Polibits

ISSN: 1870-9044

polibits@nlp.cic.ipn.mx

Instituto Politécnico Nacional

México

Chen, Mei-Hua; Huang, Chung-Chi; Huang, Shih-Ting; Liou, Hsien-Chin; Chang, Jason S.

A Cross-Lingual Pattern Retrieval Framework

Polibits, vol. 43, 2011

Instituto Politécnico Nacional

Distrito Federal, México

A Cross-Lingual Pattern Retrieval Framework

Mei-Hua Chen, Chung-Chi Huang, Shih-Ting Huang, Hsien-Chin Liou, and Jason S. Chang

***Abstract*—We introduce a method for learning to grammatically categorize and organize the contexts of a given query. In our approach, grammatical descriptions, from *general* word groups to *specific* lexical phrases, are imposed on the query's contexts aimed at accelerating lexicographers' and language learners' navigation through and *GRASP* upon the word usages. The method involves lemmatizing, part-of-speech tagging and shallowly parsing a general corpus and constructing its inverted files for monolingual queries, and word-aligning parallel texts and extracting and pruning translation equivalents for cross-lingual ones. At run-time, grammar-like patterns are generated, organized to form a thesaurus index structure on query words' contexts, and presented to users along with their instantiations. Experimental results show that the extracted predominant patterns resemble phrases in grammar books and that the abstract-to-concrete context hierarchy of querying words effectively assists the process of language learning, especially in sentence translation or composition.**

***Index terms*—Grammatical constructions, lexical phrases, context, language learning, inverted files, phrase pairs, cross-lingual pattern retrieval.**

I. Introduction

MANY language learners' queries (e.g., "play" or "role") are submitted to computer-assisted language learning tools on the Web for word definitions or usages every day. And an increasing number of Web services specifically target English as Foreign Language (EFL) learners' search questions.

Web-based language learning tools such as Sketch Engine, concordancers, and TANGO typically take monolingual single-word query and retrieve too many its collocations and example sentences such that they overwhelm and confuse users due to the amount of returned sentences and different usages therein. However, users may want to learn the context patterns, or grammatical sequences underlying contextual word strings, (e.g., 'play article adjective role') of a specific word sense of a word and submit multiple-word queries (e.g., "play role"), and users may need an index to quickly navigate through one usage to another. Besides, EFL users may prefer submitting queries in their first languages. These queries could be answered more appropriately if a tool provided grammatical categories to their contexts and understood other languages.

Manuscript received November 28, 2010. Manuscript accepted for publication January 5, 2011.

Mei-Hua Chen, Chung-Chi Huang, Shih-Ting Huang, and Jason S. Chang are with ISA, NTHU, HsinChu, Taiwan, R.O.C. 300 (e-mail: {chen.meihua, u901571,koromiko1104,Jason.jschang}@gmail.com).

Hsien-Chin Liou is with FL, NTHU, HsinChu, Taiwan, R.O.C. 300 (e-mail: liuhsienc@gmail.com).

Consider the learner query "play role". The best response is probably not the overwhelming set of sentences containing "play role". A good response might generalize and categorize its representative contexts such as: "play role" separated by "DT JJ" (common instantiation: "an important") where "DT" denotes an article and "JJ" an adjective, "play role" followed by "IN VBG" (instantiation: "in determining") where "IN" denotes a preposition and "VBG" a gerund, and "play role" preceded by "NN MD" (instantiation: "communication will") where "NN" denotes a noun and "MD" an auxiliary verb. Such generalization and categorization of the query's contexts can be achieved by part-of-speech (PoS) tagging its sentences. Intuitively, by word-class or PoS information, we can bias a retrieval system towards grammar-like pattern finder. On the other hand, by leveraging machine translation techniques, we can channel the first-language query to its English substitutes.

We present a new system, *GRASP* (grammar- and syntax-based pattern-finder) that automatically characterizes the contexts of querying collocations or phrases in a grammatical manner. An example cross-lingual *GRASP* search for the Chinese collocation "扮演角色" ("play role" or "play part") is shown in Figure 1. *GRASP* has directed the first-language query "扮演角色" to one of its probable English translations, "play role", and gathered its predominant patterns of phraseology in terms of the relative position between the query and its contexts, and the distances between the querying words, based on a balanced monolingual corpus. Take the most frequent distance (i.e., 3) where "play" and "role" are apart from each other for example. "Play" and "role" are most likely to be separated by word group "DT JJ", constituting the lexically open formal idiom or grammatical construction "play DT JJ role" what we call *GRASP* syntactic pattern. And this *GRASP* pattern's frequent idiomatic lexical realizations or phrases, or lexically filled substantive idioms[1], are "play an important role". To extract such formal or substantive idioms, *GRASP* learns translations and word-to-sentence mappings automatically (Section 3).

At run-time, *GRASP* starts with an English query or a first-language query for usage learning.

GRASP then retrieves aforementioned formal idioms lexically anchored with English query words' lemmas and their substantive counterparts/instantiations. The former are designed for quick word usage navigation and the latter for better understanding of phraseological tendencies.

[1] See (Fillmore *et al.*, 1988).

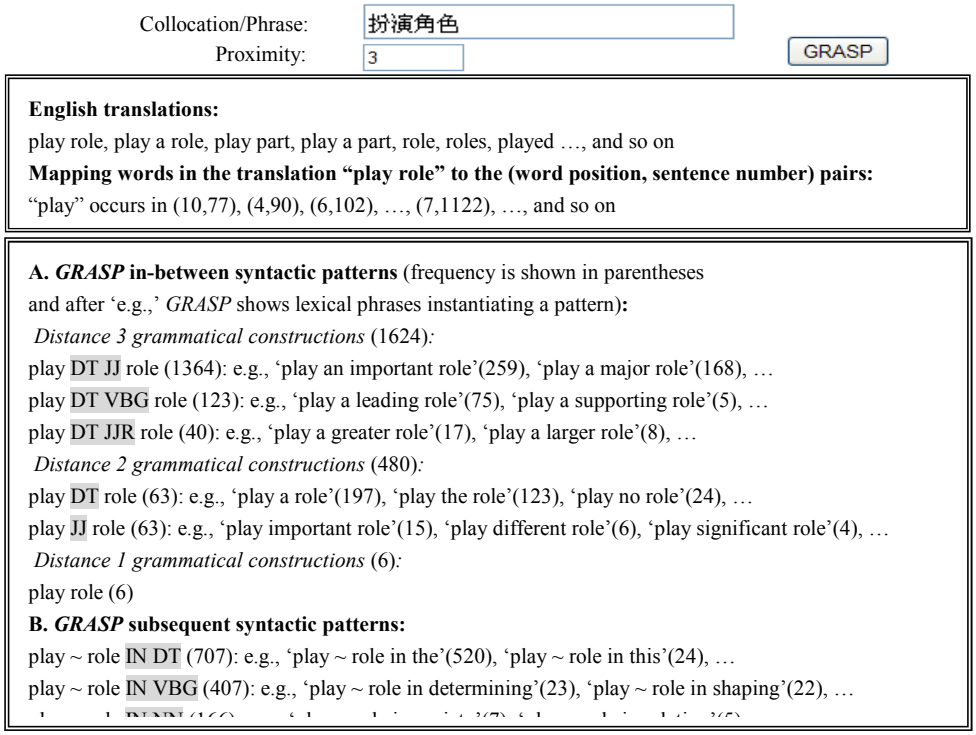

Fig. 1. An example *GRASP* response to the query "扮演角色" ("play role").

In our prototype, *GRASP* accepts queries of any length and responds with example sentences and frequencies of the formal or substantive idioms.

II. Related Work

Ever since large-sized corpora and computer technology became available, many linguistic phenomena have been statistically modeled and analyzed. Among them is collocations long been considered essential in language learning. In the beginning, collocations are manually exemplified and examined (Firth, 1957; Benson, 1985; Benson *et al.*, 1986; Sinclair 1987; Lewis, 2000; Nation, 2001). Right after a pioneering statistical analysis on collocations (Smadja, 1993), the area of research soon becomes computationally possible (Kita and Ogata, 1997) and active especially in English for academic purpose (Durrant, 2009) or second language learning (e.g., (Liu, 2002) and (Chang *et al.*, 2008)).

Recently, some collocation finders such as *Sketch Engine*, *TANGO* and *JustTheWord* have been developed and publicly available. *Sketch Engine* (Kilgarriff *et al.*, 2004) summarizes a word's collocational behavior. *TANGO* (Jian *et al.*, 2004) further provides cross-language searches while *JustTheWord* automatically clusters co-occurring words of queries. In this paper, we take note of the regularities of words' contexts and grammatically express the regularities as patterns for language learning. Such patterns go beyond the collocations from collocation finders, possibly limited to certain combinations of lexical or grammatical collocations and missing the important contextual word groups or words of the collocations.

Textual cohesion is observed in phrases as well. Therefore, phraseology and pattern grammar have drawn much attention. Phraseology can be studied via lexically fixed word sequences, i.e., n-grams (Stubbs, 2002), or totally lexical-open PoS-grams (Feldman *et al.*, 2009; Gamon *et al.*, 2009). In contrast to these two extremes, Stubbs (2004) introduces phrase-frames (p-frames) which bases on n-gram but with one variable slot. Our framework lies between n-grams and PoS-grams, and our extracted sequences of patterns consist of more-than-one variable slots featuring the contexts surrounding the querying words.

Recent work has been done on statistically analyzing the contexts, patterns, frames or constructions of words. A lexical-grammatical knowledge database, StringNet, is built and described in (Wible *et al.*, 2010). However, their work may over-generalize the querying words to word groups during database construction and does not handle multi-word or cross-lingual queries. Users of language tools may submit these two types of queries in that patterns are closely associated with meanings, senses of words, and multiple words usually restrains the senses of words (see (Yarowsky, 1995)), and users may experience problems composing queries in the language they are learning. Hence, we propose a multi-word and cross-lingual pattern-retrieval framework in which patterns are anchored with users' querying words with their contextual words generalized. In a study more related to our work, (Cheng *et al.*, 2006) describes the concept of conc-grams and how to use conc-grams to find constituency and positional variants of search words. The main difference from their work is that we give descriptions to query words' predominant contexts in a grammatical and systematic manner. The descriptions are thesaurus index structures, consisting of

constructions from lexically-open syntactic patterns to lexically fixed idioms.

III. The *GRASP* Framework

A. Problem Statement

We focus on imposing a thesaurus index structure on the querying words' contexts. This structure, formed by a hierarchy from *general* (lexically open) grammatical constructions to *specific* (lexically fixed) substantive idioms anchored with query words, provides a means for quick navigation and understanding of words' typical patterns and their instantiated lexical phrases, and is returned as the output of the system. The returned constructions, or patterns can be examined by learners and lexicographers directly or a syntax-based machine translation system. Thus, it is crucial that the set of patterns cannot be so large that it overwhelms the user. At the same time, there is a need for first-language query search among EFL learners. Therefore, our goal is to return a reasonable-sized set of recurrent grammatical patterns and their idiomatic lexical realizations for language learning or lexicography that represents queries' attendant phraseology and expected lexical items, taking both monolingual and cross-lingual query search. We now formally state the problem that we are addressing.

Problem Statement: We are given a large-scale general corpus C (e.g., British National Corpus), a parallel text T (e.g., Hong Kong Parallel Text), and a query phrase Q. Our goal is to extract and organize the contexts of the query Q lexico-grammatically and lexically based on C that are likely to assist users in navigating and learning the usages of Q. 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 lexico-grammatical patterns and their lexical instances depicting the query's context are likely to be quickly retrieved. T, on the other hand, makes cross-lingual query and learning possible.

B. Corpora Preprocessing

In the corpora preprocessing, we attempt to find transformations from words in the query into (position, sentence) pairs, collocations for single-word query for starters, and English translations for first-language query, expected to accelerate the search for *GRASP* grammatical patterns and expected to accommodate EFL learners' habits of composing a query.

Lemmatizing, PoS Tagging and Shallow Parsing. In the first stage of the preprocessing, we lemmatize each sentence in the general corpus C and generate its most probable PoS tag sequence and shallow parsing result. The goal of lemmatization is to reduce the impact of inflectional morphology of words on statistical analyses while that of PoS tagging is to provide a way to grammatically describe and generalize the contexts/usages of a collocation/phrase. Shallow parsing results, on the other hand, provide the base phrases of a sentence. And consecutive base phrases are often used for extracting collocation candidates.

Finding Collocations. In the second stage of the preprocessing process, we identify a set of reliable collocations in C based on statistical analyses. Collocations of single-word queries may be presented to language learners with, to some extent, few clues, as starters for more complete and specific queries.

The input to this stage is a set of lemmatized, PoS tagged and shallowly parsed sentences while the output of this stage is a set of statistically-suggested collocations. The method for finding reliable collocations in C consists of a number of steps, namely, determining the head words in the base phrases from shallow parser, constituting the head words as collocation candidates, calculating the pair-wise mutual information (MI) values of the head words, and filtering out the collocation candidates whose MI values do not exceed an empirical threshold.

Considering the enrichment (usually adjectives and prepositions) *GRASP* can offer and the observation that EFL learners have hard time composing sentences with verb-noun (VN) collocations and choosing right following prepositions, collocation type to bridge single-word query focuses on VN and verb-preposition (VP) collocation. Focusing on VN collocations and VP collocations, we highlight the contiguous verb phrase and noun phrase, and verb phrase and prepositional phrase in C. In the highlighted verb, noun and prepositional phrases, we intuitively consider their last verb, noun and preposition to be the head words and constitute collocation candidate of the form $<\text{word}_1,\text{pos}_1,\text{word}_2,\text{pos}_2>$ based on the two head words in the two base phrases. To examine the candidates, we compute MI values using

$$\text{MI}=\log(\mathit{freq}(\text{word}_1,\text{pos}_1,\text{word}_2,\text{pos}_2)/(\mathit{freq}(\text{word}_1,\text{pos}_1)\times\mathit{freq}(\text{word}_2,\text{pos}_2)))$$

in which $\mathit{freq}(*)$ denotes the frequency. MI values have been used to determine the mutual dependence of two events. The higher the MI values, the more dependent they are. At last, we retain only candidates whose MI values exceed threshold Θ and think of them as statistically-suggested collocations.

Constructing Inverted Files. In the third stage of preprocessing, we build up inverted files for the lemmas in the corpus C. For each lemma in C, we record the positions and sentences in which it resides for run-time query. Additionally, its corresponding surface word form, PoS tag and shallow parsing result are kept for reference in that such information gathered across lemmas is useful in grammatical pattern finding and (potentially) language learning.

Word-aligning and phrase pairs extracting. In the fourth stage, we exploit a large-scale parallel text T for bilingual phrase acquisition, rather than using a manually compiled dictionary to achieve satisfying translation coverage and variety.

We acquire phrase pairs via the following procedure. First, we word-align the bitext in T leveraging the IBM model 1 to model 5 implemented in *GIZA++* (Och and Ney, 2003). To "smooth" the saw-toothed word alignments produced by directional word alignment model of IBM and collect words with no translation equivalent in another language in phrases,

grow-diagonal-final is used for bidirectional word alignment combination. Finally, heuristics in (Koehn *et al.*, 2003) are used for bilingual phrase extracting.

Pruning unlikely phrase pairs. In the fifth and final stage of the preprocessing, we filter out less probable or insignificant translation equivalents obtained from *T*. In this paper, we apply the pruning techniques described in (Johnson *et al.*, 2007). Specifically, we use their significance testing of phrases to first prune insignificant phrase pairs and rank the English translations of the first-language search queries. For language learning, an accurate and small but diverse set of translations are especially helpful. Moreover, *GRASP* patterns will be shown for the translations, if triggered or automatically, which further provides the hierarchical index for navigation through specific usages and word associations in English for the query initially in users' mother tongue. One thing worth mentioning is that the set of translation equivalents outputted in this stage includes those in which we skip some word pairs in the phrase pairs, in order to increase the translation coverage for the first-language queries. The skipped phrase pairs are constructed as follows. For each phrase pair, we skip some number of the words on the first-language end and if the skipped words have word alignments on the English part, the aligned English words are also skipped. Then we constitute the un-skipped words in the two languages as a skipped phrase pair.

C. Run-Time Index Structure Building and Pattern Finding

Once collocates, word-to-sentence mappings, and confident phrase pairs are obtained, *GRASP* constructs the thesaurus index hierarchy for English contexts and phraseology of the query using the procedure in Figure 2.

In Step (1) of the algorithm we reformulate the user-nominated query into a set of new queries, *Queries*, if necessary. The first type of the reformulation concerns the language used for the input *query*. If *query* is in a language other than that of *C*, we translate the *query* into its statistically significant (English) translations based on the pruned and skipped phrase tables from *T*, and append each of these translations to *Queries* considering it as a search query as if it were submitted by the user. The second concerns the length of the query. Since presenting single word alone to *GRASP* is uncertain with its word sense in question and contexts or pattern grammars are typically highly associated with a word's meanings, for single-word queries, we use their reliable collocations, specifically VN and VP ones, obtained from Section 3.2 as stepping stones to *GRASP* syntactic patterns. These again are incorporated into *Queries*. Note that for these two kinds of query transformation, users may be allowed to choose their own interested translation or collocation of the *query* in implementation and presented only with its (i.e., the translation's or collocation's) *GRASP* hierarchy of word usages. The prototypes for first-language, Chinese in particular, queries and monolingual single-word or multi-word queries are at http://140.114.214.80/theSite/*GRASP*_v552/ and http://140.114.214.80/theSite/b*GRASP*_v552/ respectively. In Step (2) we initialize a set *GRASPresponses* to collect *GRASP* grammatical patterns of queries in *Queries* now in English and more-than-one words.

```
procedure GRASPindexBuilding(query,proximity,N,C,T)
(1)  Queries=queryReformulation(query)
(2)  GRASPresponses= ϕ
     for each query in Queries
(3)    interInvList=findInvertedFile(w1 in query)
       for each lemma wi in query except for w1
(4)      InvList=findInvertedFile(wi)
         //perform AND operation on interInvList and InvList
(5a)     newInterInvList= ϕ ; i=1; j=1
(5b)     while i<=length(interInvList) and j<=lengh(InvList)
(5c)       if interInvList[i].SentNo==InvList[j].SentNo
(5d)         if withinProximity(interInvList[i].wordPosi,
                 InvList[j].wordPosi, proximity)
(5e)           Insert(newInterInvList, interInvList[i],InvList[j])
             else if interInvList[i].wordPosi<InvList[j].wordPosi
(5f)           i++
             else //interInvList[i].wordPosi>InvList[j].wordPosi
(5g)           j++
           else if interInvList[i].SentNo<InvList[j].SentNo
(5h)         i++
           else //interInvList[i].SentNo>InvList[j].SentNo
(5i)         j++
(5j)     interInvList=newInterInvList
     //GRASP thesaurus index building
(6)  PatternIndex= ϕ  // a collection of patterns for this query
     for each element in interInvList
(7)    PatternIndex+= {GrammarPatternGeneration(query,element,C)}
(8a) Sort patterns and their instances in PatternIndex
     in descending order of frequency
(8b) GRASPresponse=top N patterns and instances in PatternIndex
(9)  append GRASPresponse to GRASPresponses
(10) return GRASPresponses
```

Fig. 2. Run-Time Index Building and Pattern Finding.

In Step (3) *interInvList* is initialized to contain the intersected inverted files of the lemmas in the *query*. For each lemma w_i in *query*, we obtain its inverted file, *InvList* (Step (4)) before performing an AND/intersection operation on *interInvList*, intersected results from previous iteration, and *InvList* (from Step (5a) to (5j)[2]). The AND operation is defined as follows. First, we enumerate the inverted lists, *interInvList* and *InvList* (Step (5b)) after the initialization of their respective indices (i.e., *i* and *j*) and temporary resulting list *newInterInvList* (Step (5a)). Second, we incorporate a new instance into *newInterInvList* (Step (5e)) if the sentence records of the indexed elements of *interInvList* and *InvList* in question are the same (Step (5c)) and the distance between the word positions of these elements are within *proximity* (Step (5d)). Note that, in Step (5e), a new instance of (word position, sentence record) is created based on *interInvList*[*i*] and *InvList*[*j*] and inserted into *newInterInvList*. Furthermore, taking into account the positional variations of a

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

collocation/phrase (e.g., "play role" and "role play"), function withinProximity of Step (5d) considers the *absolute* difference between word positions, to cover contexts of differently-ordered querying words. Finally, we set *interInvList* to be *newInterInvList* for the next iteration of the AND operation (Step (5j)).

After finding the legitimate sentences containing a query's words within certain distance, *GRASP* retrieves and builds the hierarchical index structure for its contexts. In Step (7) we generate grammar patterns or cases of word usages for each *element*, taking the form ([wordPosi(w_1), …, wordPosi(w_i), …], sentence number) pointing out the validated sentence record and the word positions of the query's lemmas in that sentence, in *interInvList*. In function GrammarPatternGeneration, based on *element* and *C*'s lemmas and PoS tags, we first transform the legitimate sentence by replacing its words with PoS tags except for the words in positions [wordPosi(w_1), …, wordPosi(w_i), …] and replacing these words with lemmas. Afterwards, we extract contiguous segments surrounding the query lemmas from the transformed sentence, resulting in syntax-based context of the search query (e.g., "play DT JJ role" and "play ~ role IN VBG"). Such lexically open pattern grammars representing the regularity of words' contexts are referred to as *GRASP* syntactic patterns in this paper. Very similarly, the lexically fixed realizations of these patterns could be extracted.

We collect the *N* most frequent (recurrent or potentially idiomatic) *GRASP* syntactic patterns and their *N* most frequent realizations (Step (8)), and gather them as a *GRASP* response *GRASPresponse*. At last, we return all the responses (i.e., *GRASPresponses*) that may interest our users. Figure 1 illustrates the summarized grammatical context ontology for "play role" from a Chinese query "扮演角色".

D. Further Improvement to GRASP

In this subsection, we manage to further extend the *GRASP* patterns. The extension is made in two ways: lexicalization and sub-categorization.

TABLE I
PATTERNS BEFORE AND AFTER LEXICALIZATION

Query	Before	After
play role	play ~ role IN DT (707)	play ~ role IN(in) DT (599)
	play ~ role IN VBG (407)	play ~ role IN(in) VBG (397)
	role ~ play IN DT (235)	role ~ play IN(in) DT (128)
		role ~ play IN(by) DT (89)
have effect	have ~ effect IN DT (1199)	have ~ effect IN(on) DT (887)
	have ~ effect IN VBG (644)	have ~ effect IN(of) VBG (533)
		have ~ effect IN(upon) DT (83)

In writing we observe that EFL learners often have difficulty choosing the right preposition following a collocation (e.g., VN, AN, and PN collocation). Therefore, we lexicalize on the IN PoS tag, a prepositional PoS tag, in *GRASP* patterns to present the specific prepositions to users. Table I shows example *GRASP* patterns before and after lexicalization. Note that lexicalization is indicated in parentheses and that the statistics of frequencies (numbers in parentheses) may change.

Secondly, to acquire grammar rules such as "provide SOMEBODY with SOMETHING" and "provide SOMETHING to SOMEBODY" in grammar books, we semantically subcategorize PoS tags in *GRASP* patterns. Although some current patterns may be informative enough in terms of the semantic roles of the PoS tags, some are not especially the ones with the too general PoS tags NN and NNS, standing for singular and plural nouns respectively. We thus classify the semantic roles of these tags in *GRASP* patterns.

We now describe our simple strategy for semantic role categorization, relying on a lexical thesaurus with words' semantic roles or meanings. In our implementation, we use WordNet where each sense of a word has a higher-level and more abstract supersense, or lexicographers' file. The strategy first, for each extracted pattern accompanied with words of the NN and NNS tags (e.g., "provide NNS(clients) with"), uniformly distributes the pattern's frequency among all supersenses of the NN or NNS words. Then by re-grouping and re-ranking the semantically-motivated patterns, *GRASP* finds not only the grammatical contexts but the most probable semantic roles of NN and NNS tags in these contexts. Sample of semantically subcategorized patterns is shown in Table II where semantic roles are in squared parentheses.

TABLE II
PATTERNS BEFORE AND AFTER SEMANTIC ROLE LABELING

Query	Before	After
provide with	provide NNS with (394)	provide NNS[PERSON] with (252)
		provide NNS[GROUP] with (43)
provide to	provide NN to (325)	provide NN[COMMUNICATION] to (65)
		provide NN[ACT] to (63)

IV. EXPERIMENTS

A. Experimental Settings

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. We exploited GENIA tagger developed by Tsujii Laboratory to obtain the lemmas, PoS tags and shallow parsing results of *C*'s sentences. After lemmatizing and syntactic analyses, all sentences in BNC (approximately 5.6 million sentences) were used to build up inverted files and used as examples for extracting grammar patterns. As for bilingual parallel data, we used Hong Kong Parallel Text (LDC2004T08) assuming the first language of the language learners is Chinese. We leveraged CKIP Chinese segmentation system (Ma and Chen, 2003) to word segment the Chinese sentences within.

B. Interesting Patterns GRASP Extracted

In this subsection, we examine some grammar-like patterns generated by *GRASP*. Take monolingual query "make up" for

example. *GRASP* identified its four lexico-grammatical patterns with different associated senses: "make up PRP$[3] NN[COGNITION]" (e.g., "make up his mind"), "make up IN(for) DT" (e.g., "make up for the") for the sense *to compensate*, "NNS WDT make up" (e.g., "groups that make up") and passive "make up IN(of) NNS[PERSON]" (e.g., "made up of representatives") for the sense *to constitute*, and "make up DT NN[COMMUNICATION]" (e.g., "make up the story") for the sense *to fabricate*. It is challenging for collocation finders to obtain such patterns or usages since they usually do not accommodate multi-word queries, let alone finding the prepositions following a verbal phrase like "make up". Due to *GRASP*'s flexibility in the word order of the query in extracted patterns, it tolerates mis-ordered query words. Take the Chinese-ordered query "1990 Jan. 20" for example. The grammar pattern "IN Jan. 20 , 1990 , DT" (e.g., "On Jan. 20, 1990, the") *GRASP* yielded provides not only the common way to put dates in English sentences but the right order.

As for the cross-lingual mode, *GRASP* accepted Chinese queries like "打擊犯罪" (fight crime) and returned the characteristic syntax-based patterns anchored with their confident English translations: "fight crime", "combat crime" and "crack down on crime". EFL learners would benefit from cross-lingual *GRASP* in that it helps them to learn correct and yet versatile translations of the first-language queries, bypassing the erroneous user-nominated English queries because of first-language interference, as well as those translations' grammatical contexts. Take the Chinese query "學習知識" (*acquire knowledge*) for instance. *GRASP* responded with its diverse translation equivalents "acquire knowledge", "acquire the knowledge of", "learn skills" and so on, *excluding* the miscollocation "learn knowledge" commonly seen in English writing from Chinese learners.

C. Evaluation Results

To carefully control the variables in assessing the effectiveness of the thesaurus index structure *GRASP* provides for usage learning and navigation, we introduced monolingual *GRASP*[4] *alone* to EFL learners and they were taught on how to use *GRASP* for their benefits. Two classes of 32 and 86 first-year college students learning English as second language participated in our experiments. They were asked to perform a common language learning practice: sentence translation/composition, comprising two tests of pretest and posttest. In our experiments, pretest was a test where participants were asked to complete English sentences with their corresponding Chinese sentences as hints, while posttest was a test where, after utilizing traditional tools like dictionaries and online translation systems or *GRASP* in-between pretest and posttest to learn the usages of collocations/phrases in a candidate list provided by us, participants were also asked to complete the English translations of the Chinese sentences. In both the tests, there were exactly the same 15 to-be-finished test items, English translations with Chinese sentences, only with different orders. Each test item contains one frequent collocation/phrase based on the statistics from BNC corpus.

As mentioned above, a candidate list of 20 frequent collocations and phrases in BNC was provided for learning between tests. Participants were asked to concentrate on learning the contexts of the senses of the English collocations/phrases (e.g., "place order") specified by their Chinese counterparts (e.g., "下訂單"). To evaluate *GRASP*, half of the participants 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 *Yahoo! Babel Fish*).

We summarize the averaged scores of our participants on pre- and post-test in Table 3 and 4 where *GRASP* stands for the (experimental) group using *GRASP* and *Trad* for the (controlled) group using traditional tools, and "ALL" denotes all students in the group, "UH" the upper half of the group in scores and "BH" the bottom half. As suggested by Table III and IV, the partition of the classes was quite random in that the difference between *GRASP* and *Trad* was insignificant under pretest and the index structure imposed by *GRASP* on words' contexts was helpful in language learning. Specifically, in table III *GRASP* helped to improve students' achievements on completing/composing the English sentences by 15.5% (41.9-26.4). Although students also performed better after consulting online dictionaries or translation systems by 5.6% (32.7-27.1), *GRASP* seemed to help students with more margin, almost tripled (15.5 vs. 5.6). Encouragingly, if we look closer, we find that 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 the effectiveness of *GRASP* in language learning do not confine to certain level of students but crosses from high-achieving students to low-achieving.

TABLE III
THE PERFORMANCE ON PRETEST AND POSTTEST OF THE 1ST CLASS

	pretest (%)			**posttest (%)**		
	All	UH	BH	All	UH	BH
GRASP	26.4	34.4	18.3	**41.9**	**48.0**	**35.7**
Trad	27.1	34.2	19.9	32.7	33.4	32.0

TABLE IV
THE PERFORMANCE ON PRETEST AND POSTTEST OF THE 2ND CLASS

	pretest (%)	**posttest (%)**
GRASP	43.6	**58.4**
Trad	43.8	53.4

The helpfulness of *GRASP* was observed in another class (see Table IV). Class-to-class, in spite of the fact that the pretest performance of the 2nd class was much better than that

[3] PRP$ stands for a pronoun or a possessive.
[4] The system we introduced is at http://koromiko.cs.nthu.edu.tw/*GRASP*/

of the 1st class, the *GRASP* group of this high-achieving class *still* outperformed the *Trad* group (58.4 vs. 53.4), another indicator that the assistance of *GRASP* system is across different levels of students in language learning. Even in this comparatively high-performing class, the *GRASP*'s gain (58.4-43.6=14.8) is one third of the original pretest score (i.e., 43.6) and the gain is more than 1.5 times larger than *Trad*'s gain (53.4-43.8=9.6), suggesting that *GRASP* is much more effective and efficient in language learning than traditional lookup methods, mostly attributed to *GRASP* general-to-specific categorized usages, contexts, or phraseologies of words.

V. Conclusions and Future Work

Many avenues exist for future research and improvement of our system. For example, an interesting direction to explore is the effectiveness of our fully capable *GRASP*, responding to both monolingual and cross-lingual queries, in language learning. Additionally, we would like to examine the possibility of constructing a grammar checker based on our *GRASP* lexical-grammatical patterns. Yet another direction of research is to apply the *GRASP* framework to different languages and to associate the *GRASP*-extracted patterns in different languages for syntax-based machine translation system.

In summary, we have introduced a framework for learning to impose general-to-specific thesaurus index structures, comprising recurrent grammar patterns and their predominant lexical realizations, on queries' contexts. The characterizing context index structures assist users such as lexicographers and language learners in two ways: the generalization in patterns accelerates the navigation through different usages and the instantiations of patterns, i.e., lexical phrases, provide phraseological tendencies. We have implemented and evaluated the framework as applied to CALL, especially in second language writing. Extracted syntactic patterns have been shown to go beyond the collocations from common collocation finders and resemble phrases in grammar books. And we have verified (in two separate evaluations) that our hierarchical index structures on words' contextual regularity help the process of language learning.

References

[1] M. Benson, "Collocations and idioms," in Robert Ilson (Ed.), *Dictionaries, Lexicography and Language Learning*, 1985.

[2] M. Benson, E. Benson and R. Ilson, *The BBI Combinatory Dictionary of English. A Guide to Word Combinations*, 1986.

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

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

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

[6] S. Feldman, M. Marin, J. Medero, and M. Ostendorf, "Classifying factored genres with part-of-speech histograms," in *Proceedings of NAACL,* 2009.

[7] C.J. Fillmore, P. Kay, and M.K. O'Connor, "Regularity and idiomaticity in grammatical constructions: the case of *let alone*," *Language* 64, 1988.

[8] J.R. Firth, "Modes of meaning," *Papers in linguistics*. Oxford: Oxford University Press, 1957.

[9] M. Gamon, C. Leacock, C. Brockett, W.B. Dolan. J.F. Gao, D. Belenko, and A. Klementiev, "Using statistical techniques and web search to correct ESL errors," *CALICO*, 26(3), 2009.

[10] J.Y. Jian, Y.C. Chang, and J.S. Chang, "TANGO: Bilingual collocational concordance" in *Proceedings of ACL,* 2004.

[11] J.H. Johnson, J. Martin, G. Foster, and R. Kuhn, "Improving translation quality by discarding most of the phrasetable," in *Proceedings of EMNLP,* 2007.

[12] A. Kilgarriff, P. Rychly, P. Smrz, and D. Tugwell, "The sketch engine," in *Proceedings of EURALEX,* 2004.

[13] K. Kita and H. Ogata, "Collations in language learning: corpus-based automatic compilation of collocations and bilingual collocation concordance," in *Computer Assisted Language Learning*, 10 (3), 1997.

[14] P. Koehn, F.J. Och, and D. Marcu, "Statistical phrase-based translation," in *Proceedings of NAACL/HLT,* 2003.

[15] M. Lewis, "Language in the Lexical Approach," in M. Lewis (Ed.), *Teaching Collocation: Further Development in the Lexical Approach*, 2000.

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

[17] I.S.P. Nation, *Learning Vocabulary in Another Language*. Cambridge: Cambridge Press, 2001.

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

[19] F. Smadja, "Retrieving collocations from text: Xtract," *Computational Linguistics*, 19(1), 1993.

[20] M. Stubbs, "Two quantitative methods of studying phraseology in English," *Corpus Linguistics* 7(2), 2002.

[21] M. Stubbs, 2004. http://web.archive.org/web/20070828004603/http://www.uni-trier.de/uni/fb2/anglistik/Projekte/stubbs/icame-2004.htm.

[22] D. Wible and N.L. Tsao, "StringNet as a computational resource for discovering and investigating linguistic constructions," in *Proceedings of NAACL,* 2010.

[23] D. Yarowsky, "Unsupervised word sense disambiguation rivaling supervised methods," in *Proceedings of the Annual Meeting of the ACL,* 1995.