GRASP: Grammar- and Syntax-based Pattern-Finder for Collocation and Phrase Learning

Mei-hua Chen[a], Chung-chi Huang[a], Shih-ting Huang[b], and Jason S. Chang[b]

[a]Institute of Information Systems and Applications, National Tsing Hua University, HsinChu, Taiwan 300, R.O.C.
{chen.meihua, u901571}@gmail.com
[b]Department of Computer Science, National Tsing Hua University, HsinChu, Taiwan 300, R.O.C.
{koromiko1104, jason.jschang}@gmail.com

Abstract. We introduce a method for learning to find the representative syntax-based context of a given collocation/phrase. In our approach, grammatical patterns are extracted for query terms aimed at accelerating lexicographers' and language learners' navigation through the word usage and learning process. The method involves automatically lemmatizing, part-of-speech tagging and shallowly parsing the sentences of a large-sized general corpus, and automatically constructing inverted files for quick search. At run-time, contextual grammar patterns are retrieved and presented to users with their corresponding statistical analyses. We present a prototype system, GRASP (grammar- and syntax-based pattern-finder), that applies the method to computer-assisted language learning. Preliminary results show that the extracted patterns not only resemble phrases in grammar books (e.g., make up one's mind) but help to assist the process of language learning and sentence composition/translation.

Keywords: Computer-assisted language learning, collocation, part-of-speech tagging, grammatical patterns, and inverted files.

1 Introduction

Many language learners' queries (e.g., "play" and "role") are submitted to language-learning tools on the Web every day and an increasing number of services on the Web specifically target second language learning. For example, Word Sketch Engine (www.sketchengine.co.uk) is a concordancer that automatically summarizes a word's grammatical and collocational behavior while services such as TANGO[1] and MUST[2] provide a means of collocation finding (e.g., verb-noun and adjective-noun collocations) and collocation correcting.

Language-learning tools such as Word Sketch and TANGO typically accept only one querying word and retrieve sentences with it or words it co-occurring probabilistically more frequently than usual. However, learners may attempt to learn the usage of a certain word sense of the query word.

Consider the polysemy "play". In WordNet, it has a myriad of senses including '*participating in games or sport*', '*act or having an effect in a specified way*' and '*play on an instrument*'. Learners may be aware of its senses but intend to acquire more knowledge on the context or usage of the word sense '*act or having an effect in a specified way*'. Suggested by (Yarowsky, 1995), accompanying "play" with its collocate "role" is a good way to narrow down

[1] candle.fl.nthu.edu.tw/collocation/webform2.aspx
[2] candle.fl.nthu.edu.tw/vntango/

the senses of “play” (and vice versa). Therefore, multi-word query might be as important in language learning. However, the best response to the multi-word query “play role” is probably not an overwhelming set of sentences with it which may be returned by general-purpose concordancers and search engines (e.g., Google). A good response might indicate that the collocation “play role” is frequently followed by the grammatical part-of-speech (PoS) patterns ‘preposition determiner’ (e.g., “in the” and “in this”), ‘preposition noun’ (e.g., “in society” and “in relation”) and ‘preposition gerund’ (e.g., “in determining” and “in shaping”), and preceded by the patterns ‘noun auxiliary_verb’ (e.g., “communication will” and “confidence will”) and ‘adjective noun’ (e.g., “voluntary groups” and “foreign aid”) and that “play” and “role” are usually separated by the grammatical patterns ‘article adjective’ (e.g., “an important” and “a major”), ‘determiner gerund’ (e.g., “the leading” and “the supporting”) and ‘adjective’ (e.g., “significant” and “crucial”). Intuitively, these PoS patterns provide a general idea on how the querying terms are usually used in context.

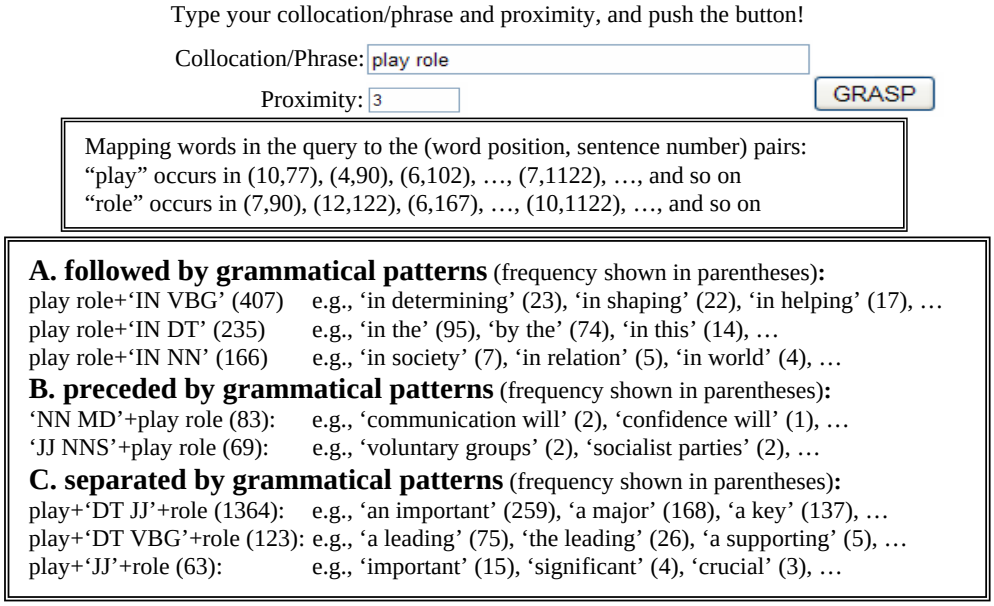

Figure 1. An example GRASP response to the query “play role”.

We present a new system, GRASP, that automatically learns to extract representative grammar-based patterns of the querying collocations/phrases. An example GRASP responses to is shown in Figure 1. GRASP has determined the sentences containing the query’s words of a specific underlying corpus (e.g., British National Corpus). GRASP learns these word-to-sentence mappings during corpus preprocessing. We describe the GRASP preprocessing process in more detail in Section 3.

At run-time, GRASP starts with a collocational/phrasal query submitted by language learners (e.g., “play role”). GRASP then identifies the sentences with the words in the query within proximity and retrieves the grammatical patterns commonly occurring before, after, within the query. In our prototype, GRASP returns patterns together with statistical analyses to users directly (see Figure 1); alternatively, the statistics returned by GRASP can be used as reference to automatically extract possible phrases regarding the query words (e.g., “make up one’s mind” concerning the query words “make up” or “make mind”).

2 Related Work

Computer-assisted language learning (CALL) has been an area of active research. Much research has been studied and developed to assist second language learners in language understanding. Research of concordancers and collocations has received most of the attention.

Concordancers provide a word's grammatical or collocational behavior by displaying example sentences. Word Sketch Engine (Kilgarriff *et al*., 2004), a famous concordancer, has been used in language learning and in the production of the Macmillan English Dictionary. On the other hand, concordancers may be implemented with cross-lingual information (e.g., translations[3]).

Researchers have long considered collocations essential and helpful in language learning and sentence composition (Benson, 1985; Benson et al., 1986; Lewis, 2000; Nation, 2001; Liu, 2002; Nesselhauf, 2003; Chen, 2009; Chen and Lin, 2009; Durrant, 2009). While services such as TANGO (Jian *et al*., 2004) and MUST (Chang *et al*., 2008) assist learners in collocation finding and collocation correcting, learners may still have problems putting a collocation or a phrase in sentences. In this paper, rather than solely returning an overwhelming chuck of sentences with the query collocation/phrase (Cheng *et al*., 2006), we impose an grammatically-motivated thesaurus structure on its context in view of speeding up the process of language learning and lexicography.

3 The GRASP System

3.1 Problem Statement

We focus on providing language learners a means to quickly grasp representative usage of their search phrase and to quickly identify the context they would like to use. The grammatically-motivated context with statistical analyses is returned as the output of the system. The returned analyses can be examined by human learners or lexicographer directly, or passed on to a phrase extraction model (extracting phrase like "make up one's mind"). Therefore, our goal is to return a reasonable-sized set of grammatical patterns that, at the same time, highly represent the context of the query phrase. We now formally state the problem that we are addressing.

Problem Statement: We are given a general corpus *C* (e.g., British National Corpus) that collects articles from a wide range of sources, and a collocation/phrase query *Q*. Our goal is to extract a set of grammatical patterns from *C* that are likely to represent the context *Q* commonly used in. For this, we transform words, $w_1,\ldots, w_m$ in *Q* into sets of (word position, sentence record) pairs such that the top *N* grammatical patterns depicting the query's context are likely to be quickly retrieved.

In the rest of this section, we describe our solution to this problem. First, we preprocess the large-sized general corpus (Section 3.2). This preprocessing includes lemmatizing, PoS tagging, shallow parsing, and constructing aforementioned (position, sentence) pairs. Finally, we show how GRASP extracts and evaluates grammatical patterns/context at run-time (Section 3.3).

3.2 Corpus Preprocessing for GRASP

We attempt to find transformations from words in the query phrase into (position, sentence) pairs expected to accelerate the search for grammatical patterns preceding, following, and separating the query. Our preprocessing process is shown in Figure 2.

(1) Lemmatize and PoS tag the sentences in *C*
(2) Shallowly parse the sentences in *C*
(3) Construct inverted files for corpus *C*
(4) Output inverted files and phrase pairs

Figure 2. Outline of the GRASP Preprocessing.

Lemmatizing and PoS Tagging. In the first stage of the preprocessing (Step (1) in Figure 2), we lemmatize the sentences and generate the most probable PoS tag sequence for each sentence in the general corpus *C*. For example, the sentence "The British Section refugee office has

[3] http://candle.fl.nthu.edu.tw/totalrecall/totalrecall/totalrecall.aspx

played a leading role in this area of work." is lemmatized and grammatically tagged as "The/DT British/JJ Section/NN refugee/NN office/NN have/VBZ play/VBN a/DT lead/VBG role/NN in/IN this/DT area/NN of/IN work/NN ./.". The goal of lemmatization is to reduce the impact of inflectional morphology of words on statistical analyses.

On the other hand, the goal of PoS tagging is to provide a way to group/classify the context/usage of a collocation/phrase. Actually, using PoS tags is quite natural: a myriad of examples of them being used for generalization can be identified in grammar books, such as the "one's" (i.e., possessive pronoun) in the phrase "make up one's mind", the "superlative adjective" (e.g., happiest) in "the most superlative_adjective", the "oneself" (i.e., reflexive pronoun) in "enjoy oneself very much", the "NN" (i.e., noun) and "VB" (i.e., base form of verb) in "insist/suggest/recommend/demand/propose that NN VB" and so on.

Shallow Parsing. In the second stage of the preprocessing process (Step (2) in Figure 2), we generate the shallow parsing result for each sentence. The input to this stage is a set of sentences (likely with PoS tag sequences) while the output of this stage is a set of parsing results of base phrases such as noun phrases, verb phrases, and prepositional phrases. Shallow parsing is also aimed to generalize the query's context.

Constructing Inverted Files. In the third and final stage of preprocessing, we build up inverted files for the lemmas in the corpus *C*. It is one thing to answer to which lemmas occur in the sentences but it is another to answer to in which sentences and positions a lemma appears. While the former is quite straightforward, the latter is not and an opposite case of the former, thus the term "inverted".

For each lemma in *C*, we record the positions and sentences it occurs for run-time search. We also keep track of its corresponding surface word form, PoS tag and shallow parsing result for reference in that they are useful in grammatical pattern finding and language learning. The lemmas' inverted files are targeted at short response time and quick search.

3.3 Run-Time Grammatical Pattern Finding

Once the word-to-sentence mappings, or inverted files, and PoS tagging and shallow parsing results are obtained, GRASP retrieves and evaluates the grammar-based context of the search phrase using the procedure in Figure 3.

In Step (1) of the algorithm we initialize a set, *interInvList*, to collect the intersected inverted lists of the lemmas in the collocation/phrase *query*. For each lemma w_i in *query*, we first obtain its inverted file, *InvList* (Step (2)) and then perform an AND/intersection operation on *interInvList* from previous iteration and *InvList* (from Step (3a) to (3j)[4]).

The AND operation is defined as follows. Firstly, we enumerate the inverted lists, *interInvList* and *InvList* (Step (3b)) after the initialization of their respective indices (i.e., *i* and *j*) and resulting list *newInterInvList* (Step (3a)). Secondly, we incorporate new instance into *newInterInvList* (Step (3e)) if the sentence records of the elements of *interInvList* and *InvList* in question are the same (Step (3c)) and the distance between the word positions of these elements are within *proximity* (Step (3d)). Otherwise, the indices (i.e., *i* and *j*) to the lists are moved accordingly (from Step (3f) to (3i)). Note that, in Step (3e), an instance of (word position, sentence record) is created based on *interInvList*[*i*] and *InvList*[*j*]. For example, if *interInvList*[*i*] is (4,90) and *InvList*[*j*] is (7,90), the newly-created instance is ([4,7],90) indicating the have-already-been-examined lemmas in the search phrase appear in the positions of 4 and 7 of the sentence number 90. Furthermore, to cover more context of the search phrase, function withinProximity of Step (3d) considers the *absolute* difference between word positions. Subsequently, the query words may not appear in order in the patterns. For instance, GRASP would also find the pattern "DT JJ+role+TO+play" (e.g., a vital role to play) for the query "play role". On the other hand, since *query* may contain more than two words (e.g., "in order to" and "as a matter of fact"), *interInvList*[*i*].wordPosi may be a list of word positions of lemmas

[4] These steps only hold for sorted inverted files.

already seen. In our prototype, the distance between *interInvList*[*i*].wordPosi and *InvList*[*j*].wordPosi is defined as the absolute value of the smallest difference between *InvList*[*j*].wordPosi and word position in *interInvList*[*i*].wordPosi. Alternatively, it may be defined as the farthest. Finally, we set *interInvList* to be *newInterInvList* for the next iteration of the AND operation.

```
procedure EvaluateGrammarPattern(query,proximity,N,C)
(1)  interInvList=findInvertedFile(w1 in query)
     for each lemma wi in query except for w1
(2)    InvList=findInvertedFile(wi)
       //perform AND operation on interInvList and InvList
(3a)  newInterInvList= ϕ ; i=1; j=1
(3b)  while i<=length(interInvList) and j<=lengh(InvList)
(3c)     if interInvList[i].SentNo==InvList[j].SentNo
(3d)        if withinProximity(interInvList[i].wordPosi,InvList[j].wordPosi, proximity)
(3e)           Insert(newInterInvList, interInvList[i],InvList[j])
            else if interInvList[i].wordPosi<InvList[j].wordPosi
(3f)           i++
            else //interInvList[i].wordPosi>InvList[j].wordPosi
(3g)           j++
         else if interInvList[i].SentNo<InvList[j].SentNo
(3h)        i++
         else //interInvList[i].SentNo>InvList[j].SentNo
(3i)        j++
(3j)  interInvList=newInterInvList
     //extract grammatical patterns
(4)  PatternCol= ϕ  // a collection of patterns
     for each element in interInvList
(5)    PatternCol += {extractGrammarPattern(element, C)}
(6)  RankedPatterns=Sort grammar patterns in PatternCol in descending order of frequency
(7)  Return the N RankedPatterns with highest frequency
```

Figure 3. Evaluating Patterns at Run-Time.

In Step (5) for all the legitimate sentence records, the query's lemmas along with their context words' PoS tags and shallow parsing results are identified and extracted from sentences. Such information is gathered to express the context of the search phrase in a syntax-based manner: querying words are shown in lemmas and the context in PoS tags (e.g., "play+DT JJ+role"). With the intuition that more frequent the patterns, more representative they are, we rank patterns according to their occurrences in *PatternCol* (Step (6)).

At last, we return the top *N* frequent grammar patterns preceding, following, and separating the query phrase. These patterns are aimed for learners' and lexicographers' quick grasp on and navigation through the usage of collocational/phrasal query. Aside from grammatical patterns, GRASP also displays some of the patterns' characteristic examples for users to better understand the contextual words to choose. The retrieved patterns and examples of a query "play role" are shown in Figure 1. For simplicity, Figure 1 does not include the patterns with reverse word order of "play role", such as "DT JJ+role+TO+play", "DT VBG+role+TO+play" and etc.

4 Preliminary Results

GRASP was designed to retrieve grammatical patterns from a large-scale monolingual corpus that are likely to impose a grammatically-motivated thesaurus structure on the context of a given collocation/phrase in the same language as the monolingual corpus. Furthermore, since the goal of GRASP is to assist users in language learning and understanding, GRASP also extracts characteristic examples for each syntax-based pattern. In this section, we first present the experimental settings in the GRASP system (Section 4.1). Then, Section 4.2 examines some interesting grammatical patterns GRASP extracted. In Section 4.3, we report the positive impact

of GRASP on language learning, especially sentence composition. Finally, we point out future improvements of our system (Section 4.4).

4.1 Experimental Setting and Data Used

We used British National Corpus (BNC) as our underlying large-sized general corpus *C*. It is a 100 million word collection of samples of written and spoken British English from a wide range of sources including newspapers, periodicals and journals, academic books and popular fiction, university essays and etc. We exploited GENIA tagger[5] developed by Tsujii Laboratory to obtain the base forms, PoS tags and shallow parsing results of *C*'s sentences. The tagger based on the Penn Treebank Tagset. After lemmatizing and tagging, all sentences in BNC (approximately 5.6 million sentences) were used to build up inverted files and used as examples for extracting grammar patterns.

4.2 Interesting Patterns GRASP Extracted

In this subsection, we examine some grammar-like patterns and their corresponding representative examples GRASP retrieved for the given collocational/phrasal queries.

For the query "play role", GRASP found that the frequent context which precedes, follows and separates it is "NN MD" (e.g., "communication will") and "JJ NNS" (e.g., "voluntary groups"), "IN DT" (e.g., "in the") and "IN VBG" (e.g., "in determining"), and "DT JJ" (e.g., "an important") and "DT VBG" (e.g., "a leading"), respectively (see Figure 1 for more examples). Also, GRASP pointed out that if "play" and "role" are to appear in reverse order in context, they would often be preceded by "DT JJ" (e.g., "a vital"), followed by "IN DT" (e.g., "in the"), and separated by "TO" (e.g., "to"), implying a common pattern to include the inverted "play role" is "DT JJ+role+TO+play+IN DT" (e.g., "a key role to play in the").

For queries "have impact", "exert influence" and "in order for", GRASP retrieved the worth-learning syntax-based patterns "have+DT JJ+impact+IN NNS" (e.g., "have a profound impact on people"), "exert+DT JJ+influence+IN DT" (e.g., "exert a significant influence on the") and "in order for+PRP TO" (e.g., "in order for us to"), respectively. For the query "make up", four frequently used phrases/patterns were extracted: "make up+PRP$ NN" (e.g., "make up his mind"), "make up+IN DT" (e.g., "make up for the"), "NNS WDT+make up" (e.g., "groups that make up") and passive "make up+IN NNS" (e.g., "made up of representatives"). Moreover, as suggested by GRASP, "as a matter of fact" was frequently surrounded by punctuation marks and was likely a transitional phrase.

Encouragingly, due to GRASP's flexibility in the word order of the query in extracted patterns, it tolerates mis-ordered[6] query words. Take the Chinese-ordered query "1990 Jan. 20" ("一九九零年一月二十日") for example. The grammar pattern "IN+Jan. 20+,+ 1990+, DT" (e.g., "On Jan. 20, 1990, the"[7]) GRASP yielded provided not only the common way to put dates in English sentences (i.e., the commas and the preposition "on") but the right order (i.e., "Jan. 20 1990" instead of "1990 Jan. 20").

4.3 Evaluation Results

We introduced monolingual GRASP to a class of 32 first-year college students learning English as second language. They were taught on how to use GRASP[8] for their benefits (finding common contexts for collocations/phrases) and asked to perform two tests: pretest and posttest.

In our experiment, pretest was a test where students were asked to complete English sentences with Chinese sentences as hints, while posttest was a test where, after using traditional methods like dictionaries or online translation systems (controlled group), or GRASP (experimental group) in-between pretest and posttest to learn the contexts of

[5] http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger/
[6] Mostly because of first-language (L1) interference.
[7] The capitalized word "On" indicates the beginning of a sentence.
[8] A prototype is at http://koromiko.cs.nthu.edu.tw/grasp/

collocations/phrases in a candidate list provided by us, students were also asked to complete the translations of the Chinese sentences.

In both the pretest and the posttest, there were exactly the same 15 to-be-finished English translations of Chinese sentences with test items reordered, in which contain one frequent collocation/phrase of BNC corpus. Below shows one of the test items:

Chinese: 環境保護對地球有深遠的影響

English: Environmental protection has ___ impact ___.

The test item aims to examine students' understanding on the contexts of the collocate "has impact". The answer to the first blank is "a profound" and the second is "on the Earth". And a candidate list of 20[9] frequent collocations and phrases in BNC was provided for learning between tests (See Table 1 for example). Half of the class used GRASP for learning and the other half used traditional learning approach such as online dictionaries or online translation system (i.e., Google Translate and Babelfish).

Table 1. Examples in our candidate list for learning.

Example	Translation
make … difference…	差異很大
… individual needs	滿足需求
place … order …	下訂單
turn … ear …	充耳不聞
in … future	不久的將來

We summarize the averaged students' scores on pre- and post-test in Table 2 where *G* stands for GRASP (experimental group) and *T* for traditional methods (controlled group), and "All" denotes all students in the group, "UH" the upper half of the group in scores, and "BH" the bottom half.

Table 2. The performance on pretest and posttest.

	pretest (%)			posttest (%)		
	All	UH	BH	All	UH	BH
G	26.40	34.48	18.31	**41.91**	**48.06**	**35.77**
T	27.10	34.23	19.95	32.75	33.49	32.01

As suggested by Table 2, the difference between *G* and *T* was insignificant under pretest, implying the partition of the class was quite random. Very encouragingly, GRASP helped to improve students' achievements on completing the English sentences (by 41.9-26.4=15.5). Although students also performed better after querying online dictionaries/translation systems (by 32.7-27.1=5.6), GRASP seemed to help students with more margin, almost tripled (15.5 vs. 5.6). More closely, both UH and BH students benefited from GRASP, from score 34.4 to 48.0 (+13.6) and from score 18.3 to 35.7 (+17.4), respectively. This suggests that GRASP is suitable for not only high-achieving students but low-achieving for language learning.

4.4 Future Improvements of GRASP

Many avenues exist for future research and improvement of our system. For example, shallow parsing results (e.g., B-NP, B-VP[10]) could be further incorporated to replace the "DT" in pattern "play role+IN DT" and the "NN" and "MD" in "NN MD+play role". Patterns "play role+IN NP" and "NP VP+play role" would be more informative and grammar-like. On the other hand, since many second language learners have difficulties choosing the right prepositions, we would like to include prepositional words into grammatical patterns. For queries "play role", "have impact", and "exert influence", the patterns lexicalized on prepositional PoS tags look like "play role+IN(in) DT", "have impact+IN(on) NN", and "exert influence+IN(on) NN" where collocating prepositions are shown in parentheses. Such lexicalized grammar patterns if accompanied with target-language translations may be helpful not only for language learners but for syntax-based machine translation systems, such as Hiero (Chiang et al., 2005).

[9] Include the 15 test items.

[10] The "B" stands for the beginning of a phrase constituent.

We would also like to construct a cross-lingual GRASP system in which it accepts query terms in second language learners' mother tongue and returns grammatical patterns of their suitable translations. For example, for the first-language query "打擊 犯罪" (fight crime), the cross-lingual GRASP would provide organized grammatical patterns (as described in this paper) of its English translations such as "fight crimes", "combat crimes", "crack down on crimes" and etc. Language learners may benefit from the cross-lingual GRASP since they may be more at ease (or prefer) submitting queries in their first language, or may have hard time translating their thoughts, i.e. collocations/phrases, into the language they are learning at the first place.

5 Summary and Future Work

We have introduced a method for extracting common grammar patterns that contain the search collocations/phrases. The method involves automatically lemmatizing, PoS tagging, and shallowly parsing the sentences of a general corpus, and building up words' inverted files for quick run-time search. We have implemented the method and preliminarily evaluated our method as applied to language learning. The promising and interesting results prompt us to better the system, GRASP, and to move the research to the next step: more extensive survey and experiment on the system in helping lexicographers and language learners monolingually or *bilingually*. Aside from future work described in Section 4.4, we would like to incorporate GRASP-extracted patterns like "play+DT JJ+role" into syntax-based MT decoders as context-sensitive/lexicalized rules. Since these patters are shown to be important to human users in sentence composition, they are likely to be vital to syntax-based decoders as well.

References

M. Benson. 1985. Collocations and idioms. In *Robert Ilson (Ed.), Dictionaries, Lexicography and Language Learning*.

M. Benson, E. Benson and R. Ilson. 1986. *The BBI Combinatory Dictionary of English. A Guide to Word Combinations*. Amsterdam/Philadelphia: Benjamins.

D. Chiang, A. Lopez, N. Madnani, C. Monz, P. Resnik, and M. Subotin. 2005. The Hiero machine translation system: extensions, evaluation and analysis. In *Proceedings of HLT/EMNLP*.

W. Cheng, C. Greaves, and M. Warren. 2006. From n-gram to skipgram to concgram. In *Corpus Linguistics*, 11 (4).

Y.C. Chang, J.S. Chang, H.J. Chen, and H.C. Liou. 2008. An automatic collocation writing assistant for Taiwanese EFL learners: a case of corpus-based NLP technology. In *Computer Assisted Language Learning*, 21 (3), 2008.

M.H. Chen. 2009. English collocation competence of college students and learner factors. In *Proceedings of the 26th International Conference on English Teaching and Learning in the R.O.C.*

M.H. Chen and M. Lin. 2009. Common collocation errors of college students in Taiwan. In *Proceedings of the International Conference on Applied Linguistics and Language Teaching*.

P. Durrant. 2009. Investigating the viability of a collocation list for students of English for academic purposes. In *English for Specific Purposes*, 28 (3).

J.Y. Jian, Y.C. Chang, and J.S. Chang. 2004. TANGO: Bilingual collocational concordancer. In *Proceedings of the Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions.*

A. Kilgarriff, P. Rychly, P. Smrz, and D. Tugwell. 2004. The sketch engine. In *Proceedings of EURALEX*.

P. Koehn, F. J. Och, and D. Marcu. 2003. Statistical phrase-based translation. In *Proceedings of NAACL/HLT*.

M. Lewis. 2000. Language in the Lexical Approach. In *M. Lewis (Ed.), Teaching Collocation: Further Development in the Lexical Approach*.

L.E. Liu. 2002. *A corpus-based lexical semantic investigation of verb-noun miscollocations in Taiwan learners' English*, PHD dissertation.

I.S.P. Nation. 2001. *Learning Vocabulary in Another Language*. Cambridge: Cambridge Press.

N. Nesselhauf. 2003. The use of collocations by advanced learners of English and some implications for teaching. In *Applied Linguistics*, 24 (3).

D. Yarowsky. 1995. Unsupervised word sense disambiguation rivaling supervised methods. In *Proceedings of the Annual Meeting of the ACL*.