Collocation Extract: A Tool for Extracting Collocation

Wirote Aroonmanakun

Abstract

Collocation knowledge is necessary for language learners if they want to produce native-like languages. Collocation Extract is a free software tool for searching and ranking word chunks that could be collocations. Three basic statistical methods, namely log-likelihood, mutual information, and chi-square, are used for weighting collocation candidates. Since these statistics can be used to measure associations only between two elements, pseudo-bigram transformation is used to estimate statistical values of word chunks with three or more words. This software can also be used to extract technical terms when applied to a corpus of specific domain.

Introduction

Collocation is a linguistic phenomenon in which two or more words tend to be used together, e.g., "*smoker*" co-occurs with the adjective "*heavy*" rather than "*strong*"; "*information*" co-occurs with "*accurate*", "*correct*", "*precise*", but not "*true*". It is now widely accepted that collocation knowledge is necessary for foreign students to produce native-like languages. Without the knowledge of collocation, foreign students may produce grammatically correct but unacceptable sentences. That is why the topic of teaching collocation has recently attracted the attention of many language teachers. Collocation is an important topic in the Lexical Approach, which has been reappeared and discussed in many recent works, such as Nation's (1990) *Teaching and Learning Vocabulary*, Sinclair's (1991) *Corpus, Concordance, Collocation*, and Lewis' (1993) *The Lexical Approach*. The basis of this approach is to emphasize the important of vocabulary. In Lewis' (2000) view, "*language is fundamentally lexical.*" Learners, then, should be taught to learn words in combination rather than words in isolation.

Types of Collocations

Though most people agree on the importance of collocations, a definition of the term varies. Firth (1957, p.181) was the first to use the term, collocation: "*collocations of a given word are statements of the habitual or customary places of that word*". Kjellmer (1987) viewed a collocation as a "*sequence of words that occurs more than once in identical form and which is grammatically well-structured.*" In other words, a sequence of word has to occur repeatedly and be grammatically well formed (Oaks, 1998, p.160). Kjellmer's definition of collocation is similar to Choeka's (1988). Choeka defined a collocation as "*a sequence of two or more consecutive words, that has characteristics of a syntactic and semantic unit, and whose exact and unambiguous meaning cannot be derived directly from the meaning or connotation of its components.*" From this point of view, only multi-words that can form a linguistic unit can be a collocation. In addition, Cowie (1986) distinguished restricted collocations from free collocations by defining restricted collocations as combinations of words that are limited and not as open as free collocations. For example[1], the combination of "*fire staff*" is considered free because the word "*fire*" can be replaced easily by any word such as "*dismiss*", "*lay off*", "*sack*", while "*staff*" can also be replaced by words like "*worker*", "*employee*", "*clerk*". But the

[1] Examples are taken from Fontenelle (1994)

combination of "*make a decision*" is restricted because the word "*make*" can be replaced by only a limited range of words. Defining collocation along this line is to observe the degree of association between word combinations. At the other end of free combinations are multi-words that are idioms or fixed phrases. In between the two ends of the scale, some authors (e.g., Benson et al., 1993) distinguish grammatical collocations from lexical collocations. Grammatical collocations would have one word from an open class and another word from a closed class, e.g., "*depend*" collocates with "*on*" not "*of*", while lexical collocations would involves two words from an open class. Moreover, since computers have been used for extracting multi-words that occur more often than by chance, the distinction between collocation and co-occurrence/association then should be noted. Church and Hanks (1989) used a statistical method (mutual information) to determine word association, or finding words that co-occur with the specified word more often than expected. They may or may not be collocations in the same sense discussed earlier. However, even though statistical co-occurrences are not exactly the same as collocations, they could be used to locate potential collocations. And the ability of computers to automatically extract significant word combinations is attractive.

Collocation Extract

"Collocation Search" is usually included as a part of concordance software. Some concordancers, such as *Concordancer for Windows*[2], or *MonoConc*[3], display the frequency of collocates in terms of raw or absolute frequency. Using absolute frequency is the easiest way to look for collocates of a specified word. However, absolute frequency might not provide us an accurate view of collocations because the high number of co-occurrence could be merely the result of the high frequency of each word. To verify whether the co-occurrence is not due to chance, some statistical methods should be used. Different statistical methods are used in different concordance applications. For example, *WordSmith*[4] and *TACT*[5] use z-score to measure collocation strength; *Sara*[6] uses z-score and mutual information. In addition to z-score and mutual information, other statistical methods are also used in many research studies related to collocation extraction. For example, Dunning (1993) proposed using log-likelihood for collocation extraction. Daille (1995) tested many statistical methods, such as simple matching coefficient, Yule coefficient, cubic association ratio, log-likelihood, etc., for term extraction.

Fortunately, some software is designed specifically for collocation search, e.g., *Collocate*[7], *Xtract*, and *Collocation Extract*. *Xtract* (Smadja 1993) is a program for retrieving collocations from a large corpus. This program uses both statistical score and syntactic information to identify collocations. But it runs on a Unix system. *Collocate* and *Collocation Extract*, on the other hand, run on a Windows system. Although *Collocate* is more flexible and more powerful than *Collocation Extract*, it is commercial software. For those who want to learn more about collocations, and do not work with a large corpus, *Collocation Extract* would be the best choice to start

---

[2] The program is developed by Zdenek Martinek from the University of West Bohemia, Pilsen, Czech Republic, in close collaboration with Les Siegrist from the Technische Hochschule Darmstadt, Germany.
[3] See http://www.athel.com/mono.html
[4] See http://www.lexically.net/wordsmith/index.html
[5] See http://www.chass.utoronto.ca/tact/index.html
[6] Concordance software written for searching British National Corpus
[7] See http://www.athel.com/colloc.html

with. *Collocation Extract* is free software. It is not designed for automatic collocation extraction like *Xtract*. Rather, it is used as a tool for locating potential collocations in the corpus. The rest of this paper will explain how to use *Collocation Extract*. First, three statistical methods used in *Collocation Extract* will be described. So, the reader will get a general idea of how to interpret the results. Then, the steps in using the software will be described in detail.

Statistical Methods

Since different statistical methods capture different characteristics of the data, three well-known statistical methods, namely mutual information, chi-square, and log-likelihood, are available in *Collocation Extract*. They are explained in detail below:

Mutual information is used to capture the degree of independence between two variables, in this case, two words, by comparing the probability of observing two words together with the probability of observing them independently. The formula is as follow:

$$I(w1;w2) = \log_2 \frac{P(w1,w2)}{P(w1)P(w2)}$$

P(w1,w2) is calculated by counting the number of times that *w1* is followed by *w2*, and dividing by the total words (the size of the corpus). *P(w1)* and *P(w2)* are word probabilities obtained by counting the number of observations of *w1* and *w2* in the corpus and dividing by the total words. If words *w1* and *w2* are not associated to each other, then the joint probability *P(w1,w2)* will be equal to the occurrences by chance *P(w1)* * *P(w2)*, and consequently, *I(w1;w2)* = 0. But if the two words are highly associated, i.e., *w2* always occurs after *w1*, the joint probability *P(w1,w2)* would equal to *P(w1)*. Then, *I(w1;w2)* would be $\log_2 1/P(w2)$, which is greater than 0. On the other hand, if the two words are negatively associated, i.e., it is rare to find *w2* after *w1*; *P(w1,w2)* will be much smaller than *P(w1)* * *P(w2)*; and consequently *I(w1;w2)* would be less than 0.

Chi-square is a commonly used statistical method to measure the relation between two variables. In this case, it is used to measure the association between two words and the likelihood that their co-occurrences are not just by chance. It does not assume a normal distribution. Thus, it is better than the t-test, which assumes the normal distribution of data. Chi-square is calculated using the formula below, where Oi,j is the observed frequency for cell (i,j), and Ei,j is the expected frequency in each cell:

freq(w1 - w2)	freq(¬w1- w2)
freq(w1 - ¬w2)	freq(¬w1 - ¬w2)

$$\chi^2 = \sum_{i,j} \frac{(O_{i,j} - E_{i,j})^2}{E_{i,j}}$$

freq(*w1- w2*) is the observed occurrences of *w1* followed by *w2* in the corpus; freq(*w1 - ¬w2*) is the observed occurrences of *w1* followed by any words that are not *w2*; and so on. Expected frequency in each cell is equal to (row total * column total) / grand total. It is the expected number of co-occurrences if the co-occurrences are due to chance. The higher the number of the chi-square value, the more significant the collocation between *w1* and *w2*.

Log-likelihood is another statistical method for measuring the association between two elements. Though it is not as widely known as the chi-square, Dunning

(1993) argued that it is more appropriate to use on sparse data than the chi-square. The test compares two hypotheses about *w1* and *w2*

Hypothesis 1 : $P(w2|w1) = p = P(w2|\neg w1)$ (hypothesis of independence)

Hypothesis 2 : $P(w2|w1) = p1 \neq p2 = P(w2|\neg w1)$ (hypothesis of dependence)

Assuming binomial distribution, the log-likelihood ratio is calculated as follows:

$$\log\lambda = \log\frac{L(H1)}{L(H2)} = \log\frac{b(c12, c1, p)b(c1 - c12, N - c1, p)}{b(c12, c1, p1)b(c1 - c12, N - c1, p2)}$$

$$= \log L(c12, c1, p) + \log L(c2 - c12, N - c1, p) - \log L(c12, c1, p1) -$$

$$\log L(c2 - c12, N - c1, p2)$$

$$where\ L(k, n, x) = x^k (1 - x)^{n-k}$$

Note that *c1* is the frequency of *w1*, *c2* is the frequency of *w2*, *c12* is the frequency of bigram *w1-w2*, *N* is the number of total words in the corpus, $p = c2/N$, $p1 = c12/c1$, and $p2 = (c2 - c12) / (N-c1)$. The higher the value, the more likely that *w1* and *w2* are a collocation. Although the value -2 log λ can be used to compare with the table of chi-square distribution to test the null hypothesis H1 (Manning and Schütze, 1999), it is suggested to look at the values relatively rather than absolutely. Log-likelihood is less sensitive to rare events. In other words, unlike chi-square and mutual information, log-likelihood does not overemphasize the significance when the frequency of co-occurrence is small.

Tables 1 – 3 show different results of collocations extracted from the same corpus[8], ranked in order of significance.

Word1	Freq1	Word2	Freq2	Freq12	log-likelihood
can	1138	be	1680	548	4211.8486
of	8045	the	14172	1992	3630.7048
such	444	as	1411	267	2270.9783
passive	347	solar	1017	226	2135.0822
in	4287	the	14172	1051	1844.8514
natural	353	ventilation	522	176	1782.591
it	753	is	3070	289	1629.7517
u	321	s	748	164	1530.5497
should	299	be	1680	184	1497.8622
the	14172	building	2488	701	1406.9085
based	262	on	1451	159	1334.0101
energy	2487	efficient	307	184	1332.3953
life	182	cycle	167	100	1307.4116
et	81	al	75	72	1249.2552
e	315	g	174	103	1196.8755
on	1451	the	14172	485	1138.1398
m	397	sup2	214	106	1120.3156
fuel	251	cell	105	84	1101.8336
ft	131	sup2	214	85	1085.7197
solar	1017	radiation	223	121	1062.0737

Table 1 : List of top 20 collocations when using log-likelihood

[8] Though not all of the extractions are true collocations, for simplicity, all the extractions are referred to as "collocation" in this paper.

Word1	Freq1	Word2	Freq2	Freq12	mi
blair	6	mccarry	6	6	15.476941
rg	6	schrade	5	5	15.476941
gustavo	7	gili	7	7	15.254549
kevin	5	lomas	7	5	15.254549
peter	7	busby	5	5	15.254549
osram	9	sylvania	9	9	14.891979
editorial	9	gustavo	7	7	14.891979
sol	9	tot	5	5	14.891979
bob	5	fox	9	5	14.891979
nick	10	merrick	10	10	14.739976
phosphoric	9	acid	9	8	14.722054
rocky	8	mountain	9	7	14.699334
oued	11	tlilat	11	11	14.602472
godrej	11	gbc	7	7	14.602472
bruce	8	fowle	10	7	14.547331
arasteh	10	fn17	7	6	14.517583
mclaren	7	sport	10	6	14.517583
mcgraw	10	hill	12	10	14.476941
merrick	10	hedrich	12	10	14.476941
giuliano	6	todesco	12	6	14.476941

Table 2 : List of top 20 collocations when using mutual information

Word1	Freq1	Word2	Freq2	Freq12	chi-square
oued	11	tlilat	11	11	273637
nick	10	merrick	10	10	273637
osram	9	sylvania	9	9	273637
gustavo	7	gili	7	7	273637
blair	6	mccarry	6	6	273637
pink	43	corpus	42	42	267272.37
strengthskey	15	weaknesses	16	15	256533.75
rights	51	reserved	53	50	253082.57
hedrich	12	blessing	13	12	252587.08
relational	41	competence	39	37	234271.57
et	81	al	75	72	233492.69
oak	31	ridge	37	31	229258.4
rg	6	schrade	5	5	228030
merrick	10	hedrich	12	10	228029.17
mcgraw	10	hill	12	10	228029.17
sri	40	lanka	33	33	225744.75
registered	14	trademark	13	12	216501.36
phosphoric	9	acid	9	8	216205.23
editorial	9	gustavo	7	7	212827.22
karan	15	grover	18	14	198636.15

Table 3: List of top 20 collocations when using chi-square

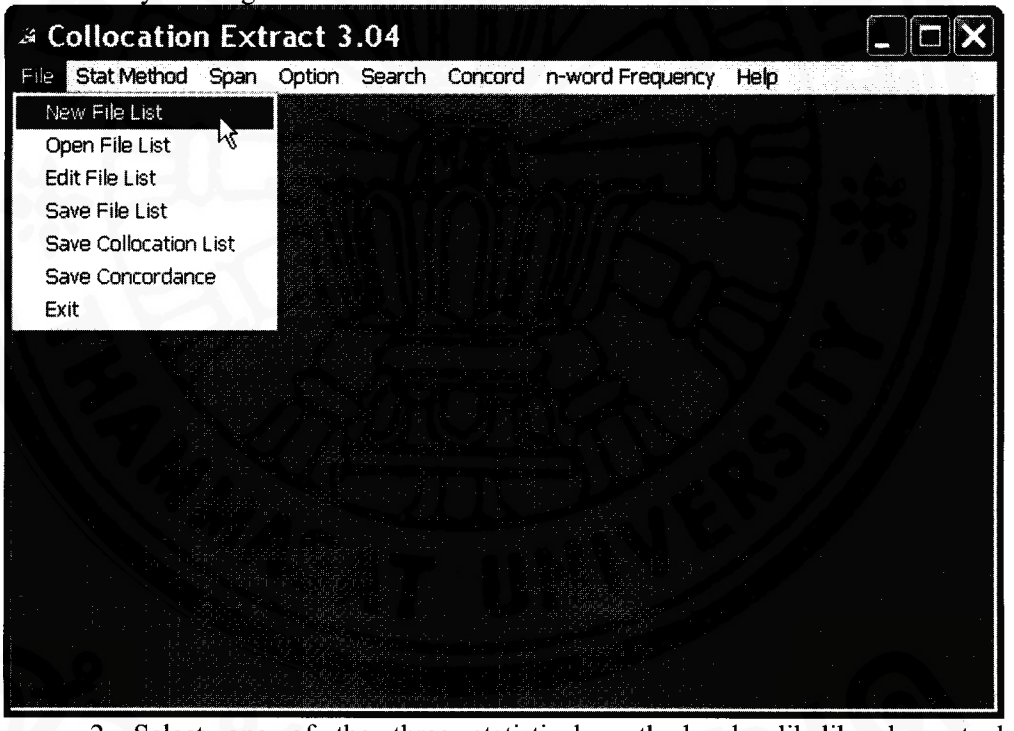
These three statistical methods can be used to measure collocation strength between two words only. They cannot be applied directly to measure collocations of three or more words. To check for collocation strengths of three or more words, we

adopted Silva and Lopes' pseudo bigram transformation (1999). The idea is to estimate n-word associations from 2-word association. For example, a sequence of four words, *w1-w2-w3-w4*, can be viewed as a combination of two parts in three ways: [*w1-w2-w3* and *w4*], [*w1-w2* and *w3-w4*], or [*w1* and *w2-w3-w4*]. Then, we apply these statistical methods to measure the collocation strengths between the two parts in each view, and average the sum of all collocation strength.[9]

Using *Collocation Extract*

Collocation Extract is designed to provide a list of potential collocations in the corpus. Users can search for collocates of a particular word in the range of 2-5 word chunks, or search for all collocations of two-word chunks. The steps in using the program are described below.

1. First, select all files in the corpus ("*File – New File List*"). These files can be either plain text or annotated files, such as html, sgml, or xml files. Select the file type that matches the data. Once the *File List* is defined, users can save the list for future use by clicking on "*File-Save File List*".



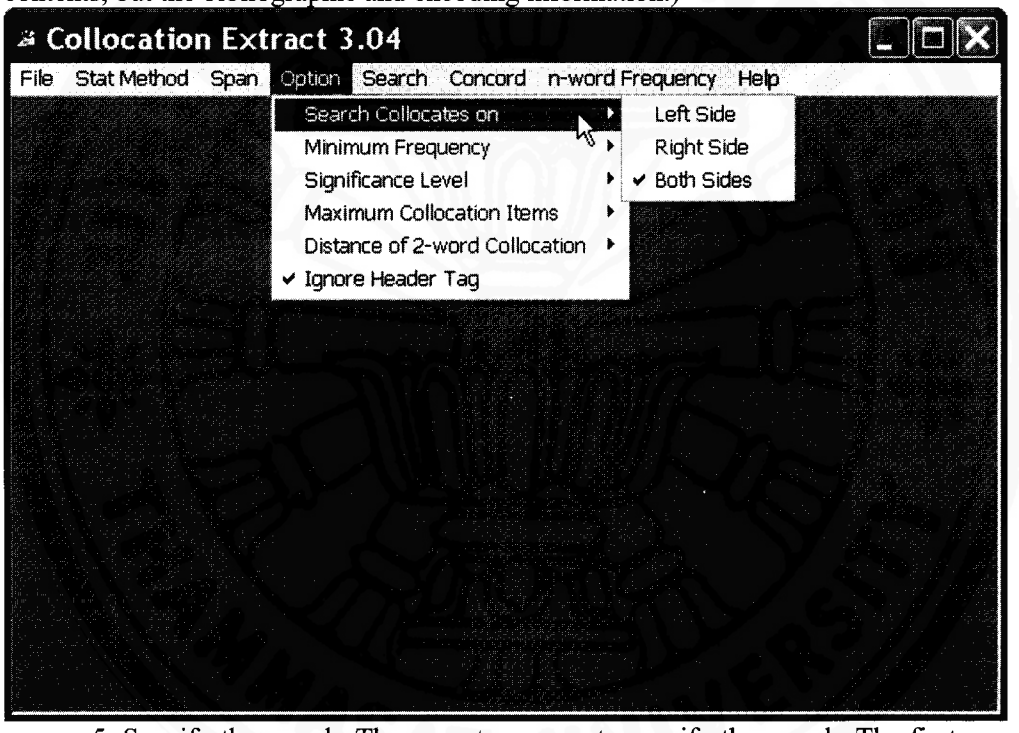
2. Select one of the three statistical methods: log-likelihood, mutual information, or chi-square, or select "*Raw Frequency*" to see only the frequency of occurrences. The default is log-likelihood.

3. Select the span ranged from 2 to 5. The number indicates the number of words to look for as collocations. For example, if "*2 words*" is selected, the program will search for collocations of two-word chunks. If "*3 words*" is selected, the program will search for collocations of three-word chunks.

4. Set all the options. First, set the direction for searching for collocations by selecting "*Option-Search Collocates on*". If "*Left Side*" is selected, the program looks for all collocates that occur before the keyword. The default is "*Both Sides*". Select the minimum frequency of n-word collocations. This will instruct the program to look

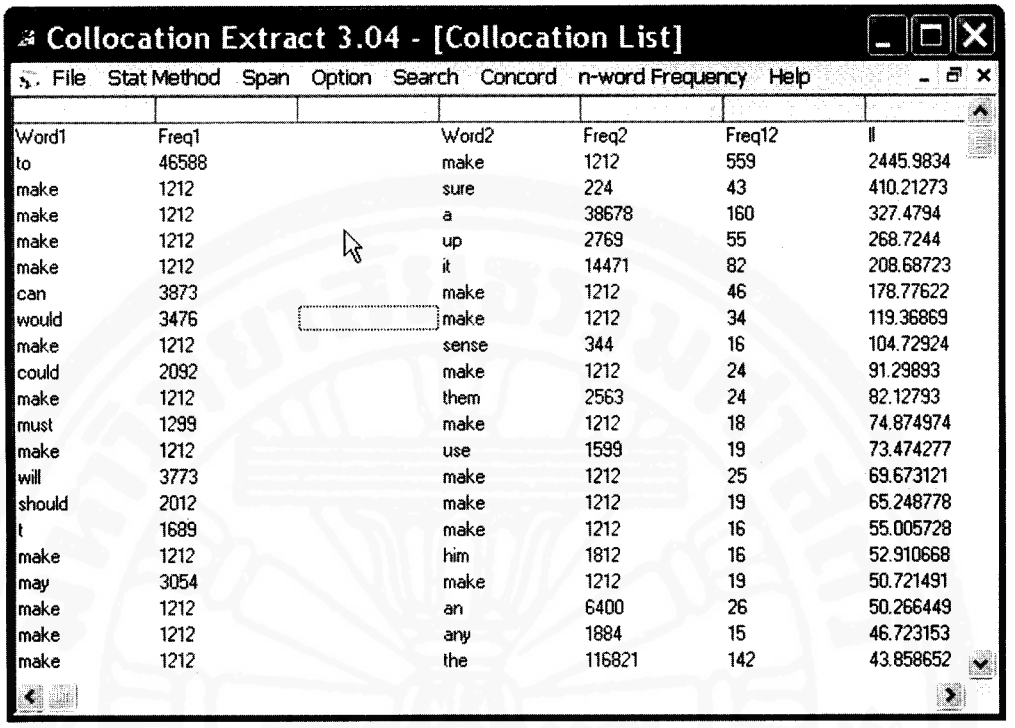
[9] For those who are interested to learn more about this technique, please see Silva and Lopes (1999).

for only collocations that occur at least *N* times, where *N* is the number specified. Then, select the statistical significance at the level of "*p > .005*", "*p > .05*", or "*all occurrences*". Set the maximum number of collocations to be extracted. The default is "*500*" items. When searching for 2-word collocations, users can specify the distance between the two words. If set at "*2*", the two words are separated by one word. This option is provided because collocations sometimes can be separated by other words, such as "*hold (oppositional) views*", "*hold (a similar) view*", etc. The last option, "*Ignore Header Tag*", is selected by default. This will instruct the program to ignore all information in the header tag <Header> ….. </Header>, which is encoded in sgml and xml files. (Information in the header tag is usually not the contents, but the bibliographic and encoding information.)

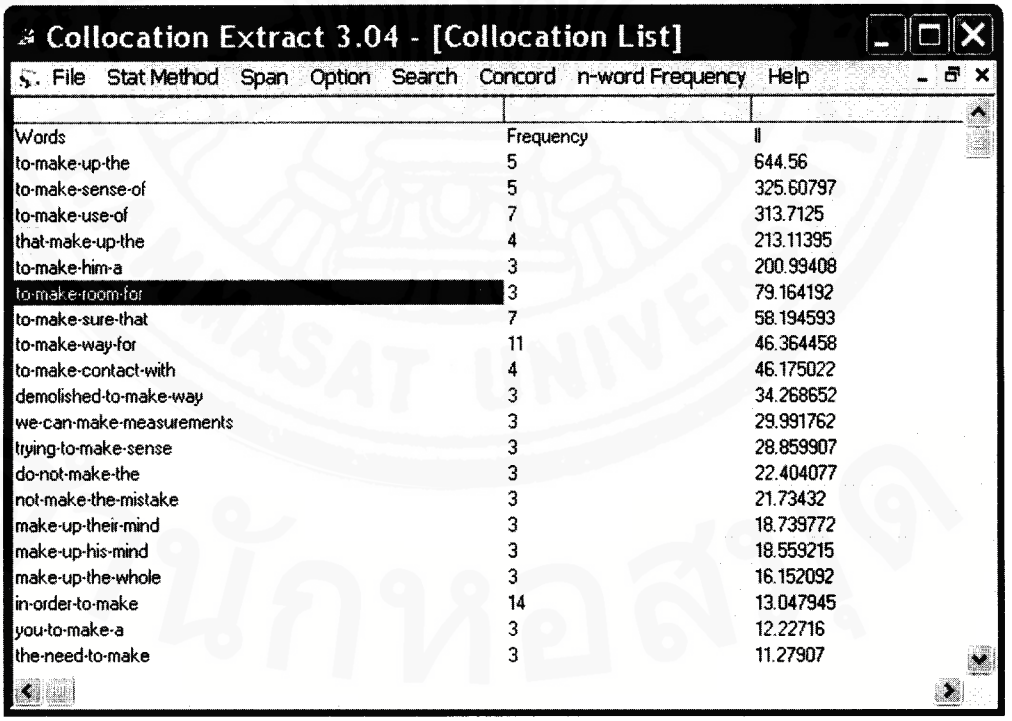


5. Specify the search. There are two ways to specify the search. The first one is to specify the keyword ("*Search – Keyword*") to be searched. The second way is to search for all 2-word collocations "*Search - All 2-word Collocations*". When searching for all 2-word collocations, users can specify the distance between the two words, as explained previously.

6. When the search is completed, the collocation windows will display the list of collocates that co-occur with the keyword, sorted by order of significance. If two-word collocation is searched and the distance of the two words is greater than one, a number of underscore symbols, "_", will be marked between the two words to indicate the distance. Users can save the collocation list by selecting "*File-Save Colloc List*". The output will be saved as a text file with tab delimited between each column. The output file then can be imported into the Excel program for further use.

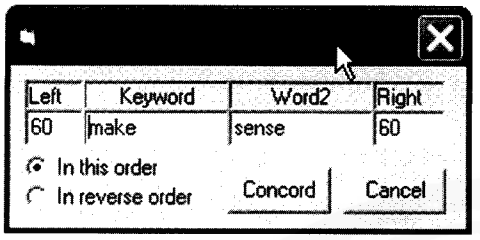


Word1	Freq1	Word2	Freq2	Freq12	ll
to	46588	make	1212	559	2445.9834
make	1212	sure	224	43	410.21273
make	1212	a	38678	160	327.4794
make	1212	up	2769	55	268.7244
make	1212	it	14471	82	208.68723
can	3873	make	1212	46	178.77622
would	3476	make	1212	34	119.36869
make	1212	sense	344	16	104.72924
could	2092	make	1212	24	91.29893
make	1212	them	2563	24	82.12793
must	1299	make	1212	18	74.874974
make	1212	use	1599	19	73.474277
will	3773	make	1212	25	69.673121
should	2012	make	1212	19	65.248778
t	1689	make	1212	16	55.005728
make	1212	him	1812	16	52.910668
may	3054	make	1212	19	50.721491
make	1212	an	6400	26	50.266449
make	1212	any	1884	15	46.723153
make	1212	the	116821	142	43.858652

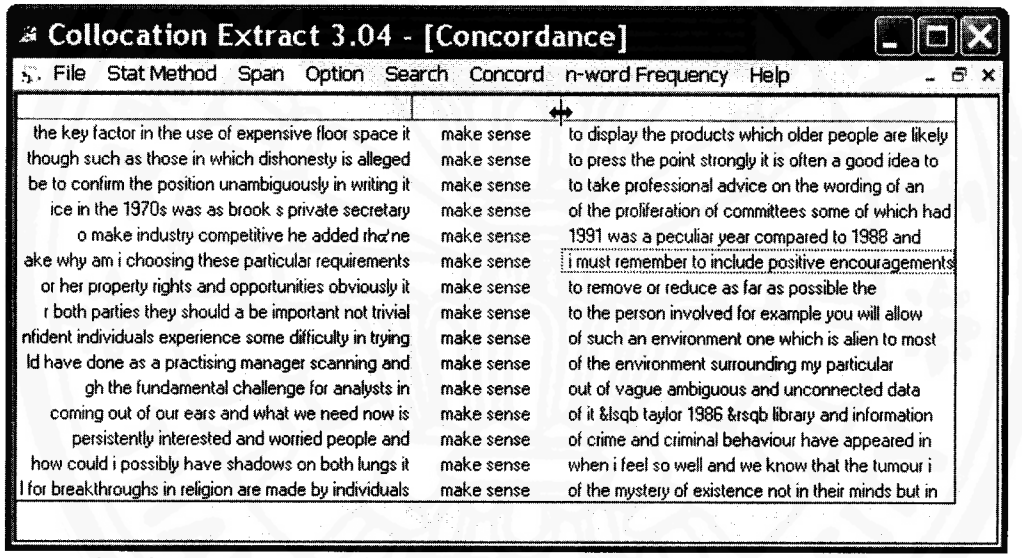


Words	Frequency	ll
to-make-up-the	5	644.56
to-make-sense-of	5	325.60797
to-make-use-of	7	313.7125
that-make-up-the	4	213.11395
to-make-him-a	3	200.99408
to-make-room-for	3	79.164192
to-make-sure-that	7	58.194593
to-make-way-for	11	46.364458
to-make-contact-with	4	46.175022
demolished-to-make-way	3	34.268652
we-can-make-measurements	3	29.991762
trying-to-make-sense	3	28.859907
do-not-make-the	3	22.404077
not-make-the-mistake	3	21.73432
make-up-their-mind	3	18.739772
make-up-his-mind	3	18.559215
make-up-the-whole	3	16.152092
in-order-to-make	14	13.047945
you-to-make-a	3	12.22716
the-need-to-make	3	11.27907

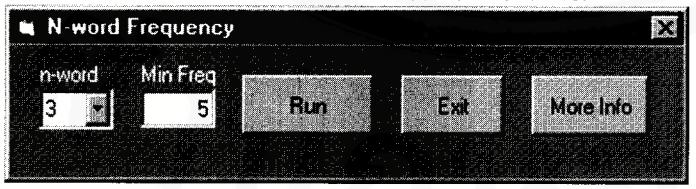
7. If users want to see the contexts of a particular word, they can click on that word, and select the menu bar "*Concord*". The specified word and its collocate will be shown. Use "*" to mark all words. Select the order of occurrences, and the number of characters for left and right contexts.



8. The concordance output will be shown in three columns. Users can save the result for further use by selecting "*File - Save Concordance*". Although this concordance feature is made available in *Collocation Extract*, it is suggested that concordance software should be used if users want to work extensively on the concordance results.[10] This feature is provided for users just to take a quick look at the contexts.



9. The program can list sequences of n-words from the corpus and the frequency of occurrences. Click on the "*n-word Frequency*" menu, and specify the number of word sequences and the minimum frequency. For example, if n-word is set as "*3*", and "*Min Freq*" as "*5*", the program will list all sequences of three-word chunks that occur at least 5 times in the corpus. The result of this search will not be shown on the screen. It will be saved as a text file.



Tips on how to use *Collocation Extract*

As stated earlier, this software is not intended to be an automatic collocation extraction tool, but it is collocation extraction aided software. Collocations extracted may or may not be true collocations. Users have to interpret the results themselves. Below are some tips on how to use this software and how to interpret the results.

[10] A number of Concordance softwares for Windows can be easily downloaded or purchased, such as Kwic (http://www.chs.nihon-u.ac.jp/eng_dpt/tukamoto/kwic_e.html), MonoConc (http://www.athel.com/mono.html).

1. The statistical values should be interpreted relatively rather than absolutely. Users should not be concerned much with the statistical values, and do not try to interpret whether a word chunk is a collocation purely on the basis of statistical values. Rather, users should see the results as a ranked list of potential collocations in accordance with the statistical significance.

2. Using different statistical methods yields different results. For example, as seen in *Tables 1-3*, when searching for all two-word chunks with mutual information and chi-square, word chunks whose components always occur together will be ranked on top of the list; within this group, mutual information will rank chunks with low frequency before those with high frequency. Because of this characteristic, mutual information and chi-square are likely to include word chunks that are proper names in the list. Log-likelihood, on the other hand, has no bias for word chunks with low frequency. As a result, it is possible that collocations in the list may be composed of function word like "*the*", "*of*", "*in.*" However, searching all two-word collocations can provide an overview of words that could be a part of collocations in the corpus. Users can do this search first. Then, pick a particular word to search for two or more collocations. When searching for a specific word, especially a high frequency word, the results from using each statistical method will not be dramatically different, like those of the "*All 2-word Collocations*" search.

3. A collocation should be a word chunk that is a linguistic unit. Collocations like "*in the*", "*of the*", can be easily omitted from the list. However, even when the extracted collocation appears to be a linguistic unit, users must not draw any quick conclusions, because that collocation could be part of a larger unit. For example, "*wastewater treatment*" might be retrieved from searching two-word collocations, but the actual collocation could be "*municipal wastewater treatment.*" It is advised that two-word collocations should be extracted first. Then, users should look for potential collocations that might be longer than two words, and specify the search for three-word collocations, and so on.

4. When working on a corpus of specific domain, collocations extracted are likely to be technical terms in that particular domain (Manning and Schütze, 1999, Daille, 1995). Actually, one of the properties of collocations discussed in Smadja (1993) is that "*collocations are domain-dependent*". In other words, in a specific field, technical jargon is often found and terms composed of common words would have different meanings in that particular field. For example, "*a wet suit*" does not refer to a suit that is wet, but clothing worn for swimming. Therefore, it is highly possible to use *Collocation Extract* to detect technical terms in a corpus of a specific domain.

5. To determine whether the extracted collocation is actually a collocation in general language or a technical term in a subject specific language, one criterion is to verify that its meaning is not exactly the composition of meanings from its parts. Thus, users have to examine the collocation and interpret its meaning from the context. *Collocation Extract* spots and reveals words, but the actual decisions are entirely based on user analysis.

6. Because collocations are determined from the corpus, the corpus is fundamentally important for collocation search. The corpus must have two basic properties. First, the corpus must be representative. It must be composed of texts that users have to study. Second, the amount of data must be sufficient for the task. When working on a subject specific language or a sub-language, the corpus size could be small, i.e. 100,000 words. But when working with general language, the corpus should be as large as possible.

Conclusions

This paper describes in detail how to use *Collocation Extract*. Three statistical methods used in the software and the steps in using the software are explained. The software applies a pseudo-bigram transformation to search for collocations with three or more words. When applied to a specialized corpus, *Collocation Extract* could be used for identifying technical terms. However, the software only provides a list of potential collocations. To determine whether those word chunks are in fact collocations, users must examine the contexts and make their own decision.

Acknowledgement

This program is further developed from *Collocation Test*, which can handle only two-word collocations. The development of *Collocation Test* was sponsored by the Development Grants for New Faculty/Researchers, Chulalongkorn University, in 1999, and a research grant from the Research Division of the Faculty of Arts in 2000.

References

Benson, M., Benson, E., and Ilson, R. (1993). The BBI Combinatory Dictionary of English. Amsterdam : John Benjamins

Choeuka, Yaacov. (1988). Looking for Needles in a Haystack or Locating Interesting Collocational Expressions in Large Textual Databases. In Proceedings of the RIAO.

Church, K.W. and Hanks, P. (1989). Word Association Norms, Mutual Information, and Lexicography. In Proceedings of 27th Annual Meeting of the Association for Computational Linguistics, Vancouver, 26-29 June, 76-83.

Cowie, Anthony P. (1986). Collocational Dictionaries - A comparative view. In Murphy (ed.) Fourth Joint Anglo-Soviet Seminar. (pp.61-69). London: British Council.

Cowie, Anthony P. (1992). Multiword Lexical Units and Communicative Language Teaching. In Arnaud & Bjoint (eds) Vocabulary and Applied Linguistics. (pp.1-12). London: Macmillan.

Daille, B. (1995). Combining Approach for Terminology Extraction: Lexical Statistics and Linguistic Filtering, UCREL Technical Papers, Vol 5., Department of Linguistic, University of Lancaster.

Dunning, Ted (1993). Accurate Methods for the Statistics of Surprise and Coincidence. Computational Linguistics, 19(1), Special Issue on Using Large Corpora: I, 61-74.

Fontenelle, T. (1994). Design and Structure of the Prototype Collocation Lexicon: Part I: What Are Collocations? - the DECIDE approach. November 1994, University of Liege.

Kjellmer, G. (1987). Aspects of English Collocations. In Computational Linguistics & Beyond (pp. 133-40). Rodopi: Amsterdam.

Lewis, M. (1993). The Lexical Approach: The State of ELT and a Way Forward. Hove, England: Language Teaching Publications.

Lewis, M. (1997). Implementing the Lexical Approach. Hove, England: Language Teaching Publications.

Lewis, M. (ed.) (2000). Teaching Collocation: Further Developments in Lexical Approach. Hove, England: Language Teaching Publications.

Manning, C., and Schütze, H. (1999). Foundations of Statistical Natural Language Processing. Cambridge: MIT Press.

Nation, I. S. .P. (1990). Teaching and Learning Vocabulary. Rowley, MA: Newbury House.

Oakes, Micheal P. (1998). Statistics for Corpus Linguistics. Edinburgh University Press.

Howarth, P. (1993) A Phraseological Approach to Academic Writing, In G Blue (ed) Language, Learning and Success: Studying through English. London: Macmillan.

Howarth, P. (1998). The Phraseology of Learners' Academic Writing. In A.P. Cowie, (ed.), Phraseology: Theory, Analysis and Application. (pp. 161-186). Oxford: Clarendon Press.

Silva, J., Lopes, G. (1999). A local Maxima Method and a Fair Dispersion Normalization for Extracting Multiword Units. In Proceedings of the 6th Meeting on the Mathematics of Language, Orlando, July 23-25.

Sinclair, J. (1991). Corpus, Concordance, Collocation. Oxford: Oxford University Press.

Smadja, F (1993). Retrieving Collocations from Text: Xtract. Computational Linguistics, 19(1), 143-177.