Computational Linguistics and Chinese Language Processing
Vol. 14, No. 3, September 2009, pp. 257-280

A Thesaurus-Based Semantic Classification of English Collocations

Chung-Chi Huang*, **Kate H. Kao**+, **Chiung-Hui Tseng**+ **and**

Jason S. Chang+

Abstract

Researchers have developed many computational tools aimed at extracting collocations for both second language learners and lexicographers. Unfortunately, the tremendously large number of collocates returned by these tools usually overwhelms language learners. In this paper, we introduce a thesaurus-based semantic classification model that automatically learns semantic relations for classifying adjective-noun (A-N) and verb-noun (V-N) collocations into different thesaurus categories. Our model is based on iterative random walking over a weighted graph derived from an integrated knowledge source of word senses in *WordNet* and semantic categories of a thesaurus for collocation classification. We conduct an experiment on a set of collocations whose collocates involve varying levels of abstractness in the collocation usage box of Macmillan English Dictionary. Experimental evaluation with a collection of 150 multiple-choice questions commonly used as a similarity benchmark in the TOEFL synonym test shows that a thesaurus structure is successfully imposed to help enhance collocation production for L2 learners. As a result, our methodology may improve the effectiveness of state-of-the-art collocation reference tools concerning the aspects of language understanding and learning, as well as lexicography.

Keywords: Collocations, Semantic Classification, Semantic Relations, Random Walk Algorithm, Meaning Access Index and *WordNet*.

* CLCLP, TIGP, Academia Sinica, Taipei, Taiwan

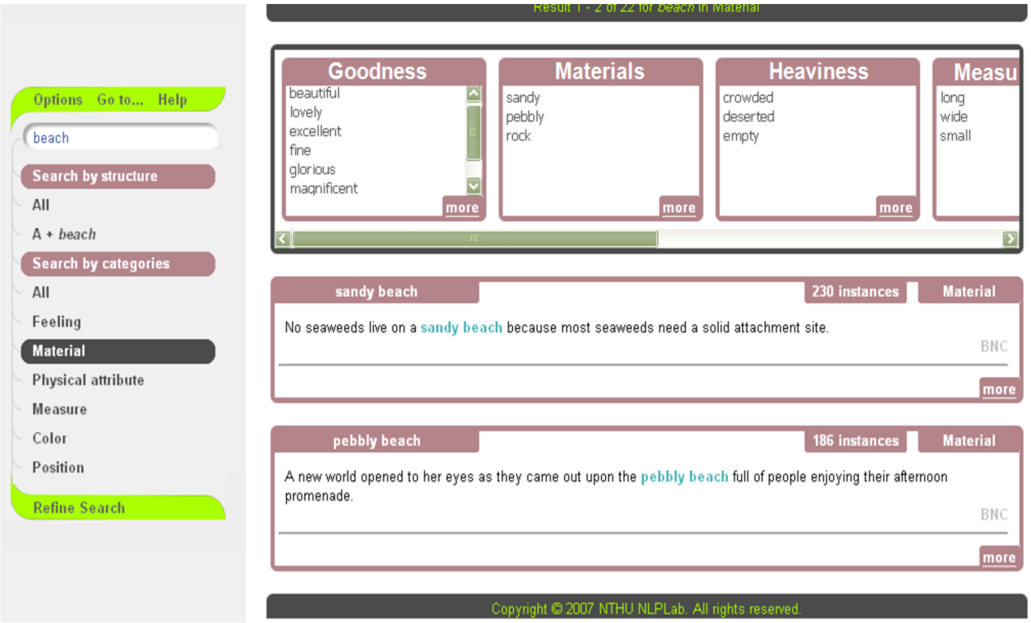
+ Institute of Information Systems and Applications, NTHU, Hsinchu, Taiwan
E-mail: {u901571, msgkate, smilet, jason.jschang}@gmail.com

1. Introduction

Researchers have developed applications of computational collocation reference tools, such as several commercial collocation dictionary CD-ROMs, Word Sketch (Kilgarriff & Tugwell, 2001), *TANGO* (Jian *et al*., 2004), to answer queries (*e.g*., a search keyword “beach” for its adjective collocates) of collocation usage. These reference tools typically return collocates (*e.g*., adjective collocates for the pivot word “beach” are “rocky,” “golden,” “beautiful,” “raised,” “sandy,” “lovely,” “unspoiled,” “magnificent,” “deserted,” “fine,” “pebbly,” “splendid,” “crowded,” “superb,” *etc*.) extracted from a corpus of English texts (*e.g*., *British National Corpus*).

Unfortunately, existing tools for language learning sometimes present too much information in a batch on a single screen. With corpus sizes rapidly growing to Web scale (*e.g*., Web 1 Trillion 5-gram Corpus), it is common to find hundreds of collocates for a query word. The bulk of information may frustrate and slow L2 learners’ progress of learning collocations. An effective language learning tool also needs to take into consideration second language learners’ absorbing capacity at one sitting. To satisfy the need for presenting a digestible amount of information at one time, a promising approach is to automatically partition collocations of a query word into various categories to support meaningful access to the search results and to give a thesaurus index to collocation reference tools.

Consider the query “beach” in a search for its adjective collocates. Instead of generating a long list of adjectives like the above-mentioned applications, a better presentation could be composed of clusters of adjectives inserted into distinct semantic categories such as: {*fine, lovely, superb, beautiful, splendid*} assigned with a semantic label “*Goodness,*” {*sandy, rocky, pebbly*} assigned with a semantic label “*Materials,*” *etc*. Intuitively, by imposing a semantic structure on the collocations, we can bias the existing collocation reference tools towards giving a thesaurus-based semantic classification as one of the well-developed and convincingly useful collocation thesauri. We present a thesaurus-based classification system that automatically groups collocates of a given pivot word (here, the adjective collocates of a noun, the verb collocates of a noun, and the noun collocates of a verb) into semantically related classes expected to render highly useful applications in computational lexicography and second language teaching for L2 learners. A sample presentation for a collocation thesaurus is shown in Figure 1.



No seaweeds live on a sandy beach because most seaweeds need a solid attachment site.

A new world opened to her eyes as they came out upon the pebbly beach full of people enjoying their afternoon promenade.

Figure 1. Sample presentation for the adjective collocate search query "beach".

Our thesaurus-based semantic classification model has determined the best semantic labels for 859 collocation pairs, focusing on: (1) A-N pairs and clustering over the adjectives (*e.g.*, "fine beach"); (2) V-N pairs and clustering over the verbs (*e.g.*, "develop relationship"); and (3) V-N pairs and clustering over the nouns (*e.g.*, "fight disease") from the specific underlying collocation reference tools (in this study, from *JustTheWord*). Our model automatically learns these useful semantic labels using the Random Walk Algorithm, an iterative graphical approach, and partitions collocates for each collocation types (*e.g.*, the semantic category "*Goodness*" is a good thesaurus label for "fine" in the context of "beach" along with other adjective collocates such as "lovely," "beautiful," "splendid," and "superb"). We describe the learning process of our thesaurus-based semantic classification model in more detail in Section 3. At runtime, we assign the most probable semantic categories to collocations (*e.g.*, "sandy," "fine," "beautiful," *etc.*) of a pivot word (*e.g.*, "beach") for semantic classification. In this paper, we exploit the Random Walk Algorithm to disambiguate word senses, assign semantic labels, and partition collocates into meaningful groups.

The rest of the paper is organized as follows. We review the related work in the next section. Then, we present our method for automatic learning to classify collocations into semantically related categories, which is expected to improve the presentation of underlying collocation reference tools and support collocation acquisition by computer-assisted language learning applications for L2 learners (Section 3). As part of our evaluation, two metrics are designed with very little precedent of this kind. One, we assess the performance of resulting collocation clusters by a robust evaluation metric; two, we evaluate the conformity of semantic labels by a three-point rubric test over a set of collocation pairs chosen randomly from the classifying results (Section 5).

2. Related Work

Many natural language processing (NLP) applications in computational lexicography and second language teaching (SLT) build on one part of lexical acquisition emphasizing teaching collocation for L2 learners. In our work, we address an aspect of word similarity in the context of a given word (*i.e.*, collocate similarity), in terms of use, acquisition, and ultimate success in language learning.

This section offers the theoretical basis on which recommendations for improvements to the existing collocation reference tools are made, and it is made up of three major sections. In the first section, an argument is made in favor of collocation ability being an important part of language acquisition. Next, we show the need to change the current presentation of collocation reference tools. The final section examines other literature on computational measures for word similarity versus collocate similarity.

2.1 Collocations for L2 Learners

The past decade has seen an increasing interest in the studies on collocations. This has been evident not only from a collection of papers introducing different definitions of the term "collocation" (Firth, 1957; Benson, 1985; Nattinger & DeCarrico, 1992; Nation, 2001), but also from the inclusive review of research on collocation teaching and the relation between collocation acquisition and language learning (Lewis, 1997; Hall, 1994).

New NLP applications for extracting collocations, therefore, are a great boon to both L2 learners and lexicographers alike. SLT has long favored grammar and memorization of lexical items over learning larger linguistic units (Lewis, 2000). Nevertheless, several studies have shown the importance of acquisition of collocations; moreover, they have found specifically that the most important is learning the right verbs in verb-noun collocations (Nesselhauf, 2003; Liu, 2002). Chen (2004) showed that verb-noun (V-N) and adjective-noun (A-N) collocations were found to be the most frequent error patterns. Liu (2002) found that, in a study of English learners' essays from Taiwan, 87% of miscollocations were attributed to the misuse of V-N collocations. Of those, 96% were due to the selection of the wrong verb. A simple example will suffice to illustrate: in English, one writes a check and also writes a letter while the equivalent Mandarin Chinese word for the verb "write" is "kai" (開) for a check and "xie" (寫) for a letter, but absolutely not "kai" (開) for a letter.

This type of language-specific idiosyncrasy is not encoded in either pedagogical grammars or lexical knowledge but is of utmost importance to fluent production of a language.

2.2 Meaning Access Indexing in Dictionaries

Some attention has been paid to the investigation of the dictionary needs and reference skills of language learners (Scholfield, 1982; Béjoint, 1994), and one important cited feature is a structure to support users' neurological processes in meaning access. Tono (1984) was among the first attempts to claim that the dictionary layout should be more user-friendly to help L2 learners access desired information more effectively. According to Tono (1992) in his subsequent empirical close examination of the matter, menus that summarize or subdivide definitions into groups at the beginning of entries in dictionaries would help users with limited reference skills to access the information in the dictionary entries more easily. The *Longman Dictionary of Contemporary English*, 3rd edition [ISBN 0-582-43397-5] (henceforth called *LDOCE3*), has just such a system called "**Signposts**". When words have various distinct meanings, the *LDOCE3* begins each sense anew with a word or short phrase which helps users more effectively discover the meaning they need. The *Cambridge International Dictionary of English* [ISBN 0-521-77575-2] does this as well, creating an index called "**Guide Word**" which provides similar functionality. Finally, the *Macmillan English Dictionary for Advanced Learners* [ISBN 0-333-95786-5], which has "Menus" for heavy-duty words with many senses, utilizes this approach as well.

Therefore, in this paper, we introduce a classification model for imposing a thesaurus structure on collocations returned by existing collocation reference tools, aiming at facilitating concept-grasping of collocations for L2 learners.

2.3 Similarity of Semantic Relations

The construction of practical, general word sense classification has been acknowledged to be one of the most difficult tasks in NLP (Nirenburg & Raskin, 1987), even with a wide range of lexical-semantic resources such as *WordNet* (Fellbaum, 1998) and *Word Sketch* (Kilgarriff & Tugwell, 2001).

Lin (1997) presented an algorithm for word similarity measured by its distributional similarity. Unlike most corpus-based word sense disambiguation (WSD) algorithms, where different classifiers are trained for separate words, Lin used the same local context database as the knowledge source for measuring all word similarities. Approaches presented to recognize synonyms have been studied extensively (Landauer & Dumais, 1997; Deerwester *et al.*, 1990; Turney, 2002; Rehder *et al.*, 1998; Morris & Hirst, 1991; Lesk, 1986). Measures of recognizing collocate similarity, however, are not as well developed as measures of word similarity.

The most closely related work focuses on automatically classifying semantic relations in noun pairs (*e.g.*, mason:stone) and evaluation with a collection of multiple-choice word analogy question from the SAT exam (Turney, 2006). Another related approach, presented in Nastase and Szpakowicz (2003), describes how to automatically classify a noun-modifier pair, such as “laser printer,” according to the semantic relation between the head noun (printer) and the modifier (laser). The evaluation is manually conducted by human labeling. For a review of work to a more fine-grained word classification, Pantel and Chklovski (2004) presented a semi-automatic method for extracting fine-grained semantic relations between verbs. VerbOcean (http://semantics.isi.edu/ocean/) is a broad-coverage semantic network of verbs, detecting similarity (*e.g.*, transform::integrate), strength (*e.g.*, wound::kill), antonymy (*e.g.*, open::close), enablement (*e.g.*, fight::win), and temporal happens-before (*e.g.*, marry::divorce) relations between pairs of strongly associated verbs using lexico-syntactic pattern over the Web. Hatzivassiloglou and McKeown (1993) presented a method towards the automatic identification of adjectival scales. Based on statistical techniques with linguistic information derived from the corpus, the adjectives, according to their meaning based on a given text corpus, can be placed in one group describing different values of the same property. Their clustering algorithm suggests some degree of adjective scalability; nevertheless, it is interesting to note that the algorithm discourages recognizing the relationship among adjectives, *e.g.*, missing the semantic associations (for example a semantic label of “time associated”) between *new-old*. More recently, Wanner *et al*. (2006) sought to semi-automatically classify the collocations from corpora via the lexical functions in dictionary as the semantic typology of collocation elements. While there is still a lack of fine-grained semantically-oriented organization for collocation, *WordNet* synset (*i.e.*, synonymous words in a set) information can be explored to build a classification scheme for refinement of the model and develop a classifier to measure the distribution of class for the new tokens of words set foot in. Our method, which we will describe in the next section, uses a similar lexicon-based approach for a different setting of collocation classification.

3. Methodology

3.1 Problem Statement

We focus on the preparation step of partitioning collocations into categories for collocation reference tools: providing words with semantic labels, thus, presenting collocates under thesaurus categories for ease of comprehension. The categorized collocations are then returned in groups as the output of the collocation reference tool. It is crucial that the collocation categories be fairly consistent with human judgment and that the categories of collocates cannot be so coarse-grained that they overwhelm learners or defeat the purpose of users’ fast access. Therefore, our goal is to provide semantic-based access to a well-founded collocation thesaurus. The problem is now formally defined.

Problem Statement: We are given (1) a set of collocates $Col = \{C_1, C_2, \ldots, C_n\}$ (*e.g*., “sandy,” “beautiful,” “superb,” “rocky,” *etc*.) with corresponding parts-of-speech $P=\{p| \ p \in Pos$ and $Pos=\{\text{noun,adjective,verb}\}\}$ for a pivot word *X* (*e.g*., “beach”); (2) a combination of thesaurus categories (*e.g*., *Roget’s Thesaurus*), $TC = \{(W, P, L)\}$ where a word *W* with a part-of-speech *P* is under the general-purpose semantic category *L* (*e.g*., feelings, materials, art, food, time, etc.); and (3) a lexical database (*e.g*., *WordNet*) as our word sense inventory *SI* for semantic relation population. *SI* is equipped with a measure of semantic relatedness: REL(*S*, *S*’) encodes semantic relations holding between word sense *S* and *S*’.

Our goal is to partition *Col* into subsets of similar collocates by means of integrated semantic knowledge crafted from the mapping of *TC* and *SI*, whose elements are likely to express related meanings in the same context of *X*. For this, we leverage a graph-based algorithm to assign the most probable semantic label *L* to each collocation, thus giving collocations a thesaurus index.

For the rest of this section, we describe our solution to this problem. In the first stage of the process, we introduce an iterative graphical algorithm for providing each word with a word sense (Section 3.2.1) to establish integrated semantic knowledge. A mapping of words, senses, and semantic labels is thus constructed for later use of automatic collocation partitioning. In the second stage (Section 3.2.2), to reduce out-of-vocabulary (OOV) words in *TC*, we extend word coverage of limited *TC* by exploiting a lexical database (*e.g*., *WordNet*) as a word sense inventory, encoding words grouped into cognitive synonym sets and interlinked by semantic relations. In the third stage, we present a similar graph-based algorithm for collocation labeling using the extended *TC* and Random Walk on a graph in order to provide a semantic access to collocation reference tools of interest (Section 3.3). The approach presented here is generalizable to allow construction from any underlying semantic resource. Figure 2 shows a comprehensive framework for our unified approach.

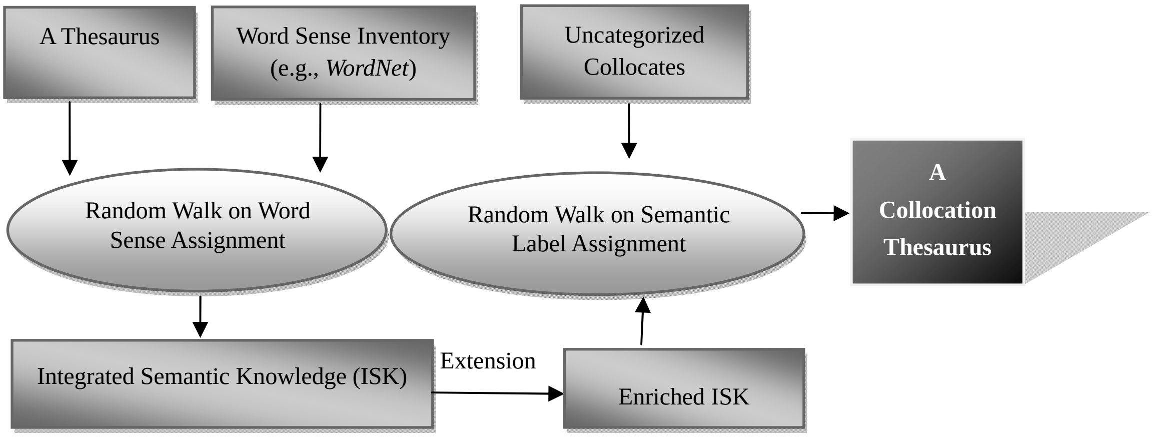

A Thesaurus
Word Sense Inventory (e.g., *WordNet*)
Uncategorized Collocates
Random Walk on Word Sense Assignment
Random Walk on Semantic Label Assignment
A Collocation Thesaurus
Integrated Semantic Knowledge (ISK)
Extension
Enriched ISK

Figure 2. A comprehensive framework for our classification model.

3.2 Learning to Build a Semantic Knowledge by Iterative Graphical Algorithms

In this paper, we attempt to provide each word with a semantic label and attempt to partition collocations into thesaurus categories. In order to partition a large-scale collocation input and reduce the out-of-vocabulary (OOV) encounters for the model, we first incorporate word sense information in *SI*, into the thesaurus, *i.e.*, *TC*, and extend the former integrated semantic knowledge (*ISK*) using semantic relations provided in *SI*. Figure 3 outlines the aforementioned process.

(1) Build an Integrated Semantic Knowledge (*ISK*) by Random Walk on Graph (Section 3.2.1)

(2) Extend Word Coverage for Limited *ISK* by Lexical-Semantic Relations (Section 3.2.2)

Figure 3. Outline of the learning process of our model.

3.2.1 Word Sense Assignment

In the first stage (Step (1) in Figure 3), we use a graph-based sense linking algorithm which automatically assigns appropriate word senses to words under a thesaurus category. Figure 4 shows the algorithm.

Algorithm 1. Graph-based Word Sense Assignment

Input: A word list, *WL*, under the same semantic label in the thesaurus *TC*; A word sense inventory *SI*.

Output: A list of linked word sense pairs, {(*W*, *S**)}

Notation: Graph *G* = {*V*, *E*} is defined over admissible word senses (*i.e.*, *V*) and their semantic relations (*i.e.*, *E*). In other words, each word sense *S* constitutes a vertex $v \in V$ while a semantic relation between senses *S* and *S'* (or vertices) constitutes an edge in *E*. Word sense inventory *SI* is organized by semantic relations *SR* and REL(*S*,*S'*) identifies the semantic relations between sense of *S* and *S'* in *SI*.

PROCEDURE AssignWordSense(*WL,SI*)

Build weighted graph *G* of word senses and semantic relations

```
        INITIALIZE $V$ and $E$ as two empty sets
        FOR each word $W$ in $WL$
            FOR each of the $n(W)$ admissible word senses, $S$, of $W$ in $SI$
(1)             ADD node $S$ to $V$
        FOR each node pair $(S,S')$, where $S$ and $S'$ belong to different words, in $V \times V$
(2)         IF ( REL$(S,S') \neq$ NULL and $S \neq S'$ THEN ADD edge $E(S,S')$ to $E$ and $E(S',S)$ to $E$
        FOR each word $W$ AND each of its word senses $S$ in $V$
(3)         INITIALIZE $P_S = 1/n(W)$ as the initial probability
```

```
(3a)    ASSIGN weight (1-$d$) to matrix element $M_{S,S}$
(3b)    COMPUTE $e(S)$ as the number of edges leaving $S$
        FOR each other word $W' \neq W$ in $WL$ AND each sense $S'$ of $W'$
(3c)        IF there is an edge between $S$ and $S'$ THEN ASSIGN Weight $d/e(S)$ to $M_{S,S'}$
            OTHERWISE ASSIGN 0 to $M_{S,S'}$
```

Score vertices in *G*

```
        REPEAT
            FOR each word $W$ AND each of its word senses $S$
(4)             INTIALIZE $Q_S$ to $P_S \times M_{S,S}$
                FOR each other word $W' \neq W$ in $WL$ AND each sense $S'$ of $W'$
(4a)                INCREMENT $Q_S$ by $P_{S'} \times M_{S',S}$
            FOR each word $W$, SUM $Q_S$ over $n(W)$ senses as $N_w$
            FOR each word $W$ AND each of its word senses $S$
(4b)            REPLACE $P_S$ by $Q_S/N_w$
        UNTIL probability $P_S$‘s converge
```

Assign word sense

```
(5)     INITIALIZE $List$ as NULL
        FOR each word $W$ in $WL$
(6)         APPEND ($W$,$S^*$) to $List$ where $P_{S^*}$ is the maximum among senses of $W$
(7)     OUTPUT $List$
```

Figure 4. Algorithm for Graph-based Word Sense Assignment.

The algorithm for the best sense assignment S^* for W consists of three main parts: (1) construction of a weighted word sense graph; (2) sense scoring using the iterative Random Walk algorithm; and (3) word sense assignment.

In Step 1 of the algorithm, by referring to *SI*, we populate candidate $n(W)$ senses for each word W in the word list, *WL*, under the same semantic category as vertices in graph *G*. In *G*, directed edges $E(S,S')$ and $E(S',S)$ are built between vertex S and vertex S' if and only if there exists a semantic relation between the word sense S and S' in *SI*. Figure 5 shows an example of such a graph.

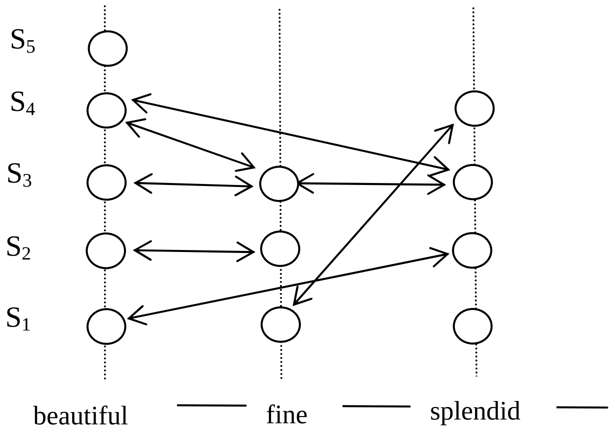


Figure 5. Sample graph built on the admissible word senses (vertical axis) for three words (horizontal axis) under the thesaurus category of "Goodness". Note that self-loop edges are omitted for simplicity.

We initialize the probability concerning the sense S of a word W, P_s, to $1/n(W)$, uniform distribution among the senses of W (Step (3)). For example, in Figure 5, the probability of the fourth sense of the word "beautiful" is initialized to 0.2. Then, we construct a matrix, whose element $M_{x,y}$ stands for the proportion of the probability P_x , that will be propagated to node y. Since $M_{x,y}$ may not be equal to $M_{,y,x}$, the edges in G are directed. In matrix M, we assign 1-d to $M_{x,x}$ where $x \in V$(Step (3a)) while the rest of the proportion (*i.e.*, d) is uniformly distributed among the outgoing edges of the node x (Step (3c)). Take the fourth sense (Node 4 for short) of the word "beautiful" and the third sense (Node 8 for short) of the word "fine" in Figure 5 for example. $M_{4,8}$ is $d/2$ since there are two outgoing edges for Node 4. On the other hand, $M_{8,4}$ is $d/3$ in that there are three edges leaving Node 8. d is the damping factor and was first introduced by PageRank (Brin & Page, 1998), a link analysis algorithm. The damping factor is usually set around 0.85, indicating that eighty-five percent of the probability of a node will be distributed to its outbound nodes.

In the second part of the algorithm, probabilities will be iteratively re-distributed among the senses of words until convergence of probabilities. For each sense S of a word W, first, (Step (4)) Q_s is assigned to $P_s \times M_{s,s}$ (*i.e.*, some proportion, $M_{s,s}$, of the probability of P_s is propagated to the node s), then (Step (4a)) Q_s is incremented by $P_{s'} \times M_{s',s}$, the ingoing probability propagation from node s', whenever there is an edge between s' and s. In Step (4b), we re-calculate the probability of the sense S, P_s, by dividing Q_s by $\sum_{s' \in sense(W)} Q_{S'}$, where S and S' are different word senses of the same word W and $sense(W)$ is the set of admissible senses of W in SI for the next iteration. $\sum_{s' \in sense(W)} Q_{S'}$, or N_w in the algorithm, is the normalization factor. The propagation of probabilities at each iteration in this graph-based algorithm, or Random Walk Algorithm, ensures that if a node is *semantically*[1] linked to another node with high probability, it will obtain quite a few probabilities from that node, indicating that this node may be important[2] in that probabilities converse and tend to aggregate in senses (*i.e.*, nodes) of words that are semantically related (*i.e.*, connected).

Finally, for each word, we identify the most probable sense and attach the sense to it (Step (6)). For instance, for the graph in Figure 6, the vertex on the vertical axis represented as the *sense #3* of "fine" will be selected as the best sense for "fine" under the thesaurus category "*Goodness*" with other entry words, such as, "lovely," "superb," "beautiful," and "splendid". The output of this stage is a set of linked word sense pairs (W, S^*) that can be utilized to extend the coverage of thesauri via semantic relations in SI.

[1] Edges only exist when there is a semantic relation between vertices, or senses.

[2] As probable.

Theoretically, the method of PageRank (Brin & Page, 1998) distributes more probabilities or more scores through edges to well-connected nodes (*i.e.*, well-known web pages) in a network (*i.e.*, the Web). That is, more connected nodes tend to collect scores, in turn propagating comparatively more significant scores to their connected neighboring nodes. Consequently, the flow or re-distribution of probabilities or scores mostly would be confined to nodes in groups and the convergence of the probabilities over the network is to be expected normally. In this stage of our method, an edge is added if and only if there are some semantic relations, in the sense inventory, existing between two word senses (*e.g.*, one is the immediate hyponym/hypernym of the other), to differentiate semantically-related senses from those that are not. The PageRank-like algorithm in Figure 4 is exploited to determine the most well-connected or more semantically related (sense) group. Additionally, the senses in the group are assumed to be the most suitable senses of words for the given semantic category or semantic topic. This assumption is more likely to be correct if the number of given words in a category is big enough (it is usually easier to uniquely determine the sense of words given more words). Moreover, empirically, the number of iterations needed for probabilities to converge is less than ten (Usually, six is enough. It took only three iterations for words in Figure 6 to converge.); a quick scan of the results of this sense-assigning step reveals that the aforementioned assumption leads to satisfying sense analyses.

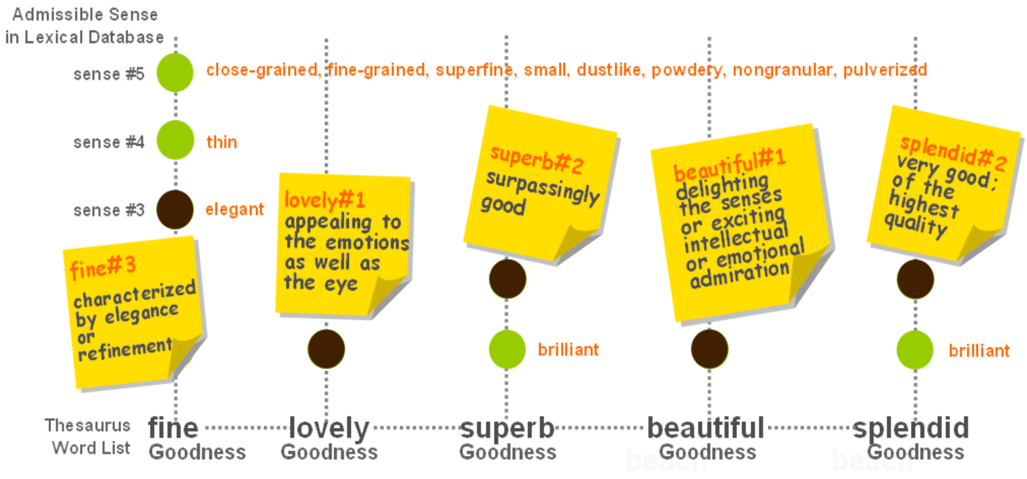

Admissible Sense
in Lexical Database
sense #5
close-grained, fine-grained, superfine, small, dustlike, powdery, nongranular, pulverized
sense #4
thin
sense #3
elegant
fine#3
characterized by elegance or refinement
lovely#1
appealing to the emotions as well as the eye
superb#2
surpassingly good
brilliant
beautiful#1
delighting the senses or exciting intellectual or emotional admiration
splendid#2
very good; of the highest quality
brilliant
Thesaurus
Word List
fine
Goodness
lovely
Goodness
superb
Goodness
beautiful
Goodness
splendid
Goodness

Figure 6. Highest scoring word sense in the stationary distributions for thesaurus word list under category "Goodness" assigned automatically by Random Walk on graph.

3.2.2 Extending the Coverage of Thesaurus

Automating the task of constructing a large-scale semantic knowledge base for semantic classification imposes a huge effort on the side of knowledge integration. Starting from a widespread computational lexical database, such as *WordNet*, overcomes the difficulties of building a knowledge base from scratch. In the second stage of the learning process (Step (2) in Figure 3), we attempt to broaden the limited thesaurus coverage in view of reducing encounters of unknown words in collocation label assignment in Section 3.3. The sense-annotated word lists generated as a result of the previous step are useful for enlarging and enriching the vocabulary of the thesaurus.

Take the sense-annotated result in Figure 6 for example. “Fine” with other adjective entries “beautiful,” “lovely,” “splendid,” and “superb” under the semantic label “*Goodness*” is identified as belonging to the word sense *fine#3* “*characterized by elegance or refinement or accomplishment*” rather than other admissible senses (as shown in Table 1). After knowing the sense of the word “fine” under the semantic category “*Goodness*,” we may now add its similar words via feasible semantic operators (as shown in Table 2) provided in the word sense inventory (*e.g.*, *WordNet*). Its similar word, as suggested in Table 1 and 2, elegant#1 can be acquired by applying the operator “syn operator” on fine#3. Then, elegant#1 is incorporated into the knowledge base (*e.g.*, *ISK*) under the semantic category of fine#3, “*Goodness*”.

Table 1. Admissible senses for adjective “fine”

Sense Number	Definition	Example	Synsets of Synonym
fine #1	(being satisfactory or in satisfactory condition)	*“an all-right movie”; “everything’s fine”; “the passengers were shaken up but are all right”; “dinner and the movies had been fine”; “things are okay”*	all right#1, o.k.#1, ok#1, okay#1, hunky-dory#1
fine #3	(characterized by elegance or refinement or accomplishment)	*“fine wine” ; “a fine gentleman”; “looking fine in her Easter suit”; “fine china and crystal”; “a fine violinist”*	elegant#1
fine #4	(thin in thickness or diameter)	*“a fine film of oil”; “fine hairs”; “read the fine print”*	thin#1

Table 2. Semantic relation operators for extending the coverage of thesaurus.

semantic relation operators	Description	Relations Hold for
syn operator	synonym sets for every word that are interchangeable in some context without changing the truth value of the preposition in which they are embedded	all words
sim operator	adjective synsets contained in adjective clusters	adjectives

In the end, by using semantic operators in lexical database (*e.g.*, *WordNet*), the coverage of the integrated semantic knowledge obtained from Step (1) in Figure 3 can be enlarged for assigning the semantic label of a collocation at run-time (Section 3.3).

3.2 Giving Thesaurus Structure to Collocation by Iterative Graphical Algorithms

Provided with the extended semantic knowledge obtained by following the learning process in Section 3.2, we build a thesaurus structure for the query results from online collocation reference tools. Figure 7 illustrates a thesaurus structure imposed on some adjective collocations (*i.e.*, "superb," "fine," "lovely," "beautiful," "splendid," *etc*.) of the word "beach" by our system.

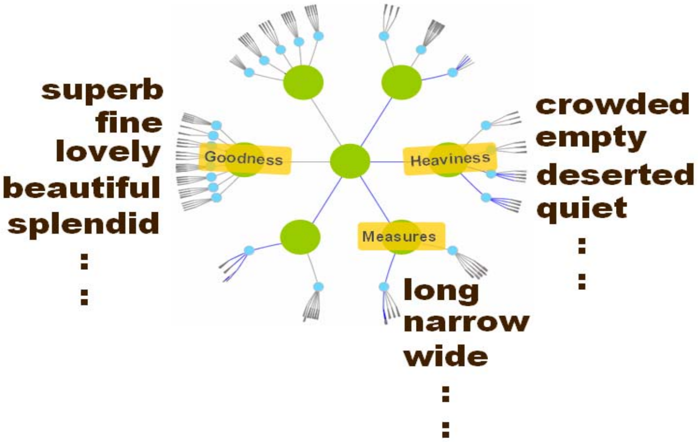


Figure 7. Sample adjective collocations of the word "beach" after being classified into some general-purpose semantic topics.

At run-time, we apply the Random Walk algorithm, which is very similar to the one in Figure 4, to automatically assign semantic labels to all collocations of a pivot word (*e.g.*, "beach") by exploiting semantic relatedness identified among these collocations. Once we know the semantic labels, or thesaurus categories, of the collocates, we partition them in groups according to their labels, which is helpful for dictionary look-up and for L2 learners to quickly find their desired collocations under some semantic meaning. The following depicts the semantic labeling procedure.

The input to this procedure is (1) a set of collocations, *Col*, for the query word *X*; (2) the integrated semantic knowledge (*i.e.*, *ISK*) from Section 3.2, {(*W*, *L*)} where a word *W* is semantically labeled as *L*. The output of this procedure is sets of collocations, each of which is classified under a semantic label and contains semantically-related collocations of the query word (see Figure 7).

At first, we construct a graph G={V,E} where a vertex in V represents a possible semantic category for a collocation C in *Col* and an edge in E represents a semantic relatedness holding between vertices. Note that we can look up possible semantic labels of a word from *ISK* and that edges in G are directed.

We use P_L to depict the probability of the candidate label, L, of a collocation in *Col*. Prior to the random-walking process, P_L is uniformly initialized over possible labels of a collocation. Once the matrix M, representing the proportions of probabilities to be propagated, is built, P_L will be iteratively changed, based upon current statistics, until convergence of probabilities. Recall that an element $M_{x,y}$ in the matrix will be set to 1-d if node x is equal to node y; will be set to $d/e(x)$ if x is different from y, there is an edge between x and y, and there are $e(x)$ edges leaving x; and will be set to zero otherwise. At each iteration, the probabilities of the candidate labels of a collocate sum to one, suggesting normalization is needed for each iteration as in the algorithm of word sense assignment in Figure 4.

Finally, we identify the most probable semantic label L^* for each collocate C, resulting in a list of (C, L^*). The procedure is designed to arrange given collocations in thesaurus categories with semantically related collocations therein, providing L2 learners with a thesaurus index for easy lookup or easy concept-grasping (see Figure 7 for an example).

4. Experimental Setting

4.1 Experimental Data

In our experiment, we applied the Random Walk Algorithm (in Section 3.2 and Section 3.3) to partition collocations into existing thesaurus categories, thus imposing a semantic structure on the raw data (*i.e.*, given collocations). In analysis of learners' collocation error patterns, verb-noun (V-N) and adjective-noun (A-N) collocations were found to be the most frequent error patterns (Liu, 2002; Chen, 2002). Hence, for our experiments and evaluation, we focused our attention particularly on V-N and A-N collocations.

Recall that our classification model starts with a thesaurus consisting of lists of semantically related words and extends the thesaurus using sense labeling in Section 3.2.1 and semantic operators in the word sense inventory in Section 3.2.2. The extended semantic knowledge provides collocates with topic labels for semantic classification of interest. Two kinds of resources required in our experiment to obtain the extended knowledge base are described below.

4.1.1 Data Source 1: A Thesaurus

We used *Longman Lexicon of Contemporary English* (*LLOCE* for short) as our thesaurus of semantic categories (*i.e.*, *TC*). *LLOCE* contains 15,000 distinct entries for all open-class words, providing semantic fields of a pragmatic, everyday common sense index for easy reference. The words in *LLOCE* are organized into approximately 2,500 semantic word sets. These sets are divided into 129 semantic categories and further organized as 14 semantic fields. Thus, the semantic field, category, and word set in *LLOCE* constitute a three-level hierarchy, in which each semantic field contains 7 to 12 categories and each category contains 10 to 50 sets of semantic related words. The *LLOCE* is based on coarse, topical semantic classes, making them more appropriate for WSD than other finer-grained lexica. Alternatively, *Roget's Thesaurus* can be used as the thesaurus.

4.1.2 Data Source 2: A Word Sense Inventory

For our experiments, we need comprehensive coverage of word senses. Word senses can be obtained easily from any definitive record of the English language (*e.g*. an English dictionary, encyclopedia or thesaurus). We used *WordNet 3.0* as our sense inventory. It is a broad-coverage, machine-readable lexical database, publicly available in parsed form (Fellbaum, 1998) and consists of 212,557 sense entries for open-class words, including nouns, verbs, adjectives, and adverbs. *WordNet* is organized by the synonymous sets, or synsets, and provides semantic operators to act upon its synsets.

4.2 Experimental Configurations

Given the aforementioned two data sources, we first integrate them into one then broaden the vocabulary of the thesaurus, the basis knowledge for assigning semantic labels to collocations.

4.2.1 Step 1: Integrating Semantic Knowledge

For each semantic topic in *LLOCE*, we attach word senses to its constituent words based on semantic coherence (within a topic) and semantic relations created by lexicographers from *WordNet*. The integrated semantic knowledge can help interpret a word by providing information on its word sense and its corresponding semantic label.

Recall that, to incorporate senses into words with semantic topics, our model applies the Random Walk Algorithm on a weighted directed graph whose vertices (word senses) and edges (semantic relations) are extracted from and are based on *LLOCE* and *WordNet 3.0*. All edges are drawn and weighted to represent the magnitudes of semantic relatedness among word senses. See Table 3 for the relations (or semantic operators) existing in edges in our experiment.

Table 3. Available semantic relations.

Relations	Semantic Relations for Word Meanings	Relations Hold for
syn	synonym sets for every word that are interchangeable in some context without changing the truth value of the preposition in which they are embedded	all words
hyp	hypernym/hyponym (superordinate/subordinate) relations between synonym sets	nouns verbs
vgp	verb synsets that are similar in meaning and should be grouped together when displayed in response to a grouped synset search.	verbs
sim	adjective synsets contained in adjective clusters	adjectives
der	words that have the same root form and are semantically related	all words

4.2.2 Step 2: Extending Semantic Knowledge

Based on the senses mapped to words with semantic labels (via the graph-based sense assignment algorithm), we further utilize the semantic operators in *WordNet* (*i.e.*, *SI*) to add new words into *LLOCE* (*i.e.*, *TC*). Depending on the part-of-speech (*i.e.*, noun, adjective, or verb) of the word at hand, various kinds of semantic relation operators (see Table 3) are available for enriching the vocabulary of the integrated semantic knowledge (*i.e.*, *ISK*) of *WordNet* and *LLOCE*. In the experiment, using the *syn* operator alone broadened the vocabulary size of *ISK* to a size more than twice as large as that of the thesaurus *LLOCE* (*i.e.*, 39,000 vs. 15,000).

4.3 Test Data

We used a collection of 859 V-N and A-N collocation pairs for testing. These collocations were obtained from the website: *JustTheWord* (http://193.133.140.102/JustTheWord/). *JustTheWord* clusters collocates into sets without any explicit semantic label. We will compare its clustering performance with our model's performance in Section 5.

In the experiment, we evaluated semantic classification of three[3] types of collocation pairs: (1) A-N pairs and clustering over the **adjectives** (**A**-N), (2) V-N pairs and clustering over the **verbs** (**V**-N), and (3) V-N pairs and clustering over the **nouns** (V-**N**). For each type, we selected five pivot words with varying levels of abstractness for L2 learners and extracted a subset of their respective collocations from *JustTheWord*, leading to a test data set of 859 collocation pairs. Table 4 shows the number of the collocations for each pivot of each collocation type. In total, 307 collocates were extracted for **A**-N, 184 for **V**-N, and 368 for

[3] We do not consider the case of A-N in that, usually, various nouns can follow an adjective.

V-**N**.

To appropriately select our testing pairs from *JustTheWord*, we were guided by research into L2 learners' and dictionary users' needs and skills for second language learning, especially taking account the meanings of complex words with many collocates (Tono, 1992; Rundell, 2002). The pivot words we selected for testing are words that have many respective collocations and are shown in worth-noting boxes in *Macmillan English Dictionary for Advance Learners* [ISBN 0-333-95786-5] (First edition, henceforth *MEDAL*).

Table 4. Statistics of our testing collocation pairs.

collocation type	pivot word	some collocations	count
A-N (N=pivot)	advice	helpful, dietary, impartial, free	36
	attitude	healthy, moral, aggressive, right	49
	description	clinical, excellent, fair, precise	47
	effect	serious, inevitable, possible, sound	114
	impact	dramatic, negative, powerful, severe	61
V-N (N=pivot)	balance	strike, maintain, achieve, tilt, tip	29
	disease	cure, combat, carry, transmit, carry	21
	issue	settle, clarify, identify, remain, avoid	38
	plan	propose, submit, accept, involve	54
	relationship	forge, alter, develop, damage, form	42
V-**N** (V=pivot)	deserve	blame, support, title, thanks, honor	51
	express	love, anger, fear, personality, doubt	82
	fight	disease, war, , enemy, cancer, duel	24
	hold	funeral, presidency, hope, knife	151
	influence	health, government, opinion, price	60

5. Results and Discussions

Two pertinent sides were addressed for the evaluation of our results. The first was whether such a model for a thesaurus-based semantic classification could generate collocation clusters correlating with human word meaning similarities to a significant extent. Second, supposing it could, would its results of semantic label assignment lead to easy dictionary lookup or better collocation understanding and production? In the following sections, two evaluation metrics are described to respectively examine our results in these two aspects, that is, the accuracy of our collocation clusters and the helpfulness of our labels in terms of language learning.

5.1 Performance Evaluation for Semantic Clusters

Traditional cluster evaluation (Salton, 1989) might not be suited to assess our model, where we aim to facilitate collocation referencing and help learners improve their collocation production. Hence, to evaluate the performance of our clustering results, an evaluation sheet made up of test items, resembling synonym test items of the Test of English as a Foreign Language (TOEFL), was automatically generated for human judgment. Landauer and Dumais (1997) first proposed using the synonym test items of TOEFL as an evaluation method for semantic similarity. Fewer fully automatic methods of a knowledge acquisition evaluation, *i.e.*, ones that do not depend on knowledge being entered by a human, have been capable of performing well on a full scale test used to measure semantic similarity. A test item provided by Landauer (1997, as cited in Padó & Lapata, 2007) is shown below where “crossroads” is the synonym for “intersection” in the context.

You will find the office at the main **intersection**.

(a) place (b) crossroads (c) roundabout (d) building

As to our experiment, we evaluated the semantic relatedness among collocation clusters according to the above-mentioned TOEFL benchmark by setting up test items out of our clustering results. Then, human judges performed a decision task similar to TOEFL test takers: deciding which one of the four alternatives was synonymous with the target word. A sample question is shown below where “rocky” is clearly the most similar word for “sandy” given the pivot word “beach”.

***sand*y** beach

(a) long (b) rocky (c)super (d)narrow

There were 150 multiple-choice questions randomly constructed to test the accuracy of our clusters, 50 questions for each collocation types (*i.e.*, **A**-N, **V**-N, and V-**N**) and 10 for each of collocation pairs. In order to evaluate the degree to which our model achieved production of good clusters, two judges were asked to choose the most appropriate answer. More than one answer was allowed if the judges found some of the distractors in the test items to be plausible answers. Moreover, the judges were allowed not to choose any of the alternatives given if they thought no satisfactory answer was provided. Table 5 shows the performance of collocation clusters generated by *JustTheWord* and the proposed system. As suggested in the table, our model achieved significantly higher precision and recall in comparison with our baseline, *JustTheWord*.

Table 5. Precision and recall of two systems

Results	**Judge 1**		**Judge 2**		**Inter-Judge Agreement**
System	**Precision**	**Recall**	**Precision**	**Recall**	
Ours	.79	.71	.73	.67	.82
JustTheWord	.57	.58	.57	.59	

With high inter-judge agreement (*i.e.*, 0.82), the influence of human judges' subjectivity on the performance evaluation of collocation clusters is not that severe and it is modest to say that our model's clustering results are thought to be better than the baseline's across human judges.

5.2 Conformity of Semantic Labels

The second evaluation task focused on whether the semantic labels would facilitate users scanning the collocation entries quickly and finding the desired concept of the collocations. The evaluation is aimed at examining the extent to which semantic labels are useful, and to what degree of reliability.

Two native speakers were asked to grade half of the labeled collocations randomly selected from our classifying results (all test data considered). A three-point rubric is used to evaluate the effectiveness, or usefulness, of the given semantic labels in terms of navigating users to the desired collocates. The three types of rubric points with their descriptions are: three points for those collocations with effective semantic labels in navigation in a collocation reference tool, two points for those with somewhat helpful assigned labels, and one point for those with misleading labels.

Table 5 shows that 77% of the semantic labels assigned as a reference guide have been judged as adequate in terms of guiding a user finding a desired collocation in a collocation learning tool and that our classification model provably yields productive performance of semantic labeling of collocates to be used to assist language learners. The results justify the thought that the move towards semantic classification of collocations is of probative value.

Table 6 shows that 76% of the semantic labels assigned as a reference guide were judged adequate in terms of guiding users to find a desired collocation in a collocation learning tool, and this suggests that our classification model yielded promising performance in semantically labeling collocates further to be used to assist language learners. The results justify that the move towards semantic classification of collocations is of probative value.

Table 6. Performance evaluation for assigning semantic labels as a reference guide

	Judge 1	**Judge 2**
Ours	.77	.75
JustTheWord	Not available	Not available

6. Conclusion and Future Work

The research seeks to create a thesaurus-based semantic classifier within a collocation reference tool without meaning access indices. We describe a thesaurus-based semantic classification for a semantic grouping of collocates with a pivot word. The construction of a collocation thesaurus is meant to enhance L2 learners' collocation production. Our classification model is based on two graph-based Random Walk Algorithms (*i.e.*, word sense assignment and semantic label assignment) to categorize collocations into semantically-related groups for easy dictionary lookup and collocation understanding and production. The limited vocabulary size of the semantic thesaurus is dealt with using the sense information and the semantic operators in the word sense inventory, *WordNet*. The evaluation shows that the thesaurus structure imposed by our model for an existing computational collocation reference tool is quite accurate and is helpful for users to navigate the collocations of a pivot word.

Many avenues exist for future research and improvement of our system. For example, semantic relations existing between word senses may take on different weights in that some may be more informative than others in determining semantic similarities. Another interesting direction to explore is to see if our model can benefit from other thesauri with semantic labels.

References

Benson, M. (1985). Collocations and Idioms. In *R. Ilson (Ed.), Dictionaries, Lexicography and Language Learning* (ELT Documents 120; Oxford: Pergamon), 61-68.

Béjoint, H. (1994). Tradition and Innovation in Modern English Dictionaries. Oxford: *Clarendon Press*.

Brants, T. & Franz, A. (2006). Web 1T 5-gram corpus version 1.1. Technical report, Google Research.

Brin, S. & Page, L. (1998). The anatomy of a large-scale hypertextual Web search engine. In *Proceedings of WWW Conference*.

Chen, Y. (2004). A corpus-based analysis of collocational errors in EFL Taiwanese High School students' compositions. California State University, San Bernardino. June.

Pantel, P. & Chklovski, T. (2004). VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. In *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 33-40.

Chen, J.-N. & Chang, J. S. (1998). Topical clustering of MRD senses based on information retrieval techniques, *Computational Linguistics*, 24(1), March 1998.

Downing, S. M., Baranowski, R. A. , Grosso, L.J., & Norcini, J. J. (1995). Item type and cognitive ability measured: the validity evidence for multiple true-false items in medical specialty certification. *Appl Meas Educ*, 8, 189-199.

Deerwester, S., Dumais, S. T., Furnas, G. W., Landauer, T. K., & Harshman, R. (1990). Indexing by latent semantic analysis. *Journal of the American Society for Information Science* (*JASIS*), 41(6), 391-407.

Firth, J. R. (1957). The Semantics of Linguistics Science. Papers in linguistics 1934-1951. London: *Oxford University Press*.

Fellbaum, C. (1998). WordNet: An Electronic Lexical Database. MIT Press, Cambridge, MA.

Hall, G. (1994). Review of The Lexical Approach: The State of ELT and a Way Forward, by Michael Lewis. *ELT Journal,* 44, 48.

Heimlich, J. E. & Pittelman, S. D. (1986). Semantic mapping: Classroom applications. Newark, DE: International Reading Association.

Hindle, D. (1990). Noun classification from predicate-argument structures. In *Meeting of the Association for Computational Linguistics*, 268-275.

Hatzivassiloglou, V., & McKeown, K. R. (1993). Towards the automatic identification of adjectival scales: Clustering adjectives according to meaning. In *Proceedings of the 31st ACL*, 172-182.

Jian, J.-Y., Chang, Y.-C., & Chang, J. S. (2004). TANGO: Bilingual Collocational Concordancer, *Post & demo in ACL* 2004, Barcelona.

Johnson, D. D., & Pearson, P. D. (1984). *Teaching reading vocabulary*. New York: Holt, Rinehart & Winston.

Kilgarriff, A. (1997). I Don’t Believe in Word Senses, In: *Computers and the Humanities*. 31(2), 91-113(23).

Kilgarriff, A. & Tugwell, D. (2001). “WORD SKETCH: Extraction and Display of Significant Collocations for Lexicography”. In *Proceedings of COLLOCATION: Computational Extraction, Analysis and Exploitation workshop*, 32-38.

Kemp, J. E., Morrison, G. R., & Ross, S. M. (1994). Developing evaluation instruments. In: *Designing Effective Instruction*. New York, NY: MacMillan College Publishing Company, 180-213.

Lewis, M. (1997). Implementing the lexical approach. Hove, England: *Language Teaching Publications*.

Lewis, M. (2000). Language in the Lexical Approach. In. M. Lewis (ed.) Teaching Collocation: Further development in the Lexical Approach. London, Language Teaching Publications.

Landauer, T. & Dumais, S. T. (1997). A solution to Plato’s problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review*, 104(2), 211-240.

Lesk, M. E. (1986). Automatic sense disambiguation using machine readable dictionaries: How to tell a pine cone from an ice cream cone. In *Proceedings of ACM SIGDOC ’86*, 24-26.

Lin, D. (1997). Using syntactic dependency as local context to resolve word sense ambiguity. In *Meeting of the Association for Computational Linguistics*, 64-71.

Liu, L. E. (2002). A corpus-based lexical semantic investigation of vernb-noun miscollocations in Taiwan learners' English. *Tamkang University*, Taipei, January.

Morris, J. & Hirst, G. (1991). Lexical cohesion computed by thesaural relations as an indicator of the structure of text. *Computational Linguistics*, 17(1), 21-48.

Nattinger, J. R., & DeCarrico, J. S. (1992). Lexical Phrases and Language Learning. Oxford: *Oxford University Press*.

Nesselhauf, N. (2003). The use of collocations by advanced learners of English and some implications for teaching. *Applied Linguistics*, 24(2), 223-242.

Nation, I. S. P. (2001). Learning vocabulary in another language. Cambridge: Cambridge Press.

Nirenburg, S. & Raskin, V. (1987). The subworld concept lexicon and the lexicon management system, *Computational Linguistics*, v. 13, December 1987.

Nastase, V. & Szpakowicz, S. (2003). Exploring noun–modifier semantic relations. In *Fifth International Workshop on Computational Semantics* (*IWCS-5*), 285-301.

Padó, S. & Lapata, M. (2007). Dependency-Based Construction of Semantic Space Models. *Computational Linguistics*, 33(2), 161-199.

Roediger, H. L., III, & Marsh, E. J. (2005). The positive and negative consequence of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 1155-159.

Readence, J. E., & Searfoss, L.W. (1986). Teaching strategies for vocabulary development. In E. K. Dishner, T. W. Bean, J. E. Readence, & D. W. Moore (Eds.), *Reading in the content areas: Improving classroom instruction* (2nd ed., pp. 183-188). Dubuque, IA: Kendall/ Hunt.

Rehder, B., Schreiner, M. E., Wolfe, M. B. W., Laham, D., Landauer, T. K., & Kintsch, W. (1998). Using latent semantic analysis to assess knowledge: Some technical considerations. *Discourse Processes*, 25, 337-354.

Scholfield, P. (1982). Using the English dictionary for comprehension. *TESOL Quarterly*, 16, 185-194.

Salton, G. (1989). Automatic Text Processing: The transformation, analysis, and retrieval of information by computer. *Addidon-Wesley*.

Sinatra, R., Beaudry, I., Pizzo, I., & Geishart, G. (1994). Using a computer-based semantic mapping, reading and writing approach with at-risk fourth graders. *Journal of Computing in Childhood Education*, 5, 93-112.

Tono, Y. (1984). On the Dictionary User's Reference Skills. Unpublished B.Ed. Thesis. Tokyo: *Tokyo Gakugei University*.

Tono, Y. (1992). The Effect of Menus on EFL Learners' Look-up Processes. LEXICOS 2 (AFRILEX Series) Stellenbosch: *Buro Van de Watt*.

Taba, H. (1967). Teacher's handbook for elementary social studies. Reading, MA: Addison-Wesley.

Turney, P. D. (2002). Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, 417-424.

Turney, P. D. (2006). Similarity of Semantic Relations. *Computational Linguistics*, 32(3), 379-416.