*Linguistic Research* 32(1), 253-280
DOI: 10.17250/khisli.32.1.201504.009

# Automatic acquisition of "noun+verb" idiomatic compounds in Korean*

**Sanghoun Song**
**(Nanyang Technological University)**

**Song, Sanghoun. 2015. Automatic acquisition of "noun+verb" idiomatic compounds in Korean.** *Linguistic Research* 32(1), 253-280. The state-of-the-art skills of computational linguistics pay attention to lexical semantics, because it has a potential to be used to improve language processing systems in terms of coverage as well as accuracy. In particular, utilizing multiword expressions is importantly regarded as one of the components to foster performance of language applications. Handling these expressions is particularly crucial in multilingual processing, such as machine translation. Amongst a variety of multiword expressions, the present study investigates "noun+verb" idiomatic compounds in Korean. These compounds are made up of a verb plus the verb's syntactic object, and what the combination of the two words conveys is not equivalent to the sum of the meanings of the parts. In order to acquire the "noun+verb" idiomatic compounds in Korean in a fully automatic way, the current work exploits a syntax-annotated corpus (i.e. treebank) and three lexical hierarchies in Korean. The current work extracts the syntactic patterns from the development corpus (the *Sejong* Korean Treebank), calculates the selectional preferences each verbal item has with its objects, and identifies the idiosyncratic items with reference to the three lexical hierarchies (CoreNet, KorLex, and U-WIN). The result includes 548 idiomatic compounds, 70% of which are evaluated as satisfactory. **(Nanyang Technological University)**

**Keywords** idiomatic compounds, multiword expressions, selectional preference, metaphor, the *Sejong* Korean Treebank, CoreNet, KorLex, U-WIN

---

* I am grateful to Prof. Jae-Woong Choe, who gave me a big hand for the present work. The earlier versions of this study have been presented at three different venues: the 16th Korea-Japan Workshop on Linguistics and Language Processing (Kyung Hee University, 2014-05-10), the 1st Conference of Korean Association for Linguistic Typology (Seoul National University, 2014-07-02), and the DELPH-IN Summit X (Tomar, 2014-07-22). I would like to express thanks to Prof. Jong-Bok Kim, Prof. Chu-Ren Hwang, Prof. Chungmin Lee, Dr. Sung-min Lee, Prof. Chai-song Hong, Prof. Geun Seok Lim, Prof. Francis Bond, Prof. Tim Baldwin, and Woodley Packard for their valuable comments on the current study. Of course, I should say that they do not necessarily agree with my argument and methodology. Comments from an anonymous reviewer are also much appreciated. Needless to say, I am solely responsible for all the remaining errors and infelicities.

## 1. Introduction

Lexical semantics attracts much attention in the recent research of computational linguistics. To my understanding, this is mainly because the state-of-the-art of computational linguistics has invented several techniques to use lexical semantics for improving performance of language applications in terms of coverage, accuracy, and speed. This paper provides a data-driven study of language computing from a standpoint of non-compositional lexical semantics. The present work delves into how to extract idiomatic expressions from a considerable size of language data in an automatic way.

The pattern of linguistic expressions the present work has an interest in is a transitive verb plus its object noun. Two well-known instances in English are provided below.

(1) a. kick the bucket: die
    b. kick the ball ('object')
    c. kick the bottle: ∦ (a), ∥ (b)
    d. *kkangthong-ul cha-ta.*
       bucket-ACC kick-DEDAL
       '(lit.) kick the bucket' ⇔ die
(2) a. hit the road: begin a journey
    b. hit the door ('object')
    c. hit the trail: ∥ (a), ∦ (b)
    d. *kil-ul chi-ta.*
       road-ACC hit-DECL
       '(lit.) hit the road' ⇔ begin a journey

These examples exhibit the basic properties of the "noun+verb" idiomatic compounds. First, idiomatic usages of human language are created by means of combination of multiple words. When two (or more) words are grouped, they can sometimes involve an extra and/or new meaning, and neither of the words is not fully responsible for the meaning.[1] Second, these expressions do not follow so-called

[1] One of the two may behave as a semantic head of the compounds, and this makes a distinction between so-called endocentric compounds and exocentric compounds (Aronoff and Fudeman,

semantic compositionality. In the idiomatic usage of (1a), the meanings that *kick* and *bucket* respectively deliver disappear, and a meaning close to an intransitive verb *die* is newly introduced. Third, the linguistic mechanism of introducing such a non-compositional meaning largely depends on metaphorical interpretation. For example, *kick the bucket* in (1a) is reminiscent of a scenario in which a person tries to commit suicide by hanging himself or herself. Fourth, the two words are quite cohesive to each other. Even if the object is replaced by other nouns that belong to the same noun class, the multiwords do not denote an idiomatic expression as shown in (1b) and (2b). Notably, there exists a variation in this mapping. While (1c) does not involve the same meaning as (1a), (2c) in which a synonym *trail* is used does. There seems to be no specific condition for this selection. Finally, the idioms are not language-universal as presented in (1d) and (2d). That is to say, different languages use different idiomatic expressions, and this property raises a necessity to pay keen attention to multiword expressions in multilingual processing in these days.[2]

All languages employ idioms, and the phrase structure consisting of a verb and its object would be one of the cross-linguistically common loci where an idiomatic interpretation arises. This paper is concerned with how to extract this kind of idiomatic expressions in Korean in a systemic way by utilizing language resources. Since such idioms contribute to meaning of the entire sentence in a non-compositional manner, they should be separately registered in the dictionary. From a viewpoint of monolingual processing, such a description in language resource is important for (i) increasing parsing accuracy and coverage, (ii) improving parse selection, and (iii) enhancing naturalness of sentence generation. From a viewpoint of multilingual processing, this lexical information also aids in (iv) producing better performance in machine translation. Thus, it is crucial to establish the list of multiword idiomatic compounds as a basis of language processing.

This paper is structured as follows. Section 2 provides some background of the present study. Section 3 gives an explanation of why the present work takes a

---

2011). What I want to lay emphasis on here is that there is no one-to-one mapping. Although an idiomatic compound is semantically endocentric, the head word is not capable of denoting the idiomatic meaning by itself.

2 Roughly speaking, multiword expressions refer to "idiosyncratic interpretation that cross word boundaries (Baldwin and Bond, 2002)". The present work defines multiword expressions as lexical units that consist of two or more words in the surface form but do not necessarily convey a compositional meaning of the individual words.

data-oriented approach to create the list of "noun+verb" idiomatic compounds. Section 4 addresses how "noun+verb" idiomatic compounds in Korean are automatically acquired by exploring two types of language resources. Section 5 discusses the properties of the result, and Section 6 concludes this paper.

## 2. Background

Table 1, created by exploring the *Sejong* Korean Treebank (*http://www.sejong.or.kr*), includes the top-10 frequent co-occurring objects of a transitive verb *mek-* 'eat' in Korean. (Notice that a word and a synset (i.e. meaning) are distinctively described in this paper: The former is italicized (e.g. *word*), and the latter is single-quoted (e.g. 'meaning').) These ten nouns can be further classified depending on the lexical semantic behaviours. Amongst them, the type that the present work is exclusively concerned with is TYPE IV, of which the instances are boldfaced in Table 1. Each type is determined by seeing which English word the verb *mek-* is translated into.

Table 1. Frequency of the co-occurring objects of *mek-* 'eat'

| rank | noun | meaning | frequency | TYPE |
|---|---|---|---|---|
| 1 | pap | 'meal' | 28 | I |
| 2 | swul | 'liquor' | 22 | III |
| 3 | cenyek | 'dinner' | 18 | II |
| **4** | **maum** | **'mind'** | **13** | **IV** |
| 5 | umsik | 'food' | 12 | I |
| **6** | **nai** | **'age'** | **9** | **IV** |
| 6 | yak | 'medicine' | 9 | III |
| **8** | **kep** | **'fear'** | **8** | **IV** |
| 8 | cemsim | 'lunch' | 8 | II |
| 10 | koki | 'meat' | 7 | I |

In TYPE I (=*eat*), the verb *mek-* 'eat' can be directly translated into *eat*, as presented below.

(3) pap/umsik/koki-(l)ul mek-ta
meal/food/meat-ACC eat-DECL
'**eat** a meal/some food/meat'

In this type, the relation between the verb *mek-* 'eat' and its objects is semantically the default. That is, the object nouns, such as *pap* 'meal', *umsik* 'food', and *koki* 'meat', belong to the noun class tagged as 'food' (or 'the edibles'), and the noun class 'food' represents the default property of an event of eating. This relation is often called Selectional Preference, which refers to the degree of correlation between two co-occurring linguistic categories (Resnik, 1996; Erk, 2007; Song and Choe, 2014). In other words, 'food' is the most preferred noun class of the verb *mek-* 'eat'.

The nouns of TYPE II (≈*eat*) are slightly different from those of TYPE I, but the verb *mek-* is still translated as *eat* in English.

(4) cenyek/cemsim-ul mek-ta
dinner/lunch-ACC eat-DEDAL
'**eat** dinner/lunch'

The two words *cenyek* 'dinner' and *cemsim* 'lunch' are not a name of the edibles, but they can be interpreted as a proper object of an eating event in a sense. Given that this relation is quite straightforward in the concept structure of human language, using *eat* as the corresponding translation in English is plausible.

In TYPE III (≥*eat*), the corresponding words are different as glossed in (5).

(5) swul/yak-ul mek-ta
liquor/medicine-ACC eat-DECL
'**drink** liquor' / '**take** a medicine'

While *mek-* in Korean can denote an action of drinking as well, English prefers *drink* to *eat* as a verb taking the drinkables as its object. Similarly, if the object has a property of 'medicine', *take* is more preferred in English. Nevertheless, the use of *eat* in this case does not necessarily result in a misinterpretation. For instance, when a non-native speaker says 'eat liquor' or 'eat a medicine', most English native speakers may understand what the speaker wants to express. Notably, the same does not go for TYPE IV (≠*eat*). When it comes to TYPE IV, *mek-* does not directly denote an action of eating, and the translations are different in English. Whilst the mistranslation in TYPE III is still understandable to English native speakers, the

infelicitous use of *eat* in this case brings about a miscommunication. The appropriate translations are provided in (6).

(6) a. maum-ul mek-ta
mind-ACC eat-DECL
'decide' / '#eat a mind'
b. nai-lul mek-ta
age-ACC eat-DECL
'get old' / '#eat an age'
c. kep-ul mek-ta
fear-ACC eat-DECL
'be frightened' / '#eat a fear'

In other words, the "noun+verb" compounds in TYPE IV are idiomatic, and thereby the meanings are conveyed in a non-compositional way.[3] As is well-known, compositionality has been regarded as one of the fundamental properties of human language: Meaning of a phrase is equivalent to the sum of each component's meaning of the phrase. In contrast, idioms are traditionally said to be non-compositional (Chomsky, 1980).[4]

The basic properties of the "noun+verb" idiomatic compounds are as follows: First, these expressions do not follow the principle of semantic compositionality. Instead, these expressions are interpreted only in a metaphorical manner. That is to say, metaphor breaks into the semantic compositionality and takes priority in meaning representation. Second, there are very few or no alternative words as exemplified in (7): Even a synonym cannot take the place.

(7) a. maum/#cengsin-ul mek-ta
mind-ACC eat-DECL ⇔ decide
b. nai/#yenlyeng-(l)ul mek-ta

---

3 There is a slightly different view to this. Nunberg, Sag, and Wasow (1994) dwell on the details of semantic properties of idioms in English. They reveal that idioms are sometimes compositional.

4 Previous literature provides several diagnostic tools to see whether meaning of idioms is preserved. These include passivization (Nunberg *et al*, 1994), relativization (Fabb, 1990), and raising and control predicates (Kim and Sells, 2008). Along the line of the studies, the current work assumes that idiomatic expressions are mostly non-compositional, but not completely.

age-ACC eat-DECL ⇔ get old

c. kep/#kongpho-(l)ul mek-ta
fear-ACC eat-DECL ⇔ be frightened

So to speak, the relation between two components in the idiomatic compounds is word-to-word. Thus, the compounds have to be registered in the dictionary as a single entry. This type of lexical entries, consisting of two or more words, is often called multiword expressions (Sag *et al*. 2002; Baldwin and Kim, 2010, among many others). Third, the relationship between the co-occurring items is language-specific. Just as idiomatic compounds in English (e.g. *kick the bucket*) do not directly correspond to the corresponding sequence of words in Korean, idiomatic compounds in Korean (e.g. *maum-ul mek-ta* 'decide') cannot be literally translated.

The research question the present work raises is how we can acquire such an idiomatic meaning each verb has with respect to its co-occurring objects in a systemic way. More specifically, how can we acquire the "noun+verb" idiomatic compounds (i) on a comprehensive scale, (ii) in an automatic way utilizing language resources, and (iii) for a practical purpose? The following section deals with the methodology.

## 3. Methodology

There are several ways to investigate idiomatic compounds. Some of them do not see language data, and some of them do not use computational methods. To take a representative instance of lexicographical studies in Korean, the *Sejong* Electronic Dictionary specifies which lexical item involves which idioms.[5] To take another instance, Kim *et al*. (2013) provide a corpus study of eating verbs and drinking verbs in Korean with reference to online texts. These studies have a significance in that each idiomatic compound can be specifically described and some novel ways of expressing metaphor can be detected in the description.[6]

---

5 The language resources built up in these days normally include multiword idiomatic expressions. For example, in WordNet 1.7 (Fellbaum, 1998), 41% of the entries are multiword (Baldwin and Bond, 2002).

6 For instance, Kim *et al.* (2013) provide an example *sarang-ul mek-ta* 'love-ACC eat-DECL' with respect to metaphorical meaning extension in Korean. Such an expressions raises an interesting

Unlike the previous studies, the present study acquires the idiomatic compounds in a corpus-based and automatic way. The merits of the current method are as follows: First, this way of extraction facilitates creating a more comprehensive generalization of idiomatic compounds. Note that the previous studies mostly look at very few idiomatic compounds, and thereby they are less likely to present the whole picture of how idiomatic compounds linguistically behave in a language. Second, since the result based on the current method is scalable, we can employ the result directly for stochastic language processing. In order for language applications to reflect on real usages of idiomatic compounds, it is essential to see a large size of naturally occurring texts. Third, a computational method to exploit the data on a large scale enables us to see how many idiomatic compounds are used in the language. In sum, the current method using language resources gives an overall explanation of which idiomatic expressions are used and which linguistic preferences are found.

## 3.1 Selectional Preferences

A list of idiomatic compounds can be acquired after identifying selectional preference strength of verbal items with respect to a lexical taxonomy. Selectional preferences (or selectional restrictions) have been studied in generative theory of grammar for a long time, but they are computationally modeled by Resnik (1996). In computational linguistics, selectional preferences are roughly defined as a relative entropy indicating how much interrelationship an entity has with another entity. They serve as a handy tactics for a number of language applications, including semantic role labeling, word sense disambiguation, syntactic disambiguation, parse reranking, recognizing text entailment, and so forth.

One of the phrase structures in which selectional preference is well-found is the syntactic combination between a transitive verb and its object noun, because the object serves to provide a clue to identify the meaning of the verb. For example, *speak* in English has a different meaning, depending on whether it takes a language name as its object or not. If *speak* is used as a transitive verb (e.g., “Kim speaks

---

research topics in the study of metaphor, but they sparsely occur in running texts. Because the present work follows data-oriented method of metaphor extraction (Mason, 2004; Shutova *et al*, 2012), such an expression is not dealt with in the current system.

English."), the corresponding translation differs in different languages. Selectional preference between a verb and its object is clearly revealed in Korean, too (Song and Choe, 2014). The verb in (8) *masi-* 'drink' has a specific selectional preference with respect to its object: The object is a kind of beverage (or something to inhale), and otherwise the phrase does not sound natural.

(8) maykcwu/#chayk-ul masi-ta
beer/book-ACC drink-DECL
'... drink beer' / '#... drink a book'

In this context, we can identify the most preferred noun class associated with a transitive verb (a.k.a. association strength). The most strongly associated class technically means the Lowest Common Subsumer that has the highest value of selectional preference with a verbal item (Resnik, 1996), and it distributionally represents the semantic properties of the verbal item. For instance, 'beverage' stands for the lexical semantics of *masi-* 'drink.'[7] In other words, selectional preference strength in the present work indicates how strongly a verb constrains its objects.

This notion of selectional preference strength is also important in the study of metaphor and idioms at least within the context of statistical approach to human language (Mason, 2004; Shutova *et al*, 2012). Technically speaking, metaphor extraction on a comprehensive scale is not realizable until selectional preferences are measured by exploring (i) a considerable size of running texts and utilizing (ii) a linguistic knowledge base such as WordNet (Fellbaum, 1998).

One question that can be raised here is why calculation of selectional preference strength has to be carried out before extracting idiomatic compounds.[8] This is mainly because the strongest selectional preference provides a clue to identify idiomatic

---

7 A Lowest Common Subsumer refers to a concept which has the shortest distance from the two or more concepts in a lexical hierarchy.

8 Both of them serve as an important component in computational approach to lexical semantics, but their properties are different. First, selectional preferences have to do with a class of nouns, while idioms are constituted when specific words are grouped. Second, two or more words with the strongest selectional preference normally convey a meaning in a compositional way, while idioms do not. Third, the cognitive process is metaphorically performed in the use of idioms, while selectional preferences have less to do with metaphor. Finally, the strongest selectional preference between two linguistic items is predictable across languages, while idioms are expressed language-specifically.

compounds. If a noun has less to do with a concept in the lexical taxonomy (i.e. WordNet) with the strongest selectional preference, then we can assume that the noun is used in an idiosyncratic manner. Recall that what is normal has to be established prior to identifying what is not normal. In other words, identifying the most meaningful relation between a verb and its co-occurring objects (i.e. association strength) has to be completed before finding a set of atypical meaning relations between them.

## 3.2 Data

For the present study, two types of language resources are utilized: One is a development corpus, from which the "noun+verb" compounds are extracted. The other is a lexical taxonomy (i.e. WordNet) in which words are classified in a hierarchical order.

On the one hand, the current work makes use of the *Sejong* Korean Treebank as the development corpus. Since the linguistic pattern the current work has an interest in consists of a verb and its syntactic object, exploring syntax-tagged data is preferable. Other language resources could be used for this purpose: We could process a raw corpus with a dependency parser and exploit the parse result. Otherwise, we could extract a sequence of object and verb from a POS-tagged corpus, using the accusative marker in Korean (i.e. (*l*)*ul*) as a pivot in search. Given that the size of syntax-tagged corpora is normally smaller than the other types of corpora, using these methods may have some merits in theory. Nonetheless, the current study does not use them for a practical reason. First, there are several ongoing projects of building up a dependency parser in Korean, but to my knowledge the systems have not yet been comprehensively tested.[9] Second, since morphological marking such as accusatives does not necessarily coincide with syntactic function such as objects, using a POS-tagged corpus is not an optimal choice for this study: (i) NPs in Korean are sometimes null-marked, (ii) An adverbial expression can appear between the object and the verb. (iii) Furthermore, there are more than a few long distance dependency constructions. Given these properties of Korean syntax, appearance of the accusative marker (*l*)*ul* is neither a necessary

[9] If the corpus is big enough to make up for the challenging parts of a dependency parser, we can try using this method. This is left to future work.

condition nor a sufficient condition for the object function. As of now, using a treebank is the most available and reliable way for identifying the "noun+verb" compounds in a systemic way. Because the size of the development corpus is still significant, the present work takes the biggest treebank amongst the readily available ones: The *Sejong* Korean Treebank (about 0.8 million words).

On the other hand, lexical hierarchies for Korean words are needed in order to discriminate whether a "noun+verb" compound is idiomatically used or not. The four types presented in Section 2 can be automatically classified by using lexical hierarchies. The nouns that belong to TYPE I sufficiently come under the noun class that can represent the default meaning of the verb. The nouns that belong to TYPE II and TYPE III are presumed to be somewhere near the representative noun class within the lexical hierarchy. In contrast, the nouns of TYPE IV are less likely to come under the representative noun class, because the meaning of the "noun+verb" compound in this case is atypical (i.e. not the same as the sum of each word). Amongst the available lexical hierarchies in Korean, the present work employs three resources, viz. CoreNet (KAIST Korterm Center, 2005), KorLex (Yoon *et al*., 2009), and U-WIN (Lim *et al*., 2008; Bae and Ock, 2013). This study explores the results from the three hierarchies and tries to find an optimal solution.

## 3.3 Computation

The model of computing selectional preferences and acquiring "noun+verb" idiomatic compounds in the current work is based on a way to achieve an optimal solution via gathering tremendous partial solutions. In a sense, this method is similar to an algorithm to find approximate solutions to a problem whose exact answer can hardly be provided. This way of calculation allows us to identify an answer very close to the ideal solution though there is no evidence for believing that the answer is perfect. In addition, this way of calculation has a practical advantage in that the running time is relatively short. For this reason, the mathematical technique used in the current work is hill-climbing. This algorithm provides an optimal solution by repeatedly changing a single element of the solution. The calculation is iterated until no further improvement can be detected, and then the end result is regarded as the optimal solution.

This mathematical technique is of great use to the study of human language,

given that a large number of language resources have been produced to date. This technique is particularly useful when it is almost impossible to find the perfect solution to a problem. As is well-known, the data-based study in linguistics can never be perfect, because no matter how big a corpus is, it is a subset of an infinite set of sentences. Although our knowledge of human language is not perfect yet and probably for ever, we can create an optimal finding of a specific linguistic phenomenon by utilizing the language resources we have produced thus far. Furthermore, it is promising to use the method if it facilitates creating a linguistic finding in a relatively short time and at a low cost.

This study is along the line of this approach: The programming skill used in the current work allows us to acquire the "noun+verb" idiomatic compounds in Korean in a way of preventing the researcher's intuition from affecting the result.

## 4. Acquisition

In order to acquire the "noun+verb" idiomatic compounds in a fully automatic way, it is necessary to extract the "noun+verb" items from the development corpus. After the basic data are collected, the selectional preferences that each verb has with noun classes are measured. Building upon the preferences, the idiosyncratic "noun+verb" items are identified with respect to the lexical hierarchies. The workflow of acquiring the "noun+verb" idiomatic compounds in Korean consists of three steps. This section describes these in detail.

### 4.1 Extraction

The first step in the workflow extracts the "noun+verb" patterns from the development corpus (i.e. the *Sejong* Korean Treebank). The verbal items this study delves into include verbs, adjectives, and verbal nouns. The first two are tagged as 'VV' and 'VA' in the corpus, and the third one co-occurs with a light verb tagged as 'XSV' or 'XSA'.[10] The object nouns are annotated as 'NP_OBJ' at the phrase level in the treebank. If a node tagged as 'NP_OBJ' is found to be dependent on a

[10] Note that adjectives sometimes take an object in surface form though its argument structure can be grammatically different from the argument structure of ordinary transitive verbs.

verbal item in the parse tree, the syntactic head of the 'NP_OBJ' node is searched. All parse trees in the *Sejong* Korean Treebank are built up thoroughly in an endocentric fashion, this search routine works quite straightforwardly. Sometimes, the head of NPs cannot be explicitly identified in the parse tree. For example, coordinated NPs may have multiple heads or no head. In this case, a heuristic is used: Given that Korean is a head-final language, the head (or head-like) item is normally in the rightmost position.[11]

After the frequency table of the "noun+verb" patterns is established, the nouns in the table are compared to the lexemes in the three lexical hierarchies, including CoreNet, KorLex, and U-WIN. If a noun extracted from the treebank does not appear in the lexeme list of a lexical hierarchy, the "noun+verb" pattern is excluded. Given that three lexical hierarchies are used, three frequency tables of the "noun+verb" compounds are separately created.

## 4.2 Calculation

The second step measures the selectional preference strength of the "noun+verb" compounds. As aforementioned, calculating selectional preferences plays the critical role in determining linguistic properties of the combination between a verbal item and its argument, given that the model assumes that there is often a semantically coherent set of concepts that can take the argument position. That is to say, selectional preferences describe linguistic knowledge of plausible fillers for a verbal item's syntactic dependents, such as objects.

The most widely used method to induce selectional preferences from a corpus is the model that relies entirely on WordNet (Fellbaum, 1998). Resnik (1996) groups noun classes into semantic clusters with reference to the noun synsets in WordNet, and then induces the selectional preference strength of a verb for a particular argument by computing the divergence between two probability distributions. In this model, the unit of a particular cluster is defined as Lowest Common Subsumer, which refers to the most specific concept which is an ancestor of two different concepts within a lexical hierarchy. If there are multiple candidates for the lowest common subsumer, the candidate that results in the shortest path is chosen.

[11] This heuristic sometimes does not work correctly. Section 5 discusses the exceptional case, such as the so-called double object constructions.

Selectional preference strength is mathematically defined as formulated in (9a), in which *S* stands for "Strength", *v* stands for "Verb", and *c* stands for "Class of nouns". On the other hand, (9b) defines the selectional association that indicates the contribution of each lowest common subsumer to the verbal item's preference strength.

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

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

The workflow is as follows: The first one is collecting the lowest common subsumers of the nouns dependent on each verbal item by means of measuring the distance between two concept nodes. The lowest common subsumers are gathered using a hill-climbing technique.[12] This data collection takes several days on computer, because a large-scale calculation is required. The second one is measuring the selectional preference strength that each noun class has with each verbal item. The strength is calculated by means of information divergence (a.k.a. Kullback-Leibler Divergence (Resnik, 1996; Jurafsky and Martin, 2008)).[13] The final one is identifying the strongest selectional preference. This means finding out the maximum value out of the selectional preference strengths with respect to each verbal item.

This calculation is along the line of Song and Choe (2014), but there is one significant difference. The calculation provided in Song and Choe (2014) does not include items whose noun type is single. Since Song and Choe (2014) is exclusively concerned with co-relationship between a verbal item and a noun class, if no class of nouns is found, the class is ignored in the calculation. In other words, if a verbal item takes only a specific single noun as its object, the noun is excluded from the

12 The hill-climbing technique normally requires two parameters. One is the number of iteration, and the other is a threshold for starting the iteration. The present work makes use of 32 for the former and 16 for the latter, replicating the basis presented in Song and Choe (2014).

13 This model is widely used in statistics and pattern recognitions to measure dissimilarity between two probability distributions. Notice that this paper is, so to speak, a study of recognizing a specific pattern in human language.

list in the way used in Song and Choe (2014). Recall that the idiomatic expressions are conditioned by individual words while selectional preferences have to do with a class of words. Therefore, the selectional preferences in the current work includes both individual nouns and the classes of the nouns.

The whole steps described thus far were implemented in a way of batch processing. In order to check whether the selectional preferences are reliably measured in a correct direction, two different computers separately ran the batch processing: One computer is a 32-bit machine whose RAM capacity is 4GB, and the other is a 64-bit machine with 16GB RAM. It took about ten days for the first computer to complete the job. A programming technique of parallel processing was applied to the procedure performed on the second computer. Consequently, it took less than three days for the second computer to finish the same job. What is noteworthy is the divergence between the outputs taken from the two different computers.[14] Some differences were found in the two sets of outputs, but the divergence is quite small.[15] In order to scale the difference between the two sets of outputs, I conducted an intrinsic evaluation, following the quantitative evaluation method presented in Song and Choe (2014): Precision, recall, and F-measure are computed, comparing the current outputs to the descriptions provided in the *Sejong* Electronic Dictionary. Note that precision (i.e., the fraction of responsive instances that are extracted), recall (i.e., the fraction of extracted instances that are responsive), and F-measure (i.e., a harmonic mean of precision and recall) are the most common measures in evaluating how good a system is. The measures are provided in Table 2, in which all the differences in F-measure are less than 0.1%.

Table 2. Comparison of two sets of outputs

| | **1st trial** | | | **2nd trial** | | |
|---|---|---|---|---|---|---|
| | **precision** | **recall** | **F-measure** | **precision** | **recall** | **F-measure** |
| **CoreNet** | 11.90% | 41.04% | 18.46% | 11.92% | 41.15% | 18.48% |
| **KorLex** | 17.71% | 37.82% | 24.13% | 17.65% | 37.71% | 24.05% |
| **U-WIN** | 13.42% | 28.44% | 18.24% | 13.42% | 28.41% | 18.23% |

14 Since three different lexical hierarchies were used when extracting the "noun+verb" items and collecting the lowest common subsumers, there are six sets of outputs (3 hierarchies × 2 computers) in total.

15 This comparison was made by using a shell command *diff* on a linux system.

One interesting point is that the two computers provide exactly the same result regarding the strongest selectional preferences: They include the same list of noun classes associated with each verbal item, and the values of the strongest selectional preference are congruous. In short, even though a trivial difference is found in the partial solutions, the optimal solution produced by incrementally associating the partial solutions is the same. Table 2 indicates that the hill-climbing algorithm works for creating a data-based finding of idiomatic compounds in a right direction.

One implication Table 2 provides is the difference between precision and recall. In all rows, the recall is higher than the precision. Notice that precision measures how well the system weeds out the unwanted items, while recall measures how well the system finds the wanted items. While the *Sejong* Electronic Dictionary, compared to the retrieved instances in the current study, is a precision-based language resource, the current study that aims to find the wanted items exhaustively has more to do with recall. It is my firm opinion that the precision-based language resources and the recall-based language resources are complementary to each other for producing better performance in language processing.

## 4.3 Identification

The final step is examining kinship relations amongst the collected lowest common subsumers with reference to the location where the association strength is detected. Figure 1 is illustrative of the process of identifying idiomatic compounds.

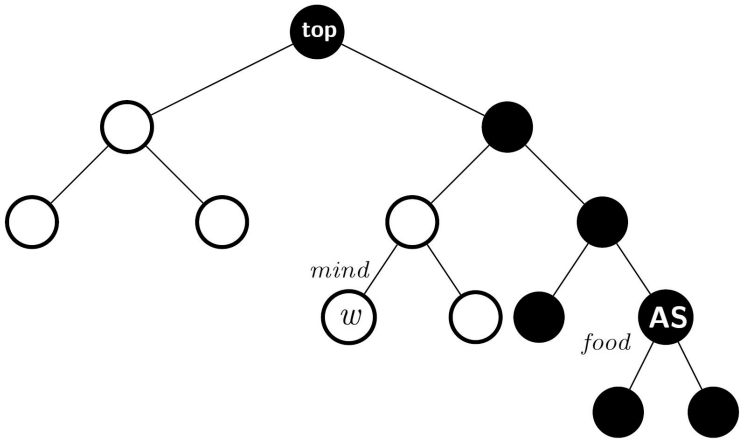


Figure 1. Kinship relations

In Figure 1, the circle marked as 'AS' (Association Strength) is assumed to be the node that has the strongest selectional preference with the verb *mek-* 'eat'. The concept node in the lexical hierarchy is named 'food'. On the other hand, the circle in which *w* is inserted represents another concept node that has a preference with the verb *mek-*. The node is named 'mind' as presented on the circle, and the lexemes under the concept node include *maum* 'mind' in Korean. The difference between the two circles is exemplified in (10).

(10) umsik/maum-ul mek-ta
food/mind-ACC eat-DECL
'eat food' / 'decide'

If we regard the 'AS' circle as the pivot to examine the kinship relation, we can classify the circles in the tree into two types: One is the direct relatives of the 'AS' circle (i.e. the black circles), and the other is its collateral relatives (i.e. the white circles). First, the children and the descendants of the 'AS' circle are direct relatives. For instance, the nouns of TYPE I presented in Section 2 are included either in the 'AS' concept node (e.g. *umsik* 'food') or its descendant nodes (e.g. *koki* 'meat'). Second, its siblings are also direct relatives though the descendants of its siblings are not. Third, its parent and its ancestors are also direct relatives. For instance, the

nouns of TYPE II (e.g. *cemsim* 'lunch') and TYPE III (e.g. swul 'liquor') are likely to belong to either one of its sibling nodes or its ancestor nodes. In contrast to the circles in black, the collateral circles in white are conceptually tangential to the circle marked as 'AS'.

In this way of classification, all lowest common subsumers collected in the previous step were examined. In order to enhance the accuracy of the result, one heuristic was additionally used. If a noun appears only once as an object of a verbal item as an object in the development corpus (i.e. the *Sejong* Korean Treebank), it was not identified as forming an idiomatic compounds with the verbal item. That is to say, the entries with an absolute frequency of 1 were filtered out. This heuristic defers to what Haugereid and Bond (2011) employ for extracting cross-lingual multiwords expressions from parallel texts.[16]

The process presented hitherto produces three sets of lexical entries consisting of a noun plus a verb with respect to three lexical hierarchies. These lexical entries are potentially evaluated as containing an idiomatic usage, but not necessarily. Each set of the entries may not be reliable in itself, mainly because none of the lexical hierarchies is necessarily seamless in terms of conceptualizing words in Korean. Notice that each lexical hierarchy has its own pros and cons, and we cannot rely entirely on any of them. For instance, the three lexical hierarchies provide the following tables with respect to the verb is *mek-* 'eat'. Note that these Tables 3 to 5 include only the top-9 entires. These tables indicate that some of the entries (i.e. non-boldfaced) have less to do with the idiomatic compounds.

[16] Haugereid and Bond (2011: 94) regard the absolute frequency number as a confidence score: "The larger, the more accurate and reliable the translation probabilities, 1 is the lowest score."

Table 3. CoreNet

| | noun | SPS | T |
|---|---|---|---|
| 1 | cenyek | .0044 | II |
| **2** | **maum** | **.0031** | **IV** |
| **3** | **nai** | **.0020** | **IV** |
| **4** | **kep** | **.0018** | **IV** |
| 5 | achim | .0015 | II |
| 6 | achimpap | .0008 | II |
| **7** | **yok** | **.0006** | **IV** |
| **7** | **ay** | **.0006** | **IV** |
| **7** | **ton** | **.0006** | **IV** |

Table 4. KorLex

| | noun | SPS | T |
|---|---|---|---|
| **1** | **maum** | **.0044** | **IV** |
| **2** | **kep** | **.0026** | **IV** |
| **3** | **sayngkak** | **.0022** | **IV** |
| **4** | **ton** | **.0009** | **IV** |
| **4** | **yok** | **.0009** | **IV** |
| **4** | **ay** | **.0009** | **IV** |
| 7 | ppwuli | .0006 | I |
| 8 | pan | .0005 | etc. |
| 8 | ocinge | .0005 | I |

Table 5. U-WIN

| | noun | SPS | T |
|---|---|---|---|
| 1 | swul | .0012 | III |
| **2** | **maum** | **.0072** | **IV** |
| **3** | **sayngkak** | **.0053** | **IV** |
| 4 | so | .0043 | etc. |
| 5 | yak | .0042 | III |
| **6** | **kep** | **.0037** | **IV** |
| 7 | cengto | .0034 | etc. |
| **7** | **nai** | **.0034** | **IV** |
| 9 | mwul | .0022 | III |

The better way to produce more reliable entries of the "noun+verb" idiomatic compounds is to draw an intersection amongst three different sets of entries. Figure 2 represents the intersection.

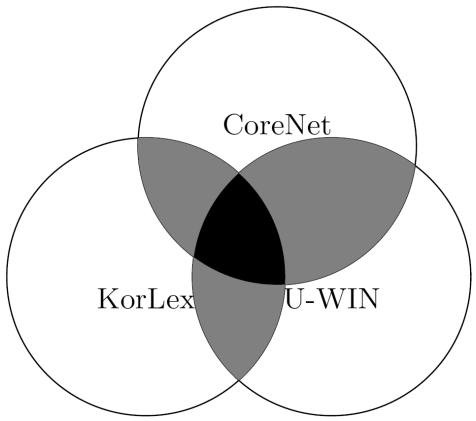


Figure 2. Intersection

maum 'mind' : 'decide'
sayngkak 'thought' : 'think'
nai 'age' : 'get old'
kep 'fear' : 'be frightened'
ton 'money' : 'be bribed'
yok 'abuse' : 'be blamed'
ay 'difficulty' : 'be troubled'

The sections in gray include the entries identified by at least two different lexical hierarchies, and the section in black in the middle includes only the entries licensed by all the three. Normally, the most significant criterion in computational linguistics is how much of the information the system returns is correct (i.e. precision). Thus, from a conservative standpoint, we might be better to take only the subset that shows a point of convergence amongst three lexical hierarchies. For instance, the

intersected entries in Table 3, 4, and 5 are the underlined ones in Figure 2: *maum* 'mind' and *kep* 'fear'.

As a result, the "noun+verb" idiomatic compounds are acquired in a fully automatic fashion. The result is discussed in the following section.

## 5. Result

Table 6 summarizes the result each step in the previous section produces. About 5% of the 2,760 verbal items occurring in the *Sejong* Korean Treebank are found to involve an idiomatic compound. Out of about 18,000 types of "noun+verb" entries, 548 idiomatic compounds (approximately, 3%) are identified with respect to all the three lexical hierarchies used in the current work. These portions increase to 9% and 4% if the compounds are less conservatively identified.

Table 6. Basic measures (#)

| | **CoreNet** | **KorLex** | **U-WIN** |
|---|---|---|---|
| verbal items | 2,760 | | |
| verbs (VV, VA) | 1,447 | | |
| verbal nouns | 1,313 | | |
| tokens of object nouns | 27,044 | 27,365 | 26,899 |
| types of object nouns | 18,189 | 18,609 | 18,144 |
| collected LCSs | 46,052 | 32,787 | 22,259 |
| verbal items with idioms | 236 | 360 | 305 |
| idiomatic compounds | 762 | 1,360 | 894 |
| verbal items with idioms (∩) | 137 (>2) / 254 (≥2) | | |
| idiomatic compounds (∩) | 548 (>2) / 724 (≥2) | | |

Table 6 indicates that the three lexical hierarchies provide different numbers of retrieved idiomatic compounds. The difference largely depends on the depth of each lexical hierarchy.[17] CoreNet is constructed in a relatively flat fashion, while KorLex has the deepest tree among the three hierarchies. As a consequence, CoreNet produces the smallest number of idioms (762), KorLex produces the biggest number (1,360), and U-WIN produces the number in the middle (894).

17 Notice that this difference has nothing to do with which lexical hierarchy is the best.

## 5.1 Analysis

The nouns analyzed as forming idiomatic compounds with a transitive verb *mek-* 'eat' are provided in (11).

(11) a. *maum* 'mind', *nai* 'age', *kep* 'fear', *yok* 'abuse', *ay* 'difficulty', *ton* 'money', *ttang* 'land'
b. *pan* 'half', *ppwuli* 'root', *mwulkoki* 'fish', *ocinge* 'squid'

The entries given in (11a) surely introduce idiomatic expressions when they are used as the object of *mek-* 'eat'.[18] For example, *ttang-ul mek-ta* (land-ACC eat-DECL) delivers a meaning like 'extort someone's estate'. On the other hand, the four entries provided in (11b) are rather controversial. First, *pan* 'half' seems to be wrongly extracted because it often participates in the so-called double object constructions in Korean. For example, in [$_{VP}$[$_{NP\_OBJ}$ *sakwa-lul pan-ul*] *mek-ta*] (apple-ACC half-ACC eat-DECL), *pan* in the rightmost position of the NP_OBJ constituent is not the genuine head of the noun phrase. Song and Song (2014) report that the frequency of the double object constructions in Korean is less than 0.06%. Following the data analysis, the present study regards the unwanted entries caused by such a special syntactic operation as marginal and exceptional ones. The other items are ambiguous in that they denote not only "a physical entity" but also "a food ingredient". Since all the lexical hierarchies do not regard them as a food ingredient, they are analyzed as a member of collateral relatives in the process of examining kinship relations.[19]

The workflow presented in the previous section produces 548 "noun+verb" compounds, in total. As with the most computational ways of acquiring lexical information from corpora, these compounds include both well-qualified ones and rather unsatisfactory ones. They are exemplified in the followings, respectively.

---

[18] It is true that they carry an idiomatic interpretation, but they are not in the same grammatical status. That is to say, their syntactic, semantic, and event structures are different. For instance, *maum-ul mek-ta* 'decide' usually requires a verbal complement. The contributions that the verb *mek-* makes to semantics are also different. For instance, *mek-* behaves differently in *nai-lul mek-ta* 'get old' and *ton-ul mek-ta* 'be bribed'. This difference should be researched more, and the data the present study provides can be of use to the study as distributional evidence.

[19] This implies that the current result can be used to detect a missing concept in a lexical hierarchy.

The compounds provided in (12) are relatively well-acquired.

(12) a. swum-ul ketwu-ta
breath-ACC withdraw-DECL
'die'
b. wenswu-lul kaph-ta
enemy-ACC repay-DECL
'revenge'
c. mokcheng-ul katatum-ta
vocal.cord-ACC clear-DECL
'clear one's throat'
d. tewui-lul ssis-ta
warmth-ACC wash-DECL
'make oneself feel cool'

The compounds presented in (12a-b) are one of the well-known idiomatic compounds in Korean. The meanings are metaphorically and non-compositionally (or partially compositional) conveyed. Because the literal translations of these expressions are not clearly understandable to speakers of other languages, they have to be individually registered into a dictionary. (12c) is slightly different from (12a-b) in that the translation in English has almost the same structure. This is because such an expression is heavily motivated by metaphor across languages. If there is a metaphorical similarity across languages, the linguistic expressions are also likely to similar to each other. This means that not all idiomatic expressions are necessarily language-specific. (12d) is one of the intriguing idiomatic expressions that the current work finds out. This seems to be a specific expression in Korean, and the event structure of (12d) is quite different from the that of the others (See Footnote 18). In this way, this data-based method of acquiring idiomatic expressions can locate the new metaphorical expressions that the previous studies have not yet dealt with. This implies that the present work makes a contribution to the theoretical study of idioms and metaphor.

On the other hand, there are some poorly acquired items, as presented in (13).

(13) a. mwunhwa-lul dayphyoha-ta

culture-ACC represent-DECL
'represent the culture'
b. cisik-ul sayngsanha-ta
knowledge-ACC produce-DECL
'produce a piece of knowledge'

The verbs in (13) are too generic to have a specific selectional preference. Recall that identifying an idiosyncratic usage in this study hinges on the location of the strongest preferences within a lexical hierarchy. In any lexical hierarchies used in the present work (viz. CoreNet, KorLex, and U-WIN), the noun class node which is most strongly associated with the verbs *dayphyoha-* 'represent' in (13a) and *sayngsanha-* 'produce' in (13b) is very close to the top node. As a consequence, too many nodes in the lexical hierarchy are regarded as a member of collateral relatives of the strongest node. Thus, the current method sometimes works poorly if the selectional preference has a weak statistical power. Further work should make up for this weak point. Nonetheless, these unwanted outputs do not mean that the current work has less significance. One of the most widely acknowledged methods in the construction of linguistic data is "annotate automatically, correct manually" (Marcus *et al*., 1993). When implementing the current result into a practical system of language processing, the unwanted items such as (13a-b) will be manually excluded.

## 5.2 Evaluation

This subsection presents a quantitative analysis of the instances retrieved as idiomatic compounds. Normally, conducting a quantitative test in the study of language processing requires a gold standard and a commonly used evaluation metric. The current study has neither of them: There is no language resource that can be said as a gold standard in terms of idiomatic expressions in Korean. No method has been provided to evaluate a set of idioms.[20] Currently, the most available method to see the feasibility of the retrieved instances would be using a

[20] There is an extrinsic way of evaluating a system. We can apply the retrieved instances into a practical system, such as syntactic parsing, semantic interpretation, and machine translation, and then see how much the system performance increases. This way of evaluation is left to future work.

machine-readable dictionary, such as the *Sejong* Electronic Dictionary. The comparison was already presented in Table 2 using the three basic measures (i.e., precision, recall, and F-measure). This way of an intrinsic evaluation also has a limitation in that we cannot say that the *Sejong* Electronic Dictionary was constructed especially focusing on idiomatic expressions. For this reason, one additional evaluation was carried out using a different type of dictionary.

I referred to the NAVER online dictionary service (http://endic.naver.com) consisting of the NeungYule Korean-English dictionary and the Dong-A Prime Korean-English dictionary. The main reason why I chose the Korean-English dictionaries as a comparable data source is that the current study ultimately aims to contribute to Korean-English machine translation. Note that the dictionary service does not provide a gold standard, either. Suffice it to say that this evaluation enables us to examine whether the result is relatively satisfactory.

The evaluation method is as follows: (i) When I search an instance retrieved by the current work into the dictionary, if the instance is registered as an idiom, the instance is tagged as 'Y'. (ii) If the instance is not registered as a single entry in the source dictionary (NeungYule and Dong-A Prime), but it is presented as a web-collected item, it is tagged as 'W'.[21] (iii) Otherwise, the instance is tagged as 'N'. The proportion table of these three is provided in Table 7, in which 'Y' and 'W' account for more than 70%.

Table 7. Evaluation

| | **number** | **proportion** |
|---|---|---|
| **Y** | 201 | 36.68% |
| **W** | 185 | 33.76% |
| **N** | 162 | 29.56% |
| **total** | 548 | 100% |

The instances tagged as 'W' are exemplified in (14).

(14) a. namphyen-ul ilh-ta
husband-ACC lose-DECL

[21] The web-collected data may be constructed in a similar way to the current work, using web documents.

'become a widow'
b. hyeythayk-ul ip-ta
benefit-ACC put.on-DECL
'benefit'
c. cwumwn-ul oy-ta
spell-ACC recite-DECL
'incant'

Although these kinds of expressions are not registered in the paper dictionaries, they meet the purpose of the current work in that they are multiword expressions. In particular, they often correspond to a single word in English, and such a lexical mapping has to be dealt with in machine translation. Thus, we can say that they are relatively well-acquired instances.

## 6. Conclusion

There are more than a few previous studies on idioms in Korean, and there are also several language data that include idiomatic expressions in Korean. Yet, the idiomatic expressions in the previous studies are mostly investigated by hand focusing on a very few items. As an alternative way, the current study makes use of a data-based method of acquiring the idiomatic expressions on a comprehensive scale, focusing on the "noun+verb" compounds in Korean.

The "noun+verb" idiomatic compounds can be enumerated by calculating selectional preference strengths. In this study, two types of language resources for Korean were utilized: namely, the *Sejong* Korean Treebank as a development corpus and three lexical hierarchies in Korean, including CoreNet, KorLex, and U-WIN. Building upon the data, the whole analysis was made in a fully automatic way, using the hill-climbing algorithm and examining kinship relations between concept nodes. As a result, 548 idiomatic compounds were acquired out of about 18,000 "noun+verb" patterns. Many of them look quite satisfactory, but if the selectional preference strength is rather weak, the result tends to be poorly acquired.

Identifying the idiomatic compounds in a systemic way aids in the production of natural-seeming translations in multilingual processing. In this context, the result of

the present study has a potential to be used to improve natural language processing systems. In particular, the result will be of great use of creating transfer rules of idiomatic multiword expressions. Since it is quite time-consuming and difficult to create transfer rules by hand for semantics-based machine translation, an automatic way of acquiring transfer rules from language resources is more preferred (Haugereid and Bond, 2011). For this further study, exploiting parallel texts is additionally required, but the result of the current study will save the time and effort to look up genuine pairs of idiomatic compounds across languages.

In addition, the current result will be of use to the theoretical studies of human language. As with other linguistic researches, a deep analysis of idioms and metaphorical meaning extension requires an analysis of language data. If the theoretical studies are supported by distributional findings like the current result, a better generalization about lexical semantics can be made.

Further studies include the followings: First, a bigger development corpus can be used to extract a list of idiomatic compounds with a higher rate of precision and recall. The main reason that the current work exploits the *Sejong* Korean Treebank is that it is the biggest syntax-tagged corpus amongst the available language resources as of now. If we can employ a dependency tagger that provides a satisfactory solution to resolve the syntactic functions (e.g. subjects, objects, etc.) in Korean, a larger development corpus is preferred to be used for this study. Second, the current work bypasses polysemy words and homonyms. For instance, a surface form *ssu-* has at least four meanings in Korean, such as 'write', 'use', 'wear (a cap)', and 'bitter'. Because these kinds of different meanings are not annotated in the development corpus of the present work, the current result does not have such a discrimination caused by polysemy and homonym. The further study has to make up for this limitation. Third, the idiomatic compounds consisting of verbs plus subject need to be researched in a similar way. For example, *son-i khu-ta* 'hand-NOM big-DECL' and *kan-i khu-ta* 'liver-NOM big-DECL' are idiomatic expressions, of which the meanings are 'generous' and 'bold' respectively.

In order for other theoretical and computational linguists to use the data constructed in this study for their own research interests, the whole dataset is readily redistributed online. All materials are downloadable on the following webpage.

(15) http://corpus.mireene.com/download/noun+verb.html

The materials include the sets of lowest common subsumers, the list of the concept nodes most strongly associated with verbal items, the evaluation tables, and the whole entries of the "noun+verb" idiomatic compounds.

## References

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.

Baldwin, Timothy and Francis Bond. 2002. Multiword expressions: Some problems for Japanese NLP. In Proceedings of the 8th Annual Meeting of the Association for Natural Language Processing (Japan), pp. 379-382, Keihanna, Japan.

Baldwin, Timothy and Su Nam Kim. 2010. Multiword expressions. Handbook of Natural Language Processing, second edition. Morgan and Claypool.

Chomsky, Noam. 1980. Rules and representations. New York: Columbia University Press.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Erk, Katrin. 2007. A simple, similarity-based model for selectional preferences. In Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics, pp. 216-223, Prague, Czech Republic.

Fabb, Nigel. 1990. The difference between English restrictive and nonrestrictive relative Clauses. *Journal of Linguistics*, 26(1): 57-77.

Haugereid, Petter and Francis Bond. 2011. Extracting transfer rules for multiword expressions from parallel corpora. In Proceedings of the Workshop on Multiword Expressions: from Parsing and Generation to the Real World, pp. 92–100, Portland, Oregon.

Jong-Bok Kim, Mi Jang, and Jooyoung Lim. 2013 Metaphorical extensions of eating and drinking in English and Korean: A usage-based approach. In Proceedings of the 2013 Metaphor Festival, Stockholm, Sweden.

Jurafsky, Daniel and James H. Martin. 2008. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Second Edition*. Prentice Hall.

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

Kim, Jong-Bok and Peter Sells. 2008. *English Syntax: An Introduction.* CSLI publications, Stanford, CA.

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.

Marcus, Mitchell P., Mary Ann Marcinkiewicz, and Beatrice Santorini. 1993. Building a large annotated corpus of English: the Penn Treebank. *Journal of Computational Linguistics* 19: 313-330.

Mason, Zachary J. 2004. CorMet: a computational, corpus-based conventional metaphor extraction system. *Computational Linguistics* 30(1): 23-44.

Nunberg, Geoffrey, Ivan A. Sag, and Thomas Wasow. 1994. Idioms. *Language* 70(3): 491-538.

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

Sag, Ivan A., Timothy Baldwin, Francis Bond, Ann Copestake, and Dan Flickinger. 2002. Multiword expressions: A pain in the neck for NLP. In Computational Linguistics and Intelligent Text Processing, pp. 1-15. Springer Berlin Heidelberg.

Shutova, Ekaterina, Simone Teufel, and Anna Korhonen. 2013. Statistical metaphor processing. *Computational Linguistics* 39(2): 301-353.

Song, Sanghoun and Jae-Woong Choe. 2014. Selectional preferences of Korean verbal items. *Linguistic Research* 31(2): 249-273.

Song, Sanghoun and Ji Young Song. 2014. A Data Compilation of Multiple Case-marking Constructions: using the *Sejong* Spoken Corpus. Language Information. 19: 57-90. [In Korean]

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**
Division of Linguistics and Multilingual Studies,
Nanyang Technological University
14 Nanyang Drive Level 3 Room 31
Singapore 637332
E-mail: sanghoun@ntu.edu.sg

Received: 2014. 11. 18.
Revised: 2015. 03. 22.
Accepted: 2015. 03. 22.