Linguistic Research 31(2), 249-273

Selectional preferences of Korean verbal items*,**

Sanghoun Song* · Jae-Woong Choe******
(Korea University)

Song, Sanghoun and Jae-Woong Choe. 2014. Selectional Preferences of Korean Verbal Items. *Linguistic Research* 31(2), 249-273. The present study computes the selectional preference the verbal items show in relation to their co-occurring subjects and objects in Korean. The selectional preference indicates how significant is the semantic corelation between a verbal item and the class of nouns that appear in its argument position. The selectional preference measurements presented in the current work is automatically measured in a bottom-up way, using two types of language resources: One is a parsed corpus (the *Sejong* Korean Treebank), and the other is lexical hierarchies or wordnets. The three Korean wordnets that are made use of are CoreNet, KorLex, and U-WIN. The acquisition model is grounded upon the Lowest Common Subsumer that represents the closest common ancestor node for the given two nodes within the hierarchy. As a way to address the issue of exponential growth in calculation time and others, this study makes use of a programming technique called hill climbing. The selectional preferences, specifically Selectional Preference Strengths and Selectional Associations proposed in Resnik (1996), are defined by Kullback-Leibler Divergence, and their values are derived based on the collection of LCSs. The results are evaluated with reference to the *Sejong* Electronic Dictionary, with human evaluation, with parsing evaluation using a stochastic parser, and finally in a qualitative manner. **(Korea University)**

Keywords selectional preference, verbal items in Korean, lowest common subsumer, the *Sejong* Korean Treebank, Korean WordNets, the *Sejong* Korean Electronic Dictionary, evaluation

---

* This is a revised version of the conference paper by the same authors presented at PACLIC 26 (Song & Choe 2012b).

** We are very grateful to Prof. Aesun Yoon, Prof. Cheol-Young Ock, and Prof. Tim Baldwin for their valuable comments. Comments and constructive criticisms from two anonymous reviewers are also much appreciated. Needless to say, all remaining errors and shortcomings are our own responsibility. This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government (MEST) (No. 2012S1A5A2A01016651).

*** first author

**** corresponding author

1. Introduction

Selectional Preference (or SP for short) refers to the degree of correlation between two co-occurring linguistic categories. The present work addresses calculation of the SP between verbal items (e.g. verbs, adjectives, and verbal nouns) and their co-occurring subjects and objects in Korean.[1] The research inquiry this paper raises is exemplified in (1).

(1) a. *chinkwu-ka* *maykcwu-lul* *masi-ess-ta.*
friend-NOM beer-ACC drink-DECL
'(My) friend drank beer.'
b. #*chinkwu-ka* *chayk-ul* *masi-ess-ta.*
friend-NOM book-ACC drink-DECL
'(lit.) (My) friend drank a book.'

The verb *masi* 'drink' in (1) imposes a selectional restriction on its subject and object: The subject should be an animate entity, and the object should be something drinkable. These restrictions have to be specified in the lexical information (e.g. the argument structure), for example, like <NP[ANIMACY +], NP[DRINKABLE +]>. While *maykcwu* 'beer' in (1a) with [DRINKABLE +] feature can be used as its object, *chayk* 'book' in (1b), whose lexical semantic information is contradictory to the feature, is inappropriate as the object of the verb. Note that these two sentences constitute a minimal pair composed of the same morphological and syntactic configuration, except for the object noun. As implied by using a boolean feature above, this relation between predicates and their dependents has been regarded as a kind of restriction. Yet, this boolean-based approach sometimes works too strictly to represent a variety of language phenomena: Not all relations are necessarily viewed as a matter of black-or-white. For this reason, the data-oriented studies, including the present work, have quantified the relation, and 'preference' has been chosen for the terminology rather than 'restriction.'

The question that we raise herein is how we can acquire such a lexical semantic

[1] In this paper we use "Selectional Preference" as a cover term for the more technically defined concepts "Selectional Association (SA)" and "Selectional Preference Strength (SPS)", which will be discussed later in Section 3.4. See (13). We would like to thank a reviewer for pointing out some inadvertant confusion regarding the use of the terms in an earlier version of this paper.

relation between a verbal item and its dependent(s) in a systemic and (semi-)automatic way. This study contends that this questions can be properly answered by Selectional Preference which defines the relationship between verbs and the elements within a noun class hierarchy. The present work, utilizing two types of Korean language resources and employing the Kullback-Leibler Divergence model formulated by Resnik (1996), calculates SP measures between verbal items in Korean and the classes of their co-occurring nouns that function either as subjects or as objects.

The outcomes of the current work result from using language resources such as treebanks, wordnets, and electronic dictionaries. The measurement offered by this study can be significant from at least three standpoints. First, this analysis makes a meaningful contribution to our understanding of the syntactic and semantic interaction between verbal items and argument structure in Korean. Second, such distributional evidence helps us improve natural language processing applications for Korean. Finally and most importantly, the present study substantiates how reliably Korean resources work: Although various language resources in Korean have been built up so far, their usage has not been thoroughly tested in terms of grammatical validity as well as feasibility in real systems. This research is an endeavour in that direction.

This paper is structured as follows. Section 2 discusses why it is important to acquire the measurements of selectional preferences within the context of natural language processing and offers an explanation of background knowledge of the present study. Section 3 delves into the computational model the current work deploys step by step. Section 4 computes SP measures using two types of language resources. They include a development corpus (e.g. the *Sejong* Korean Treebank) and three Korean wordnets: namely, CoreNet, KorLex, and U-WIN. The results are evaluated both quantitatively and qualitatively in Section 5. This paper closes in Section 6 with a brief look at our further work.

2. Background

As is well-known, the Korean language is an agglutinative language with a large number of grammatical function morphemes. The basic linguistic properties of

Korean include the right-headedness (i.e. V-final), scrambling (i.e. OSV in some circumstances), and *pro*-drop (i.e. virtually free deletion of any element from a sentence). In addition to these general properties, there are several language specific phenomena in Korean. First, regarding the morphosyntactic strategy of coordination, Korean employs asyndeton (e.g. A B C (no marking; also known as juxtaposition)) with other strategies, such as monosyndeton, polysyndeton, and omnisyndeton (Kim and Yang, 2006). Second, the genitive maker *-uy* is optionally used, and thereby two adjacently co-occurring NPs can be analysed as an adnominal case constructions in Korean (Choe *et al*., 2012). The sentence given in (2) exemplifies these properties mentioned thus far all together, and the sentence can in principle have at least six readings.[2]

(2) *chinkwu maykcwu masi-ta.*
friend beer drink-DECL
a. ‘The friend drinks beer.’ (the most common one)
b. ‘Beer drinks the friend.’ (implausible)
c. ‘The friend and the beer drink something.’ (implausible)
d. ‘Somebody drinks the friend and the beer.’ (implausible)
e. ‘Beer of the friend drinks something.’ (implausible)
f. ‘Somebody drinks beer of the friend.’ (possible, but not common)

Although the other readings do not disobey the main properties of Korean mentioned above, Korean native speakers interpret the sentence as (2a) almost invariably. The other interpretations are structurally possible, but simply do not make much sense. This tells us that choosing the most plausible parse tree heavily depends on lexical semantics of the verbal entries functioning as the semantic head of the entire sentence.

2.1 Selectional Preferences

Selectional Preference is an information theoretic concept modeled by Resnik (1996). Roughly speaking, this is defined as a kind of relative entropy indicating how much

2 Note that the interpretation of (2b) can be allowed in contemporary Korean as a kind of metaphor, meaning that the person looks like an alcoholic.

interrelationship an entity has with another entity. The basic notion is exemplified in (3-4) where two structurally similar questions are given.

(3) a. Experimenter: *Could a cow be green*?
    b. Subject: *I think they're usually brown or white.*
(4) a. Experimenter: *Could an idea be green*?
    b. Subject: *No, silly*! *They're only in your head.* (Resnik, 1996:127)

Although green cows are non-existent in the real world, we can figure them out by drawing a picture. In contrast, since we can hardly come up with *a green idea*, the question in (4) sounds bizarre. That means *cow* which is a kind of animals has a closer relationship with a colour name *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 {cow ○ green} > {idea ○ green}, given that ○ stands for degree of the relational property. Here we can define the relational property an operator ○ represents as selectional preference, and the values that each relation has can be computed as numbers; for instance, {cow ○ green = 100}, {idea ○ green = 5}. Furthermore, we can make the relationship in a more abstract fashion. If we switch one item with another with 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 sister (or cousin) relations with each other within the lexical hierarchy (i.e. WordNet), whereby they are in complementary distribution as shown in (5). Notice that both *a green purple cow* and *a green cow dove* in (5c) sound infelicitous because the two elements in complementary distribution cannot occupy the same grammatical slot.

(5) a. a green cow / a purple cow / a green dove
    b. #a green idea / #a purple idea / #a green opinion
    c. #a green purple cow / #a green cow dove

Each element in (5a) and (5b) respectively has a very similar or even the same relational values. For example, {cow ○ green} is near equivalent to both {dove ○ green} and {cow ○ purple}. With reference to the English WordNet, *cow* belongs reflexively to a hierarchy of synsets, such as ‘animals’, ‘object’, and ‘physical entity’ in turn. It is clear that the synset hierarchy differs from that starting from ‘idea’. In a nutshell, selectional preferences resort to the semantic properties that a class of words share.

2.2 Data

To our knowledge, there exists no previous study to calculate the selectional preferences of verbal items in Korean on a comprehensive scale. Now that several Korean language resources constructed on a large scale are available, the time is ripe for a systematic analysis of SP measurements in Korean and to look into which significant patterns are identified through such analysis. Basically two types of resources are required to calculate them; namely, (i) a lexical hierarchy (e.g. WordNet), (ii) a development corpus. In addition to them, the present work refers to (iii) a comparable dataset for evaluation. Let us discuss the three types of data in more detail.

First, as discussed in the previous subsection, a lexical hierarchy that represents the kinship relationships of words as a tree (or graph) structure plays an essential role in measuring SP. Several Korean lexical hierarchies have been created so far; CoreNet (KAIST Korterm Center 2005),[3] KorLex (Yoon *et al*. 2009),[4] and U-WIN (Lim *et al*., 2008; Bae and Ock, 2013).[5] This study makes use of these wordnets for Korean in order to draw the whole picture of noun classes in Korean in a comparative way. In other words, this study compares the results obtained from each hierarchy, and then selects the most appropriate one for our current purpose.

Second, a development corpus (preferably, naturally occurring texts) plays a critical part in computing SP, because a data-oriented observation that shows which verbs take which nouns as their subjects or objects is required. 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

3 http://semanticweb.kaist.ac.kr/home/index.php/CoreNet

4 http://korlex.cs.pusan.ac.kr

5 http://nlplab.ulsan.ac.kr/club/u-win

preferred. There are several available treebanks for Korean, including the *Sejong* Korean Treebank[6] and the *Penn* Korean Treebank (Han *et al.*, 2002). This study makes use of the *Sejong* Korean Treebank for two reasons. First, the corpus contains the largest number of words among the treebanks publicly available. Second, there is a readily available tool called Xavier (Song and Jeon, 2008) for processing the *Sejong* Korean Treebank.

The third type of data we make use of in this study is a electronic dictionary; this study makes a comparative analysis with the *Sejong* Electronic Dictionary for the purpose of basic quantitative evaluation.[7] The dictionary was manually compiled by linguists, and all linguistic features including selectional restriction are described 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'. By comparing the selectional preferences of the current work with the selectional restrictions given in the *Sejong* Electronic Dictionary, this study offers a quantitative evaluation using three measures; namely, precision, recall, and F-measure.

3. Model

The verbal items, including verbs, adjectives, and verbal nouns, and their argument(s) would be one of the representative pairs that clearly exhibit selectional preferences. In particular, the classes of nouns that function as subjects and objects have been studied in many ways and in many languages, because identifying grammatical functions plays 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 a series of experiments that the object nouns of *drink* cluster densely together, while those of *find* are scattered. The same holds true for Korean as exemplified in (6).

(6) a. *maykcwu/khephi/#chayk-(l)ul masi-ta*

6 http://www.sejong.or.kr
7 http://www.sejong.or.kr

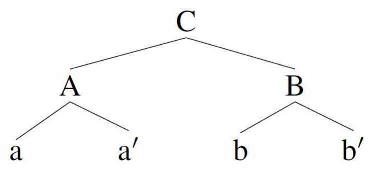
beer/coffee/book-ACC drink-DECL
b. *chayk/sinmwun/#maykcwu-(l)ul* *ilk-ta*
book/newspaper/beer-ACC read-DECL
c. *maykcwu/chayk-(l)ul* *chac-ta*
beer/book-OBJ find-DECL

A set of entities which run counter to the lexical semantic feature, such as *chayk* 'book' ([DRINKABLE −]) and *maykcwu* 'beer' ([READABLE −]) cannot be felicitously used as the object of specific types of verbs, such as *masi* 'drink' and *ilk* 'read', respectively. Yet, they can be felicitously used as the object of verbs without this kind of selectional restriction, such as *chac* 'find', as presented in (6c).

3.1 Lowest Common Subsumer

Computational models for measuring similarity between words are roughly divided into two major types. One makes use of the dictionary definitions (a.k.a. the Lesk algorithm (Lesk, 1986)), and the other employs the Lowest Common Subsumer (shortened as LCS) between two words. This study employs the latter on the basis of which more algorithms have been implemented. LCS, according to Resnik (1995), means the lowest ancestor node that simultaneously subsumes its child nodes, by which the distance between the children can be measured. For instance, in a hierarchical tree (7), the LCS of 'a' and 'a′' is 'A', that of 'b' and 'b′' is B, and that of 'a' and 'b' is C.

(7)



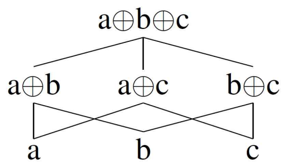
(8) is an instance of (7). Each number in parenthesis in (8) stands for the index specified in KorLex, and they refer to 'beverage', 'production', and 'entity', respectively.

(8) a. (07406270)-ACC *masi* ‘drink’
b. (03856368)-ACC *ilk* ‘read’
c. (00001740)-ACC *chac* ‘find’

3.2 Power Set

LCSs can be identified 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 (9a), and therefore the power set of the set S can be given as in (9b), ignoring the empty set.

(9) a.



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

Returning to the little ontology given in (7), if a verbal item *v* takes three elements {a, a′, b} as its object nouns, the verb involves seven mapping relations to subsets of the set as shown in (9a) with respect to a relational operator ○ that defines SP and another operator ● that represents the LCS of the operands. For instance, {a ● a′ = A, a ● b = C, a′ ● b = C}, as sketched out in (10). Notice that the operator ● satisfies the associative law (i.e. a ● b = b ● a).

(10) a. *v* ○ a
b. *v* ○ a′
c. *v* ○ b
d. *v* ○ (a ● a′) = *v* ○ A
e. *v* ○ (a ● b) = *v* ○ C
f. *v* ○ (a′ ● b) = *v* ○ C
g. *v* ○ (a ● a′ ● b) = *v* ○ 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 (11).

(11) 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)

Note the difference in notation we use in (11) between italicized words and the words in single quotations. The former represents a word, and the latter a synset. The numbers in parenthesis are the same as the ones given before. Notice that a single word can be included in different synsets. For example, *coffee* is associated with two synsets: One is a kind of 'grains', and the other is a kind of 'beverages'. Thus, some words (given in italicized fonts) can be associated with multiple synsets.

3.3 Hill Climbing

The cardinality of the power set of a set that includes *n* elements is represented as $2^n - 1$, excluding the empty set (i.e. Ø). That implies the cardinality of power sets grows exponentially. For example, some frequently used verbs such as *mek* 'eat' take more than 100 nouns as their objects. In that situation, such a verbal item would involve more than $2^{100} - 1$ subsets to be examined. These numbers are too huge to calculate within a common development environment. Thus, it is highly necessary to devise a means to overcome the problem in calculation.

As a way to address the issue of exponential growth in processing time, this study makes use of a programming technique called hill climbing.[8] This technique refers to a computational algorithm to solve the whole problem by incrementally finding partial solutions. Although it sounds like an ad-hoc method *prima facie*, if

8 Prof. Tim Baldwin commented that we could use a grid search algorithm, instead of Hill Climbing. We agree that there exist other options to deal with the issue of the exponential rate in calculation (e.g. parallel computing, etc.). We chose Hill Climbing as it can be considered as one of the most economical ways. Recall the three virtues of data-driven approach: namely, 'cheap', 'fast', and 'easy'.

we are able to repeat it until no further improvements can be found, a better solution to the problem can be obtained. In particular, this method works when the ultimate conclusions are not likely to be drawn with an ordinary approach.

Our model to compute LCS starts hill climbing with two parameters *m* and *n*, if the number of object nouns for a given verbal item 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 $\{a_1, a_2, ..., a_{100}\}$, (12) is one of the instances that our model can create, given that *m*=4, *n*=3.

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

If we use parameters big enough to cover the greater part of the whole elements, we can expect to obtain fairly plausible results. In this study, we set the parametric values at *m*=32 and *n*=16. The first value (i.e. *m*=32) was chosen as a result of a preliminary experiment: We conducted a series of tests to calculate precision, recall, and F-measure, and each test was carried out with different numbers of iteration (i.e. different *m* values). We found that *m*=32 is the threshold: If the search was iterated more than 32 times, no significant increase was found any longer. The second value (i.e. *n*=16) was selected considering the average number of types of verbal items' objects in the *Sejong* Korean Treebank is approximately 16.

3.4 Selectional Association

The basic algorithm the present work makes use of is largely adapted from the Kullback-Leibler Divergence model presented in Resnik (1996), which plays a part in figuring out which LCS is the most significantly relevant to the given verbal item. (13a) defines the Selectional Preference Strength (SPS) 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. It is called the relative entropy, which measures "the amount of extra information one would need to add to the approximation in order to make it fit the truth perfectly (Resnik, 1996: 135).” That is, SPS can be characterized as a relationship between a predicate and the entire conceptual space of its arguments. However, our primary concern is such that given a particular conceptual (sub-)class, what contribution the class can make to SPS. Resnik calls it the Selectional Association (SA), whose calculation relies on SPS as shown in (13b).

(13) a. $$S(v) = \sum_{c} P(c|v) \log \frac{P(c|v)}{P(c)}$$

b. $$A(v,c) = \frac{P(c|v) \log \frac{P(c|v)}{p(c)}}{S(v)}$$

Now, for each LCS of a predicate, we can calculate its SA using the formula (13b), thereby obtaining all the SA values for the given verbal item.

4. Calculation

This study establishes the following guidelines to conduct an experiment of calculating Selectional Preference Strength and Selectional Association. First, the calculation is performed in a bottom-up way. Note that there already exists a resource constructed in a top-down way (e.g. the *Sejong* Electronic Dictionary). The outcome of the current study is compared to the manually constructed data in the evaluation process. Second, we try to measure SPS and SA 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. Our attempt provides one fine-grained way to utilize the currently available constructed data for better understanding of linguistic phenomena in Korean. Third, the system is implemented with an eye towards running in an automatic way, which facilitates applying the whole procedure to some similar future work.

4.1 Workflow

The first step of the current work (Step I) is to make a list of verbal items with reference to the development corpus. In the *Sejong* Korean Treebank, there are two groups of verbal items in terms of annotation formats. The first one is tagged with 'VV' or 'VA', which includes common verbs and common adjectives respectively. The second one is formatted as [NNG/XR + *ha*], in which NNG and XR are verbal nouns, and *ha* functions as a light verb (tagged as 'XSV' or 'XSA').

Notice that there are two types of combinations between verbal nouns and light verbs, such as *ha*. The first type does not have any white space between the two items; for example, *kongpwu-ha-ta* 'study-LV-DECL'. In other words, the verbal nouns are selected via a lexical rule, and thereby the verbal nouns are adjoined to the light verb *ha* with no intervening space. In the second type, there is a white space between them; for example, *kongpwu-*(*lul*) *ha-ta* 'study-(ACC) LV-DECL'. This means that the two items are combined with each other by a phrase structure rule. The current work focuses on the first type, and does not regard the second type as a genuine light verb construction. There are several theoretic and distributional reasons for this view (Chae, 1996). In particular, we took notice of the fact that modifiers can be relatively freely inserted between the two items in the second type. For instance, *kongpwu-lul yelsimhi ha-ta* 'study-ACC hard LV-DECL', meaning 'study hard', is a perfect expression in Korean. (For more information about this decision, see Song and Choe (2012a).) In the *Sejong* Korean Treebank, there are 2,892 verbal items which involve object nouns, which consist of 1,447 verbs (VV), 79 adjectives (VA), and 1,366 verbal nouns (NNG or XR). On the other hand, there are 3,701 verbal items which take nouns as their subject. Amongst them, 1,628 items are verbs (VV), 295 items are adjectives (VA), and 1,778 items are other verbal items (NNG or XR).

The second step (Step II) is to extract nouns which are dependent on the verbal items. Xavier (Song and Jeon, 2008) extracts subjects (tagged as 'NP_SBJ') and objects (tagged as 'NP_OBJ') from the *Sejong* Korean Treebank. After that, nouns that do not appear on the wordnet (e.g. CoreNet, KorLex, and U-WIN) are excluded, because it is almost impossible to calculate their SPS/SA without reference to the lexical hierarchy. The next step (Step III) is to collect the Lowest Common Subsumers of each verbal item, building upon the model presented in the previous

Table 1. Basic measures (subjects)

Step	Item	CoreNet	KorLex	U-WIN
I	# of verbal entries (A)	3,701		
	# of verbs	1,923		
	# of verbal items	1,778		
II	# of tokens (B)	27,108	27,459	27,042
	# of types (C)	14,984	15,345	15,015
	tokens per verbal entry (B/A)	7.3245	7.4193	7.3067
	types per verbal entry (C/A)	4.0486	4.1462	4.0570
	type/token ratio (C/B)	55.28%	55.88%	55.52%
III	# of collected LCSs (D)	48,169	33,060	22,390
	LCSs per verbal entry (D/A)	13.0151	8.9327	6.0497
IV	SPS: mean(σ)	.0158(.0006)	.0214(.0008)	.0364(.0013)
	SA: mean(σ)	.0598(.0036)	.0723(.0041)	.1059(.0051)

Table 2. Basic measures (objects)

Step	Item	CoreNet	KorLex	U-WIN
I	# of verbal entries (A)	2,760		
	# of verbs	1,447		
	# of verbal items	1,313		
II	# of tokens (B)	27,044	27,365	26,899
	# of types (C)	18,189	18,609	18,144
	tokens per verbal entry (B/A)	9.7986	9.9149	9.7460
	types per verbal entry (C/A)	6.5902	6.7424	6.5740
	type/token ratio (C/B)	67.26%	68.00%	67.45%
III	# of collected LCSs (D)	46,052	32,787	22,259
	LCSs per verbal entry (D/A)	16.6855	11.8793	8.0649
IV	SPS: mean(σ)	.0125(.0005)	.0165(.0007)	.0263(.0011)
	SA: mean(σ)	.0561(.0041)	.0667(.0045)	.0944(.0057)

section. The final step (Step IV) is to measure SPS and SA, and find the top one among the SAs for each verbal item.

4.2 Basic Measures

The basic measures of the outcomes produced in the workflow presented in the previous subsection are provided in Table 1 and 2. These basic measures are analyzed as follows.

First, in Step II of Table 1 and 2, there is no significant difference in values across the three lexical hierarchies. All these three hierarchies include almost the same tokens and types of subject and object nouns with reference to the *Sejong* Korean Treebank, and the type/token ratios are also almost the same (about 55% for subjects and about 67% for objects). This implies that the three lexical hierarchies contain almost the same size of words in Korean.

Second, in contrast to the very similar results from Step II, the values in Step III are quite different from each other depending on the lexical hierarchy adopted. The numbers of LCSs obtained with CoreNet are the largest, those with U-WIN are the smallest, and those with KorLex are in-between. This difference basically reflects the characteristics of the three WordNets, namely, how each lexical hierarchy forms the tree structure. The tree structure of CoreNet is fairly flat, whereas the other two have relatively deep tree structures. The numbers of LCSs in Step III indicate how narrowly each lexical hierarchy captures, from the conceptual point of view, the distribution of arguments derived from the running texts in Korean. Roughly speaking, the smaller the value is, the more representative each node is in a lexical hierarchy. The results from Step III given in Table 1 and 2 show that U-WIN captures the semantic relationship between verbal items and their dependents using the least number of concepts.[9]

Third, the difference in Step III also has an influence on the measures in Step IV: U-WIN has the highest SPS and SA values. This is largely because U-WIN has the smallest branch nodes applied for calculating the Kullback-Leibler Divergence (presented in (13)). In other words, the smaller the number of LCSs is, the higher the mean of SAs is.

4.3 Frequency

This subsection looks at the relevance of SA to frequency. This analysis is made in terms of four factors that can potentially be correlated with each other. The first two are concerned with verbal items; one is (i-a) the frequency of verbal items themselves, and the other is (i-b) the type/token ratio of subject/object nouns of

---

9 It should be noted that we do not intend any overall comparison between the three hierarchies in this paper, and, needless to say, the relevant discussion in this paper is valid only in the confines of the data covered so far.

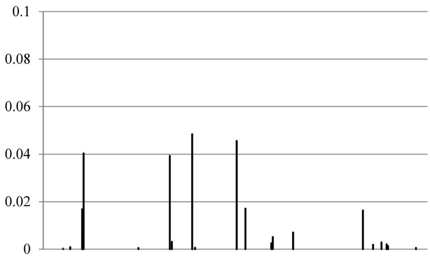


Figure 1. SAs of *masi* ‘drink’

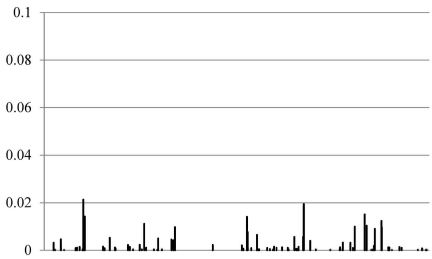


Figure 2. SAs of *ilk* ‘read’

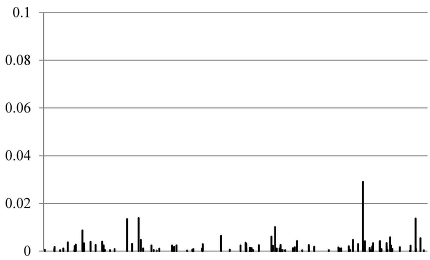


Figure 3. SAs of *chac* ‘find’

Table 3. SA

verb	SA	synset/index
masi ‘drink’	.0484	liquid/256583
ilk ‘read’	.0213	object/144215
chac ‘find’	.0291	UWIN/435936

verbal items. The other two include (ii-a) the size of LCSs and (ii-b) the value of each SA. These four factors were evaluated with reference to all three lexical hierarchies. The findings are as follows:[10] First, as is expected, the high, middle, and low frequency items (i-a) are proportionally related to the sizes of LCSs (ii-a). Second, regarding the correlation between (i-a) and (ii-b), the result shows that verbal items that appear less frequently have full range of values, whereas the frequently used verbal items tend to be associated with relatively low values of SA. Finally, there seems to be no clear correlation 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.

4.4 Selectional Associations

Figure 1 to 3 indicate the distributional properties of SAs of verbal items exemplified in (5): namely, *masi* ‘drink’, *ilk* ‘read’, and *chac* ‘find’. These charts show the results from the distribution of object nouns with respect to U-WIN. Recall that Resnik (1996) also exemplifies *drink* and *find* in English for capturing the

10 For a detailed discussion, see Song and Choe (2012b).

Table 4. Other SAs of *masi* 'drink'

SA	synset/index
.0456	beverage/301563
.0403	material/145678
.0393	liquor/232403
.0171	recognition/310701
.0169	object/144215

Table 5. Other SAs of *chac* 'find'

SA	synset/index
.0139	living things/202068
.0137	animal/88506
.0135	human/191137
.0102	space/34205
.0087	object/144215

contrastive distribution in SA, and U-WIN involves the highest SA values in Table 1 (subjects) and Table 2 (objects). The verbal items exemplified in (6) have their highest SA values as given in Table 3.

Figure 1 stands in stark contrast to Figure 3, and Figure 2 is somewhere between them. In each figure, the number of bars is the same as the number of LCSs, which represents how many synsets are associated as object arguments with respect to the verbal item. The more bars a chart has, the more LCSs were collected with respect to the verbal item. On the other hand, the height of bars stands for the value of SA. This means that 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 in Figure 1, but they are relatively taller than those in Figure 2 and Figure 3. 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 3, mostly short, which implies *chac* 'find' can co-occur with a wide variety of nouns but their relationships are quite looser.

The difference between *masi* 'drink' and *chac* 'find' can also be found in the list of candidates that are ranked below the top one, which are given in Table 4 and Table 5, respectively. The closely associated synsets with *masi* are relatively concrete and are cognitively relevant to an action of drinking (e.g. 'beverage' and 'liquor'), whereas those with *chac* are relatively abstract (e.g. 'living things', etc.).

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 contains lexical information at various linguistic levels, including selectional restrictions of verbal items. The comparative analysis of this study checks out how well the Selectional Association 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. They are defined as shown in (14), respectively.

(14) a. $precision = \frac{tp}{tp+fp}$

b. $recall = \frac{tp}{tp + fn}$

c. $F-measure = \frac{2 \times precision \times recall}{precision + recall}$

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 Lowest Common Subsumers 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 6 for the ease of exposition.

Table 6. True/False Positive/Negative

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

Figure 4 and Table 7 summarize our quantitative evaluation. On the one hand, Figure 4 indicate the growth of the three evaluation measures depending on using different values in hill climbing (i.e. *m* for iteration and *n* for threshold). Precision does not show such a difference depending on using different values, whereas recall

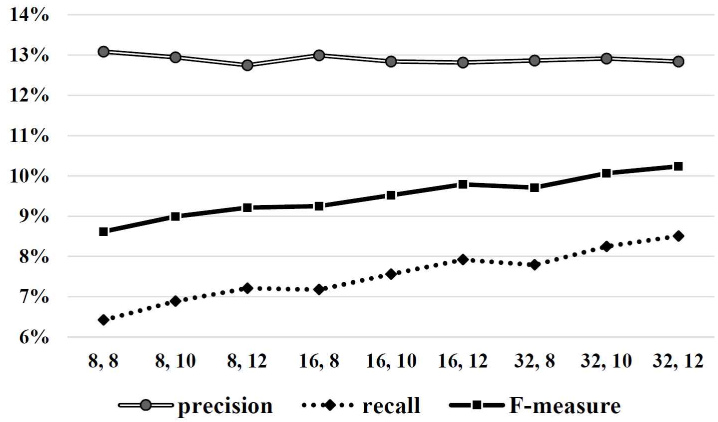


Figure 4. Measures with different values of *m* and *n*

Table 7. Quantitative evaluation

		CoreNet	KorLex	U-WIN
subjects	precision	16.01%	22.12%	8.67%
	recall	16.43%	13.98%	4.74%
	f-measure	16.21%	17.13%	6.12%
objects	precision	8.50%	13.02%	12.98%
	recall	9.86%	9.08%	8.99%
	f-measure	9.13%	10.70%	10.62%

grows up. Thus, the bigger the values are, the higher the f-measures are. Nonetheless, the gap between the highest one and the lowest one does not look statistically significant (about 1.62%). Recall that the current study makes use of m=32 and n=16 for hill climbing, which are chosen following the result of this preliminary experiment.

On the other hand, Table 7 gives the evaluation measurement conducted by the formulae in (14) and the definition in Table 6. It turns out the measures are pretty low, the f-measures being less than 20%, 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 the Korean wordnets and the *Sejong* Electronic Dictionary in the first place. It is true that the lexical hierarchies can be built upon different theoretical assumptions. Take, for example, KorLex and see how it matches with the *Sejong* Electronic Dictionary. The ontologies in the *Sejong*

Electronic Dictionary and KorLex are much different from each other (Bae *et al.*, 2010), so a comparison and evaluation should be done after the mapping between the two heterogeneous ontologies is properly established.[11] Another reason for the poor evaluation results, which is basically the same problem as the first, is that the terms used in the KorLex and *Sejong* 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.[12] 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 accordingly.

5.2 Human Evaluation

In order to check whether the outcomes calculated in the current work sound plausible to Korean native speakers, we conducted an experiment with human subjects. We randomly selected 50 verbal items for subjects and objects, respectively (100 items, in total). Out of the final result with the SA values taken from each lexical hierarchy, we extracted the word associated with the value and then created examples. Thus, the experiment exhibits 300 stimulus items to the human subjects. There were eleven participants for the experiment, and they were asked to respond to a set of a noun plus a verbal item. For this experiment, we made use of a five-point Likert scale, and the experiment was carried out using PsychoPy (Peirce 2007; 2009). This scale was meant to capture a five-way distinction in acceptability, with 1 being labeled 'least natural', 5 being labeled 'most natural', and the midpoint 3 being labeled 'so-so'. The result of the experiment is presented in Table 8. The mean value is 3.776, and the standard deviation is 0.0198. The three lexical hierarchies do not show such a significant difference from each other. While a more rigorous experiment may be necessary for a more thorough evaluation, these initial results indicate that the created outcomes sound fairly plausible to native speakers.

---

[11] See Bae *et al*. (2010) which directly addresses the issue.

[12] KorLex takes the former, and the *Sejong* Electronic Dictionary takes the latter.

Table 8. Human evaluation

		mean	σ
subjects	CoreNet	3.4964	0.0501
	KorLex	3.5247	0.0503
	U-WIN	3.6435	0.0487
	subtotal	3.5548	0.0287
objects	CoreNet	3.9344	0.0472
	KorLex	4.0418	0.0458
	U-WIN	4.0218	0.0459
	subtotal	3.9994	0.0267
total		3.7776	0.0198

Table 9. Parsing evaluation

	baseline	**KorLex**	**U-WIN**
coverage	87.4%	87.4%	87.4%
precision	43.2%	43.11%	43.2%
accuracy	95.59%	95.52%	95.59%

5.3 Evaluation with a Stochastic Parser

As an extrinsic test to verify if these outcomes (in particular, SA values) can be used to improve some language processing systems, we applied the outcomes to a stochastic parser. The parser that we used for this purpose is based on the CKY algorithm and PCFG rules acquired from the *Sejong* Korean Treebank. We added a lexical hierarchy-based weight to the PCFG rules and then looked at the evaluation measures in comparison with the baseline. The evaluation method used in this work is Parseval.[13] The experimental result is provided in Table 9. This result indicates that using a lexical hierarchy does not contribute to a better performance of stochastic parsing. In fact, this was also pointed out by several previous studies that delve into the correlation between word sense disambiguation and statistical human language processing (Carpuat and Wu 2005; 2007a). The main reason is that the language model used for stochastic processing inherently involves information related to word sense disambiguation. Therefore, adding information taken from a lexical hierarchy to the language model ends up as a kind of redundancy. The current experiment also confirms that this redundant information does not improve stochastic processing for Korean.

Nonetheless, within the context of statistical processing, some previous studies report that word sense disambiguation improves statistical language applications

13 One reviewer correctly pointed out that Dependency Evaluation would be more appropriate for testing selectional preferences. But of course choosing a different method presupposes different kind of parsers, and it would be an interesting topic for further research.

when a more elaborate strategy is applied (Carpuat and Wu 2007b; Chan *et al*. 2007). So for now it is suffice to say that it remains to be seen how lexical hierarchies and the SA values for the verbal items can contribute to building natural language systems like rule-based parsers and generators, supertagger, etc.[14]

5.4 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 some 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 to assume that ‘roll up’ is more connected to ‘arm’ rather than ‘eight’ as shown in ‘roll up one’s sleeves’, the outcomes provide only *phal* ‘eight’ as the SA 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 stay in situ in the constructions. We had tried to get rid of the form ‘*-key/tolok ha*’ from the observed data and repeated the experiment, 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. We tentatively concluded that the variation in form-meaning mapping in Korean causatives requires a further systematic inquiry in a corpus-oriented way.[15]

Finally, two closely related words sometimes are positioned far from each other within the hierarchy, which eventually creates a problem. For example, michi ‘exert’

---

[14] Thanks are due to Prof. Aesun Yoon (p.c.) for pointing out this possibility.

[15] For more information regarding this issue, see Oh (2013).

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 related to each other in terms of their senses, but they are not in the sister relation with each other in, for example, KorLex; the former is specified as an action, while the latter is classified as a kind of abstract concept. Since the two words cannot form an integrated concept at an appropriate lower level for the verbal item *michi* 'exert' in the current processing model, we cannot construct the pattern like 'exert an influence on' from our results.[16]

6. Summary

In this paper, we calculated the Selectional Preference measurements between verbal items and the classes of their co-occurring nouns. The Selectional Preference Strength and the Selectional Association was automatically measured with reference to two types of Korean language resources; (i) three lexical hierarchies (CoreNet, KorLex, and U-WIN), and (ii) the *Sejong* Korean Treebank as the development corpus. The acquisition model is grounded upon the Lowest Common Subsumer that represents the closest common ancestor node for the given two nodes within the hierarchy. SPS and SA are defined by Kullback-Leibler Divergence, and its value is derived based on the collection of LCSs. The results were evaluated with reference to the *Sejong* Electronic Dictionary which has been manually constructed. The results were also tested with human evaluation and parsing evaluation using a stochastic parser. In addition, the current work examined the outcomes in a qualitative manner from several points of view.

16 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 (Prof. Aesun Yoon, p.c.)

References

Alsina, Alex, Joan Bresnan, and Peter Sells. 1996. Complex predicates: structure and theory. In Alex Alsina, Joan Bresnan, and Peter Sells (eds.), Complex Predicates: 1-12. CSLI Publications, Stanford.

Bae, Sun-Mee, Kyoungup Im, and Aesun Yoon. 2010. Mapping heterogenous ontologies for the HLP Applications - *Sejong* semantic classes and KorLex noun 1.5. *Korean Journal of Cognitive Science* 21: 95-126.

Bae, Young-Jun and Cheol-Young Ock. 2013. Semantic Analysis of Korean Compound Noun using Lexical Semantic Network(U-WIN). *Journal of KIISE*: *Software and Applications* 40(12): 833-847.

Carpuat, Marine and Dekai Wu. 2005. Word sense disambiguation vs. statistical machine translation. In *Proceedings of the 43rd Annual Meeting on Association for Computational Linguistics*: 387-394.

Carpuat, Marine and Dekai Wu. 2007a. Context-dependent phrasal translation lexicons for statistical machine translation. In *Proceedings of Machine Translation Summit XI*: 73-80.

Carpuat, Marine and Dekai Wu. 2007b. Improving Statistical Machine Translation Using Word Sense Disambiguation. In *Proceedings of EMNLP-CoNLL 2007*: 61-72.

Chae, Hee-Rahk. 1996. Properties of *ha-* and light predicate constructions. *Language Research* 32: 409-476.

Chan, Yee Seng, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of 45th Annual Meeting of the Association for Computational Linguistics*: 33-40.

Choe, Jae-Woong, Sachiko Shudo, and Yasunari Harada. 2012. A contrastive study on the adnominal case constructions in Japanese and Korean based on relative frequency of '-no' vs. '-ui'. *Language Information* 14: 151-178.

Han, Chung-hye, Na-Rare Han, Eon-Suk Ko, and Martha Palmer. 2002. Development and evaluation of a Korean treebank and its application to NLP. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation*: 1635-1642.

KAIST Korterm Center. 2005. *CoreNet*: *Core Multilingual Semantic Word Net*. Korterm Series 13-2. KAIST Press.

Kim, Jong-Bok and Jaehyung Yang. 2006. Coordination structures in a typed feature structure grammar: formalization and implementation. *Lectrue Notes in Artificial Intelligence* 4139: 194-205.

Lesk, Michael. 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*: 24-26.

Lim, Jihui, Hoseop Choe, and Cheolyoung Ock. 2008. Automatic Construction of Syntactic

Relation in Lexical Network(U-WIN). *Journal of KIISE*: *Software and Applications* 35(10): 627-635.

Oh, Eunjeong. 2013. Transitivity of benefactive verbs: evidence from a typological and corpus-based study. *Language and Linguistics* 60: 133-158.

Park, Heum, 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*: 14-21.

Peirce, Jonathan W. 2007. PsychoPy - Psychophysics software in Python. J*ournal of Neuroscience Methods* 162(1): 8-13.

Peirce, Jonathan W. 2009. Generating stimuli for neuroscience using PsychoPy. *Frontiers in Neuroinformatics* 2. (doi: 10.3389/neuro.11.010.2008.)

Resnik, Philip. 1995. Using information content to evaluate semantic similarity. In *Proceedings of the 14th International Joint Conference on Artificial Intelligence*: 448-453.

Resnik, Philip. 1996. Selectional constraints: an information-theoretic model and its computational realization. *Cognition* 61(1-2): 127-159.

Song, Sanghoun and Jae-Woong Choe. 2012a. Automatic Construction of Verbal Type Hierarchy for Korean. *Language and Linguistics* 54: 201-238.

Song, Sanghoun and Jae-Woong Choe. 2012b. Calculating selectional preferences of transitive verbs in Korean. In *Proceedings of the 26th Pacific Asia Conference on Language, Information and Computation*: 89-98.

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

Yoon, Aesun, Soonhee Hwang, Eunryoung Lee, and Hyuk-Chul Kwon. 2009. Construction of Korean Wordnet「KorLex 1.5」. *Journal of KIISE*: *Software and Applications* 36(1): 92-108.

Sanghoun Song
Dept. of Linguistics, Korea University.
145, Anam-ro, Seongbuk-gu, Seoul,
136-701 Korea.
E-mail: sanghoun@uw.edu

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

Received: 2014. 07. 25
Revised: 2014. 08. 20
Accepted: 2014. 08. 20