequency of the relation R between $w_1$ and, $w_2$. A wild card * can occurs in place of $w_1$, $R$, or $w_2$ to represent the all cases. Hence MI between $w_1$, and $w_2$ given a relation R is given below (Kilgarriff & Tudgell 2002):

$$(5) \qquad I(w_1, R, w_2) = \log\left(\frac{\|*,R,*\| \times \|w_1,R,w_2\|}{\|w_1,R,*\| \times \|*,R,w_2\|}\right)$$

With Salience ranking, SkE gives a one page summary of the most significant grammatical behaviors of any given word. The report includes SUBJ, OBJ, modifier, coordination, etc. SkE is also able to calculate Sketch differences between two sketches, and create automatic thesauri that underline the comparisons between the synonym pairs based on sketch similarity.

3.2 Chinese Word Sketch I: Naïve Adoption

A crucial claim of the SkE is that this methodology can be easily adapted to new languages. That is, each language would require a different set of collocational patterns for relation extraction. SkE has been successfully ported to Czech and Irish (Kilgarriff et al. 2004). And work has been done to produce a prototype of Chinese Word Sketch (called CWS I hereafter for easy reference, Kilgarriff et al. 2005).

One issue not addressed in previous literature on SkE or similar work on automatic extraction of grammatical information is how much can existent grammatical knowledge help. While SkE requires only simple collocational information, it was not clear if more sophisticated grammatical information will help or hurt the result of the SkE. Three previous adaptation of the SkE, including Kilgarriff et al.'s (2005) adaptation of CWS I, relies heavily on transferring the original BNC-based templates to a different language and achieved reasonable results. However, there have been observations that they seem to miss some language-specific grammatical behaviors.

Word Sketch uses regular expressions over POS-tags to formalize rules of collocation patterns. CWS I utilizes 11 collocating patterns to extract all grammatical relations and only one pattern for the simplest verb-object relation is shown as (6).

(6) Collocating Pattern for Object from CWS I

1:"V[BCJ]" "Di"? "N[abc]"? "DE"? "N[abc]"? 2: "Na" [tag!= "Na"]

("XXX" represents XXX is a regular expression, "XXX"? represents XXX appears zero or one time, "XXX"{a,b} represents XXX appears a~b times.)

In (6), the 1: and 2: identify the two collocated components. Between the components, zero or one particle may appear (denoted by "Di"?), zero or one processor may appears (denoted by any_noun? "DE"?), and zero or one noun-modifier may appears (denoted by "N[abc]"?)

Huang et al. (2005) pointed out that the prototype version of CWS I did not deal with the prevalent non-canonical word orders in Chinese (7). In addition, we also noticed that it fails to identify grammatical relations

when an argument lies some distance away from a verb because of internal modification (8). Chinese objects often occur in pre-verbal positions in various pre-posing constructions, such as topicalization.

(7) a. 全穀麵包，吃了很健康。
*quan.gu mian.bao, **chi** le hen jian.kang*
whole-grain bread, **eat** LE very healthy
'Eating whole-grain bread is very healthy.'
b. 有人嘗試要將這荷花分類，卻越分越累。
*you ren chang.shi yao **jiang** zhe he.hua **fen.lei**, que yue fen yue lei*
someone try to **JIANG** the lotus **classify**, but more classify more tired
'People have tried to decide what category the lotus belongs in, but have found the effort taxing.'

(8) 他 只 吃了 一 口 飯 …
T*a zhi **chi** let yi kou **fan***
s/he only **eat** ASP one mouthful **rice**

Such examples led to the question of whether the simple collocation rules adapted in Kilgarriff et al. (2005) was sufficient and if a knowledge-rich approach would yield better results.

### 3.3 Chinese Word Sketch II: A Knowledge-Rich Approach

The important design criteria of SkE is that salience statistics is compiled based on relational tuples such as $\{w_1, R, w_2\}$. This is a crucial decision since word-based lexical statistics itself does not offer enough grammatical information, while it is hard to obtain enough information-rich parsed trees for statistic studies. It is interesting to observe that Kilgarriff et al. (2002) obtained only 70 million tuples based on the 100 million words BNC. In terms of elements that need to be traced, this is indeed comparable to a general bi-gram model and definitely less complex than models that allows any lexical bi-gram without adjacency conditions. The reason for the reduction in complexity is because the collocational patterns serve as filters that disregard

non-significant relations. Based on this model, a set of collocational patterns that contains richer grammatical information will enable the sketch engine to better identify grammatical relation tuples and render more precise grammatical information. Ideally, the most effective collocational patterns are those with explicit annotations of the targeted grammatical relations. Hence we propose to port a lexical grammar with argument annotation as CWS collocational patterns.

In this knowledge-rich approach, we adopt the Information-based Case Grammar (Chen & Huang 1990), a unification-based formalism proposed specifically for Chinese language processing. ICG is a head-driven lexical grammar in the sense that all grammatical information is encoded on the verb. Each verb is encoded with a set of basic patterns (BP) which stipulate the possible structural instantiations of that verb as well as the positions of participant roles (called Case) for each verb. There are over 100 templates of patterns corresponding to each verb sub-class. In the Academia Sinica CKIP lexicon, over 40,000 verbs are annotated with ICG information. Each verb starts with a default assignment according to its verbal sub-class, with the template information manually corrected based on corpus data and linguistic analysis. Obviously, not unlike the Levin classes for English (Levin 1993), each BP is repeated and shared by a number of verb sub-classes. Both the BP information and the Verb sub-classes information will be utilized in our adaptation of Chinese Word Sketch (referred to as CWS II hereafter).

After incorporating lexico-grammatical knowledge from ICG, the patterns for identification of objects for Chinese become fine-grained and more accurate. Comparing CWS II's definition of objects, given in (9) below, to the simplistic rule (6) of CWS we given above:

(9) *Collocating patterns of Object/Object_of*

```
*DUAL
=Object/Object_of

  1:"V[ACFJKL].*" (particle|prep)? NP not_noun
  1:[tag="VH12"|tag="VH14"|tag="VH16"|tag="VH17"|tag="VH2
  2"] (particle|prep)? NP not_noun
```

[word="把"|word="將"|word="向"] NP adv_string 1:"VB.*" [tag!="DE"]
[word="把"|word="將"] NP adv_string 1:"VC.*" [tag!="DE"]
NP_without_NcNd time1 location1 time1 adv? passive_prep adv_string 1:"V[BCJ].*" [tag!="DE"]
NP_without_NcNd time1 location1 time1 adv? passive_prep NP1 adv_string 1:"V[BCJ].*" [tag!="DE"]
begin time1 location time1 adv? passive_prep adv_string 1:"V[BCJ].*" [tag!="DE"]
begin time1 location time1 adv? passive_prep NP1 adv_string 1:"V[BCJ].*" [tag!="DE"]
1:"V[BD].*" (particle|prep|[word="給"])? NP not_noun
1:"VG.*" [word="為"|word="作"]? NP not_noun
[word="對"|word="以"] NP adv_string 1:"VI.*" [tag!="DE"]
1:"VI.*" [word="自"|word="於"]? NP not_noun
[word="對"] NP adv_string 1:"VJ.*" [tag!="DE"]
1:"VD.*" (particle|prep|[word="給"])? NP1 NP not_noun
[word="把"|word="將"|word="向"] NP adv_string 1:"VD.*" (particle|"Ng"|"Ncd.*")? end
NP_without_NcNd time1 location1 time1 adv? passive_prep adv_string 1:"V[DE].*" (particle|"Ng"|"Ncd.*")? end
NP_without_NcNd time1 location1 time1 adv? passive_prep NP1 adv_string 1:"V[DE].*" (particle|"Ng"|"Ncd.*")? end
1:"VE.*" (particle|prep)? NP (particle|"Ng"|"Ncd.*")? end
1:"VE.*" (particle|prep)? NP1 NP (particle|"Ng"|"Ncd.*")? end
[word="向"] NP adv_string 1:"VE.*" (particle|"Ng"|"Ncd.*")? End

CWS II defines 32 relations based on 80 patterns. For instance, the object relation alone has 20 patterns. For CNA data alone, 59,183,238 tuples are defined for over 510 million words. The knowledge incorporated is about 5 times richer than CWS I (11 relations on 11 patterns). The magnitude of knowledge extracted is at roughly the same scale as knowledge extracted from BNC. SkE extracted about 70 million tuples from 939,028 word types based on BNC. CWS, the Chinese

version of SkE with ICG grammatical knowledge incorporated, extracted nearly 60 million tuples from 1,917,093 word types.

3.4 Evaluation and Analysis

Overall evaluation is still being conducted and results will be shown in the final paper. The spot-checking so far does show clear and evident improvements over CWS I.

(10) Object Recall Comparison

| | CWS I | CWS II |
|---|---|---|
| 紅 hong2 (red) | 0 | 0 |
| 跑 pao3 (run) | 0 | 8,704 |
| 看 kan4 (look) | 32,350 | 64,096 |
| 打 da3 (hit) | 26,016 | 47,182 |
| 送 song4 (give) | 0 | 76,378 |
| 説 shuo1 (say) | 0 | 20,350 |
| 相信 xiang1xin4 (believe) | 0 | 52,373 |
| 勸 quan4 (persuade) | 0 | 3,852 |

The recall data compared in (10) underlines the drastic improvement of CWS II over CWS I. For simple transitive verbs (the state verb kan4 and the activity verb da3), CWS II recall almost twice as many objects as CWS I. For more complex verb (ditransitive song4, as well as all types of clause taking verbs xiang1xin4, and quan4), CWS I fails to identify any of their objects, while CWS II correctly extracts their objects. On the other hand, for intransitive verbs, CWS I and CWS II both correctly extract no object relations for the state verb 紅 *hong2*. The fact that CWS II extracted some object relations for the activity verb 跑 *pao3*, although with relatively low frequency, is worth noting. Upon further examination, we found that many of the objects extracted have habitual readings, such as 跑馬拉松 *pao3 ma3la1song1* 'runs marathon' or idiomatic reading 跑白帖 *pao3 bai2tie3* '(of a politician) runs from one funeral to another'. These are additional senses of the lemma *pao3* that to take objects. In sum, the recall comparison data shows improvement of both quality and quantity.

In order to contrast the quality of the extracted grammatical

knowledge, we take the verb *chi1* 'to eat' for a more in-depth analysis. For *chi1*, only 23,421 objects were identified by CWS I, while we identified 33,038 objects with the richer grammar patterns in CWS II. This is an improvement of over 42% in terms of recall and a substantial quantitative gain. In terms of quality improvement, we observed that the following three objects are among the top 20 collocates identified by CWS II, but not by CWS I.

| (11). | Frequency | Salience | Ranking |
|---|---|---|---|
| a. 飯 fan4 rice | 802 | 70.96 | (4) |
| b. 虧 kui disadvantage | 329 | 59.24 | (12) |
| c. 苦頭 ku3tou2 suffering | 194 | 58.71 | (14) |

Note that the three numbers following each object is its frequency (as object of *chi1*), its saliency in this relation, and its saliency ranking (in parentheses). Note that both 吃虧 *chi1-kui1* 'to be taken advantage of' and 吃苦頭 *chi1-ku3tou2* 'to suffer' are both idiom chunks, and expected to be among the most salient collocating objects of *chi1*. However, since they both allow frequent internal modification (e.g. 吃張三的**暗**虧 *chi1 zhang1san1 de **an4** kui1*, 'been taken advantage of **in the dark** by Zhangsan'), a simple collocation pattern such as adopted by CWS I fails to identify them. Our adaptation in CWS II took internal modification into consideration and successfully identified them. The case with 飯 *fan4* 'rice' is even more general and potentially more interesting in terms of extracting basic collocation. Rice is undoubtedly the most typical conceptual object of *chi1* 'to eat' and it occurs frequently in the corpus. However, CWS I only identified 266 instances of *fan4* as object of *chi1*, even less than the 427 instances of 檳榔 *bin1lang2* 'beetlenut'. This is because *fan4* represents a basic and generic concept and is rarely used along without modification. Since it often does not occur in concatenation with the verb, the simple collocation pattern of CWS I cannot identify it. We can see in (10) that CWS II identifies 802 instances of *fan4* as object of *chi1*, a recall improvement of over 200%. In addition, CWS II shows that *fan4* as object of *chi1* is almost twice as frequent as *bin1lang2* (450). This fact is more consistent with our knowledge of the Chinese language and a clear indication that our adaptation successfully

corrected the bias introduced by the incomplete grammatical knowledge of in CWS I. Nevertheless, a recall of instance *fan4* as an object improves over 200% in terms of its identification, misplacement of instance *fan4* as a subject still remains. As CWS II shows, 718 instances of *fan4* as a subject require us to modify our grammar adaptation. In fact, instance *fan4* will never serve the grammatical relation of a subject, hence collocation patterns of object/object_of ought to be adapted according to its sub-classes. In view of 718 instances of *fan4* as a subject, we found that both the negative marker *mei2* and possessive marker *you3* that precede a POS "Na" play a significant role in marking an object and identifying topicalization.

(12) 保證 災民 有 飯 吃、有 衣 穿、有 住處。
baozheng zaimin you fan chi 、yao yi chuan 、yao zhuchu
ensure victims YOU rice eat 、YOU clothes wear 、have dwelling place
'We ensure that the victims will have rice to eat, clothes to wear and have dwelling places.'

(13) 他 相信 水利處 工作 人員 不會 沒有 飯 吃。
ta xiang xin shuilichu gongzuo renyuan buhui meiyou fan chi
he believe department of irrigation and engineering staff won't MEI rice eat
'He believes that the staff in department of irrigation and engineering will have rice to eat.'

The examples above reveal that an object is likely to be identified between *mei2/you3* and "VC.*". In that case, collocating pattern for object in CWS II can be altered and added to extract the very collocation of verb_object like this,

[word="沒"|word="沒有"|word="有"]NP adv_string 1:"VC.*" [tag!="DE"]

Although this collocating pattern cannot capture all the topicalized objects (e.g. 我**飯**吃完就走了 *wo3* ***fan4*** *chi1-wan2 jiu4 zou3-le* 'I will leave as soon as I finish eating.'), it seems to help identify instance *fan4*

as an object as illustrated in CWS II, or rather, it helps to mark the object in another collocation of verb_object indeed. In addition to the collocating pattern illustrated above, there exists a sentence pattern that helps to point out the topicalized objects,

(14) 他 經常 是 一頭 扎進 實驗室 就 連 飯 都 顧不上 吃 。
ta jingchang shi yitou zhajin shi yan shi jiu lian fan dou gubushang chi
he often SHI completely invest laboratory jiu LIAN rice DOU unconcernedly eat
'He often dives right into the laboratory and become so focused that he forgets to eat.'

Example (14) represents a predication of *lian-dou* pattern and the topicalized object fan is in-between. Therefore, we may extract the collocation of verb_object stated as below,

[word="連"] NP [word="都"| adv_string] 1:"VC.*" [tag!="DE"]

Hereby, we still are confronted with one problem as below, though *lian-dou* construction seems to help extract all the topicalized objects:

(15) 這種 飯 就 連 乞丐 都 不 吃。
zhezhong fan jiu lian qigai dou bu chi
This sort rice jiu LIAN beggar DOU not eat
'Even a beggar won't eat this sort of rice. '

In the light of the sentence (15), we are certain to come up with more refined grammar adaptation to capture the real topicalized object that instantiates in the natural language realization. Identifying an object to be a topicalization is really a thorny problem in terms of grammatical knowledge; even though the above suggested collocating patterns advance the identification of object as a sub-class, the goal is aimed to extract all sorts of topicalized objects in CWS II.

## 4. LINGUISTIC STUDIES USING CWS

In this section, we present a lexical semantic study using CWS to illustrate how the enhanced corpus tool can also expedite new discoveries in linguistic studies. Table 3 shows words associated with the two ingestion verbs. The object collocation data extracted through CWS clearly show that actual usage of the two verbs 吃 *chi1* 'to eat' and 喝 *he1* 'to drink' do not follow selectional restrictions based on their lexical meanings.

**Table 3** The common patterns for *chi1'eat'* and *he1'drink'*

| | **More usage for *chi1*** | **Common usage** | **More usage for *he1*** |
|---|---|---|---|
| **object** | yao4 (medicine)、dong1 xi1 (foodstuff)、shi2 wu4 (foodstuff)… | xi1 fan4 (porridge)、xi3 jiu3 (wedding banquet) 、nai3 shui3 (milk)、leng3 yin3 (cooling drink)… | jiu3 (wine)、cha2 (tea)、ku3 shui3 (complaints)… |

### 4.1 Neutralized Selectional Restrictions

In terms of eventive verbal semantics, this data suggest that the general event of ingestion is classified according to the nature of the patients involved. However, one set of challenging facts for selectional restrictions involves cases where they are neutralized (Hong et al. 2008). The example, when an object has both solid and liquid attributes, objects will be selected both by "吃" and "喝" such as below:

(16) 吃 稀飯
to eat porridge

(17) 喝 稀飯
to drink porridge

These neutralization effects can also be found with metaphoric uses. For instance, both wedding banquet (喜酒) and afternoon tea time (下午茶) can be selected by both verbs 吃 and 喝. Therefore, although the widely shared intuition that the two verbs of ingestion 吃 and 喝

select solid and liquid food respectively is supported by various dictionaries and preliminary observation of corpora, corpus data show that there are significant counter-examples such as below:

(18) 吃/喝 稀飯
to eat/drink porridge

(19) 吃/喝 喜酒
to eat/drink wedding banquet

(20) 吃/喝 奶水
to eat/drink milk

The Module-Attribute Representation of Verbal Semantics Theory (MARVS, Huang et al. 2000) offers a straightforward way to account for the three different types of event coercion. The two types of modules in MARVS, event modules and role modules, allow the description of two sets of attributes, event-internal attributes and role-internal attributes. Intuitively, *吃* and *喝* will respectively select the [+solid] feature and the [+liquid] feature for the role-internal attributes of their patients..

4.1.1 Xi1-Fan4 (Porridge)

When a noun such as *稀飯* (porridge) appears as the patient, it satisfies the event representation requirements of both *吃* and *喝* simultaneously since it has both [+solid] and [+liquid] attributes. For example, in *稀飯*, the patient involved contains both [+solid] and [+liquid] substances. Hence the event type is neutralized.

Within our lexical knowledge, porridge contains two primary ingredients: rice (solid) in soup (liquid). Conventionally, it is a kind of rice diluted by water (*稀-飯*, where *飯* is rice) by lexical combination. Hence, in a merged lexical ontology, it should inherit properties from both solid and liquid materials, depending on whether the focus is on the rice or the soup. This kind of representation not only accounts for the fact that both verbs are allowed, but also picks up the subtle focus on the liquid type.

4.1.2 Xi3-Jiu3 (Wedding Banquet)

The metaphorical meaning of *喜酒* 'wedding banquet', literally 'lucky + wine', is eventive and coerced the event representation to shift from entity-type patient to event-type patient, with the [+solid] and [+liquid] attributes inherited. However, since a wedding banquet necessarily includes sub-events of eating solid food and drinking wine, it can select both verbs *吃* and *喝*.

As for *喜酒*, the metaphorical extension of the patient refers to a complex event type which contains separate sub-events that involve ingestion of solid and liquid foods. We use the wine drunk at the wedding banquet to refer to the event. Eating and drinking are the most salient activities at a wedding banquet. The activity involves both eating and drinking, so both verbs *吃* and *喝* are allowed. Since two verbs of ingestion *吃* and *喝* are allowed for *喜酒*, there are [+food] and [+liquid] at this event.

4.1.3 Nai3-Shui3 (Milk)

"*奶水* (milk)" presents the most intriguing situation. The patient involved is clearly liquid. Milk is liquid food in the literal interpretation yet this reading allows substitution of the two verbs of ingestion only when the agent is an infant or young child. Again, when metaphoric uses are involved, "*奶水*" refers to nourishment for either the body or the soul.

We observe that the differentiation of liquid and solid food is significant only for adults as infants and young children can ingest only liquid food. In other words, when the agents are infants or young children, the liquid/solid classification of ingestion events is not applicable since either type of event is sufficient to meet the ingesting goal of nutrition.

Finally, the example of "*奶水* " shows that the classification of events is dependent on the intention of the subject. Even though the solid/liquid contrast does exist in the physical/scientific world, the contrast is not significant for ingestion events involving infants and young children. Hence these two types of events are coerced and neutralized in the intentional context of these subjects.

In this case, a speaker uses both two verbs of ingestion. This is because the metaphor involves nurture and nourishment. Therefore, "*吃 / 喝 奶水*" can be easily represented in MARVS by subject-internal attributes specifying that it allows the liquid/solid contrast to be

neutralized. That is, the Agent Role has the feature [INGEST OBJECT [+/- Solid]].

## 5. CONCLUSION

This study presented our recent research on automatic annotation of the Chinese Gigaword corpus as well as the preparation of grammatical relation extraction rules for Chinese Word Sketch based on the Sketch Engine platform. The first is a breakthrough in Chinese corpus linguistics in terms of construction of gargantuan size corpora, such as web as corpus. The second tool, dependent on the availability of annotated gargantuan corpus prepared by the first task, is a breakthrough in terms of automatic extraction of grammatical information from Chinese corpora as well as for deeper linguistic analysis of corpora. It is important to note that both studies rely crucially on previously acquired grammatical information. The extensive coverage as well as high quality of previously established lexico-grammatical information of Chinese at Academia Sinica is the key to the success of both studies. This bodes well for the future of Chinese corpus linguistics, as we show that corpus is not only a quick and convenient tool to get coarse information. The results reported here demonstrate that corpus supports a paradigm of research such that disciplinary knowledge can be accumulated and researchers are able to stand on the shoulder of giants to gain some insights and make some breakthroughs.

## NOTES

1. This research is partially supported by the “Aim for the Top University Project” and “Center of Learning Technology for Chinese” of National Taiwan Normal University (NTNU), sponsored by the Ministry of Education, Taiwan, R.O.C. and the “International Research-Intensive Center of Excellence Program” of NTNU and Ministry of Science and Technology, Taiwan, R.O.C. under Grant no. NSC 103-2911-I-003-301. In addition, this research is supported by GRF grant no. PolyU 5435/12H and GRF grant no. PolyU 5440/11H.

2. Sketch Engine (SkE, also known as Word Sketch Engine) http://www.sketchengine.co.uk/ (accessed before 02/19/2014)
3. The required corpus size was not specified in SkE literature. However, we estimate from existing work that for SkE to be efficient, corpus scale must be 100 million words or above.

## REFERENCES

CHEN, Keh-jiann, and Chu-Ren Huang. 1990. Information-based Case Grammar. Paper presented at The 13th International Conference on Computational Linguistics, Helsinki (Finland), in its proceedings (COLING '90), Vol. ii, 54 - 59.

CHEN, Keh-Jiann, Chu-Ren Huang, Feng-Yi Chen, Chi-Ching Luo, Ming-Chung Chang, Chao-Jan Chen, and Zhao-Ming Gao. 2003. Sinica treebank: Design criteria, representational issues and implementation. In *Building and Using Parsed Corpora*. Text, Speech and Language Technology: Volume 20, ed. by Anne Abeille, 231-248. Dordrecht: Kluwer.

CHIANG, Tung-Hui, Jing-Shin Chang, Ming-Yu Lin, and Keh-Yih Su. 1992. Statistical models for word segmentation and unknown word resolution. Paper presented at The 22nd Conference on Computational Linguistics and Speech Processing, Taipei (Taiwan), in its proceedings (ROCLING 1992), 121-146.

CHURCH, Kenneth W., and Patrick Hanks. 1989. Word association norms, mutual information and lexicography. Paper presented at The 27th Annual Meeting of the Association for Computational Linguistics, Vancouver(Canada), in its proceedings, 76 - 83.

CKIP *See* Zhongwenci zhishiku xiaozu

DOUGHERTY, Ching-Yi. 1969. A pragmatic approach to machine translation from Chinese to English. Paper presented at The 13th International Conference on Computational Linguistics, Stockholm (Sweden), in its proceedings (COLING '69), a preprint (No. 30).

FILLMORE, Charles. 1992. 'Corpus Linguistics' or 'Computer-Aided Armchair Linguistics.' In *Directions in Corpus Linguistics – Proceedings of Nobel Symposium 82, 4-8 August 1991*, Trends in

Linguistics Studies and Monographs 65, ed. by J. Svartvik, 35-60. Berlin: Mouton.

GRAFF, David, and Ke Chen. 2003. Chinese Gigaword First Edition. Pennsylvania: LDC. https://catalog.ldc.upenn.edu/LDC2003T09

GRAFF, David, Ke Chen, Junbo Kong, and Kazuaki Maeda. 2005. Chinese Gigaword Second Edition. Pennsylvania: LDC. https://catalog.ldc.upenn.edu/LDC2005T14

HONG, Jia-Fei, Chu-Ren Huang, and Kathleen Ahrens. 2008. Event selection and coercion of two verbs of ingestion: A MARVS perspective. *International Journal of Computer Processing of Oriental Language* 21(1):29-40.

HUANG, Chu-Ren. 1994. Corpus-based studies of Mandarin Chinese: Foundational issues and preliminary results. In *In Honor of William S-Y. Wang: Interdisciplinary Studies on Language and Language Change*, ed. by M. Y. Chen and O.J.-L. Tzeng, 165-186. Taipei: Pyramid.

_____. 2007. Tagged Chinese Gigaword. Pennsylvania. LDC Catalog No: LDC2007T03. Philadelphia: Linguistic Data Consortium. https://catalog.ldc.upenn.edu/LDC2007T03, access February 19, 2014.

_____. 2009. Tagged Chinese Gigaword Version 2.0. Pennsylvania: LDC Catalog No: LDC 2009T14. Philadelphia: Linguistic Data Consortium. https://catalog.ldc.upenn.edu/LDC2009T14, access February 19, 2014.

HUANG, Chu-Ren, Kathleen Ahrens, Li-Li Chang, Keh-Jiann Chen, Mei-Chun Liu, and Mei-Chih Tsai. 2000. The Module-attribute representation of verbal semantics: From semantics to argument structure. *Computational Linguistics and Chinese Language Processing* 5(1):19-46, http://www.aclclp.org.tw/clclp/v5n1/v5n1a2.pdf (access February 19, 2014)

HUANG, Chu-Ren, and Keh-Jiann Chen. 1992. A Chinese corpus for linguistic research. Paper presented at The 15th International Conference on Computational Linguistics, Nantes (France), in its proceedings (COLING '92), 1214 - 1217.

HUANG, Chu-Ren, Keh-Jiann Chen, and Zhao-Ming Gao. 1998. Noun class extraction from a corpus-based collocation dictionary: An integration of computational and qualitative approaches. In *Quantitative and Computational Studies on the Chinese Language*,

ed. by B.K. T'sou, T.B.Y. Lai, S.W.K. Chan and W.S-Y. Wang, 339-352. Hong Kong: City University of Hong Kong.

HUANG, Chu-Ren, Adam Kilgarriff, Yiching Wu, Chih-Ming Chiu, Simon Smith, Pavel Rychly, Ming-Hong Bai, and Keh-Jiann Chen. 2005. Chinese sketch engine and the extraction of grammatical collocations. Paper presented at The 4th SIGHAN Workshop on Chinese Language Processing, Jeju (Korea), in its proceedings (SIGHAN '05), 48-55.

KILGARRIFF, Adam, and David Tugwell. 2002. Sketching words. In *Lexicography and Natural Language Processing: A Festschrift in Honour of B.T.S. Atkins*, ed. by Marie-Hélène Corréard. Euralex, 125-137. Grenoble: Euralex.

KILGARRIFF, Adam, Pavel Rychlý, Pavel Smrz, and David Tugwell. 2004. The Sketch engine. Paper presented at the Eleventh EURALEX International Congress, Lorient (France), in its proceedings (EURALEX 2004), 105-116.

KILGARRIFF, Adam, Chu-Ren Huang, Pavel Rychlý, Simon Smith, and David Tugwell. 2005. Chinese word sketches. Paper presented at ASIALEX 2005: Words in Asian Cultural Context, Singapore.

KUCERA, Henry, and W. Nelson Francis. 1967. *Computational Analysis of Present-day American English*. Providence: Brown University Press.

LEVIN, Beth. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. Chicago: University of Chicago Press.

LIN, Dekang. 1998. Automatic retrieval and clustering of similar words. Paper presented at The 36th Annual Meeting of the Association for Computational Linguistics and The 17th International Conference on Computational Linguistics, Montreal (Canada), in its proceedings (COLING '98-36 ACL), 768-774.

MA, Wei-Yun, and Keh-Jiann Chen. 2005. Design of CKIP Chinese word segmentation system, *Journal of Chinese Language and Computing* 14(3):235-249.

REDINGTON, Martin, Nick Chater, Chu-Ren Huang, Li-Ping Chang, Steve Finch, and Keh-Jiann Chen. 1995. The Universality of simple distributional methods: Identifying syntactic categories in Mandarin Chinese. Paper presented at the International Conference on Cognitive Science and Natural Language Processing, July 7-11, Dublin (Ireland).

SINCLAIR, John M. (ed.) 1987. *Looking Up: an account of the*

*COBUILD project in lexical computing*. London: Collins.

TSENG, Huihsin, and Keh-Jiann Chen. 2002. Design of Chinese morphological analyzer. Paper presented at The 1st SIGHAN Workshop on Chinese Language Processing, Taipei (Taiwan), in its proceedings (SIGHAN '02), 49-55.

TSENG, Huihsin, Pichuan Chang, Galen Andrew, Daniel Jurafsky, and Christopher Manning. 2005. A conditional random field word segmenter. Paper presented at The 4th SIGHAN Workshop on Chinese Language Processing, Jeju (Korea), in its proceedings (SIGHAN '05), 168-171.

T'SOU, Benjamin K., Tom B.Y. Lai, Samuel W.K. Chan, and William S-Y. Wang. (eds.) 1998. *Quantitative and Computational Studies on the Chinese Language*. Hong Kong: City University of Hong Kong.

WANG, William S.-Y. 1970. Project DOC: Its methodological basis. *Journal of the American Oriental Society*, 90(1):57-66

XIA, Fei, Martha Palmer, Nianwen Xue, Mary Ellen Okurowski, John Kovarik, Fu-Dong Chiou, Shizhe Huang, Tony Kroch, and Mitch Marcus. 2000. Developing guidelines and ensuring consistency for Chinese text annotation. Paper presented at The 2nd International Conference on Language Resources and Evaluation, Athens (Greece), in its proceedings (LREC'00): no page number.

XUE Nianwen, Fu-Dong Chiou, and Martha Palmer. 2002. Building a large-Scale annotated Chinese corpus. In Paper presented at The 19th International Conference on Computational Linguistics, Taipei (Taiwan), in its proceedings (COLING '02): no page number.

Zhongwenci zhishiku xiaozu 中文詞知識庫小組 (Chinese knowledge information processing group (CKIP)). 1995. Zhongyang yanjiuyuan pingheng yuliaoku de neirong yu shuoming 中央研究院平衡語料庫的內容與說明 (Content explanations of Sinica Corpus). Jishu baogao di95-02hao 技術報告第 95-02 號 (Technical Report no.95-02). Taipei 台北：Zhongyang yanjiuyuan zixunsuo中央研究院資訊所.

_____. 1998. *Cipin cidian* 詞頻詞典 (Word Frequency Dictionary). Jishu baogao di98-01hao 技術報告第 98-01 號 (Technical Report no.98-01). Taipei 台北：Zhongyang yanjiuyuan zixunsuo 中央研究院資訊所.

# 語料與語法：標記語料中語法關係的自動抽取

**黃居仁** 香港理工大學
**洪嘉馡** 臺灣師範大學
**馬偉雲** 中央研究院,臺灣
**石穆** 香港理工大學

提要
從標記語料庫中自動抽取語法知識，一直是語料庫語言學的終極目標挑戰。本研究的研究方法，是透過已經標示詞性的中文語料庫，使用搜尋引擎(Sketch Engine, SkE)平台進行自動抽取中文詞彙，以及語言的綜合詞彙語法的訊息。一方面，語料庫的大小攸關著語法關係自動抽取時，所需要的各種關係的足夠實例，這是需要從千萬字語料庫規模才能觀察得到。另一方面，詞語語法訊息是極為重要的，這是基於所屬語境的潛在關係組的辨識。自動抽取的技術品質是依靠可用詞語語法訊息的品質。我們呈現廣泛詞語語法，基於信息語法(Chen and Huang 1990)和覆蓋率超過40000個動詞，才能有效幫助句法關係偵測，進行檢測的準確度和召回率。最後，本研究強調整合現有的合理語法信息，以滿足從大型語料庫自動抽取語法知識的挑戰的重要性。

關鍵詞
**漢**語 **語**法知識 **自**動擷取 **詞**彙語法 **素**描引擎