Journal of Culture and Information Science, 2(1), 1 — 15. (March 2007)

研究論文

Using Simple Computational Linguistic Techniques for Teaching Collocations

Tomonori Nagano and Kenji Kitao

This article examines possible applications of collocation extraction techniques to second/foreign language (especially, English) instruction. We will employ four simple collocation extraction measures – *t-statistic*, *chi-square*, *Mutual Information*, and *log likelihood* – and demonstrate how those collocation measures help language teachers identify important collocations in authentic L2 reading. We will also examine several typical collocation-related mistakes by Japanese-speaking English language learners. We suggest that some collocation errors can be explained by the influence of learners' native language. With this background in mind, we developed two pilot programs (*automatic collocation exercise generation* and *automatic collocation error detection*) using the aforementioned four collocation measures.

1. Introduction

In this paper, we will discuss the notion of *collocations* and its possible contribution to the language education (particularly focusing on the English-language instruction in Japan).

In the field of second language acquisition and English as a Second/Foreign Language (ESL/EFL) education, little attention has been paid to collocations compared with other domains of language, such as vocabulary, grammar, and phonetics/phonology. It is, however, widely acknowledged that learning collocations is a challenge for ESL/EFL learners. There is often no apparent reason why one collocation is better than another, but still substituting a synonym for a component word (in this paper, we call each word that makes up a collocation a *component word*) in a collocation may result in an ill-formed phrase. Teaching collocations to non-native speakers is also a challenge because few teaching resources focus on teaching collocations.

On the other hand, collocations have received a considerable amount of attention in computational linguistics, especially since the early 1990's. A fair amount of research has been conducted in various domains of computational linguistics, some of which has taken advantage of linguistically idiosyncratic nature of collocations, that is, their *non-compositionally* and *non-substitutability* (which will be explained in the following section).

The goal of this article is to address the gap between the two fields and to consider possible applications of computational approaches to the teaching of collocations for ESL/EFL learners. We are especially interested in the application of techniques developed in computational linguistics to second/foreign language pedagogy. In the second half of the article, we will demonstrate that very simple computational linguistic techniques can make a considerable contribution to second/foreign language education.

2. Definition of collocations

In spite of the familiarity of the term *collocation*, its definition is rarely discussed in the second/foreign language education and pedagogy literature. Language educators often have different definitions of collocations and some use the term collocation as a synonym of *idioms* or *phrasal verbs*. Thus, the treatment of collocations varies among language teachers, and there has not been any agreed-upon

definition of collocations in literature.

There is, however, a clear intuitive distinction of the types of collocations. Collocations are constrained by two broad kinds of constraints – part of speech and lexical constraints. The part-of-speech constraint is a grammatical restriction on collocations. For instance, the combination of adjective + noun (e.g., *powerful computer*) or verb + preposition (e.g., *pitch in*) is extremely frequent, but there are few collocations consisting of adverb + verb (e.g., *fiercely fight*) or (underivational) noun + preposition (e.g., *bag of*)[1]. Thus, it can be said that well-formedness of collocations is to some extent restricted by specific part-of-speech sequences.

The other kind of collocation, lexical constraint, is, on the other hand, independent of the grammatical constraint. The lexical collocation is lexically specific and each individual word plays a significant role in forming a collocation. The lexical constraint is used to explain the contrast between two collocations that are comparable in terms of part of speech, such as *powerful computer* and **strong computer* (both are in the adjective + noun sequence, but one phrase is far better than the other).

It is unfortunate that such characteristics of collocations have rarely gained attention in the language classroom. While most language teachers are aware of the importance of collocations and know some apparent properties of collocations, the importance of collocations is at best emphasized as part of vocabulary learning, and the collocation is rarely the topic of the language lesson.

In contrast with second/foreign language acquisition and language education, there is rich research on collocations in lexicography and computational linguistics. For example, *BBI Dictionary* by Benson, Benson, and Ilson (1997) categorizes collocations into *syntactic collocations* (e.g., prepositional phrases, the verb + complement phrase combination etc.; equivalent to *the part-of-speech constraint* discussed above) and *lexical collocations* (e.g., adjective + noun, verb + adverbial phrase, etc.; equivalent to *the lexical constraint*). Both categories are analyzed in depth and several sub-categories are proposed in both kinds of collocations. Benson, Benson, and Ilson also suggest that second/foreign language speakers typically have problems in lexical collocations. Therefore, we will chiefly discuss the lexical collocations in this study.

Benson (1989) proposes a functional definition of collocations and attempts to define collocations by their unique functional properties. We adopt Benson's functional definition of collocations that, in effect, makes the term *collocation* an umbrella term that includes a wide variety of co-occurrence phrases known as *idioms*, *fixed combinations*, *prepositional phrases,* etc. We prefer Benson's definition because it is independent of the traditional collocation terms and less prone to cause conceptual misunderstanding due to the biases of each individual language teacher.

Benson's definition consists of the following three functional properties of collocations.

- Non-compositionality
  A collocation typically generates extra semantic information that is not available from individual words that make up the collocation. For example, the exact meaning of the expression *to follow the instructions to the letter* (which means "follow the instructions exactly") is not predictable from the meanings of the component words.[2]

- Non-substitutability
  It is not possible to substitute synonyms for component words in a collocation. Phrases like **high building* (rather than *tall building*), **perpetrate suicide* (rather than *commit suicide*) and **make an estimation* (rather than *make a guess*) are awkward for this reason.

- Non-modifiability
  Collocations are not easily modified with

[1] We consider the collocation of derivational noun + preposition is a sub-type of the verb + preposition collocation. For instance *preparation for* is structurally identical to *prepare for* in spite of the different parts of speech.

[2] Light verbs (e.g., make, get, have, etc.) are also characterized by non-compositionality, but we will put aside the distinction between collocations and light verb phrases.

additional lexical modifiers such as adjectives and adverbs. For example, ??*I have annoying butterflies in my stomach* sounds odd because the modifier *annoying* has been inserted into *have butterflies in my stomach*.

In this paper, we will take the position that *non-compositionality* is the primary property of collocations. Since words that can collocate with each other are highly specific, extra semantic information (*non-compositionality*) can be generated only for a combination of limited kinds of lexical items. In this sense, the property of *non-substitutability* can be considered as a by-product of *non-compositionality*. Although *non-substitutability* subsumes *non-compositionality*, its property has an enormous potential for second/foreign language education, as we will discuss below. Finally, we will assume that only strongly fixed collocations (e.g., *idioms*) have the property of *non-modifiability*. It is evident since some collocations are readily modifiable with an adverbial phrase or adjective. For example, *it went without a major hitch* (a modified collocation derived from *without a hitch*) is an acceptable phrase (in contrast with *having annoying butterflies in my stomach*).

3. Second/foreign language learners and collocation mistakes

As mentioned above, collocations have attracted very little attention in second/foreign language education, despite their crucial role in determining one's fluency in the second language (L2) production. For instance, (1) – (4) are typical sentences by beginner/intermediate L2 English speakers that show apparent characteristics of non-nativeness. (A better collocation is listed in parentheses after the sentence.)

(1) *There are many high buildings in Tokyo. (*tall buildings*)
(2) *I took a business journey to London. (*business trip*)
(3) *A stiff wind rustles the tree. (*stiff breeze*)
(4) *The wrestler faced a powerful challenge. (*strong challenge*)

Most native speakers of English and advanced learners of English as a second/foreign language will find the above sentences awkward. Such awkwardness is, however, rarely dealt with in English language education. We believe that the following help explain the underemphasis on collocation-related mistakes.

- **Collocation-related mistakes are grammatical and meaningful**
  The central problem in the teaching of *collocations* is the fact that collocation-related mistakes are often grammatical (with respect to the traditional descriptive grammar), and they often do not obscure the intended meaning. In fact, sentences (1) – (4) are not ungrammatical, and their intended meanings are apparent. Although the better collocation is preferable from a communication perspective, it makes it hard for language teachers to justify why one phrase (e.g., *a business trip*) is better than the other phrase (e.g., *a business journey*).

- **No clear measure to compare collocations**
  Another problem surrounding collocation instruction is that the selection of collocations is arbitrary and, in many cases, the correct choice depends on the speaker's preference. For example, while (3) does not seem to be a correct collocation, it is hard to tell what the best alternative collocation is among phrases like *stiff breeze*, *strong breeze*, and *strong wind*. Even among native speakers, the judgments may not be consistent in such a case.

- **Dictionaries are not helpful**
  Crucially, traditional dictionaries do not help language learners learn collocations. The problem is that the number of possible collocations becomes so large that the traditional paper-based dictionary cannot include all useful collocations. Although dictionaries often list several sample sentences using the headword (which are very often good collocations), they by no means cover all collocations. It is easy to understand why traditional paper dictionaries are not suitable for collocations when we consider that the number of possible collocations grows exponentially as the size of a learner's vocabulary increases. For instance, it is possible to present a list of 8000 words to learn, but its possible collocations (that are derived from all possible combinations of 8000

words) are practically impossible to list.

- **Frequency is sometimes not reliable**
  The most common approach to detecting good collocations is to find frequent combinations of the target word. For instance, when we search for *respectable person* on Google, more than 120,000 hits are reported. On the other hand, *respectable individual* has only 8000 hits, which suggest *respectable person* is a far more frequent word sequence than *respectable individual*. The question is, however, whether we can conclude *respectable person* is a better collocation than *respectable individual*. Later in this article, we will argue that frequency is not as reliable as it is assumed. In fact, at least to us, *respectable individual* is as a good collocation as *respectable person*.

- **Too many collocations to focus on in class**
  In addition to the subtlety of collocation misuse, the volume of collocations makes it difficult to focus on them in the language classroom. There are such a large number of collocations in reading materials that the instructor cannot cover them in the limited class time. In addition, very few study materials for collocations are available because it is extremely time-consuming (even for professional material developers and publishers) to detect collocations and collocation errors in the language education materials.

In summary, the acquisition of collocations is a very important aspect of second/foreign language education. The collocation is, however, undervalued in classroom instruction due to its own nature as described above. In the following section, we will argue the possibility of the first language influence in the misuse of collocations.

4. Insights from Second Language Acquisition Research

Why is it so hard for non-native speakers to use collocations correctly? It is, of course, in part a matter of fluency – if ESL/EFL learners do not have sufficient vocabulary, they will have trouble using collocations – but the picture is not as simple as it may look.

First of all, collocations remain difficult for advanced ESL/EFL speakers. The misuse of collocations is still obvious in production by advanced ESL/EFL speakers and, in fact, bad collocations (along with accent) often appear as a subtle indication of the non-nativeness of near-native ESL/EFL speakers.

One of the obvious influences on the non-nativeness of L2 utterances is the influence of their first language. The influence of L1 is termed *language transfer* and has been a major topic in second language acquisition (SLA) research. Generally speaking, *language transfer* research is concerned about what role the native language (L1) plays in the SLA process. While it is acknowledged that the L1 is not the sole factor in L2 learners' errors, and some universal cognitive mechanism governs second/foreign language learning, it is generally accepted that the L1 plays a crucial role when considering whether an ESL/EFL learner will succeed in language learning. Many researchers argue that the lexical influence of L1 is far greater than the transfer of L1 grammar; thus, according to this view, L2 speakers have more difficulty with collocations due to the L1 influence (see Epstein, Flynn, and Martohardjono [1996] for a comprehensive review of language transfer issues). For instance, it is anecdotally supported (and probably true) that speakers of a Germanic language have an advantage over Japanese speakers in learning English as a second/foreign language.

One hypothesis of *language transfer* claims that only lexical items (vocabulary) transfer to L2, but not the functional items (Vainikka & Young-Scholten, 1996). (A simple example is the case of Japanese speakers, who have no problem with the English word order SVO, in spite of the fact that Japanese has the SOV word order. See Flynn [1987] for the parameter re-setting hypothesis of the ESL of Japanese-speakers.) Odlin (1989) proposes that, in the process of lexical item transfer, L2 learners overextend the senses of L2 words due to the influence of L1. For instance, a Japanese learner of English may produce sentences such as:

- I've seen the tallest building in New York.
- ??I've seen the highest building in New York.

In Japanese, both *tall* and *high* are translated into the same word たかい (takai), and there is no sense distinction between *tall* and *high* as there is in English. Therefore, Japanese ESL/EFL learners often overextend the senses of *tall* and *high* and may produce an unconventional word sequence as above. It is important to note that the overextension of word senses can take place even if there is a one-to-one word correspondence. For example, an English word *name* has a direct translation in Japanese, なまえ (namae). なまえ in Japanese, however, lacks the sense of *a well-known or notable person*, which exists in English as in *a big name*. Thus, it is expected that Japanese ESL/EFL learners have difficulty in using phrases like *his name is widely acknowledged*.

The differences between two languages might look insignificant at the individual word level, but if we consider that our lexicon consists of a semantic network (as assumed in *WordNet* [Miller, Beckwith, Fellbaum, Gross, and Miller, 1993]), the lack or abundance of senses will result in a huge distortion of the whole semantic network for second/foreign language learners.

In the following section, we will present a brief survey of research on collocations in computational linguistics that sheds new light on the problems in the teaching of collocations in the language classroom.

5. Collocations in computational linguistics

The recent upsurge of collocation studies in computational linguistics has grown out of the proposal made by Church and Hanks (1989a; 1989b), who argued that semantic and syntactic word relationships are automatically computable from machine-readable corpora. Using an information-theoretic measure *Mutual Information* (MI), Church and Hanks demonstrated that the association between words could be numerically computable with an electronic corpus of a reasonable size.

Following this tradition, several alternative measures have been proposed. Church and Hanks (1989a; 1989b) suggest the application of hypothesis testing (i.e., *t-test*) to the extraction of collocations. Church and Mercer (1993) propose using non-parametric statistics, such as *chi-square* instead of parametric measures. Dunning (1993) argues that the statistical methods unjustifiably violate the fundamental assumptions of statistics theories (e.g., independence in parametric statistics and skewed data in non-parametric statistics) and, instead, proposes *log-likelihood ratio* as an alternative measure. In our study, we employed basic four association measures: *t-test*, *chi-square*, *Mutual Information*, and *log likelihood*. Further discussion of the statistical approach to collocation discovery can be found in Appendix.

Next, we will briefly explain how these statistics apply to analyzing collocations.

When applied to collocation discovery, the *t-test* is assumed to measure how (un)likely word co-occurrence is above chance. A high *t-statistic* is considered an indication of fixed placement of words and thus more likely to be a good collocation, whereas a low *t-statistic* suggests the words are scattered throughout the corpus (therefore, not forming collocations).

Chi-square is another statistical measure, but unlike *t-test*, the *chi-square* does not assume the normal distribution of the population. Since the distribution of words is highly constrained by grammar, the assumption of the normal distribution is undoubtedly violated.

Mutual Information (MI) is an information theoretic measurement. The MI we adopted in our study is very simple one; that is, log of the ratio of a joint probability (actual frequency) to an independent probability (expected frequency). However, it is pointed out that MI is not very reliable when the actual frequency of the collocation is fewer than 10 (Manning and Schutze, 1999).

Finally, *log likelihood* is a measure to evaluate the degree of dependence between words in a collocation phrase. In computing *log likelihood*, two hypotheses of extreme cases are assumed. H_1 assumes independence of word co-occurrence (thus, no chance of a collocation) and H_2 assumes full dependence of words, which means that when one word appears the

other word must appear in the context. *Log likelihood* is simply a degree of dependency between two words measured by the ratio between hypothesis 1 (independence) and hypothesis 2 (dependence).

6. Collocation extraction for language instruction

In this section, we will demonstrate that collocation candidates are easily extracted from text by using raw frequencies and a large corpus. We will argue, however, that mere raw frequencies are not a reliable measure for determining the strength of collocations. We will argue that the association measures discussed in the previous section are more reliable than raw frequency. To demonstrate how efficiently those collocation measures can extract *collocations*, we developed two pilot programs. For the pilot experiment, we used two corpora (*The American National Corpus first release (ANC)* (Ide, Reppen, and Suderman, 2002) and *The Wall Street Journal Corpus (WSJ)* collection from the ACL/DCI corpus. After deleting non-words (i.e., punctuation and non-ASCII symbols), the total number of tokens was 54 million words (10 million words from ANC and 44 million words from WSJ). The collocations are limited to 2-word sequences (bigrams) in this study.

6.1 Raw Frequencies and collocations

One of the most intuitive facts about collocations is that good collocations tend to appear more frequently than bad collocations or non-collocation phrases. This intuition is true to some extent – in a corpus, good collocations tend to have higher frequencies whereas non-collocation phrases do not appear at all or have very low frequencies.

There are several online tools that take advantage of this strong correlation between collocations and frequencies. For example, *VIEW: Variation in English Words and Phrases* by Mark Davies (Davies, 2006) lists frequencies of bigrams (two-word phrases) from the 100-million-word British National Corpus. *VIEW* has a powerful search function that enables the user to list phrases in a certain syntactic context (e.g., only adjective + noun phrases) and a keyword-in-context (KWIC) function that can show exactly in what contexts the collocation is used.

VIEW is a useful tool to discover good collocations. For example, if the user wants to know what adjectives can form good collocations with the word *coffee*, he/she can get a list all frequent phrases that match the "adjective + *coffee*" context. A sample output for this search condition on VIEW is listed below.

phrase	frequency
black coffee	97
instant coffee	58
hot coffee	46
international coffee	22
fresh coffee	21
strong coffee	19
cold coffee	14
empty coffee	14
ground coffee	13
real coffee	13
good coffee	11
decaffeinated coffee	10
milky coffee	10

Table 1: Output for "[ad*] *coffee*" on VIEW (phrases with a frequency of less than 10 are omitted)

The results include many phrases that we intuitively judge as good collocations. For example, *black coffee* fulfills our definition of collocations – first, its semantic interpretation (*coffee without sugar and milk*) is different from its literal meaning (*black-color coffee*), meeting the non-compositional definition. It also meets the non-substitutability condition because *black* cannot be replaced with its synonym; for instance, **inky coffee* and **dusky coffee* do not mean *black coffee*. We believe some other collocations in the list (e.g., *strong coffee*) also meet our definition of collocations.

Therefore, we think the tools like *VIEW* are quite useful for detecting collocations. However, we also think they are not the optimal approach to collocation detection. While we find quite a few collocations in the results of *VIEW*, we also find a lot of non-collocation phrases (e.g, *hot coffee*, *cold coffee*, *empty coffee*, *real coffee*, *good coffee*, and *milky coffee*).[3] (The

[3] As mentioned above, *collocations* are defined as non-compositional phrases in our paper. By using such a semantic judgment, we intend to prevent the influence of individual preferences of collocations.

bigram *empty coffee* only occurs in phrases like *empty coffee cups*.) In fact, a frequency-based collocation list often includes a lot of non-collocation phrases in its output. It is because the frequency is not an absolute measure but a relative measure. Thus, the high frequency of *black coffee* (97) and the relatively low frequency of *strong coffee* (19) do not directly indicate the strength of collocations, but rather they are mostly accounted for the different frequencies between *black* and *strong*.

Given that result, statistical association measures are considered a far better measure to determine the strength of collocations (Manning & Schutze, 1999; and many others). To test this claim, we computed the association measures of each of the phrase in Table 1.

	freq	t-test	chi	MI	LL
decaffeinated coffee	11	4.43E-04	2.34E+05	1.44E+01	365
milky coffee	1	1.34E-04	2.74E+03	1.14E+01	557
instant coffee	18	5.66E-04	1.72E+04	9.90E+00	360
hot coffee	9	3.97E-04	8.48E+02	6.59E+00	507
ground coffee	8	3.72E-04	4.87E+02	5.97E+00	521
fresh coffee	3	2.27E-04	1.48E+02	5.68E+00	554
black coffee	12	4.53E-04	5.17E+02	5.49E+00	503
empty coffee	1	1.30E-04	3.43E+01	5.18E+00	566
international coffee	13	4.68E-04	4.36E+02	5.15E+00	504
cold coffee	1	1.26E-04	1.52E+01	4.10E+00	568
strong coffee	1	1.25E-04	1.35E+01	3.95E+00	568
good coffee	1	1.25E-04	1.35E+01	3.95E+00	568
others	0	-	-	-	-

Table 2: Association measures for collocation candidate phrases from Table 4, using our experimental (54-million-word) corpora (in descending order of MI). (LL values are after the subtraction of 2.0E+09)

The results in Table 2 clearly show that the association measures produce a different order of collocation phrases that was not captured by the frequency-based model like *VIEW*. For instance, in Table 2, phrases like *decaffeinated coffee* and *instant coffee* are ranked higher than other high-frequency phrases.

Thus, we conclude that the mere frequency-based collocation extraction is not the only approach. We can clearly have an alternative approach by using the association measures. In the following sections, we will present further analyses of the collocation association measures.

6.2 Collocation candidates

The first set of collocation candidates are given in Table 1. Those whose native language is not English (or even native speakers of English) are encouraged to try to rank those collocations before reading the results.

stiff breeze	strong breeze	strong wind
broad daylight	strong coffee	strong challenge
stiff wind	bright daylight	narrow daylight
powerful coffee	powerful challenge	strong computer
strong drugs		

Table 3: Collocation Candidate Set 1

The results are sorted in ascending order of *t-statistic* and *MI*. There are several interesting facts in the results.

	freq	t-test	chi	MI	LL
broad daylight	20	5.97E-04	7.91E+04	1.20E+01	275
strong challenge	11	4.07E-04	1.15E+02	3.63E+00	536
powerful challenge	4	2.53E-04	6.59E+01	4.20E+00	556
stiff wind	3	2.30E-04	5.00E+02	7.40E+00	547
strong coffee	3	1.96E-04	1.38E+01	2.69E+00	565
strong breeze	2	1.85E-04	1.04E+02	5.76E+00	559
strong wind	2	1.52E-04	6.69E+00	2.36E+00	568
stiff breeze	1	1.33E-04	5.89E+02	9.21E+00	561
strong drugs	0	-	-	-	-
narrow daylight	0	-	-	-	-
bright daylight	0	-	-	-	-
powerful coffee	0	-	-	-	-
strong computer	1	-3.82E-04	2.12E+00	-1.95E+00	568

Table 4: Results of Collocation Candidate Set 1 (descending order of *t-test* scores) (LL values are after the subtraction of 2.0E+09)

	freq	t-test	chi	MI	LL
broad daylight	20	5.97E-04	7.91E+04	1.20E+01	275
stiff breeze	1	1.33E-04	5.89E+02	9.21E+00	561
stiff wind	3	2.30E-04	5.00E+02	7.40E+00	547
strong breeze	2	1.85E-04	1.04E+02	5.76E+00	559
powerful challenge	4	2.53E-04	6.59E+01	4.20E+00	556
strong challenge	11	4.07E-04	1.15E+02	3.63E+00	536
strong coffee	3	1.96E-04	1.38E+01	2.69E+00	565
strong wind	2	1.52E-04	6.69E+00	2.36E+00	568
strong drugs	0	-	-	-	-
narrow daylight	0	-	-	-	-
bright daylight	0	-	-	-	-
powerful coffee	0	-	-	-	-
strong computer	1	-3.82E-04	2.12E+00	-1.95E+00	568

Table 5: Results of Collocation Candidate Set 1
(descending order of MI scores)
(LL values are after the subtraction of 2.0E+09)

First, it appears that different collocation measures rank collocations in different manners – within our data and examples, *t-test* and *log likelihood* seem to be sensitive to raw frequencies of collocations (although this does not mean that the ranking of collocations in those measures is solely determined by the raw frequency, since with other sample sets, those measures ranked less frequent words higher than more frequent words) whereas *chi-square* and *MI* are not as dependent on frequencies as *t-test* and *log likelihood*. Second, with the exception of *strong computer*, very few bad collocations appear in our corpus (i.e., a frequency of 0). Since our corpus is moderately large (54 million words), it might be the case that mere raw frequencies can eliminate the bad collocations. In other words, if the frequency of a collocation candidate is 0, it can be concluded that the collocation is likely to be a bad or misused one. However, it should be pointed out that all four measures successfully distinguished *strong computer* from other collocations. This suggests that bad collocations do appear sometimes (*strong computer* probably appeared in a context such as *strong computer skills* in which *computer* is inserted into the collocation *strong skills*). Thus, frequency-based collocation detection may work in most cases, but it will fail to exclude bad collocations that appear in the corpus by chance.

The additional data (as given in Table 6) support our analyses.

constructive criticism	evasive answer
expensive tests	impressive results
impulsive behavior	inventive plot
oppressive heat	permissive behavior
confidential information	critical review
economical buy	fanatical supporters
hysterical reaction	magical movement
mysterial experience	personal relationships
terminal illness	contestable statement
groundbreaking ceremony	preemptive right

Table 6: Collocation Candidate Set 2

	freq	t-test	chi	MI	LL
confidential information	301	2.32E-03	3.78E+05	1.03E+01	-3231
personal relationships	36	7.98E-04	7.56E+03	7.73E+00	256
impressive results	7	3.42E-04	1.97E+02	4.91E+00	537
expensive tests	6	3.12E-04	1.19E+02	4.45E+00	546
terminal illness	5	2.98E-04	3.35E+03	9.39E+00	516
constructive criticism	3	2.30E-04	5.14E+02	7.44E+00	546
compulsive gambler	3	2.31E-04	6.74E+04	1.44E+01	517
critical review	3	2.05E-04	2.06E+01	3.13E+00	564
groundbreaking ceremony	1	1.34E-04	3.93E+03	1.19E+01	557
oppressive heat	1	1.33E-04	3.54E+02	8.48E+00	562
all others	0	-	-	-	-

Table 7: Results of Collocation Candidate Set 2
(descending order of t-test scores)
(LL values are after the subtraction of 2.0E+09)

	freq	t-test	chi	MI	LL
compulsive gambler	3	2.31E-04	6.74E+04	1.44E+01	517
groundbreaking ceremony	1	1.34E-04	3.93E+03	1.19E+01	557
confidential information	301	2.32E-03	3.78E+05	1.03E+01	-3231
terminal illness	5	2.98E-04	3.35E+03	9.39E+00	516
oppressive heat	1	1.33E-04	3.54E+02	8.48E+00	562
personal relationships	36	7.98E-04	7.56E+03	7.73E+00	256
constructive criticism	3	2.30E-04	5.14E+02	7.44E+00	546
impressive results	7	3.42E-04	1.97E+02	4.91E+00	537
expensive tests	6	3.12E-04	1.19E+02	4.45E+00	546
critical review	3	2.05E-04	2.06E+01	3.13E+00	564
all others	0	-	-	-	-

Table 8: Results of Collocation Candidate Set 2
(descending order of MI scores)
(LL values are after the subtraction of 2.0E+09)

The results of the data set 2 also suggest that the mere raw frequency may not be a good indicator for

collocations. The results show that some of low-frequency collocations (e.g., *compulsive gambler* and *oppressive heat*) are ranked high, indicating that these results evaluate the well-formedness of collocations, independent of the frequency of occurrence. It is important to point out that those collocations are as good as some high-frequency collocations such as *personal relationships*.

To summarize, the four collocation measures appear to be effective in detecting correct collocations. They are generally better indicators than the raw frequency, which is otherwise often considered as a sole determinant of the collocation.

Given the findings above, in the following section we will propose some possible applications of the collocation extraction methods.

7. Applications of collocation detection measures

7.1 Automatic generation of collocation exercises

One of the obvious applications of collocation extraction to second/foreign language education is to automatically generate exercises on collocations.

As suggested above, collocations will be ill-formed when a component word is replaced with its synonym (non-substitutability). For instance, **business journey* is not a good collocation; *business trip* is preferred. Such ill-formed collocations are extremely difficult for non-native speakers of English to detect. In spite of the obvious need for exercises on collocations, very few instructional resources are available on the market. As discussed above, it was because making exercises on collocations is difficult due to the lack of clear-cut measures for collocations.

We argue that the computational linguistic technique for collocations may help solve this problem. As described above, collocation measures such as *t*-test can assign numeric values for collocations and rank them in a certain order. While the ranking varies among collocation measures, it seems clear that most collocation measures can successfully detect ill-formed collocations from better ones. In addition, high-spec computers, which are ubiquitously available now, can compute collocation measures very rapidly. Thus a computational approach to collocation exercises is not only possible but also an optimal approach to developing materials on collocations.

Keeping this in mind, we developed a pilot program that automatically generates multiple-choice exercises on collocations. The program generates information to develop traditional 4- to 6-items multiple-choice questions in which all of the items (answer and distracters) share the same lexical item that collocates with other words. The distracters use synonyms of the correct collocation, since we assume that L2 speakers will have trouble those synonym collocations due to L1 transfer.

We employed *WordNet* (Miller, Beckwith, Fellbaum, Gross, and Miller, 1993) to list synonyms of target words. *WordNet* is an electronic dictionary in which word meanings are hierarchically structured. Unlike traditional dictionaries, the headwords are *word sense* (meaning) rather than lexical forms. (Thus, for example, the lexically identical word *bank* has several entries, including a financial institution and sloping land, especially along side a body of water.) Our pilot program extracts a word's synonym (called *synset* in *WordNet*) and its immediate *hyponym* set (sub-ordinate words) and *hypernym* set (higher-order words). Those synonyms are replaced with words in collocations to make ill-formed collocations (which are used as distracters in collocation exercises).

The outline of this program is as follows:

- The program extracts the synonym set for each component word in a collocation. (Since the window of words is limited to 2 in this study, only two-word collocations are considered.) In the case of *business trip*, synonym sets for both *business* and *trip* are collected.
- A component word is replaced with its synonym, forming a new collocation. (*Journey* is in the synonym set of *trip*; therefore, *journey* replaces *trip* and forms a new phrase *business journey*. Note that this process repeats as many times as the number of synonyms.)
- The new collocation is evaluated with the collocation measures (that is, the association measures for *business journey* are computed.

Depending on the value of the association measure, the new collocation is classified either as a "good collocation (correct answer)" or a "bad collocation (distracter)"

- The list of good and bad collocations is produced.

A few sets of sample results are listed as below. The ill-formed collocations are marked with an asterisk and the questionable collocation is marked with ??.

a. collocation exercise generated for *business trip*

business travel	*business journey	??business tour
*business voyage	*business sail	*job journey

b. collocation exercise generated for *indisputable fact*

indisputable fact	*indisputable case
*indisputable point	*indisputable specific
??indisputable truth	*indisputable reason
*indisputable record	*undisputable fact
*sure fact	

Table 9: Results of Collocation Exercise Generation (distracters are marked with asterisks)

We believe the output is extremely useful in developing materials. If the program can automatically generate collocation exercises, language teachers can use collocation exercises that are extracted from the reading materials for his/her class. Such exercises would be impossible (due to the time and resource constraints) without the help of the computer program.

We have a few caveats, however. Teachers need to edit the results before using them in the classroom. First, not all the exercises do exhibit the same level of difficulty. Some exercises contain extremely unlikely (or nonsense) items that need to be removed by manual check. In some cases, questions have only very unlikely distracters (e.g., *job journey* and *business sail*), resulting in an extremely easy question. On the other hand, some questions are very difficult because they have several good distracters (e.g., *business journey*). Second, when a distracter's frequency in the corpus is zero, it may produce a false negative. Although our program can tell whether a collocation with non-zero frequency is bad (based on the value of collocation measures), it may rate a collocation with zero frequency as being bad when, in fact, it is not. For example *indisputable case*, *indisputable truth, indisputable reason*, and *indisputable point* are all good collocations. In other words, there is a chance that a good collocation that just didn't appear our sample corpus could be judged as a bad collocation. Thus, the classroom instructor needs to check each distracter before using it in his/her classroom.

We hope that these problems will be solved in the future as we improve our program.

7.2 Automatic collocation error detection

Another possible application of the collocation detection technique is automatic collocation error detection. Using several collocation association measures and large-size corpora, it may be possible to detect bad collocations from the writings of learners of English as a second/foreign language.

In the simplest case, all bigrams (two-word sequences) that do not appear in the corpus data can be considered as misused collocations. As mentioned in the previous section, it is not always the case that zero-frequency collocations are bad collocations. Some good collocations may not appear in a particular corpus merely due to the size of the corpus. In order to prevent such cases, we employed an extra assessment step to identify bad collocations and try to find replacements for them.

As stated above, collocation errors by L2 learners are often due to the L1 transfer and most misused collocations are semantically equivalent to the correct collocation (e.g., *business trip* vs. *business journey*). Thus, we postulated that it is a very strong sign of a collocation mistake if there is a good collocation that is semantically equivalent for the misused collocation. In other words, if our program detects a good collocation candidate (e.g., *business trip*) for a bad collocation (e.g., *business journey*) in the L2 writing, the bad collocation is most likely a collocation mistake due to the L1 transfer.

Based on this logic, we developed another program that extracts all bad bigrams (that have either a zero-frequency or low collocation measure) from the input (i.e., L2 writing) and search for a better collocation candidate. The program replaces each word in a collocation with its synonyms and re-computes the

collocation measures. The program identifies a bad collocation if the synonym collocation has a high collocation value (thus, it's most likely a misused collocation).

The outline of this program is as follows:

- Detecting misused collocations: The association measures of the input collocation are computed.
- If the association measure indicates that the collocation is ill-formed, synonyms of each component word in the collocation is extracted.
- Each component word in the ill-formed collocation is replaced with its synonyms.
- If any of the combinations bears a high association value, it is listed as a possible correct collocation.

Although the program does not frequently provide good alternative collocations, when it does, its decision on ill-formed collocations seems somewhat reliable. In most cases, the program cannot find a better synonymous collocation. We assume that that is because of the limitation of our sample corpus and the limited number of synonyms in *WordNet*.

We believe that like the collocation exercise generation program, the basic logic of this program is highly effective, and better engineering application will improve the usefulness of the program.

8. Demonstration Websites

For those who are interested in trying out our pilot programs, we have made them available at the URLs below. We also list Perl scripts that are used in the programs online.

- Collocation (error) detection program
  http://www.slacorpus.com/programs/jpn.html

The collocations or collocation errors are listed when the original text is put in the textbox and is sent to our server. This program uses same corpora as our pilot study (10-million-word ANC and 44-million-word WSJ corpus).

- Collocation exercise generation program
  http://www.slacorpus.com/programs/jpn.html

When a collocation is sent to the program, distracters for an exercise are automatically generated.

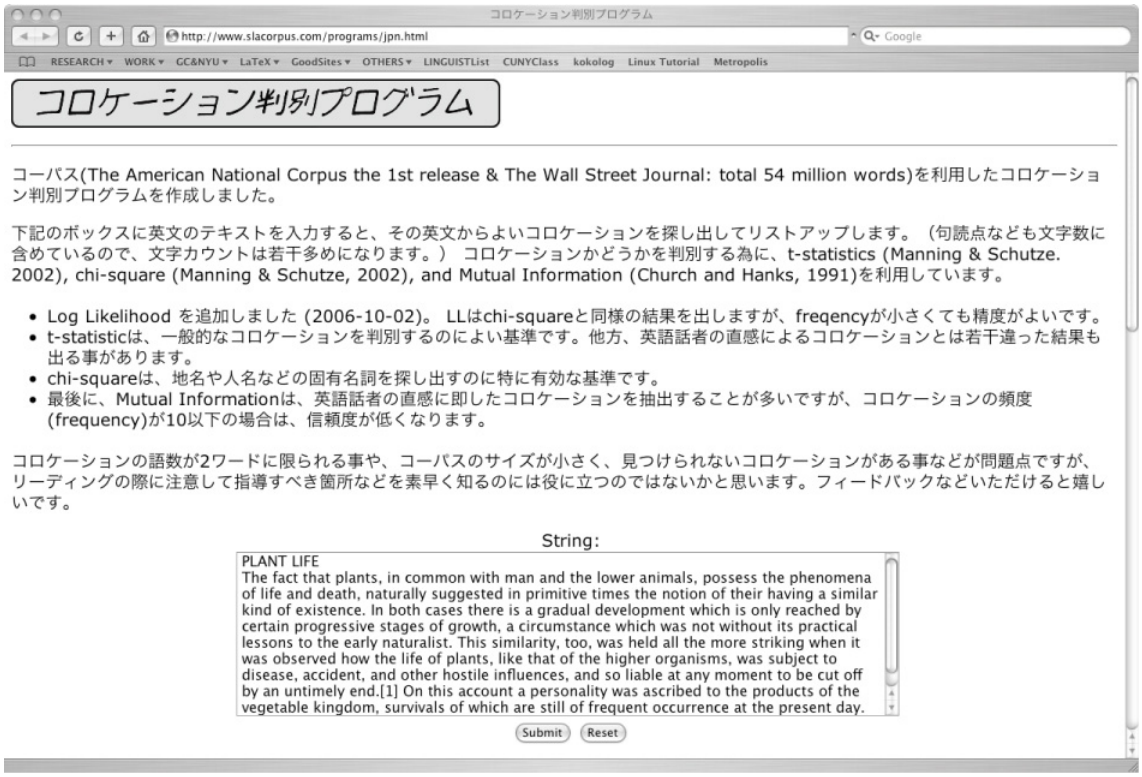

Fig1. Collocation exercise generation program (http://www.slacorpus.com/programs/jpn.html)

Fig2. Introduction to Perl (http://www.slacorpus.com/programs/introPerl.html)

• Introduction to Perl
http://www.slacorpus.com/programs/introPerl.html
Basic Perl scripts are listed on this page. The list includes modules used for programming the two pilot programs in this study, such as modules to compute t-statistic, chi-square, MI, and log-likelihood of bigrams.

9. Conclusion

In this paper, we argued that collocations pose a huge problem for ESL/EFL learners, but instructional materials are crucially lacking in this area. We discussed several underlying problems that make it difficult for language teachers to focus on collocations in the classroom. We proposed that language corpora could be a solution to those problems in the teaching of collocations. The raw frequency can successfully extract good collocations, but better yet, we presented several collocation measures that have been developed in the last two decades in computational linguistics. Finally, we presented two pilot programs that are potentially useful for second/foreign language instruction.

Needless to say, our pilot programs are in too early a stage to draw definitive conclusions. The analyses of the outputs of our pilot programs, however, seem to be very promising, given that even very simple programs produced interesting results. We hope that engineering innovations will help improve the concept of our pilot programs and will achieve results that can be used in the language classroom.

Acknowledgement

The authors would like to express their appreciation to Dr. S. Kathleen Kitao, who read this manuscript and made valuable comments.

References

Benson, M. (1989). The structure of the collocational dictionary. *International Journal of Lexicography*, *2*, 1-14.

Benson, M., Benson, E., and Ilson, R. (1986). *The BBI combinatory dictionary of English: A guide to word combinations*. Amsterdam, Netherlands: John Benjamins.

Church, K. W., and Hanks, P. (1989a). Word association norms, mutual information, and lexicography. In *The 27th annual conference of the association for computational linguistics,* 76-83.

Church, K. W., and Hanks, P. (1989b). Word association norms, mutual information and lexicography (rev). *Computational Linguistics*, *16* (1), 22-29.

Church, K.W., and Mercer, R.L. (1993). Introduction to the special issue on computational linguistic using large corpora. *Computational Linguistics*, *19*, 1-24.

Davis, M. (2006). VIEW: Variation in English Words and Phrases. Visited on October 22, 2006. http://view.byu.edu/

Dunning, T. (1993). Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, *19*, 61-74.

Epstein, S.D., Flynn, S., and Martohardjono, G. (1996). Second language acquisition: Theoretical and experimental issues in contemporary research. *Behavioral and Brain Science, 19*(4), 677–758.

Flynn, S. (1987). *A Parameter-Setting Model of L2 Acquisition. Studies in theoretical psycholinguistics*. Norwell, MA: D. Reidel Publishing Company.

Ide, N., Reppen, R., and Suderman, K. (2002). The American national corpus: More than the web can provide. In *The Third Language Resources and Evaluation Conference (LREC)*, 839–844.

Manning, C. D., and Schutze, H. (1999). *Foundations of statistical natural language processing*. Cambridge, MA: MIT Press.

Miller, G., Beckwith, R., Fellbaum, C., Gross, D., and Miller, K. (1993). *Introduction to WordNet: An on-line lexical database*. Cambridge: MIT Press.

Odlin, T. (1989). *Language transfer: Cross-linguistic influence in language learning*. Cambridge; New York: Cambridge University Press.

Vainikka, V., and Young-Scholten, M. (1996). Gradual development of L2 phrase structure. *Second Language Research*, *12,* 7–39.

Appendix

t-test

The *t-test* (a.k.a. *Student's t-test*) is a robust statistic used for hypothesis testing. The *t-test* produces a statistic value called *t-statistic* by looking at the mean x and variance s^2 of a sample and evaluates the null hypothesis (H_0) that the sample is collected from a distribution with the mean of μ. The standard formula for *t-test* is

$$t-statistic = \frac{\bar{x}-\mu}{\sqrt{\frac{s^2}{N}}} \quad (1)$$

When applied to collocation discovery, the *t-test* is assumed to measure "how (un)likely word co-occurrence are above chance." A high *t-statistic* is considered an indication of fixed placement of words and thus more likely to be a good collocation, whereas a low *t-statistic* suggests the words are scattered throughout the corpus (therefore, not forming collocations). In our study, we employed the computation of *t-statistic* proposed by Manning and Schutze (2002) as presented as Equation (2).

$$t-statistic = \frac{O_{freq}-E_{freq}}{\sqrt{\frac{s^2}{N}}} \quad (2)$$

Where O_{freq} is the observed frequency of n-grams, E_{freq} is the expected frequency (the product of the probabilities of individual words), s^2 is the binominal variance (i.e., *p(1-p)*), and N is the number of tokens in a corpus.

chi-square

The application of *t-test* to collocation discovery is common, but is a theoretical nightmare because the underlying assumptions are indisputably incorrect (e.g., the distribution of words in a corpus is not random). Given the theoretical flaw of *t-test*, some researchers propose to use the non-parametric statistic such as *chi-square*. We will not go into details of the application of *chi-square* in this paper, but interested readers may refer to Manning and Schutze (2002) and Dunning (1993).

The computation formula for *chi-square* statistic that we employed in this study is given as Equation (3)

$$\chi^2 = \sum \frac{(O_{freq}-E_{freq})^2}{E_{freq}} = N\frac{(O_{W_1W_2}-E_{W_1W_2})^2}{E_{W_1W_2}E_{\neg W_1\neg W_2}} \quad (3)$$

Where O_{w1w2} is the observed frequency of word1 and word2 in a collocation, E_{w1w2} is the expected frequency of collocation words, and $O_{\neg w1\neg w2}$ is the frequency of bigrams that do not include any words in the target collocation, and N is the number of tokens in a corpus.

Log likelihood

Log likelihood is a measure to evaluate the degree of dependence between words in a collocation phrase. In computing *log likelihood*, two hypotheses of extreme cases are assumed. H_1 assumes independence of word co-occurrence (thus, no chance of a collocation) and H_2 assumes full dependence of words, which means that when one word appears the other word must appear in the context. *Log likelihood* is simply a degree of dependency between two words measured by the ratio between hypothesis 1 (independence) and hypothesis 2 (dependence). The employed formula is shown below as Equation (4) (equation in the second line is a computational form).

$$\log\lambda = \log\frac{L(H_1)}{L(H_2)}$$
$$-2\lambda = \sum_{ij} O_{ij}\log\frac{O_{ij}}{E_{ij}} \quad (4)$$

Dunning (1993) argues that *log likelihood* is a theoretically sound approach that does not necessarily assume independence of each word in a corpus. It is also argued to be superior to other non-parametic measures (e.g., *chi-square*) because *log likelihood* produces reliable results even with small sample corpora.

Mutual Information (MI)

Finally, we employed an information theoretic measure *(point-wise) Mutual Information (MI)* in this study. The application of *MI* to collocation detection is owed to Church and Hanks (1989a; 1989b) in which *MI* is simply defined as a log of the ratio of a joint

probability to an independent probability.

$$MI = \log_2 \frac{p(x,y)}{p(x)p(y)} \quad (5)$$

In spite of its simple computation, *MI* produces interesting possible collocations when the MI value is high ($MI > 10$) and the frequency of the collocation is larger than 5.