Automatic Extraction of English Collocations and their Chinese-English Bilingual Examples: A Computational Tool for Bilingual Lexicography*

Zhao-Ming Gao
National Taiwan University

This paper describes the procedures involved in developing EXEC, a web-based system which can automatically extract English collocations and their Chinese-English bilingual examples from parallel corpora. The system draws on statistics, dependency parsing, and Chinese-English parallel corpora of more than 13 million English words and 27 million Chinese characters. By taking a word as well as the parts-of-speech of the word and its collocate as input, the system can automatically generate collocation candidates based on syntactic dependency relations as well as statistical information regarding mutual information, t-scores, and log likelihood ratios. In conjunction with a Chinese-English bilingual concordancer, it can further extract English sentences containing identified collocations along with their Chinese translations. Our evaluations suggest that the proposed system performs reasonably well in terms of accuracy and efficiency. EXEC can be used in facilitating automatic compilation of bilingual collocation dictionaries as well as in overcoming the L2 language barrier for Chinese learners of English.

Key words: collocation, dependency relation, computational lexicography, parallel corpora, mutual information, t-score, log likelihood ratio

1. Introduction

It is widely recognized that collocations present special difficulties for second language learners. Take the collocation ‘heavy smoker’ for instance. The combination of heavy and smoker is arbitrary and unpredictable. Unfortunately, general dictionaries do not give detailed treatments of collocations. With the ripening of corpus-based computational linguistics, the automatic extraction and compilation of collocations has become a state-of-the-art technology (cf. Church & Hanks 1990, Kilgarriff 2004, among others). However, since most collocation extraction programs are based on monolingual corpora, learners often cannot make sense of the example collocations because of the language barrier. We address this problem by making use of bilingual corpora and a bilingual concordancer.

The definitions of collocation are quite diverse. Firth (1968:182) refers to a collocation as “actual words in habitual company”. Cowie (1978:132) takes...
collocations to be “the co-occurrence of two or more lexical items as realizations of structural elements within a given syntactic pattern.” Martin et al. (1983:84) maintain that “a significant collocation is one in which the two items co-occur more often than could be predicted on the basis of their respective frequencies and the length of the text under consideration.” Kjellmer (1987:133) suggests that “a collocation is a sequence of words that occurs more than once in identical form and which is grammatically well-structured”, while Sinclair (1991:170) defines collocations as “the occurrences of two or more words within a short space of each other in a text.” Benson (1990:23) proposes that “a collocation is an arbitrary and recurrent word combination.” However, according to Svensén (1993:101), a collocation is “a group of words with a certain meaning which tend to occur together.”

Based on the definitions above, two approaches to the study of collocations emerge, namely, a frequency-based approach (Sinclair 1987a, Kjellmer 1987) and a semantically-based or phraseological approach (Benson 1989).

Taking the phraseological approach, Cowie (1981) and Benson (1989) consider collocations a type of word combination. Benson et al. (1997) classify collocations into two types, namely, grammatical collocations and lexical collocations. A grammatical collocation is a phrase consisting of a dominant word (noun, adjective, verb) and a preposition or grammatical structure such as an infinitive or clause, as in depend on. In contrast, lexical collocations do not contain prepositions, infinitives, or clauses, as in deepest regret.

Smadja (1993:146-147) summarizes the following four properties about collocations. (1) Collocations are arbitrary. (2) Collocations are domain-dependent. (3) Collocations are recurrent. (4) Collocations are cohesive lexical clusters. He further classifies collocations into three types, namely, predicative relations (e.g. subject-verb, verb-object, adjective-noun), rigid noun phrases, and phrasal templates.

As Firth (1957:11) aptly observes, “You shall know a word by the company it keeps.” Collocations can be identified using statistical methods such as mutual information and t-scores (Church et al. 1991) as well as log likelihood ratios (Dunning 1993). Drawing on natural language processing tools, researchers have proposed automated procedures to retrieve collocations from corpora (Church & Hanks 1990, Church et al. 1991, Basili & Pazienza 1992, Smadja 1993).

There are several online systems available which can extract English collocations from corpora. However, very few systems support English collocation extraction with bilingual Chinese-English examples. In this paper, we present the design of EXEC, a collocation retrieval system that combines statistics, dependency relations, and a Chinese-English bilingual concordancer. EXEC retrieves English collocations from parallel corpora by taking a keyword and the parts-of-speech of the input keyword and
its collocate as input. The user interface of our system is more in line with the lexicographical practice and search behavior of language researchers. Rather than getting an unorganized and unsystematic list of collocations irrelevant to users’ needs, EXEC ensures that the output of the system returns straightforward query results.

The organization of the paper is as follows. The first section outlines the motivation behind this research and summarizes the definitions of collocations. Section two reviews collocation research in applied linguistics. Section three discusses the statistical and computational aspects of collocation extraction. Section four describes the combination of an English collocation extraction system and a Chinese-English bilingual concordance. Section five evaluates our system against existing dictionaries and online systems. Section six summarizes our research and suggests directions for future research.

2. Collocations in applied linguistics

Sinclair was the first linguist to study collocations using large corpora and computational tools. The product of his project is the widely-acclaimed *Collins Cobuild English Language Dictionary* (Sinclair 1987b), which not only set a new standard for English dictionaries but also opened a new epoch of lexicographical and corpus research.

The task of compiling a collocation dictionary is daunting without the support of a team of lexicographers and computational linguists. This is why there have been only a few dictionaries of English collocations available. *A New Dictionary of English Collocations* by Katsumata (1958) is a monumental work for several reasons. It was probably the first dictionary of English collocations with its first edition dating back to 1939. This comprehensive dictionary with 1525 pages in its second edition was compiled by a Japanese scholar. The dictionary was innovative in providing English examples and Japanese translations. This is invaluable for Japanese learners of English who have difficulties in understanding the meanings of the examples. In addition, the dictionary adopted a new format which later dictionaries followed. The example in (1) contains information on the entry “exploration” with the Japanese translations removed. The organization of the examples was innovative, as the collocates are arranged by their grammatical functions and parts-of-speech. For instance, the labels V, Q, Q^2, P, P^2 refer to verbs, adjective modifiers, noun modifiers, predicative prepositions, and prepositions modifying a noun, respectively. Note that the same parts-of-speech are further classified using superscripts to reveal different syntactic functions. For example, the prepositions ‘under’ and ‘of’ function differently in “The answer to this question is now under exploration” and “the exploration of the
sea depths”. In the first example the prepositional phrase ‘under exploration’ is predicative, whereas in the second example the prepositional phrase ‘of the sea depths’ is modifier. Therefore, the prepositions ‘under’ and ‘of’ in these two examples are given different labels P and P^2.

(1) exploration (Katsumuata 1958:448)

V make explorations
Q aerial exploration; Arctic exploration under Nansen; give long hours of toil to the patient exploration of written records
Q^2 mountain exploration
P The answer to this question is now under exploration.
P^2 the exploration of the sea depths

The *BBI Dictionary of English Word Combinations* by Benson et al. (1997) is another influential work on English collocations. Designed for learners of English, the meaning of each collocation is provided. This is of invaluable help to learners, because the meanings of collocations are often unpredictable and different from words used in isolation (e.g. heavy smoker vs. heavy box). Its layout, however, is not very user-friendly, as parts-of-speech and grammatical functions are not provided and users need to go through an entry from beginning to end in order to find a collocate.

Two recent dictionaries of English collocations are the *Oxford Collocations Dictionary for Students of English* and the *Macmillan Collocation Dictionary*. These two dictionaries are products of corpus-based computational lexicography. They were compiled using large corpora and computational linguistic techniques. Like Katsumata (1958), the collocates of a keyword in these dictionaries are arranged based on the part-of-speech, which facilitates the process of looking up an item.

While the first dictionary of English collocations was published as early as 1939, it was not until the publication of Michael Lewis’s *The Lexical Approach* in 1993 that ELT/TESOL experts began to stress the importance of collocations in language learning. Lewis (1993), following the British linguistic tradition from Firth to Halliday and Sinclair on lexis and grammar, holds that language consists of prefabricated chunks, of which collocation is one of them. Lewis notes that collocations are arbitrary and gives interesting examples (e.g. high/tall building, tall boy, but not high boy). Lewis encourages teachers and students to explore corpora by using a concordance, which can display examples containing a keyword.

In the early 1990s, a school under the name of data-driven language learning began to emerge. Scholars of this school, represented by Tim Johns and his colleagues, take the position that language learners are like researchers. They need to observe
language, make hypotheses about the rules underlying linguistic phenomena, and test their hypotheses by looking at authentic examples from corpora. The tool they use is a concordancer, which can display a keyword in context. Unfortunately, a concordancer is not well-suited to studying collocations, because it takes too much time to find a collocation. As stated earlier, collocations are words that tend to co-occur more often than by chance. In other words, collocations can be identified by statistical means. Once a collocation program is developed, it has wide-ranging applications in computer-assisted language learning (CALL).


3. Collocations from the viewpoint of computational linguistics

Since collocations are words that frequently co-occur, several researchers have proposed using statistical methods to extract collocations from corpora. Church & Hanks (1990) and Church et al. (1991) use mutual information (MI) and t-scores to identify collocations. MI is computed using the following formula.

\[
I(x,y) = \log_2 \frac{P(x,y)}{P(x)P(y)}
\]

Church & Hanks (1990:23)

\(P(x,y)\) in the formula in (2) is the probability of the co-occurrence of the two words \(x\) and \(y\), whereas \(P(x)\) and \(P(y)\) are the probabilities of occurrence of \(x\) and \(y\), respectively. From the formula in (2), one can infer that MI becomes large when \(P(x)\) and \(P(y)\) are small and \(P(x,y)\) is large. In other words, if two words which rarely occur in the corpus frequently co-occur with each other, their MI value becomes very large. Theoretically, the threshold of MI is 0 for a large corpus. However, to avoid pure coincidence, some researchers suggest a minimal frequency of 10 for the word and its collocate. In addition, some scholars propose that the threshold of MI be higher than
3.

T-score, a statistical significance test, is used to filter out word pairs which occur by chance. T-score can be approximated with normalization as in (3) (cf. Church 1994).

\[
T = \frac{f(x,y) - \frac{f(x)f(y)}{N}}{\sqrt{f(x,y)}}
\]

(3)

In (3), \(f(x)\), \(f(y)\), and \(f(x,y)\) are the frequency counts of \(x\), \(y\), and \(x\) co-occurring with \(y\), respectively; \(N\) is the number of occurrences of all the tokens in the text. Like MI, the higher the t-score, the more likely it is for the word pair to be a genuine collocation. The threshold of a t-score is 1.65 when the corpus is large enough.

Although MI and t-scores can identify collocations when used together, there is a serious limitation. Most notably, they require a large corpus and are not suitable for use with small texts. Several alternatives to MI and t-scores have been proposed. Some of these methods require a contingency table like (4):

\[
\begin{array}{cc}
\text{a} = f(A \, B) & \text{b} = f(\neg \, A \, B) \\
\text{c} = f(A \, \neg B) & \text{d} = f(\neg A \, \neg B)
\end{array}
\]

(4)

where \(A\) and \(B\) are the words in question, and \(f\) is the frequency count of the co-occurrence of \(A\) and \(B\). The \(\neg\) sign means ‘not’. For instance, \(c\) is the frequency count of the word pair where \(A\) is followed by a word other than \(B\).

Dunning (1993) notes that MI is subject to overestimation when the counts are small and thus proposes using log likelihood ratio \(G^2\) as a significance test for estimating surprise and coincidence of a rare event. \(G^2\) is computed by the formula in (5). The definitions of \(a\), \(b\), \(c\), and \(d\) are in (4).

\[
G^2 = f(a) + f(b) + f(c) + f(d) - f(a+b) - f(a+c) - f(b+d) - f(c+d) + f(a+b+c+d)
\]

(5)

\[\text{where } f(x) = x \log(x)\]

In fact, syntactic dependency information is as important as statistical information and has been used by Church & Hanks (1990), Church et al. (1991), Smadja (1993), Lin (1998), Kilgarriff (2004), and Seretan (2011) to extract English collocations. Traditionally, in a frequency-based approach, collocations are identified in a window size, which is normally a span of five words on either side of a keyword based on
traditional frequency-based approach. Following Benson (1989) and Cowie (1981),
we take the position that collocations involve certain syntactic relations such as
subject-verb, verb-object, and modifier-modifiee. These relations are also referred to
as dependency relations. Compared with the frequency-based approach which
computes the statistics of potential collocations in a window size of five words around
a keyword, we focus on words which form dependency relations with the keyword.
However, unlike Church & Hanks (1990), Church et al. (1991), we do not employ
statistical information first and use dependency relations as a filter. Instead, our
approach is more in line with Kilgarriff (2004) and Seretan (2011) in treating
dependency relations as a perquisite to significant collocations and use statistical
information such as MI, t-scores, and log likelihood ratios as supportive evidence.

The procedures involved in developing our system are as follows. The system first
retrieves all the examples in the corpus which contain the keyword. It then checks the
parts-of-speech of the keyword and its collocates as well as the dependency relations
between them. If the information meets the search conditions, the system then
computes and ranks the mutual information, t-score, and log likelihood ratio of the
two words. As an English word (more precisely a lemma) may have various forms, e.g.
take, took, taken, taking, takes, both the frequency information and the dependency
relations we use in our system are based on lemmas (i.e. the basic form of a word).
Each word in the corpus was converted into its lemma via WordNet. In addition, the
dependency relations in each sentence of the corpus were stored in the database to
speed up the query process. The user interface of the system is shown in Figure 1.
Users input the parts-of-speech of both the keyword and the collocates they are
looking for. The parts-of-speech include noun, verb, adjective, adverb, and preposition.
Noun is further divided into noun in the object position and noun in the subject
position. This distinction is important in lexicography but is ignored in most
collocation extraction systems. Prepositions are included, because there are
collocational patterns which involve a verb and a preposition (e.g. rely on, or a
preposition and a noun (e.g. under attack).
Figure 1. The user interface of our collocation program

Essential to our algorithm is a sophisticated English parser capable of identifying dependency relations in a sentence as well as the head of a noun phrase. There are a number of parsers available which can identify dependency relations such as Minipar (Lin 1998) and the Stanford parser released after version 1.5 in 2005 (cf. Klein & Manning 2002). (6b) is the output of the sentence in (6a) by the Stanford parser. The output of the Stanford parser consists of three parts. The first part is the parts-of-speech of the words in a sentence. The second part is the parse tree of the sentence. The third part outputs the dependency relations in the sentence. The Stanford parser can identify several dependency relations such as nsubj (i.e. subject-verb), dobj (i.e. verb-object), modifier-noun, and verb-preposition. In (6b), the Stanford parser correctly identifies Egypt as the subject of the verb criticizes. It also correctly identifies decision and aid as the objects of the verbs criticizes and halt, respectively. The performance of the Stanford parser varies with the complexity of the sentence.

(6)  a. Egypt criticizes U.S. decision to halt aid.
    b. Egypt/NNP criticizes/VBZ U.S./NNP decision/NN to/TO halt/VB aid/NN ./.
c. (ROOT
(S
  (NP (NNP Egypt))
  (VP (VBZ criticizes)
    (NP (NNP U.S.) (NN decision))
  )
(S
  (VP (TO to)
    (VP (VB halt)
     (NP (NN aid))))))
( . ))
nsubj(criticizes-2, Egypt-1)
nn(decision-4, U.S.-3)
dobj(criticizes-2, decision-4)
aux(halt-6, to-5)
infmod(decision-4, halt-6)
dobj(halt-6, aid-7)

With the incorporation of the Minipar and the Stanford parser, our collocation retrieval system can couple dependency relations with statistics. We extracted all the dependency relations, stored them in a dependency relation database (similar to the one extracted by Minipar parser as shown in Figure 2), and combined the intersections of the output of the Minipar and the Stanford parser. Use of the intersections of the output of the two parsers is to reduce the errors made by either one of the parsers. The tables in the dependency relation database include the ID of the dependency relation, the headword, its part-of-speech, the dependency relation between the headword and its collocate, the collocate of the headword, the part-of-speech of the collocate, the ID of the sentence containing the collocation, and the corpus from which the collocation was extracted. We then added up the frequency count of each collocation of the same word pair, parts-of-speech, and dependency relation and computed its mutual information (MI), t-score, and log likelihood ratio. This information was stored in the collocation database. When the collocation retrieval system is executed, it will search the collocation database, find all the collocations of a headword in accordance with the conditions input by the user, and show the statistics of MI, t-score, and log likelihood ratio. In addition to the statistical summary of potential collocations, the system allows users to view sentences containing a collocation by clicking a hyperlink.
Figure 2. Tables in a dependency relation database extracted by Minipar

Figure 3 is the output of our English collocation retrieval system for the collocation patterns of V + brake. Note that the potential collocations are ranked based on the log likelihood ratio. Our system extracts significant verb collocates of the noun brake such as: put, apply, jerk, step on, slam, hold to, jam, hit, and handle. By clicking the hyperlink in the last column, learners can see examples of the collocation.

There are, however, some complications with the collocation extraction system. As with other collocation extraction programs, the system is error-prone. Most of the collocation errors are due to errors in dependency relations identified by the Minipar and the Stanford parser. In addition, some collocations occur only once or twice in the corpora. With such low frequency of co-occurrence, it is difficult to identify genuine collocations. Moreover, each statistical method has its own advantages and disadvantages. Although theoretically the thresholds for identifying collocations based on the mutual information, t-score, and log likelihood ratio are 0, 1.65, and 7.88, respectively, in practice one cannot reliably conclude if a word combination is a significant collocation simply by looking at the log likelihood ratio, mutual information, and t-score. For instance, ‘hit the brake’ is a significant collocation, whereas ‘invent a brake’ is not. Nevertheless, both the log likelihood ratio and mutual information of the latter are higher than those of the former in Figure 3. Similarly, in terms of the t-score, only put in Figure 3 is statistically significant, while all the other
verbs are below the threshold of 1.65 required by the t-score. However, apply, jam, jerk, hit, hold to, slam, step on are all genuine collocations. In fact, both the t-score and log likelihood ratio are statistical significance tests. If the log likelihood ratio exceeds 7.88, it means that we can reject the null hypothesis that there is no relation between them on a confidence level of $\alpha = 0.005$. In other words, the greater the value of a word pair over the thresholds of the t-score and log likelihood, the more likely it is to be a significant collocation. This also applies to mutual information, which measures the strength of association between two words. To summarize, the thresholds of the mutual information, t-score, and log likelihood cannot be taken as absolute indicators of whether a word pair is a significant collocation. This means that users and learners cannot determine if a word pair is a significant collocation simply by looking at these statistical measures. They need to inspect examples to determine their collocational strength especially in marginal cases.

### Figure 3. Output of our collocation retrieval system

4. Combining an English collocation extraction system with a parallel Chinese-English concordancer

For nonnative speakers of English, understanding the meanings of examples from authentic corpora is a big challenge. If their vocabulary and grammatical knowledge is not good enough, they probably cannot make sense of the example sentences. Since...
one of the main purposes of our collocation retrieval system is to provide Chinese learners of English with a tool to consult when writing, it will only be useful if the example sentences are understood and they can determine which collocation to use in a given context. The most effective way to help Chinese learners of English overcome the language barrier and make sense of difficult English sentences is to present such sentences together with Chinese translations. This is done by incorporating our English collocation identification system with a subset of the CERT Chinese-English parallel corpora reported in Gao (2011). The parallel corpora used in the current study contain bilingual texts from different sources. These include the *Concise Encyclopedia of Britannica*, *Scientific American*, *Sinorama Magazine*, Environmental News, an English Idiom book, and dozens of novels, totaling over 13 million English words and 27 million Chinese characters. The source language is English except for the Sinorma Magazine, whose articles are written in Chinese and translated into English by native speakers of English.

A bilingual concordancer consists of a parallel corpus with bilingual sentence pairs and a retrieval program. The bilingual sentence pairs are identified by sentence alignment programs, which can find the correspondence of a sentence in one language and its translation in another language. We use the Champollion tool kit by Ma (2006) to align our bilingual corpora. Figure 4 shows the interface invoking the Champollion tool kit to align Chinese-English bilingual texts. Champollion uses hybrid algorithms integrating information such as sentence length, dictionary lookup, numerals, and foreign terms to improve the precision of sentence alignment. Figure 5 shows the output of the sentence aligner. The symbol 1 $\leftrightarrow$ 1, 2 indicates that the first English sentence corresponds to the first and the second Chinese sentences.

As all sentence alignment programs are error-prone, it was necessary to design an interface with which users could search for the correct sentence correspondence easily if the automatically identified sentence alignment was inaccurate. We have designed a method to help users find the correct correspondence by allowing them to inspect the neighboring sentences of a proposed sentence alignment using Ajax. Our experiments show that this approach is quite effective and efficient, since the correct sentence correspondence usually lies around the proposed pair even if the correspondence is incorrect. In order to expedite the search process, a search engine was adopted.
We employ Lucene, an open source Java-based tool kit for search engines to
index and retrieve the sentence-aligned bilingual texts. With the Lucene search engine, we are able to retrieve examples containing the input query as well as their translations efficiently. Figure 6 shows the output of a search for bilingual examples containing the collocation *assume the responsibility*. From the output of the system, we can establish a correspondence between *assume the responsibility* and the synonymous Chinese phrases 負起責任, 承擔職責, 承擔責任, 出任該職.

**Figure 6. Output of the search for bilingual examples containing a collocation**

As the program is designed for higher-intermediate learners of English, we can reasonably assume that users are capable of detecting incorrect sentence alignment. When errors in sentence alignment occur, they can usually find the correct alignment in the neighboring sentences of the proposed sentence pair. By clicking a link to the previous sentence and the next sentence, the system can retrieve neighboring sentences.

5. Evaluations of the system

We made quantitative and qualitative evaluations of our system based on the *Oxford Collocations Dictionary for Students of English* (henceforth the OCD), which embodies the corpus-based approach to lexicography. The two online collocation extraction systems under comparison with our system are the Corpus of
Contemporary English (henceforth COCA) and the Sketch Engine. COCA and the Sketch Engine are by far the most widely used free and commercial corpus query systems. They also represent different approaches to collocation extraction. COCA takes the frequency-based approach, whereas the Sketch Engine takes the phraseological approach.

Table 1 and Table 2 show the collocation patterns of nouns and verbs that can be extracted by the three collocation extraction systems using the information in the OCD as the basis of comparison. COCA is more flexible than the Sketch Engine and our system in that it can handle most of the queries of collocation patterns involving nouns and verbs. However, this comes with a price, as it is also the least efficient and straightforward of the three systems. Note that none of the systems can handle phrases of indefinite length such as *act as a brake on something* or *a screech of brakes*. While the Sketch Engine can extract phrasal verbs if the keyword is a verb, it cannot extract phrasal verbs if the keyword is a particle or a preposition.

Since the precision and recall rate typically used in information retrieval tasks do not fit our applications, we proposed two measures, i.e. accuracy and coverage, for evaluating the performance of our system. The accuracy rate is the percentage of the number of correct collocations vs. the number of all the collocations identified by the system, whereas the coverage rate is the percentage of the number of shared collocations by the system and the dictionary to the number of collocations listed in the dictionary.

We randomly chose *responsibility* for evaluation. Table 3 shows the accuracy and coverage rate of the collocations identified. Note that the numeral figures outside the brackets in Table 3 indicate the number of collocations extracted by our system, whereas the first and second numeral figures inside the brackets represent the number of correct collocations extracted by our system and its intersection with the collocations listed in the OCD. Note that the accuracy and coverage rate for the collocations are over 80%.

Figure 7 shows the partial output of the verb collocates of *responsibility* in the object position identified by our system. Table 4 shows that our system can extract more collocates than those listed in the BBI.
### Table 1. Collocation patterns of nouns

<table>
<thead>
<tr>
<th>Patterns: Keyword= N</th>
<th>OCD</th>
<th>COCA</th>
<th>Sketch Engine</th>
<th>Our system</th>
</tr>
</thead>
<tbody>
<tr>
<td>A+ Keyword</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Keyword+ V</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>V + Keyword</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Keyword + N</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>N+ Keyword</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>P + Keyword</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Quantifier + N</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>-</td>
</tr>
<tr>
<td>Phrases</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 2. Collocation patterns of verbs

<table>
<thead>
<tr>
<th>Patterns: Keyword= V</th>
<th>OCD</th>
<th>COCA</th>
<th>Sketch Engine</th>
<th>Our system</th>
</tr>
</thead>
<tbody>
<tr>
<td>ADV + Keyword</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Keyword + N</td>
<td>-</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>N + Keyword</td>
<td>+</td>
<td>+</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Keyword + P</td>
<td>+</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>Phrasal Verbs</td>
<td>+</td>
<td>+</td>
<td>-/+</td>
<td>+</td>
</tr>
<tr>
<td>Phrases</td>
<td>+</td>
<td>-</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

### Table 3. Comparisons of collocations in the Oxford Collocation Dictionary and our system

<table>
<thead>
<tr>
<th>Patterns: Keyword = N</th>
<th>OCD</th>
<th>RESPONSIBILITY (N)</th>
</tr>
</thead>
<tbody>
<tr>
<td>A + Keyword</td>
<td>43</td>
<td>86{81, 33}</td>
</tr>
<tr>
<td>Keyword + V</td>
<td>3</td>
<td>58{44, 3}</td>
</tr>
<tr>
<td>V + Keyword</td>
<td>32</td>
<td>99{96, 27}</td>
</tr>
<tr>
<td>Keyword + N</td>
<td>0</td>
<td>15{10, 0}</td>
</tr>
<tr>
<td>N + Keyword</td>
<td>0</td>
<td>5{0, 0}</td>
</tr>
<tr>
<td>Keyword + P</td>
<td>2</td>
<td>23{19, 2}</td>
</tr>
<tr>
<td>Total</td>
<td>80</td>
<td>286{250, 65}</td>
</tr>
</tbody>
</table>

Statistics
- Accuracy=87.4%
- Coverage =81.2%
Table 4. A comparison of the verb collocates of responsibility identified by our system with those listed in the BBI and the OCD

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>accept</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>assume</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>shoulder</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>take</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>bear</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>exercise</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>share</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>dodge</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>evade</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>lay</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>admit</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>claim</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>disclaim</td>
<td>+</td>
<td>+</td>
<td></td>
</tr>
<tr>
<td>have</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>take on</td>
<td></td>
<td></td>
<td>+</td>
</tr>
</tbody>
</table>
Figure 8 shows some of the verb collocates of responsibility in the subject position. Our system retrieved 58 verbs which take responsibility as a subject. However, of the 58 verbs, only 44 of them are indeed verbs that take responsibility as a subject, achieving an accuracy of 75.86% for subject-verb relation. Identification of subject-verb relations is, in most cases, harder than the verb-object relation. For instance, as shown in Table 3, the accuracy rate for identifying verbs which take responsibility as object is 96/99=97%. The errors arise from errors in parsing. The parsers sometimes make an incorrect analysis of the part-of-speech of a word or the structure. As shown in (7), the verb saving in (7a) and remains in (7b) are misanalysed by the parser as the verbs of the noun responsibility due to ellipsis.
Gao: A Computational Tool for Bilingual Lexicography

(7) a. Sun Hsiao-chih, an associate professor of philosophy at NTU, points out that **saving** or taking responsibility for someone else's life requires great sacrifice. So doing represents a moral ideal, not a moral duty.

b. Even if one of a company's taxis is involved in an accident or dispute, at most the company will help by mediating, but any liability **remains** the responsibility of the driver himself.

<table>
<thead>
<tr>
<th>collocation</th>
<th>co-occurrences</th>
<th>probability of collocation</th>
<th>t score</th>
<th>log likelihood ratio</th>
<th>examples</th>
</tr>
</thead>
<tbody>
<tr>
<td>lie</td>
<td>5</td>
<td>3.06e-07</td>
<td>1.14e-04</td>
<td>5.248</td>
<td>2.177</td>
</tr>
<tr>
<td>protect</td>
<td>3</td>
<td>2.20e-07</td>
<td>1.16e-04</td>
<td>4.495</td>
<td>1.655</td>
</tr>
<tr>
<td>maintain</td>
<td>3</td>
<td>2.20e-07</td>
<td>1.47e-04</td>
<td>4.153</td>
<td>1.355</td>
</tr>
<tr>
<td>cry out for</td>
<td>1</td>
<td>7.32e-08</td>
<td>1.46e-05</td>
<td>9.215</td>
<td>0.998</td>
</tr>
<tr>
<td>secure</td>
<td>2</td>
<td>1.46e-07</td>
<td>8.49</td>
<td>6.22e-05</td>
<td>4.808</td>
</tr>
<tr>
<td>fall upon</td>
<td>1</td>
<td>7.32e-08</td>
<td>9.55e-06</td>
<td>7.429</td>
<td>0.994</td>
</tr>
<tr>
<td>distress</td>
<td>1</td>
<td>7.32e-08</td>
<td>9.39e-06</td>
<td>7.197</td>
<td>0.993</td>
</tr>
<tr>
<td>intervene</td>
<td>1</td>
<td>7.32e-08</td>
<td>9.83e-06</td>
<td>6.998</td>
<td>0.992</td>
</tr>
<tr>
<td>liberate</td>
<td>1</td>
<td>7.32e-08</td>
<td>9.17e-06</td>
<td>6.923</td>
<td>0.992</td>
</tr>
<tr>
<td>diminish</td>
<td>1</td>
<td>7.32e-08</td>
<td>201</td>
<td>1.47e-05</td>
<td>5.886</td>
</tr>
<tr>
<td>include</td>
<td>5</td>
<td>3.66e-07</td>
<td>1.12e-03</td>
<td>1.956</td>
<td>1.660</td>
</tr>
<tr>
<td>broaden</td>
<td>1</td>
<td>7.32e-08</td>
<td>226</td>
<td>1.65e-05</td>
<td>5.717</td>
</tr>
<tr>
<td>improve</td>
<td>2</td>
<td>1.46e-07</td>
<td>2202</td>
<td>1.61e-04</td>
<td>3.433</td>
</tr>
<tr>
<td>stipulate</td>
<td>1</td>
<td>7.32e-08</td>
<td>240</td>
<td>1.76e-05</td>
<td>5.630</td>
</tr>
<tr>
<td>draw up</td>
<td>1</td>
<td>7.32e-08</td>
<td>256</td>
<td>1.87e-05</td>
<td>5.537</td>
</tr>
<tr>
<td>formulate</td>
<td>1</td>
<td>7.32e-08</td>
<td>271</td>
<td>1.98e-05</td>
<td>5.455</td>
</tr>
<tr>
<td>educate</td>
<td>1</td>
<td>7.32e-08</td>
<td>331</td>
<td>2.42e-05</td>
<td>5.167</td>
</tr>
<tr>
<td>look after</td>
<td>1</td>
<td>7.32e-08</td>
<td>441</td>
<td>3.23e-05</td>
<td>4.753</td>
</tr>
<tr>
<td>take care of</td>
<td>1</td>
<td>7.32e-08</td>
<td>443</td>
<td>3.24e-05</td>
<td>4.746</td>
</tr>
<tr>
<td>rest</td>
<td>1</td>
<td>7.32e-08</td>
<td>455</td>
<td>3.33e-05</td>
<td>4.708</td>
</tr>
<tr>
<td>revive</td>
<td>1</td>
<td>7.32e-08</td>
<td>462</td>
<td>3.38e-05</td>
<td>4.686</td>
</tr>
<tr>
<td>shift</td>
<td>1</td>
<td>7.32e-08</td>
<td>490</td>
<td>3.39e-05</td>
<td>4.601</td>
</tr>
<tr>
<td>lower</td>
<td>1</td>
<td>7.32e-08</td>
<td>670</td>
<td>4.90e-05</td>
<td>4.149</td>
</tr>
<tr>
<td>range</td>
<td>1</td>
<td>7.32e-08</td>
<td>682</td>
<td>4.99e-05</td>
<td>4.124</td>
</tr>
</tbody>
</table>

**Figure 8. The output of the verb collocates of responsibility in the subject position**

Comparing our system with COCA and the Sketch Engine, we find that it has some unique features. Take COCA for example. It can only identify collocations in a given window size (e.g. five words both sides) of the input keyword. Collocations outside the window size cannot be identified. Furthermore, words identified as collocates are not necessarily genuine collocates. They might just happen to co-occur with the keyword without having any syntactic relation with it. Searching for collocations in COCA is more time-consuming and less straightforward, as the users need to decide the window size and go through the list of the output laboriously before finding what they want. The design of our interface and the output of our system are more in line with the lexicographical practice. The collocates of a keyword are distinguished based on their part-of-speech information and syntactic functions. Therefore, users of our system can quickly find a VN collocation or NV collocation...
by specifying a keyword in the object or subject position. Unlike COCA, the results of a query in our system are straightforward and unambiguous and users do not have to spend a lot of time filtering out irrelevant information from the data output.

Like the Sketch Engine, our system also makes use of syntactic dependency relations such as subject-verb, verb-object, and modifier-noun. However, it differs from the Sketch Engine in that it can provide bilingual English-Chinese examples and retrieve phraseological units, which the Sketch Engine cannot. For instance, in Figure 7 and Figure 8, our system is able to identify phrasal verbs such as *take over, carry out, turn over, draw up, and look after*. Phrasal verbs are multi-word units consisting of a verb followed by a particle or a preposition. Many existing collocation extraction system cannot extract phrasal verbs simply by specifying a particle and its part-of-speech. For instance, the Sketch Engine can only accept three parts-of-speech, namely, verb, noun, and adjective. It cannot accept queries involving the parts-of-speech of prepositions, adverbs, and particles. In other words, the Sketch Engine does not allow users to extract collocates of an adverb or a particle such as *over, up, or off*.

While Tango, the advanced collocation extraction system reported in Jian et al. (2004) supports Chinese-English bilingual searches, it is limited to four types of collocation pattern, namely, A N, V N, V N P and V P N. It cannot extract patterns such as V ADV. In contrast, our system can accept queries for both V P and V ADV. It should be noted that particles are treated as adverbs in our system. Figure 9 shows the query for phrasal verbs involving the particle *down*. Figure 10 shows the output of the 16 most frequent phrasal verbs involving *down* as extracted by our system. In total, our system extracted 411 instances of *V down*, about 80% of them phrasal verbs.

Existing online systems are like pre-cooked food in the supermarket. While they are ready-made, we still need to cook ourselves if we want something different. For instance, very few online English collocation extraction systems provide bilingual Chinese-English examples. As shown in Figure 11, our system is able to automatically identify phrasal verbs and corresponding bilingual Chinese-English examples from parallel corpora, thus greatly facilitating the compilation of a bilingual dictionary of collocations.
Figure 9. Query for phrasal verbs involving the particle *down*

<table>
<thead>
<tr>
<th>verb</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>come</td>
<td>154</td>
</tr>
<tr>
<td>look</td>
<td>97</td>
</tr>
<tr>
<td>settle</td>
<td>94</td>
</tr>
<tr>
<td>tear</td>
<td>79</td>
</tr>
<tr>
<td>bring</td>
<td>72</td>
</tr>
<tr>
<td>pass</td>
<td>62</td>
</tr>
<tr>
<td>put</td>
<td>47</td>
</tr>
<tr>
<td>shut</td>
<td>46</td>
</tr>
<tr>
<td>close</td>
<td>45</td>
</tr>
<tr>
<td>slow</td>
<td>44</td>
</tr>
<tr>
<td>kneel</td>
<td>35</td>
</tr>
<tr>
<td>take</td>
<td>30</td>
</tr>
<tr>
<td>set</td>
<td>28</td>
</tr>
<tr>
<td>let</td>
<td>27</td>
</tr>
<tr>
<td>get</td>
<td>25</td>
</tr>
<tr>
<td>throw</td>
<td>24</td>
</tr>
</tbody>
</table>

Figure 10. The most frequent phrasal verbs involving *down*
Figure 11. Examples of *set down* and their Chinese translations

6. Concluding remarks

In this paper, we have described the procedures involved in implementing an English collocation retrieval system by integrating dependency relations derived from the dependency parsers with statistical measures such as mutual information, t-scores, and log likelihood ratios. The performance of our system is affected by a number of factors such as corpus size, frequency of the keyword, and the accuracy of different dependency relations recognized by the parsers. Overall, the accuracy and coverage rate of our system is over 80%. Despite some errors, our collocation retrieval system works reasonably well and demonstrates the potential of corpus-based computational linguistics in language research. The statistical measures of MI, t-score, and log likelihood ratio can summarize collocation patterns in a straightforward manner, allowing researchers and learners to find the most salient collocations easily. Compared with the frequency-based approach, our approach is more straightforward and in line with the general practice of lexicography and phraseology. Our system can directly be used in compiling a dictionary of English collocations. Combining collocation extraction with a Chinese-English bilingual concordancer is a significant step towards automatic compilation of Chinese-English collocation dictionaries. Our next step is to align English collocations and their Chinese translations at the phrase level automatically (cf. Wu et al. 2003). In addition, we plan to improve the accuracy
of the parsers, expand the corpus size, and make the system fully automatic. We also plan to investigate the algorithms of automatically choosing appropriate examples for different levels and purposes (cf. Svensén 2009). Without doubt, the age has come when more interdisciplinary research will change the methodology and practice of lexicography. The impact of computational linguistics on lexicography and other branches of applied linguistics will be greater than ever, as evidenced by similar researches reported in Granger & Paquot (2012).

References


The corpus of contemporary American English. [http://corpus.byu.edu/coca/](http://corpus.byu.edu/coca/).

The Sketch engine. [http://www.sketchengine.co.uk/](http://www.sketchengine.co.uk/).


Department of Foreign Languages and Literatures
National Taiwan University
Taipei, TAIWAN
Zhao-Ming Gao: zmcao@ntu.edu.tw
自動擷取英文搭配語及中英文例句：
雙語辭典編纂學的計算工具

高照明
國立臺灣大學

本文描述英中雙語搭配語自動編纂線上系統 EXEC 的設計流程。EXEC 由一千三百萬英文詞及二千七百萬中文字的中英雙語平行語料庫建立而成，結合英語搭配語檢索和中英雙語檢索功能。EXEC 利用統計以及具有依存關係的英文句法剖析器擷取英文搭配語。使用者在查詢時輸入關鍵詞和關鍵詞的詞性以及所搜尋的搭配語的詞性，程式依據英文句法剖析器的依存關係和 mutual information、t-score、log likelihood ratio 等統計訊息自動擷取可能的英文搭配語，並連結包含英文搭配語的英文例句及中文翻譯。實驗顯示 EXEC 在擷取的正確率和辭典的涵蓋率都超過 80% 且可以很有效率地自動從平行語料擷取英文搭配語、例句，及中文翻譯。

關鍵詞：搭配語、依存關係、計算辭典編纂學、雙語平行語料庫、mutual information、t-score、log likelihood ratio