Calculating Selectional Preferences of Transitive Verbs in Korean

Sanghoun Song
Department of Linguistics
University of Washington
Box 354340 Seattle, WA 98195-4340, USA
sanghoun@uw.edu

Jae-Woong Choe
Department of Linguistics
Korea University
145 Anam-ro Seongbuk-gu Seoul, 136-701 Korea
jchoe@korea.ac.kr

Abstract

This study calculates the selectional preference strength between transitive verbs and their co-occurring objects, and thereby investigates how much they are co-related to each other in Korean. The selectional preference strength is automatically measured in a bottom-up way, and the outcomes are evaluated in comparison with a manually constructed resource that indicates which verb takes which class(es) of nouns as its dependents. The measurement offered by this study not only can be used to improve NLP applications, but also has a theoretic significance in that it can play a role as distributional evidence in the study of argument structure.

1 Introduction

Selectional Preference Strength (henceforth, SPS) refers to the degree of correlation between two co-occurring linguistic items. This study, exploiting some Korean language resources and employing the Kullback-Leibler Divergence model formulated by Resnik (1996), aims to calculate SPS between transitive verbs and the classes of co-occurring nouns that function as objects.

As far as we know, there has been no previous study to calculate SPS in Korean. Now that several Korean resources constructed on a comprehensive scale are currently available, it would be very interesting to conduct a systematic analysis of SPS in Korean and to see what kind of significant patterns and results can be found through such analysis. This research is an endeavour in that direction, and reports some results of our analysis of SPS between predicates and their object argument, which is based on language resources like treebanks, wordnets, and electronic dictionaries. We also expect that our analysis would make a meaningful contribution to our understanding of the semantic interaction between verbal items and argument structure in Korean.

This paper is structured as follows. Section 2 discusses why it is necessary to look into SPS in NLP, and offers a brief explanation of the background knowledge. Section 3 covers the computational model that this study employs, and Section 4 measures SPS using a Korean wordnet (i.e. KorLex) and a development corpus (i.e. the *Sejong* Korean Treebank). The results are evaluated quantitatively as well as qualitatively in Section 5. This paper closes in Section 6 with a brief look at our further work to help NLP systems perform better.

2 Background

The Korean language, as is well-known, is an agglutinative language with a large number of grammatical function morphemes. It also has features like the right-headness, scrambling, and virtually free deletion of any element from a sentence. On the more semantic side, Korean shows the usual restriction between a predicate and its selection of arguments. The sentence pair in (1) exemplifies the syntactic and semantic behaviours in Korean. The verb *masi* 'drink' can take as its object only a small set of nouns which can roughly defined as the 'drinkable', while rejecting a whole lot of other nouns. While *maykcwu* 'beer' would be a typical object, *chayk* 'book' is inappropriate as the object of the verb.

26th Pacific Asia Conference on Language,Information and Computation pages 89–98

(1) a. *maykcwu-ul masi-ta*
beer-OBJ drink-DECL
‘ ... drink beer.’

b. #*chayk-ul masi-ta*
book-OBJ drink-DECL
‘# ... drink book.’

Notice that the two sentences are of the same morphological and syntactic configuration. It is thus clear that parsing sentences depends heavily on lexical semantics of the words involved. The major question addressed in this study is how we can capture the preferences that hold between a predicate and its arguments in Korean in a systematic way. Following Resnik (1996), this study contends that the questions can be properly answered by SPS, which defines the relationship between a verb and the entire noun class hierarchy.

2.1 Selectional Preference Strength

SPS, an information theoretic concept modeled by Resnik (1996), can be defined as a kind of relative entropy, which indicates how much interrelationship an entity has with another entity. The basic notion of SPS is exemplified in two structurally similar Q/A pairs (Resnik, 1996, pp. 127).

(2) a. Experimenter: Could a cow be green?

b. Subject: I think they’re usually brown or white.

(3) a. Experimenter: Could an idea be green?

b. Subject: No, silly! They’re only in your head.

Green cows do not necessarily exist in the real world, but we can figure them out by drawing a picture. In contrast, since we can hardly come up with ‘a green idea’, the question in (3) sounds strange.[1] That means ‘cow’ which is a kind of animals has a closer relationship with ‘green’ than ‘idea’ that comes under an abstraction. If we use a scale to represent the difference between the two relational pairs, we can say $\{\text{cow} \circ \text{green}\} > \{\text{idea} \circ \text{green}\}$, given that $\circ$ stands for the relational property. Here we can define the relational property that an operator $\circ$ represents as Selectional Preference, and the values that each relation has can be computed as numbers; for example, $\{\text{cow} \circ \text{green} = 100\}$, $\{\text{idea} \circ \text{green} = 5\}$.

Furthermore, we can make the relationship more abstractive. If we switch one item with another which conveys a similar meaning, almost the same preference goes for the other pair. For instance, elements in {green, purple}, {cow, dove}, and {idea, opinion} respectively are in the sister relations with each other within the lexical hierarchy (i.e. WordNet), whereby they are in complementary distribution as shown in (4).

(4) a. a green cow / a purple cow / a green dove

b. #a green idea / #a purple idea / #a green opinion

That means each element in each (4a-b) has the very similar or even the same relational values; for example, $\{\text{cow} \circ \text{green}\}$ is near equivalent to both $\{\text{dove} \circ \text{green}\}$ and $\{\text{cow} \circ \text{purple}\}$. With reference to the English WordNet, ‘cow’ belongs reflexively to ‘animals’, ‘object’, and ‘physical entity’, whose hierarchy differs from that of ‘idea’. In a nutshell, the so-called Selectional Preferences hinges on the semantic properties that a class of words shares.

2.2 Data

Basically three types of resources are required to calculate SPS: (i) a lexical hierarchy (e.g. WordNet), (ii) a development corpus, and (iii) comparable data for evaluation.

As discussed in the previous subsection, a lexical hierarchy that represents the kinship of words as a tree (or graph) structure plays an essential role in measuring SPS. Several Korean lexical hierarchies have been created so far, which include KorLex[2], U-WIN[3], CoreNet[4], etc. This study, among them, makes exclusive use of KorLex for two reasons. First, KorLex contains a table that connects each synset with the corresponding synset on the English WordNet. This mapping table would be of great merit, when we plan to extend the current work to multilingual studies in the future. Second, there exists a table that links lexical items in the

[1] This paper does not take metaphorical expressions into consideration. For example, ‘green’ sometimes refers to a social issue related to the protection of the environment as exemplified as ‘the green movement’. The current work is not concerned with those kinds of expressions.

[2] http://korlex.cs.pusan.ac.kr

[3] http://nlplab.ulsan.ac.kr/club/u-win

[4] http://semanticweb.kaist.ac.kr/home/index.php/CoreNet

Sejong electronic dictionary with each corresponding meaning of the synsets on KorLex (Park et al., 2010). Given that the *Sejong* electronic dictionary consists of a wide coverage of lexical items with a fine-grained linguistic description, if we take advantage of the table, we can systematically design further studies on the syntax/semantics interfaces.

A development corpus (preferably, naturally occurring texts) also play a critical part in computing SPS because there should be a data-oriented observation that shows which verbs take which nouns as the objects. A more in-depth and accurate analysis of the corpus can be expected to result in a better understanding of the syntax and semantics of the language. In particular, because the linguistic generalization of this study has to be drawn relying on the occurrence of functional tags (e.g. SBJ, OBJ), texts annotated at the syntactic layer (i.e. treebanks) are much more preferred. There are two available treebanks for Korean; one is the *Sejong* Korean Treebank, and the other is the Penn Korean Treebank. This study takes the former, mainly because the former is about three times larger than the latter. This study uses Xavier (Song and Jeon, 2008) as a tool to exploit the *Sejong* Korean Treebank.

This study makes a comparative analysis with the *Sejong* electronic dictionary for the purpose of evaluation, which has been manually encoded by linguists. The dictionary specifies the linguistic features of each argument in the XML format. For example, the second argument of *masi* 'drink', playing the theme role, has the selectional restriction (tagged within '<sel_rst ... >') as 'beverages'. Comparing the selectional preferences of the current work with the selectional restrictions given in the *Sejong* electronic dictionary, this study offers a quantitative evaluation (i.e. precision, recall, and f-measure).

3 Model

The verb and its argument(s) would be one of the representative categorical pairs that display Selectional Preferences clearly. Particularly, the classes of nouns that function as objects have been studied in many ways and in many languages because resolving objects performs a significant role in ambiguity resolution as well as syntactic parsing. For instance, Resnik (1995), who conducts several experiments using WordNet and English corpora such as BNC, compares the semantic characteristics of object nouns of 'drink' and 'find'. It is borne out by the experimental result that the object nouns of 'drink' cluster densely together, while those of 'find' are very scattered. The same goes for Korean as presented in (5).

(5) a. *maykcwu/khephi/#chayk-(l)ul masi-ta*
beer/coffee/book-OBJ drink-DECL

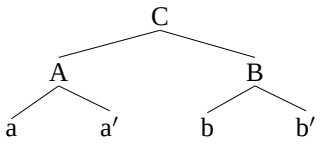
b. *chayk/sinmwun/#maykcwu-(l)ul ilk-ta*
book/newspaper/beer-OBJ read-DECL

c. *maykcwu/chayk-(l)ul chac-ta*
beer/book-OBJ find-DECL

3.1 Lowest Common Subsumer

Computational models for measuring similarity between words are roughly divided into two major types. One makes use of the definition of dictionaries (a.k.a. Lesk algorithm (Lesk, 1986)), and the other employs the Lowest Common Subsumer (hereafter, LCS) between two words. This study employs the latter because more algorithms have been implemented on the basis of it. LCS, according to Resnik (1995), means the lowest ancestor node that simultaneously subsumes its children nodes, by which the distance between the children can be measured. For instance, in a hierarchical tree (6), the LCS of 'a' and 'a′' is 'A', that of 'b' and 'b′' is B, and that of 'a' and 'b' is C.

(6)



With reference to KorLex, (5) can be abstractly converted into (7). Each number in parenthesis in (7) stands for the index of LCS of the words given in (5), which denotes 'beverage', 'production', and 'entity', respectively.

(7) a. (07406270)-OBJ *masi* 'drink'

b. (03856368)-OBJ *ilk* 'read'

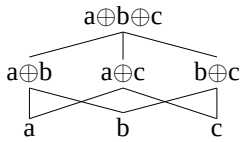
c. (00001740)-OBJ *chac* 'find'

3.2 Power Set

LCS is virtually located by creating a power set for each verbal item. A power set means a set whose elements are all the subsets of a given set, which can

be conceptualized as a lattice structure. Given that a set S consists of three elements such as {a, b, c}, the lattice structure which represents the power set is sketched out in (8a), and thereby the power set of the set S is calculated as (8b), ignoring an empty set.

(8) a.



b. {{a}, {b}, {c}, {a,b} {a,c}, {b,c}, {a,b,c}}

If it is observed that a verbal item *v* takes three elements {a, a′, b} as its object nouns, the verb involves seven mappings to subsets of the set as shown in (9) with respect to a relational operator ∘ that defines SPS and another operator • that represents the LCS of the operands.[5] Note that {(a•a′) = A, (a•b) = C, (a′•b) = C}, as sketched out in (6).

(9) a. $v \circ a$

b. $v \circ a'$

c. $v \circ b$

d. $v \circ (a \bullet a') = v \circ A$

e. $v \circ (a \bullet b) = v \circ C$

f. $v \circ (a' \bullet b) = v \circ C$

g. $v \circ (a \bullet a' \bullet b) = v \circ C$

If we make an assumption that the verb *v* is *masi* 'drink' and the three elements (i.e. a, a′, and b) are *maykcwu* 'beer', *khephi* 'coffee', and *chayk* 'book' respectively, we can obtain five relations as given in (10).[6] The numbers in parenthesis are the same as the ones given before.[7]

(10) a. *masi* 'drink' ∘ *maykcwu* 'beer' (07411192, 07411517)

b. *masi* 'drink' ∘ *khephi* 'coffee' (07452170, 14434748)

c. *masi* 'drink' ∘ *chayk* 'book' (02768681, 02769059)

d. *masi* 'drink' ∘ beverage (07406270)

e. *masi* 'drink' ∘ entity (00001740)

[5]The operator • satisfies the associative law.

[6]Note the different usages between 'w' and just w. The former represents a word, while the latter does a synset.

[7]A single word can be included in different synsets. For example, 'coffee' has two meanings; one is a kind of beans, and the other is a kind of beverages. Thus, words (i.e. 'w') can have multiple synsets as shown in (10a-c).

3.3 Hill Climbing

The cardinality of a power set of a set that includes *n* elements is represented as 2^n–1, excluding ϕ. That implies the cardinality grows exponentially. For example, if a verbal item takes 100 different nouns as its objects, 2^{100}–1 subsets will be examined, which is too huge to calculate within a common development environment.[8] Thus, it is highly necessary to devise a means to overcome the problem in calculation.

This study, for this purpose, makes use of hill climbing, which refers to a computational technique that attempts to solve the whole problem by incrementally associating the partial solutions. Though it sounds like an ad-hoc method, if we are able to repeat it until no further improvements can be found, the better solution to the problem can be offered.[9]

Our model to compute LCS starts hill climbing with two parameters *m* and *n*, if the number of object nouns is more than *n*. Our model randomly chooses *n* elements out of the whole elements, and calculates LCS of the subset consisting of *n* elements. This procedure is iterated *m* times whereby the set of LCSs grows incrementally. For example, if a verbal item takes 100 nouns such as $\{a_1, a_2, ..., a_{100}\}$, (11) is one of the instances that our model can create, given that *m*=4, *n*=3.

(11) $\{a_3, a_{29}, a_{71}\}$
$\{a_{14}, a_{55}, a_{86}\}$
$\{a_{26}, a_{49}, a_{90}\}$
$\{a_{13}, a_{65}, a_{77}\}$

If we use parameters big enough to cover the greater part of the whole elements (for this study, *m*=30, *n*=16), we can obtain fairly plausible results.

3.4 Kullback-Leibler Divergence

The algorithm that this study makes use of is largely adapted from the Kullback-Leibler Divergence model presented in Resnik (1996), which plays a part to discriminate which LCS is the most significantly relevant to the given verbal item. (12)

[8]Actually, it is observed that some frequently used verbs such as *mek* 'eat' take more than 100 nouns.

[9]In particular, it is merited in the cases in which the ultimate conclusions are not likely to be drawn with an ordinary approach.

measures each strength that a verbal item has, in which *S* means 'strength', *v* stands for a 'verb', and *c* is short for a 'class' of nouns in the given lexical hierarchy.

(12)

$$S(v, c_i) = \frac{P(c_i|v) \log \frac{P(c_i|v)+1}{P(c_i)}}{\Sigma\, P(c|v) \log \frac{P(c|v)+1}{P(c)}}$$

Consequently, LCSs acquired in the previous two subsections can be ordered by SPSs the formula (12) defines. The top-ranked one among them (i.e. the LCS that has the strongest Selectional Preference with *v*) is called the Association Strength (hereafter, AS), which distributionally represent the semantic properties of the verbal item.

4 Calculation

This study establishes the following guidelines to conduct an experiment of calculating SPS. First, the calculation is performed in a bottom-up way (i.e. a data-oriented approach), mainly because there already exists a resource constructed in a top-down way (i.e. the *Sejong* electronic dictionary). Second, we try to measure SPS on a large scale exploiting as much data as we can. Korean, as aforesaid, already has various types of linguistic resources, but there are few secondary products based on the resources. Third, the system is implemented with an eye towards running in an automatic way, which facilitates applying the whole procedure to the future work that deals with other resources or other relational pairs (e.g. verbs and subjects).

4.1 Procedures

The first step of the current work is to make a list of verbal items with reference to the development corpus. In the *Sejong* Korean treebank, there are two types of verbal items in terms of annotation formats. The first one is tagged with 'VV', which includes common verbs. The second one is formatted as [NNG + *ha*], in which NNG belongs to verbal nouns and *ha* functions as a light verb. The first one contains 1,447 verbal entries, the second one does 1,313 entries; thus in total 2,760 verbal entries are included on the list.

The second step is to extract nouns which are dependent on the verbal items. The Xavier module extracts object nouns of the verbal entries from the *Sejong* Korean treebank, which are tagged as 'NP_OBJ'. After that, nouns that do not appear on KorLex are excluded, because it is not possible to calculate their SPS without any information from the lexical hierarchy. In this way, a total of 6,948 types and 42,099 tokens of nouns are acquired. Then the type/token ratio is 16.5%, and each verbal item takes 2.52 types of 15.25 nouns as its objects on average.[10]

Table 1: Basic Measures

# of verbal entries	2,760
# of verbs	1,447
# of verbal nouns	1,313
# of tokens of objects	42,099
# of types of objects	6,948
# of collected LCSs	32,557

The next step is to collect LCSs of each verbal item, building upon the model presented in the previous section. 2,561 verbal items have one or more LCS(s). 32,557 LCSs are collected, which means each verb involves 11.8 LCSs on average. The statistical measures presented so far are summarized in Table 1.

The final step is to measure SPS, and determine the strongest one (i.e. AS) for each verbal item, whose average and standard deviation are .0667 and .0756 respectively.

4.2 Outcomes

The outcomes acquired thus far are analyzed from two viewpoints. The first one is about whether frequency has a distributional effect on the outcomes or not. The second one is to look at the representative cases in which SPS can be obviously *vs*. hardly captured, and to set up a working hypothesis building upon the findings.

4.2.1 Frequency

This subsection deals with the relevance between frequency and Selectional Preferences. The analysis will be made in terms of four factors that can potentially have a correlation with each other. The first two are concerned with verbal items; one is (i-a) the frequency of verbal items themselves and (i-b) the type/token ratio of object nouns of verbal items. The other two include (ii-a) the size of LCSs and

[10] For this reason, we use *n*=16 in hill climbing.

Figure 1: frequency (i-a) *vs*. # of LCSs (ii-a)

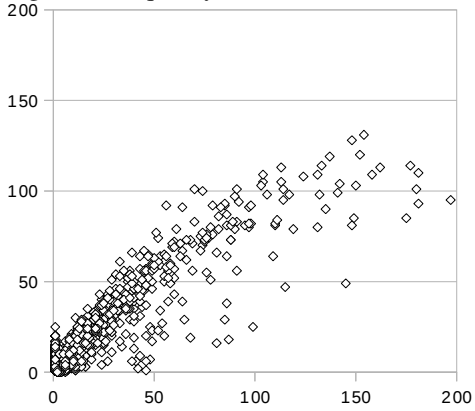


Figure 2: frequency (i-a) *vs*. AS (ii-b)

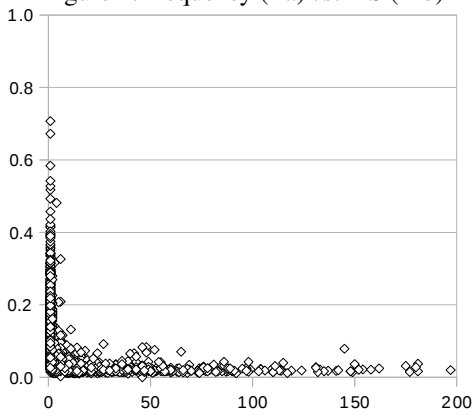


Figure 3: t/t (i-b) *vs*. # of LCSs (ii-a)

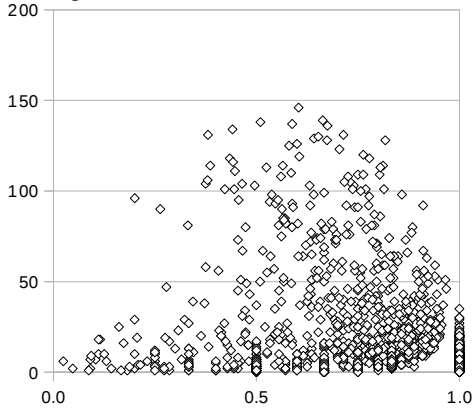
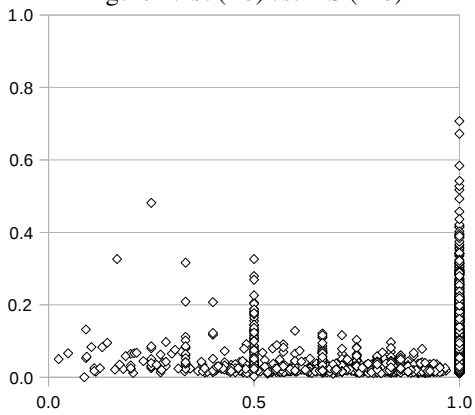


Figure 4: t/t (i-b) *vs*. AS (ii-b)



(ii-b) the value of each AS.

Figure 1, first, indicates the correlation between (i-a) the frequency on the X-axis and (ii-a) the number of LCSs on the Y-axis, in which each diamond represents (i-a, ii-a) on the coordinates. As can be expected, the high frequent items also show a high size of LCSs. Table 2 contains cases of the high, middle, and low frequent items that also show the corresponding sizes of LCSs.

Table 2: frequency *vs*. LCSs

verbs	**freq**	**LCSs**	**synset (index)**
ilwu 'achieve'	181	110	status (00024568)
ilk 'read'	180	101	production (03856368)
cwucangha 'claim'	46	48	knowledge (00020729)
ssis 'wash'	44	45	body parts (04924211)
koylophi 'bother'	6	5	human (00006026)
sunginha 'accredit'	3	1	action (00026194)

Figure 2 stands for the correlation between (i-a) frequency and (ii-b) the value of AS, which implies that verbal items that very less frequently appear can have full range of values, whereas the ASs of most other items, namely the higher frequent ones, are under .1.

Next, Figure 3 and Figure 4 illustrate the correlation between (i-b) the type/token ratio of object nouns and (ii-a) plus (ii-b), respectively. At a glance, Figure 3 and Figure 4 imply that there seems to be no clear relevance between (i-b) and (ii-a/b), except that the smaller the type/token ratio is, the less variety of nouns are used as the objects.

4.2.2 Strengths

Figure 5 to 7 indicate the distributional properties of SPSs of verbal items in (5). Figure 5 stands in stark contrast to Figure 7, and Figure 6 is somewhere between them. In each figure, the number of bars is the same as the number of LCSs, which represents how many synsets have SPS with the verbal item. The more bars a chart has, the more LCSs are collected with respect to the verbal item. On the other hand,

Table 3: SPS

verb	SPS	AS (index)
masi 'drink'	.04	beverage (07406270)
ilk 'read'	.028	production (03856368)
chac 'find'	.0218	entity (00001740)

the height of bars stands for SPS, which means the taller a bar is, the more preferably the class of nouns (on the X-axis) co-occur with the verbal item. There are not so many bars on Figure 5, but they are relatively taller than those on Figure 6 and Figure 7. That means *masi* 'drink' has a tighter relation with only a few number of synsets (i.e. LCSs). In contrast, there are quite a number of bars on Figure 7, mostly short, which implies *chac* 'find' can co-occur with a wide variety of nouns but their relationships are quite looser.

The verbal items exemplified in (5) have Association Strengths as given in Table 3. Among the verbal items that occur more than 10 times, the most typical *masi*-like items (i.e. high SPSs with few LCSs) and the most typical *chac*-like items (i.e. low SPSs with many LCSs) are exemplified in Table 4 and Table 5, respectively. The difference between *masi* 'drink' and *chac* 'find' can also be found in the list of candidates that are not selected as the AS, which are given in Table 6 and Table 7, respectively. The closely associated synsets with *masi* are relatively concrete and specific, whereas those with *chac* are the higher ones in the lexical hierarchy, namely, more abstractive and comprehensive.

Table 4: high SPSs with fewer LCSs

verbs	t/t	LCSs	SPS	AS (index)
kkwul 'kneel'	.09	2	.132	kneel (02375920)
chwu 'dance'	.13	6	.083	dance (00498636)
sso 'shoot'	.57	6	.082	arms (04387884)

Table 5: low SPSs with many LCSs

verbs	t/t	LCSs	SPS	AS (index)
tul 'carry'	.38	131	.014	linguistic unit (05901081)
phiha 'avoid'	.75	100	.012	entity (00001740)
pwuthi 'stick'	.53	94	.011	mentality (00020333)

Figure 5: SPSs of *masi* 'drink'

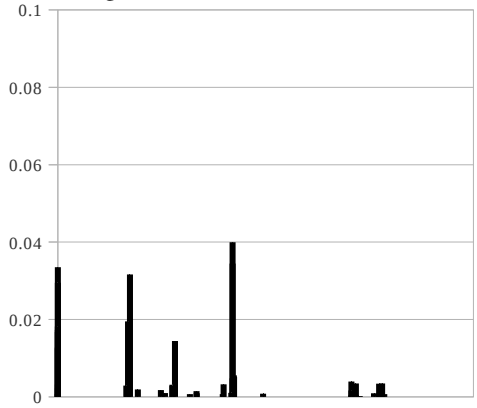


Figure 6: SPSs of *ilk* 'read'

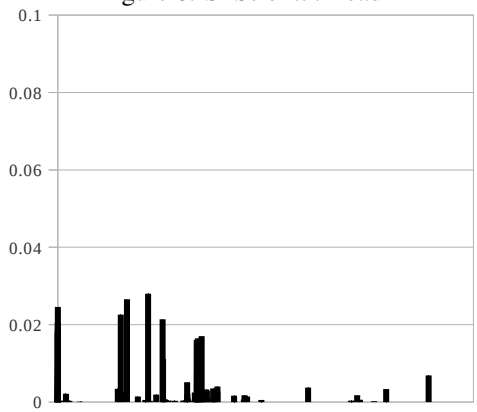


Figure 7: SPSs of *chac* 'find'

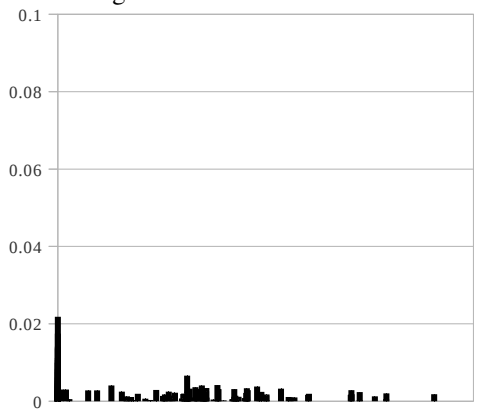


Figure 8 indicates the relationship between the number of LCSs and the value of SPSs of the corresponding verbal items. For example, the diamond corresponding to masi 'drink', whose LCSs are small but whose SPS values are relatively high, lies around the upper left area. In contrast, the mark for *chac* 'find', which has many LCSs and small values of SPSs, lies on the lower right corner. Figure 8 implies that verbal items that yield more than about ten LCSs show a tendency not to have so strong preference with co-occurring nouns.

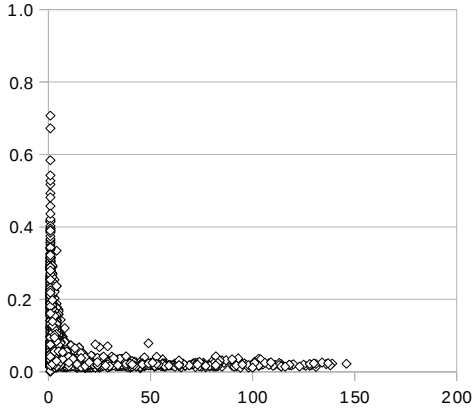
Table 6: Other SPSs of *masi*

synset (index)	SPS
alcoholic drinks (07408983)	.0345
nutrient (00018827)	.0335
medicine (03129572)	.0316
ingredient (00017572)	.0295
ornament (03054637)	.0195

Table 7: Other SPSs of *chac*

synset (index)	SPS
object (00016236)	.0175
abstraction (00020486)	.0141
mentality (00020333)	.0136
knowledge (00020729)	.0112
relation (00027929)	.0107

Figure 8: # of LCSs *vs.* SPS



5 Evaluation

5.1 Quantitative Evaluation

The quantitative evaluation in this study is based on the comparison of the results with the *Sejong* electronic dictionary, which consists of 32,714 verbs plus 6,998 adjectives. The dictionary covers various linguistic levels, including selectional restrictions of verbal items. The comparative analysis of this study checks out how well the SPS values of this study matches with the lexical information.

The quantitative measurements that this study uses are precision, recall, and f-measure, which are respectively formulated as follows. Precision means the fraction of extracted instances which has a relevance with the corresponding item, whereas recall means the fraction of relevant instances which are extracted. F-measure associates these two measures simultaneously to show the compatibility.

(13) a.

$$\textit{precision} = \frac{tp}{tp + fp}$$

b.

$$\textit{recall} = \frac{tp}{tp + fn}$$

c.

$$\textit{f-measure} = \frac{2 \times \textit{precision} \times \textit{recall}}{\textit{precision} + \textit{recall}}$$

If a certain class of nouns is specified for the object position of a predicate in the *Sejong* electronic dictionary, and is also computed as one of the LCSs of the corresponding verbal items, the value tp (i.e. true positive) increases. If a class of nouns appears in the results of this study but not in the dictionary, the value fp (i.e. false positive) increases. Finally, if a class of nouns is specified only in the *Sejong* electronic dictionary, the value fn (i.e. false negative) becomes greater by that much. The distinction among them is presented in the Table 8 for the ease of exposition.

Table 9 gives the evaluation measurement conducted by formula (13) and Table 8. It turns out the measures are pretty low, the f-measures being around 10%, which means that the two resources match with each other rather poorly. We suspect the poor results are mainly due to the difference in the lexical hierarchies assumed in KorLex and the *Sejong* electronic dictionary in the first place. It is true that the lexical hierarchies can be built upon different theoretical assumptions. The ontologies in the *Sejong* electronic dictionary and KorLex are much different from each other (Bae et al., 2010), so a proper comparison and evaluation should be done after the mapping between the two heterogeneous

Table 8: True/False Positive/Negative

	Sejong	***¬Sejong***
LCSs	tp	fp
¬LCSs	fn	tn

Table 9: Quantitative Evaluation

precision	12.98%
recall	8.99%
f-measure	10.62%

ontologies is properly established. Bae et al. (2010) is an endeavour in that direction, but we could not include it in the current work. Another reason for the poor evaluation results, which is basically the same problem as the first, is that the terms used in both ontologies are different from each other in many cases. For instance, the concept 'abstraction' can be specified as an 'abstractive concept' in one resource and as just an 'abstraction' in the other; actually, KorLex takes the former, and the *Sejong* electronic dictionary takes the latter. The evaluation in this study was based on the surface match, and thus could not accommodate the mismatch in the terms used, which means when the mismatches are well taken care of, the f-measures would increase that much. Suffice it to say at the moment that the results given in Table 9 can be taken as a baseline values for the future studies.

5.2 Qualitative Evaluation

For a qualitative evaluation of this study, a manual checkup was done on some of the results of this study. We point out three issues that are found in the process, which need to be properly addressed in the future study.

First, it is discovered that homonyms sometimes have an adverse effect on the outcomes. For example, it is reported that *ketepwuthi* 'roll up' has a strong preference with a homonym *phal*, which can convey a meaning of either 'eight' or 'arm' in Korean. Although it is much more natural that 'roll up' has a relevance to 'arm' rather than 'eight' in the sense of 'roll up one's sleeves', the outcomes provide only *phal* 'eight' as the AS of *ketepwuthi*. This problem would be solved, if some sense-tagged texts are available as the development corpus, which has been partially studied by Park et al. (2010).

Second, causative forms which often bring about argument alternations are not taken into account in the process of extracting object nouns from the development corpus (i.e. the *Sejong* Korean Treebank). The causative forms in Korean, which are in the format of '*-key/tolok ha*', need to be analyzed from a fine-grained syntactic standpoint (Alsina et al., 1996), because NPs with theme-roles may not be in situ in the constructions.[11] The variation in form-meaning mapping in Korean causatives needs to be deeply explored in a corpus-oriented way, which we would like to reserve for another inquiry.

Finally, two closely relevant words sometimes exist far from each other within the hierarchy, which eventually causes a problem. For example, *michi* 'exert' takes two major types of nouns; one is *yenghyang* 'influence' and the other is *yenghyanglyek* 'power of influence'. It is obvious that these two words are closely relelated to each other, but they are not in the sister relation with each other in KorLex; the former is specified as an action, while the latter is a kind of abstractive concept. Since the verbal item *michi* 'exert', for this reason, cannot be preferably associated with these two words in the current processing model, we cannot construct the pattern like 'exert an influence on' from our results.[12]

6 Conclusion

In this paper, we calculated the SPS between verbal items and the classes of their co-occurring nouns. The SPS has been automatically measured with reference to two Korean language resources; (i) KorLex as the lexical hierarchy of noun classes, and (ii) the *Sejong* Korean Treebank as the development corpus. The acquisition model is grounded upon the LCS that represents the closest common ancestor node for the given two nodes within the hierarchy. The SPS is defined by Kullback-Leibler Divergence, which depends on the collection of LCSs. The results are evaluated with reference to the *Sejong* electronic dictionary which has been manually constructed.

This study, on the other hand, has certain limitation, especially in the evaluation process. It needs to

[11] We had tried to get rid of the form '*-key/tolok ha*' from the observed data and conducted the experiment from the beginning again, but we learned that there were more causative forms that involve argument alternations, other than '*-key/tolok ha*'. For example, an auxiliary *cwu*, whose original meaning comes from 'give', sometimes behaves like a causative marker and alters the argument structure.

[12] The two words, of course, are not always in the same distributional condition. For example, a verb *cwu* 'give' does not tend to co-occur with *yenghyanglyek* 'power of influence', while it does with *yenghyang* 'influence'. Given that KorLex has been constructed with some reference to those kinds of relational properties (i.e. collocations), it is not unusual that two or more words apparently related to each other sometimes come under different nodes in the hierarchy (Aesun Yoon, personal communication).

be done on the basis of resources that would overcome some clear limitations of the evaluation process adopted in this study. However, in spite of the limitations, we believe the results reported in this study can have some implications for future studies, including extending the results to other grammatical functions like subject, or making use of other Korean ontologies like U-WIN or CoreNet.

Acknowledgments

We thank Professor Aesun Yoon for her valuable comments and suggestions, though it should be noted that we could not fully accommodate them in this paper. We also thank three anonymous reviewers for helpful feedback. After the final version was submitted, we found out that there was a critical mistake in our handling of a key file name during the calculation, which resulted in incorrect figures and tables in that version. Therefore, in this minimally corrected version, as we did in the actual presentation of our paper during the conference, we substituted the corrected figures (Fig. 1-4, 8) for the wrong ones, and made due changes in some related tables and discussion. We thank the Program committee for allowing us to remedy our earlier mistake at the final moment. Of course, all remaining errors and infelicities are our own.

References

Alex Alsina, Joan Bresnan, and Peter Sells. 1996. Complex Predicates: Structure and Theory. In Alex Alsina, Joan Bresnan, and Peter Sells, editors, *Complex Predicates*, pages 1–12. CSLI Publications, Stanford.

Sun-Mee Bae, Kyoungup Im, and Aesun Yoon. 2010. Mapping Heterogenous Ontologies for the HLP Applications – Sejong Semantic Classes and KorLexNoun 1.5 –. *Korean Journal of Cognitive Science*, 21:95–126.

Michael Lesk. 1986. Automatic Sense Disambiguation Using Machine Readable Dictionaries: How to Tell a Pine Cone from an Ice Cream Cone. In *Proceedings of the 5th annual international conference on Systems documentation*, pages 24–26. ACM.

Heum Park, Aesun Yoon, Woo Chul Park, and Hyuk-Chul Kwon. 2010. Considerations on Automatic Mapping Large-Scale Heterogeneous Language Resources: Sejong Semantic Classes and KorLex. In *Proceedings of the Eighth Workshop on Asian Language Resouces*, pages 14–21, Beijing, China, August. Coling 2010 Organizing Committee.

Philip Resnik. 1995. Using Information Content to Evaluate Semantic Similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*, pages 448–453.

Philip Resnik. 1996. Selectional Constraints: An Information-theoretic Model and its Computational Realization. *Cognition*, 61(1-2):127–159.

Sanghoun Song and Jieun Jeon. 2008. The Xavier Module - Information Processing of Treebanks. In *Proceedings of the International Conference of Cognitive Science (ICCS 2008)*, Seoul, South Korea.