The Relationship between Semantic Similarity and Subcategorization Frames in English: A Stochastic Test Using ICE-GB and WordNet*

Sanghoun Song, Jae-Woong Choe

Dept. of Linguistics, Korea University,
Anam-dong, Sungbuk-gu, Seoul, KOREA
{yooseon21, jchoe}@korea.ac.kr

Abstract. In this paper, we test a working hypothesis that there is a significant relationship between semantic similarity and subcategorization frames in English. This paper is under the assumption that if a group of verbs form a cluster sharing a similar meaning, they tend to share subcategorization frames. In the process, we propose a statistical method to test this assumption, making use of two language resources, namely, ICE-GB and WordNet. We come to the conclusion that the proposed hypothesis holds true to the degrees of 42~73%, based on our stochastic analysis.

Keywords: semantic similarity, subcategorization frames, ICE-GB, WordNet, statistical method, clustering, dendrogram, selectional preference strength

1. Introduction

The following examples taken from Lasnik and Uriagereka (2005) bring out an issue concerning the relationship between semantic similarity of verbs and the corresponding subcategorization frames.

(1) a. She asked what time it was.
b. She asked the time.
c. She wondered what time it was.
b. *She wondered the time.

Sentences in (1) show that verbs like *ask* and *wonder*, which are similar in meaning, show different ways of argument realization. Although it is true that arbitrariness is at the core of the form-meaning mapping in any language, there are other aspects of language where formal or grammatical realization seems to be largely determined by meaning. Levin (1993) provides many

* Acknowledgments: We owe special thanks to Jieun Jeon for her assistance and partial participation in the preparation of this paper, especially for extracting relevant data from ICE-GB. We also would like to express our thanks to Prof. Seok-Hoon You for his support for our project. Finally, comments from three anonymous reviewers for PACLIC 22 are much appreciated.

examples of semantic classes of English verbs that lead to the same or similar pattern of argument realizations. For example, the *cut* verbs (e.g. *chip*, *cut*, *scrape*, *snip*, etc.), a semantically identifiable group of words in the sense that they all involve notions of motion, contact, and effect, do not allow causative alternations. That is, their subcategorization frames are affected by their meaning.

(2) a. Carol cut the bread.
b. *The bread cut.

The question we raise in this paper is if the meaning restriction applies in a significant way to the realization of argument structure, if it does, to what extent it holds. In other words, how much is a syntactic pattern tied up with the semantic properties of the English verbs? In order to test this research question in a rather comprehensive way, we make use of large sets of language resources, in particular, ICE-GB and WordNet. In the process, we propose a more articulated methodology to take advantage of the language resources statistically and computationally. Also introduced in this paper are a software package to measure similarity of lexicons, an algorithm to cluster words in natural languages, and a kind of stochastic model to calculate selectional preference strength.

2. Methodology

There are two main methods used in this study; one is clustering, and the other is the use of frequency value.

In order to measure semantic relatedness between verbs, clustering method was used to divide verbs into various subsets in accordance with semantic similarity. As a way to ensure consistency in defining semantic similarity, WordNet (ver. 2.1) has been adopted as the basis for similarity measurement. The measurement itself was carried out by a module called WordNet::Similarity[1] that calculates similarity between words in WordNet (Pedersen et al. 2004). The results from the module, then, went through a clustering process on the basis of the hierarchical bottom-up clustering algorithm, adopted from Manning and Schütze (1999).

This study also required a method that can calculate the selectional preference strength of subcategorization frames. Selectional preference, in this paper, refers to the degree of relationship between a lexicon and its relevant items. In other words, this study seeks to figure out not merely whether a subcategorization frame can be realized or not, but how much relevance a verb has with the subcategorization frame.

3. Data

The initial set of English verbs for this study was extracted from the Collins Cobuild English Dictionary (henceforth CCED), which divides English words into five level groups according to each word's frequency. CCED includes approximately 1,600 words that belong to the highest or the second highest level groups. We then compared the list with the verb list in WordNet, and only those that appear in both lists were selected, resulting in 799 verbs. Starting with these 799 verbs, we tried to build up the hierarchical cluster and as a result obtained 61 meaningful clusters that cover 157 verbs out of 799. The next section provides an explanation of the whole procedure in detail.

4. The Analysis

4.1.Semantic Similarity

[1] The WordNet::Similarity module coded in the Perl language is freely available on the Internet (http://www.d.umn.edu/~tpederse/similarity.html).

There are three steps in the clustering process. The first step measures semantic similarity between the initial 799 verbs, using the WordNet::Similarity module and the Parse-Tree algorithm. The second step is to draw a dendrogram that indicates the hierarchy of semantic similarity, on the basis of the matrix algorithm. The final step distinguishes significant clusters from insignificant ones, after finding a critical value on the basis of the Z-score.

4.1.1. WordNet::Similarity

The WordNet::Similarity module offers several algorithms to measure similarity between words. Among them, the Lesk algorithm was adopted in this study, which measures incorporate information from WordNet glosses. That is, this algorithm 'finds overlaps between the glosses of concepts A and B, as well as concepts that are directly linked to A and B (Pedersen et al. 2004)', as shown in the following formula (Banerjee and Pedersen 2003).

$$\begin{aligned} realatedness(A,B) = {} & score(gloss(A), gloss(B)) \\ & + score(hype(A), hype(B)) + score(hypo(A), hypo(B)) \\ & + score(hype(A), gloss(B)) + score(gloss(A), hype(B)) \end{aligned} \quad (3)$$

The reason we chose the Lesk algorithm in this study is that the resulting value of the Lesk algorithm is relatively bigger than those of the others, which is of great advantage to clustering. The process of measuring is given below, where sim(A,B) is equal to sim(B,A).

```
V = {v1,v2,…,vn}
S = {sim(v1,v2),sim(v2,v3),…,sim(vn-1,vn)}
for i = 1 to n-1:
  for j = i+1 to n:
    sim(vi,vj)=lesk(vi,vj)
```

4.1.2. A Dendrogram

The above algorithm gives a similarity matrix as below.

	v_2	**v_3**	**v_4**	**…**	**v_n**
v_1	$sim(v_1,v_2)$	$sim(v_1,v_3)$	$sim(v_1,v_4)$	…	$sim(v_1,v_n)$
v_2		$sim(v_2,v_3)$	$sim(v_2,v_4)$	…	$sim(v_2,v_n)$
v_3			$sim(v_3,v_4)$	…	$sim(v_3,v_n)$
…				…	$sim(v\ldots,v_n)$
v_{n-1}					$sim(v_{n-1},v_n)$

Figure 1: A similarity matrix.

Using this matrix, this section shows the process to draw a dendrogram that represents the hierarchy of semantic similarity. Since a dendrogram is a kind of tree diagram to illustrate the arrangement of clusters, it can be built up by the so-called Parse-Tree algorithm. Data structure of the Parse-Tree algorithm is composed of three elements; the mother node, the left daughter node, and the right daughter node. Each node in the tree diagram is connected with another, forming an operator-operand relation. For example, a node can be both an operator of its daughter nodes and an operand of its mother node.

This is how the Parse-Tree algorithm is applied to draw a dendrogram: A set *S* is given as below, its elements being defined on the basis of a relation °.

```
S = {a°b=7, a°c=8, a°d=15, b°c=12, b°d=100, c°d=3}
```

The dendrogram will be produced by a similarity matrix as follows.

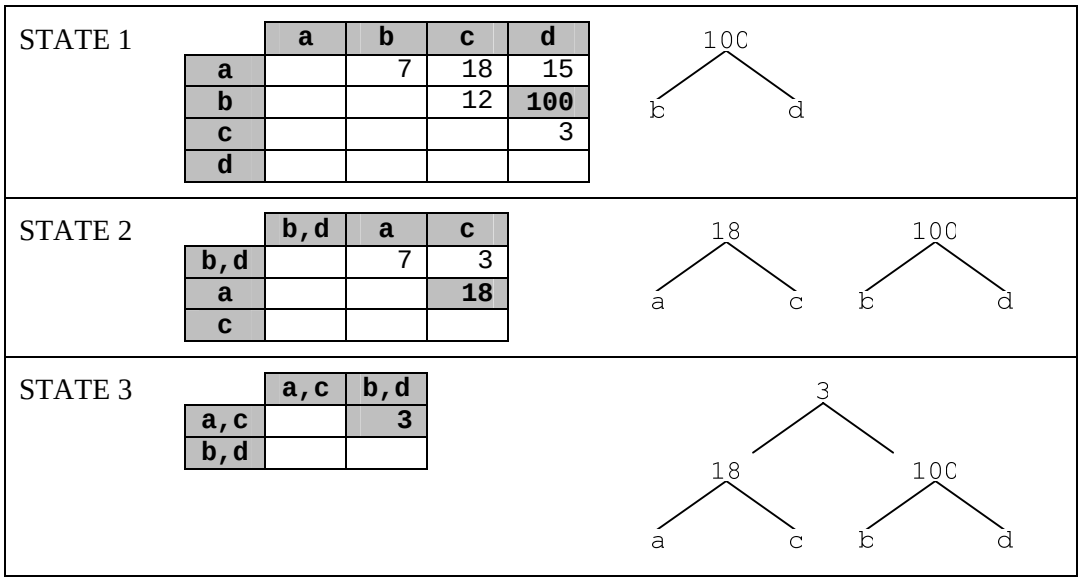

Figure 2: The procedure of drawing a dendrogram.

STATE 1 in Figure 2 is the initial matrix. The maximum value on the matrix, which is 100 at this state, will be selected, and since the value is given by the relation between **b** and **d,** they would form an initial dendrogram. Now that the relational value between **b** and **d** has been defined, those two elements are treated as a single element in the following process. Thus, in STATE 2, the pair, **b** and **d**, moves to the first place on the matrix, and each cell will be filled with the minimum value in each relation. For instance, in relation with '$a^{\circ}b=7$' and '$a^{\circ}d=15$', 7 will be allotted into the first cell in STATE 2. After all the relevant cell values have been filled up, the maximum value 18 from the matrix will be selected, introducing a new pair node **a** and **c**. In the same way, STATE 3 builds up the final dendrogram.

The algorithm to draw a dendrogram in the above is as follows.

$V_i = \{v_i\}$
$M = \{v_1, v_2, \ldots, v_{n-1}, v_n\}$
while n > 1:

$V_1 = V_{pos\_i} \oplus V_{pos\_j}$

```
for i = 1 to n-1:
    for j = i+1 to n:
        sim(Vi,Vj) = min(Vi,Vj)
        if max < sim(Vi,Vj):
            max = sim(Vi,Vj)
            pos_i = i
            pos_j = j
parse_tree(pos_i, pos_j)
n = n – 1
```

Repeating the process described above, we could draw a dendrogram for the initial 799 verbs.

4.1.3. Clusters

On the basis of the dendrogram found in the previous section, we could distinguish significant clusters from others. To begin with, we set up a critical value for discrimination. That is, if a value of a node is over the critical value, we assume the node and its descendant nodes are of significance.

The critical value in this study is the so-called Z-score. (4) stands for Z-score, in which *x* is for a value of a node, *m* is for the mean value of whole nodes, and *s* is for the standard deviation. Using this formula, we could obtain 61 clusters consisting of 157 verbs in total.

$$\frac{x-m}{s} > 1 \quad (4)$$

4.2. Subcategorization Frames

In order to get basic data which include subcategorization frames of English verbs, we employed the ICECUP III program, a search program for ICE-GB. To begin with, we listed up the inflectional paradigm of the 157 verbs found as a result of the process described in the previous section, such as *take*, *takes*, *took*, *taken*, and *taking*. With this paradigm, we could take advantage of the Text searching function. ICECUP III provides a function to extract concordances from the corpora in the form of set. (e.g. {*ask*, *asks*, *asked*, *asking*}, {*take*, *takes*, *took*, *taken*, *taking*}, or {*cut*, *cuts*, *cutting*}), which makes it easier to extract all the sentences that include at least one of the 157 verbs. We saved the search result as a separate file one by one in the form of treebanks (e.g. 'take.tre'). Plate 1 illustrates the searching and saving processes.

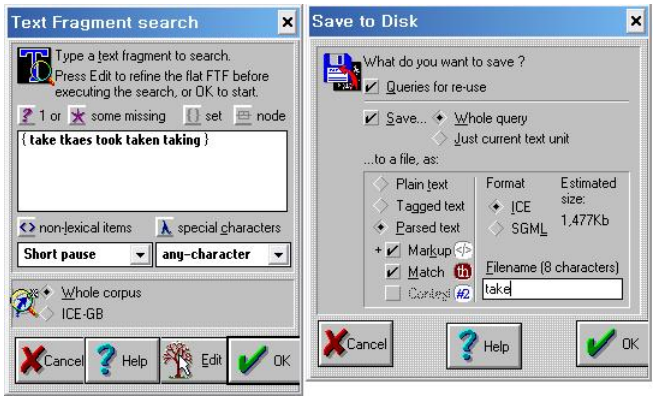

Plate 1: ICECUP III – Text searching / saving

Now that we have 157 files that contain concordances of verbs in the form of treebanks, we can calculate frequency of each verb's subcategorization frames. However, there are some issues and potential problems that need to be settled regarding the subcategorization information from ICE-GB. The next section discusses these issues and shows how they were handled in this study.

4.2.1. Some Issues

4.2.1.1 Subjects

English verbs subcategorize for a subject almost invariably; hence, subjects can be omitted from subcategorization frames. Besides, no subject appears on the surface in some VPs, such as infinitival constructions or gerundives. For these reasons, this study excludes subjects from the subcategorization frames.

4.2.1.2 Relative Clauses

Relative clauses can raise a troublesome issue in terms of extracting subcategorization frames from corpora, because one of the arguments appears outside of the relative clauses. However, the way relative pronouns are tagged in ICE-GB allows us to be able to retrieve the missing arguments in a systematic way. ICE-GB annotates pertinent information to relative pronouns as shown in Plate 2. The relative pronoun 'which' is tagged as 'OD,NP', meaning it is the direct object of the verb 'had',.

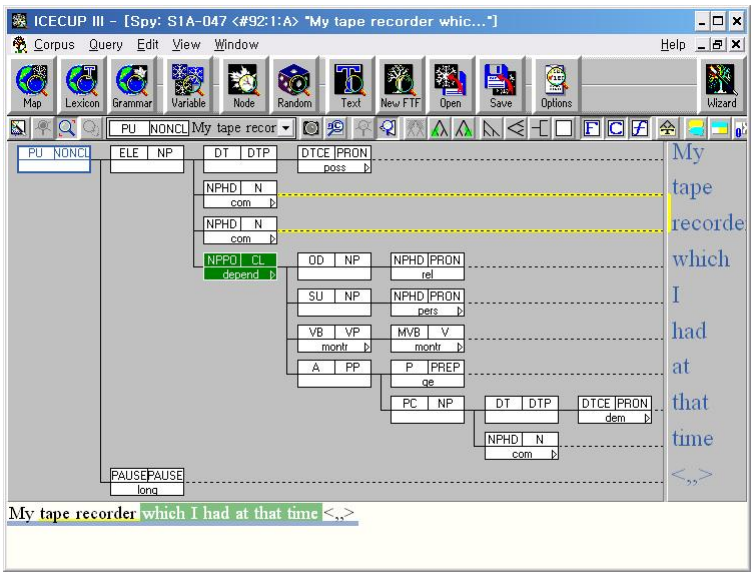

Plate 2: My tape recorder which I had at that time. (S1A-047 #92)

4.2.1.3 Oblique Complements

Let us consider the *Put* class verbs in English (Levin 1993). (5b-c) are ungrammatical because the obligatory arguments are missing. That is, an adverbial phrase 'on the desk' in (5a) functions as an oblique complement.

(5) a. John put the book on the desk.
b. *John put on the desk.
c. *John put the book.

The way ICE-GB deals with this kind of complements helps us to avoid any potential problem concerning this kind of constructions. They are straightforwardly marked as such. For example, in Plate 3, 'on me' is tagged as 'CO,PP'.

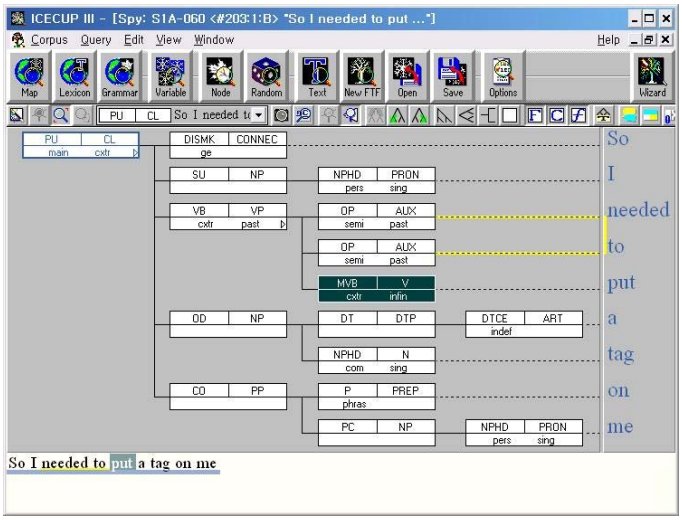

Plate 3: So I needed to put a tag on me. (S1A-060 #203)

4.2.1.4 Passives

The passive construction shows argument alternations in most languages. These alternations should be taken into consideration, because there may be a moved argument or a deleted argument in passive constructions. ICE-GB lays focus on the surface form, and we found it is rather difficult to reconstruct the 'deep, active' counterpart for all the passive constructions found in ICE-GB. Hence, we chose to exclude passive constructions from our consideration. There is also an empirical reason for our decision; some typical transitive verbs tend to be used as passive forms. If we would not get rid of these cases, the result would be biased. For instance, if we include 'be regarded' in our analysis, the most frequent subcategorization pattern of 'regard' will be something other than the most typical 'regard NP as NP'.

4.2.2. Extracting Subcategorization Frames

This section provides an explanation of how to extract subcategorization frames from the result in the form of treebanks. To take an instance, one of the sentences in which 'regard' is used is shown in the following Plate 4, and its parsed text version is as in the below.

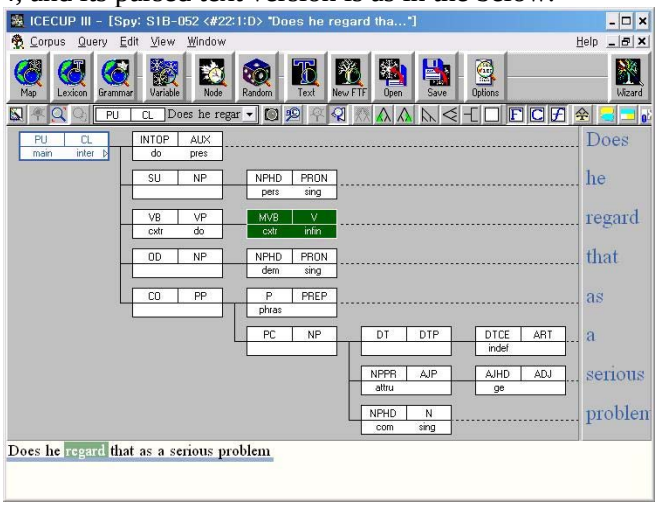

Plate 4: Does he regard that as a serious problem? (S1B-052 #22)

```
<ICE-GB:S1B-052 #022:1:D>
[<#22:1:D> <sent>]
 PU,CL(main,inter,cxtr,pres)
  INTOP,AUX(do,pres) {Does}
  SU,NP
   NPHD,PRON(pers,sing) {he}
  VB,VP(cxtr,do)
*  MVB,V(cxtr,infin) {regard} **
  OD,NP
   NPHD,PRON(dem,sing) {that}
  CO,PP
   P,PREP(phras) {as}
   PC,NP
    DT,DTP
     DTCE,ART(indef) {a}
    NPPR,AJP(attru)
     AJHD,ADJ(ge) {serious}
    NPHD,N(com,sing) {problem}
```

From the above parsed text, we have to extract ‘OD,NP’ corresponding to ‘that’ and ‘CO, PP’ corresponding to ‘a serious problem’. Note in particular the depth of each node in the above. ‘OD,NP’ and ‘CO, PP’ underlined in the above lie at the second depth, which is tantamount to the previous depth of ‘regard’ marked with asterisks. That means it is necessary to design the algorithm to collect relevant items while traversing inside the maximal projection of a verb. The algorithm is given below.

```
ARGS = ['OD', 'OI', 'CO', 'CS', 'CI', 'CT', 'PROI', 'NOOD']
N_i = {PARSED_TEXT_i, DEPTH_i}
T = {N_1, N_2, ...}
num = 1
for nd in T:
  if nd has *:
      tmp = num – 1
      do until N_tmp is the maximal projection of nd
        if PARSED_TEXT_tmp is relevant to ARGS and DEPTH_tmp == DEPTH_num - 1
          SF += PARSED_TEXT_tmp
        tmp = cnt + 1
      do until Ntmp is the maximal projection of nd
        if PARSED_TEXT_tmp is relevant to ARGS and DEPTH_tmp == DEPTH_num – 1:
          SF += PARSED_TEXT_tmp
        tmp = num – 1
  num = num + 1
```

Each subcategorization item includes functional information as well as categorical information (e.g. ‘OD,NP’, ‘CS,AJP’, or ‘CO,AJP / OD,NP’). On the other hand, if there is no subcategorization item (i.e. intransitive verbs that need only the subject), the representation of the frame is as ‘–’. Now that we get the whole subcategorization frames of 157 verbs with their frequency values, we are ready to measure selectional preference strength.

4.2.3. Selectional Preference Strength

This study makes use of the Kullback-Leibler Divergence model in order to discriminate which subcategorization frame is significantly relevant to a verb. In this study, the formula (6) from Resnik (1996) was applied to measuring each strength, in which *S* stands for ‘strength’, *v* a ‘verb’, and *sf* a ‘subcategorization frame’.

$$S(v, sf_i) = \frac{P(sf_i \mid v) \log \frac{p(sf_i \mid v) + 1}{q(sf_i)}}{\sum P(sf \mid v) \log \frac{p(sf \mid v) + 1}{q(sf)}} \quad (6)$$

Using the above formula, we could sort out the subcategorization frames of a verb ordered by strength. For instance, 'regard' takes the following as its major frames.

Table 1: The major subcategorization frames of 'regard'.

	frame	strength	Σ
1st	CO,PP / OD,NP	0.448312089	0.448312089
2nd	CO,PP	0.33378163	0.782093719
3rd	CO,PP / NOOD,CL	0.07252358	0.854617299

4.3. The Relationship

The section investigates the relevance between semantic similarity discussed in Section 4.1 and subcategorization frames dealt with in Section 4.2. In order to test the relevance, we figured out how much each cluster shares subcategorization frames. According to how much each verb is compatible with the other verb(s) in the same cluster, 157 verbs were classified into four subgroups; 'Y', 'N', 'M', and '?'. First, 'Y' means the coincidence accounts for more than 0.8. Even if a verb does not belong to 'Y', if the sum of the largest value and the second largest value is over 0.5, the verb comes under 'M' groups. Third, '?' represents the situation where the word does not appear in ICE-GB as a verb (e.g. 'queen'). Finally, the others that do not belong to any of them will be labeled as 'N'. Table 2 shows the distributional property.

Table 2: The relevance between clusters and selectional preference.

	num	%	Σ
Y	66	42.04%	42.04%
M	49	31.21%	73.25%
N	39	24.84%	98.09%
?	3	1.91%	100%

Table 2 indicates that the significant relevance between semantic similarity and subcategorization frames accounts for approximately 42%, and the likelihood that verbs in a cluster share subcategorization frames is more than 73%. These measures imply that there is a considerable relationship between them, and the selection of subcategorization frames depends on semantic characteristics to some degree.

Another interesting observation we were able to make is that the bigger the Z-score is, the more significant the cluster is. The proportion of 'Y' in the high-ranking 10 clusters ordered by Z-score is about 63%, whereas that in the low-ranking 10 clusters is about 22%.

Table 3: The comparison between high-ranking and low-ranking clusters.

	high-ranking 10		low-ranking 10	
	num	%	num	%
Y	17	62.69%	6	22.22%

M	7	25.93%	8	29.63%
N	3	11.11%	13	48.15%

Table 3 shows that two or more verbs whose meanings are much similar to each other have a tendency to use the same subcategorization frame. In a nutshell, semantic properties of the verbal lexical items constrain their choice of syntactic patterns to a significant degree.

5. Conclusion

In this study, we tested a working hypothesis that there is a positive relationship between semantic similarity and subcategorization frames in English. This paper is under the assumption that if a group of verbs form a cluster sharing a similar meaning, they tend to share subcategorization frames. In the process, we proposed a statistical method to prove this assumption. We made use of two language resources, namely, ICE-GB and WordNet. We came to the conclusion that the proposed hypothesis holds true to the degrees of 42~73%. This generalization is based on our statistical analysis of two results; one is a list of 61 verbal clusters built up on the basis of the WordNet::Similarity module and the hierarchical bottom-up clustering algorithm, the other is a set of subcategorization frames gathered from the one million word corpus of ICE-GB and ordered by the Kullback-Leibler Divergence model.

The implication of this study for some similar future studies can be summarized as follows: First, this paper provides a well-structured method to test the validity of a linguistic hypothesis based on large-scale language resources, such as ICE-GB, WordNet, or CCED. The stochastic approach introduced in this paper can also be applied to other linguistic research. Second, we believe this kind of descriptive and inductive approach complements the more theoretically oriented approaches, based mostly on intuition.

References

Banerjee, S. and T. Pedersen. 2003. Extended gloss overlaps as a measure of semantic relatedness. *Proceedings of the 18th International Joint Conference on Artificial Intelligence*, pp. 805-810.

Fellbaum, C. 1998. *WordNet: An Electronic Lexical Database*. Cambridge: The MIT Press.

Lasnik, H. and J. Uriagereka. 2005. *A Course in Minimalist Syntax*. Malden: Blackwell Pub.

Lesk, M. 1986. Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone. *Proceedings of the 1986 Special Interest Group on Design of Communication Conference*, 24-26.

Levin, B. 1993. *English Verb Classes and Alternations: a Preliminary Investigation*. Chicago: University of Chicago Press.

Manning, C. D. and H. Schütze. 1999. *Foundations of Statistical Natural Language Processing*. Cambridge: The MIT Press.

Nelson, G., S. Wallis and B. Aarts. 2002. *Exploring Natural Language: Working with the British Component of the International Corpus of English*. Philadelphia: John Benjamins Pub. Co.

Pedersen, T., S. Patwardhan and J. Michelizzi. 2004. WordNet::Similarity - measuring the relatedness of concepts. *Proceedings of the 19th National Conference on Artificial Intelligence*, pp. 1024-1025.

Resnik, P. 1996. Selectional constraints: an information-theoretic model and its computational realization. *Cognition*, 61, 127-159.