A representational system of idiomatic constructions: For the building of computational resources*

Hee-Rahk Chae
(Hankuk University of Foreign Studies)

Chae, Hee-Rahk. 2014. A representational system of idiomatic constructions: For the building of computational resources. *Linguistic Research* 31(3), 491-518. We will provide a theory/platform-neutral framework for representing idiomatic expressions. Previous analyses on idioms, including those for some systems implemented on the basis of specific theories and platforms, are not flexible enough to deal with the gradational nature of syntactic and semantic anomalies of idioms, and/or are not explicit enough to be useful in natural language processing. We will overcome these problems by introducing some theory-neutral notational mechanisms, which apply to regular phrases, to capture the peculiarities of idioms. The present framework will become a very useful guide in building effective computational resources for a variety of languages, although we are focusing on the analysis of English and Korean idioms. One of the strong points of our approach is that we can formalize precisely the properties of potentially numerous different types of idioms, which is possible because we are using regular phrases built from phrase structure rules to represent the peculiarities of idiomatic expressions. The resulting resources will enhance the parsing accuracy and speed significantly, and will play an important role in disambiguation. (Hankuk University of Foreign Studies)

Keywords idioms, constructions, structure of the lexicon, representation of idioms, computational resources, English idioms, Korean idioms

1. Introduction

In general, grammar is considered to consist of two parts: the lexicon and the rule set. The lexicon is a repository of lexical items, which are unpredictable pairings of

---

* I appreciate the constructive comments and suggestions of anonymous reviewers of this journal.
This work was supported by the 2014 research fund of Hankuk University of Foreign Studies.
form and meaning. The rule set, on the other hand, comprises rules, which represent predictable combinatory patterns of lexical items. Under this view, the meaning of a complex expression is thought to be derivable from its parts compositionally. However, there are so many multiword expressions like idioms which are not compositional. These expressions are difficult to deal with not only in theoretical descriptions but also in applications in such areas as natural language processing and language teaching.

The purpose of this paper is to provide a theory/platform-neutral representational framework for non-compositional expressions. We think this framework can capture the peculiarities of idioms effectively, at least, from a computational point of view, without assuming the mechanisms of particular theories. Although it may not be very elegant theoretically, it will be very useful in building an effective resource of non-compositional expressions in a language. The resource will enhance the parsing accuracy and speed significantly, and will play an important role in disambiguation not only on the lexical level but also on the phrasal level. The framework will work for a variety of different languages, but we will be mainly interested in providing a guide for English and Korean idioms. In this paper, however, we are not interested in implementing the results built from the framework into particular computational systems/platforms. Even though we assume that any resulting resources will be implemented into different computational systems, our focus is on a theory-neutral representational framework of English and Korean idiomatic expressions.

In chapter 2, we will briefly review previous analyses on idioms. There are largely two groups of approaches, depending on whether distinct classes of idioms are assumed or not. In chapter 3, we will lay out theoretical background for our framework by considering three issues: the relation between syntax and semantics, the definition of “constructions,” and the structure of the lexicon. Then, on the basis of this background, we will provide our framework for representing idioms in chapter 4. Its basic mechanisms will be presented in section 4.1, and the system will be tested with reference to some English idioms in section 4.2. In section 4.3, our framework will be compared with the endeavor under the tradition of “the LKB system.” In chapter 5, we will provide an analysis of some representative examples of Korean idioms.

We will see that syntactic and semantic phenomena do not necessarily go parallel and that idiomatic expressions show a continuous degree of syntactic and/or semantic
A representational system of idiomatic constructions ... 493

anomalies. Hence, we need a system which captures syntactic and semantic properties separately and which is flexible enough to deal with the continuous properties. We will also see that the traditional view on grammar, especially on the lexicon, is not adequate for describing idioms.

2. Previous Analyses on Idioms

Although idiomatic expressions are not easy to deal with, the sheer number of them prevents us from just neglecting them. There have been many different approaches to account for their idiosyncratic properties (Quirk et al. 1985, Huddleston & Pullum 2002; Moon 1998; O’Grady 1998, Titone & Conine 1999, Horn 2003, etc.). Since Nunberg (1978) and Nunberg et al. (1994), idiomatic expressions are divided into two classes: “decomposable idioms”/ “idiomatically combining expressions (ICEs)” and “non-decomposable idioms”/ “idiomatic phrases (IPs)” (Jackendoff 1997, Sag et al. 2002, Copestake et al. 2002, etc.). Grammatical tests for the distinction are based on the possibility of modification, quantification, topicalization, ellipsis and anaphora (Nunberg et al. 1994: 500-3). Idioms like *bury the hatchet* ‘reconcile disagreement’ belong to the former and idioms like *kick the bucket* ‘die’ belong to the latter. It is argued that the former have meanings distributed among their parts, while the latter do not.

Let us begin with Jackendoff’s (1997: 168-9) analysis of idioms, which maintains the distinction between ICEs and IPs basically.

(1) to *bury the hatchet*

\[
V_x \quad N \quad \text{Det} \quad \text{NP}_y
\]

\[\text{[RECONCILE} \text{[DISAGREEMENT]_y \text{]}_x\text{]}_x\]
In the entry for *bury the hatchet* (1), the two component parts *bury* and *the hatchet* are assumed to have their own (metaphorical) meanings, i.e. ‘reconcile’ and ‘disagreement,’ respectively. Syntactically, the two parts are separated from each other because the NP can be “moved” around. In the entry for *kick the bucket* (2), on the other hand, the components parts do not have separate meanings and the two parts are connected because the NP cannot be moved around.

Espinal & Mateu (2010: 1397), however, argue that the distinction is “not as clear-cut and uniform as has been assumed.” They show, on the basis of *V one’s head off* idioms (cf. Jackendoff 2002: 173-4), that the five tests mentioned above do not provide consistent evidence for the distinction.

(3) a. John laughed his head off.
   b. We laughed our heads off.

(4) Bill cried his eyes out on Wednesday, and he cried them out again on Sunday.

(5) a. *Whose/which heart did Bill eat out?
   b. *His heart, Bill ate out.

(6) a. *Bill ate his [own/inner heart] out.
   b. *We were laughing our [two heads] off.

The examples in (3) and (4) show ICE-like properties of the idioms. The determiner of the NP varies depending on the subject, as we can see in (3). In addition, a part of the idiom can be pronominalized, as we can see in (4). The examples in (5) and
(6), on the other hand, show their IP-like properties. We can see from the data in (5) that the NP cannot be questioned or topicalized. We can also see from the data in (6) that the head noun of the NP cannot be modified or quantified.

Actually, Jackendoff’s (2002: 173) analysis cannot account for the intermediate properties of *V one’s head off* idioms properly.

(7) \([vP \text{ } v \text{ } NP \text{ } PRT] \text{ } V \text{ } pro’s \text{ } head/butt \text{ } off\]

‘V excessively’

According to him, the idioms concerned have a VP structure and have the meaning of ‘V excessively.’ The NP and PRT in the VP “are lexically fixed and the V is a free variable.” As the analysis in (7) is basically the same as that for *kick the bucket* in (2), it cannot account for the ICE-like properties in (3-4).\(^1\)

In addition to pointing out the intermediate nature of *V one’s head off* idioms in (3-6), Espinal & Mateu (2010: 1403-4) also criticize, from a point of view of cognitive linguistics, that the analysis of Jackendoff (1997, 2002) uses “unmotivated lexical stipulations” and “fails to recognize the systematic syntax-semantics correspondences provided by Talmy’s (1985, 2000) typology of motion events.” That is, the syntax-semantics associations in such analyses as in (1-2) and (7) are said to be “merely stipulated.” They argue that the interpretation ‘V excessively’ in (7) is “not part of the relevant semantics that interacts directly with syntax” and that the interpretation “results from a metaphorical mapping from a source domain to a target domain.” They also argue that all the internal elements in (7), i.e. NP and PRT, are “as meaningful as the ones that can be found in literal/non-idiomatic examples…”

It is true that Jackendoff’s analysis contains lexical stipulations and the system is not flexible enough to deal with all the idiosyncratic properties of idioms. However, Espinal & Mateu’s (2010) analysis does not seem to be very reasonable, either. They argue all the internal elements of idioms have metaphor/non-literal meanings and the meanings of the whole idioms can be derived from them compositionally.\(^2\)

---

1. Of course, the analyses in (2) and (7) are not exactly the same. All the lexical items are analyzed to be fixed in the former, while some positions for lexical items are open in the latter. However, this difference cannot capture the syntactic differences between the two groups of idioms.

2. Broadly speaking, approaches dealing with idioms can be categorized into two groups (Espinal & Mateu 2009: 1397): “non-compositional” and “compositional” approaches. The former regards idiomatic phrases as units showing arbitrary syntactic and semantic properties, i.e. as a sort of
of all, it is not clear how the metaphoric meanings of the internal elements can be obtained. It might be possible to devise such a system from a cognitive linguistics point of view. However, we think we have to wait more for the system to be explicit enough to be useful computationally. Next, although stipulations are not recommended in linguistic descriptions, the “lexical stipulations” in Jackendoff (1997, 2002) are natural consequences of the very nature of idiomatic expressions themselves, which carry (some degree of) irregular/unpredictable properties necessarily. Hence, we will need a framework which is formal enough to be computationally useful, and which is flexible enough to handle all the idiosyncratic properties of idioms. In other words, we will need an enriched version of Jackendoff’s system which can handle a variety of idiomatic expressions.

3. Theoretical Background

Before providing a new framework for the representation of idioms, we are going to lay out theoretical background for it. We will consider three issues: the relation between syntax and semantics, the definition of “constructions” and the structure of the lexicon. We assume that formal/syntactic and semantic phenomena do not necessarily go parallel. In addition, we adopt the views of the construction and the structure of the lexicon in Culicover (2009: 33).

Some syntactic frameworks like those following the tradition of Chomsky’s ideas are under the assumption that syntactic and semantic phenomena go parallel. Hence, in such frameworks, all syntactic facts are to be accounted for on the basis of semantic facts. However, there are many syntactic phenomena which cannot be attributed to semantics.3 One of the most popular examples is the difference between *eat* and *devour*. Even though they have the same meaning, the object of *eat* is optional while that of *devour* is obligatory. We can see a more conspicuous example from Korean. The predicates *pwucok-ha-* and *mocala-* have the meaning of ‘be

---

3 Jackendoff (2002: 138-49) discusses the relationship between syntactic argument structure and semantics, and Jackendoff (2013: 79-83) talks about “the degree of independence of syntax from semantics,” with reference to “Heterogeneous Construction Grammar.”
scarce.’ However, as they combine with different set of inflectional affixes, the former is classified as an “adjective/ descriptive verb” and the latter as a “verb/ action verb.” We can conclude from these considerations that syntactic and semantic phenomena do not necessarily go together and, hence, that they are under different set of rules and/or principles.  

The term “construction” is used in a variety of senses. According to Goldberg & Casenhiser (2006), the term is generally used to refer to “certain grammatical patterns that have unusual quirks in either their formal properties or their semantic interpretation (or both) that make them ill-suited for universal status” (p. 344). However, to some linguists, especially to some constructionists, constructions are “any conventional pairings of form and function, including individual morphemes and root words along with idioms” and even those patterns which are fully predictable/compositional “as long as they occur with sufficient frequency” (p. 349). In addition, most constructionists consider constructions as part of a “lexicon-syntax continuum” rather than assuming a clear-cut division between them (Hoffmann & Trousdale 2013: 1).

Among the various definitions of constructions, we will adopt the definition provided in Culicover (2009: 33) in this paper:

(8) A construction is a syntactically complex expression whose meaning is not entirely predictable from the meanings of its parts and the way they are combined in the structure.

According to this definition, only “syntactically complex” expressions, i.e. only phrasal and clausal expressions, are constructions. Morphemmes and words are not regarded as constructions. In addition, only non-compositional expressions are constructions. Fully compositional expressions are not regarded as constructions. Constructions defined this way have syntactic idiosyncrasies as well as semantic idiosyncrasies. We assume that the syntactic anomalies of constructions are due to

---

4 According to Culicover & Jackendoff (2005: 15), “the combinatorial principles of syntax and semantics are independent; there is no “rule-to-rule” homomorphism.”

5 According to Jackendoff (2002: 425), lexical items “include not just words but also productive morphological affixes and idioms.” Under this view, lexical items are treated as interface rules.

6 The term “meaningful constructions” is used in a similar sense to the constructions defined in (8) (cf. Jackendoff 2013: 78).
their semantic anomalies.7

Under the definition of constructions in (8), Culicover (2009: 33) proposes that the lexicon has the following structure:

(9) The Structure of the Lexicon

Lexical entries
- Form
- Morphological structure
- Syntactic properties
- Meaning

Constructions
- Form
- Syntactic structure
- Meaning

Unlike the general conception of the lexicon, constructions as well as lexical entries are in the lexical component, because they have non-compositional properties. The “morphological structure” part is connected to the outside “word formation” component, and the “syntactic structure” part of the constructions to the outside “syntax” component. The view on constructions as lexical items is explicitly defended in Jackendoff (2002). He uses the term “lexical items” to denote only those items which are “stored in the lexicon, i.e. in long-term memory.” Lexical items, then, “may be larger or smaller than grammatical words” (pp. 153-4).

From our point of view, however, it is not essential that the constructions part should be in the lexicon. It is more important for us that the constructions part is separate from the part of lexical entries and from syntax. We need just to capture the fact that constructions interact both with lexical entries and with syntactic rules and, hence, have to be kept separate from both of them.

7 The assumption here does not seem to be in accordance with our basic tenet that syntax and semantics do not necessarily go together. However, as we still believe that the degrees of syntactic and semantic anomalies are not necessarily parallel, the tenet is not weakened by the assumption. We only assume that there would not be any phrases which have non-compositional meanings but have no syntactic anomalies.
There is one more issue to be clarified before we go on. Culicover (2009: 33) defines idioms as “the constructions that are completely or highly opaque in meaning.” However, this definition does not seem to be reasonable. Most of all, it is not clear what “completely or highly opaque in meaning” means. We do not know how to measure the meaning opaqueness of an expression. Furthermore, we do not know how much opaqueness an expression should have for it to be regarded as an idiom. As we can see from the data in section 2, different kinds of idioms have different degree of opaqueness (cf. Jackendoff 2002: 172-177). Hence, it would be better to assume that idiomatic opacity is just a matter of degree, rather than a matter of categorical distinction. At this point, we need to pay attention to the definition of constructions in (8). According to this definition, all constructions are more or less opaque in their meanings. Hence, idioms can better be defined with reference to constructions themselves rather than to a subset of constructions. Henceforth, we will assume that all idioms are represented as constructions and all constructions represent idioms.

Under our view of constructions and idioms, we can divide phrasal/clausal expressions into two different types: those which show “full productivity” and those which show “semipродuctivity” or no productivity (Jackendoff 2013: 84). The former are dealt with phrase structure (PS) rules and the latter with constructions. This view is in line with one of the major “architectural hypotheses” shared by many non-Chomskyan/alternative generative theories (Culicover & Jackendoff 2005: 14-5): “There is a continuum of grammatical phenomena from idiosyncratic (including words) to general rules of grammar.”

4. The Representational Framework

In this chapter, we will introduce the framework for representing idioms/constructions we are proposing. We will provide basic mechanisms of the framework in section 4.1, and will show that some representative types of English idioms can be analyzed effectively under the framework in section 4.2. Then, we will look at a similar endeavor under the tradition of “the LKB system” and related resources, and will

---

8 It would not be very unreasonable to assume that no productivity is a special case of semiprodutictivity.
compare it with our framework in section 4.3.

4.1 Our Framework: Synopsis

In general, idioms contain semantic and syntactic idiosyncrasies. Semantically, the meaning of the whole phrase cannot be obtained from the meanings of its parts, i.e. the phrase is not compositional. Syntactically, some parts of the idiom are restricted in their behavior externally and/or internally, which will be shown below.

Our formalism is largely the same as those in Jackendoff (2002: 167-78), Culicover & Jackendoff (2005) and Culicover (2009). To represent the fact that an idiom has a meaning not predictable from its parts, we will assign the meaning directly to the whole phrase, just as in (1-2) and (7), However, we will introduce some new syntactic mechanisms to capture various types of syntactic anomalies of idiomatic expressions.

The data in (5) and (6) represent two types of typical syntactic behavior. The former are examples examplifying “external syntax.” In general, a syntactic unit in a phrase can be extracted out of it or can undergo other processes with reference to some elements outside of the phrase. In (5), an NP is extracted out of a VP by way of question formation or topicalization. Next, please look at the following sentences:

(10) a. John is a taller man than Bill.
    b. John is so tall that he cannot take a taxi.

(11) a. John can solve the problem, but Bill cannot.
    b. John likes apples and Mary oranges.

In (10), the PP than Bill and the clause introduced by that are licensed by the external elements taller and so, respectively (Chae 1992). In (11), an element in the second clause is deleted with reference to a corresponding element in the first clause.

The data in (6) represent another type of syntactic behavior: “internal syntax.” In general, the head of a phrase can be extended by such elements as its complements, adjuncts or specifiers within the phrase. The examples in (6) show that the head Ns heart and heads cannot be extended by modifiers or quantifiers.

Normally syntactic phrases do not have restrictions in their external or internal syntax. However, there are cases where their behavior is restricted externally and/or
internally. Idiomatic expressions are the most typical examples showing such restrictions more or less. Now, we will introduce four notations/conventions to indicate these restrictions: <..., >/..., {...} and capitalization. The former two are used to restrict syntactic behavior and the latter two to regulate lexical behavior. The first one indicates that the phrase enclosed by <...> is a syntactic “island.” That is, no external processes like extraction will be allowed from the phrase. The second one indicates that the phrase enclosed by /.../ cannot be further expandable by any internal elements like complements, adjuncts or specifiers. That is, no syntactic expansion is allowed inside the phrase if not explicitly specified already. Expressions enclosed by <...> and/or /.../ are “syntactically frozen” to some degree. As we can see below, the notation {...} is used to indicate that only the lexical items listed inside these brackets are allowed to occur in the position concerned. Lastly, we will use capital letters to represent those lexical items which have to be inflected for their specific forms.

4.2 English (Idiomatic) Constructions

Here we will apply the mechanisms introduced above in analyzing English idioms. We have seen that idioms like *bury the hatchet* in (1) are rather free in their syntactic behavior, while those like *kick the bucket* in (2) are very restricted in their behavior. Somewhere in between stand idioms like *V one’s head off* (cf. the examples in (3-6)).

Among the two popular classes of idioms, let us first analyze the “non-decomposable (IP)” idiom *kick the bucket* in (2) under our framework.9

---

9 Such “fixed expressions” like *in line, by and large and in short* in Sag et al. (2002: 2, 4) can also be analyzed the same way. The whole phrases are enclosed by both <…> and /…/. The only difference between these idioms and those like (12) is that the former do not have any lexical items written in capital letters. In this sense, they are more fixed than the latter.
(12) to kick the bucket ‘to die’

Only the whole phrase has a specified meaning.\(^{10}\) We just disregard the meanings of its component parts when a higher node has a designated meaning. In addition, the VP is enclosed by both `<...>` and `/.../`, which means that its external and internal behavior is restricted. That is, the NP inside it cannot be extracted (e.g., *The bucket was kicked by John*) and no elements inside it can be expanded (e.g., *John kicked the big/plastic bucket*). In addition, the three lexical categories involved are connected with the brackets `{ }` which have only a single member. No other lexical items are allowed.

We need to add an explanation for the lexical item *KICK*, which is written in capital letters. Those in capital letters are the bases for inflectional variation. That is, they have to appear in specific inflectional forms in sentences. Please realize that inflectional forms are determined by morpho-phonological principles.\(^{11}\) For example, if *KICK* carries the information [PERSON 3] and [NUMBER sg], its form will be realized as kick-s. Syntactically, we need only the information it carries not its form, and the information will flow through the tree regardless of whether the phrase is restricted in its syntactic behavior or not. Hence, they are outside of the (pure) syntactic realm we are concerned here.

Secondly, we will examine the syntactic behavior of such “decomposable (ICE)”

---

\(^{10}\) We need to assume that the phrase which has a specified meaning carries its own subcategorization frame as well. That is, in (12), the whole phrase kick the bucket rather than kick will carry its subcategorization frame [SUBJ NP].

\(^{11}\) We can also assume that {KICK} is an abbreviation of {kick, kick-s, kick-ed, kick-ing}, which comprises all the possible forms of the lexeme kick.
idioms as *bury the hatchet* in (1).

(13) to *bury the hatchet* ‘to reconcile disagreement’

```
/VP/
  \   /  \\
  V /   NP\
   \     \\
   {BURY} Det N
       \  \  \\
       {the} {hatchet, axe}
```

VP: ‘[reconcile disagreement] (x)’

The VP as a whole has the meaning of ‘to reconcile disagreement.’ It is not enclosed by <…> because *the hatchet* can be extracted (e.g., *The hatchet was buried after years of arguing*). However, it is enclosed by /…/ because no expansion is allowed inside (e.g., ‘They buried the large/expensive hatchet’).\(^{12}\) Here the word *hatchet* can be replaced with the word *axe*.

Lastly, we will consider *V one’s head off* idioms,\(^{13}\) which are argued to be problematic to deal with to those approaches assuming a clear distinction between ICEs and IPs (cf. (3-6), (7)).

---

\(^{12}\) An anonymous reviewer pointed out that expressions like *bury the official hatchet* are possible with the idiomatic reading. If it is possible, the NP in the construction should be analyzed as $\{\text{Det}\{\text{the}\}\} \{\text{Adjp}\{\text{(official)}\}\} \{\text{NP}\{\text{(hatchet, axe)}\}\}$. This NP contains *official* as an optional Adjp. The /…/ on the VP, however, should not be removed, because the NP still does not allow adjectives in general. It allows only those adjectives which are listed in the construction.

\(^{13}\) There are similar types of idioms to the *V one’s head off* type (Jackendoff 2002: 173): *V one’s heart out* and *V up a storm*. All these have the meaning of ‘V excessively.’
(14) to $V$ one’s head off ‘to $V$ excessively’

As the verb concerned is “totally free” (Jackendoff 2002: 173), $V$ is not connected with any lexical items. The specific form of $ONE’S$ will be determined with reference to the subject by way of agreement. The nouns $HEAD$ and $BUTT$ are in capital letters, because they also vary in their forms, as we can see in (3b) and (4).

The IP-like properties shown in (5-6) can be easily treated. The NP cannot be extracted because the whole VP is enclosed by <$...$>, and cannot be expanded because the NP is enclosed by /.../. Turning to the ICE-like properties in (3-4), the fact that the determiner and its head noun can have various inflectional forms is accounted for with reference to our mechanism of using capital letters. The possibility of anaphoric expressions in (4) can also be accounted for with this mechanism. As the VPs in (12) and (14) are both externally closed, i.e. they are marked with <$...$>, the elements in them have restrictions in their external syntactic behavior. However, unlike in (12), pronominalization of the NP is allowed in (14). This difference is due to the fact that the lexical items in the NP of (14) are capitalised while the ones of (12) are not. For the realization of capitalized lexical items, we have to allow the flow of inflectional information through the tree regardless of the external enclosure by <$...$>. Here we need to realize that extraction and pronominalization are different in their syntactic characteristics. As the former is a typical syntactic phenomenon, it does not interact either with morphosyntactic properties or with semantic properties. On the other hand, the latter interacts both with morphosyntactic properties and with semantic properties. Hence, we can safely assume that the latter is less sensitive to the existence of <$...$> than the former.
One of the most peculiar properties of *V one’s head off* idioms is that “the verb must have no syntactic argument of its own aside from the subject” (e.g. *read mysteries one’s head off*) (Jackendoff 2002: 173). We can account for this property of them naturally within our framework. We just need to assume that, when analyzing sentences, the constructions component takes priority over the “lexical entries” component in (9) and over the rules in syntax. Namely, we are assuming that constructions are accessed before lexical entries and syntactic rules are accessed. Under this assumption, the analysis of the VP in the idioms will be taken care of by the construction in (14). In the construction, the verb is supposed to take only the NP *one’s head/butt*. It does not take the “normal” object which is specified in the lexical entries component (cf. footnote 10). The lexical entries component and the syntactic rules component will be responsible for the analysis of the remaining parts of the sentences concerned. The assumption here is not arbitrary because it is a general principle of linguistic descriptions that more specific entities take priority over more general ones.

As has been discussed thus far, idioms are not classified into distinct classes in our approach. We assume that they show a continuous degree of rigidity/flexibility from the most rigid (or the least flexible) expressions to the least rigid (or the most flexible) ones. Except for the most flexible ones, which are just regular phrases, every expression in between can be regarded as an idiom. Assuming the “X-bar Theory,” one of the most rigid idioms can be analyzed as follows:14

(15)

```
Specifier
   {a} X Complement
   X'
   {b} {c}

<XP/>
```

---

14 Such idioms as *in line* and *by and large* are more rigid than the one represented in (15), because not only modifiers but also complements and specifiers are not allowed in them (cf. footnote 9).
The \(<...>\) on XP indicates that no elements inside XP can be extracted outward, and the \(\ldots/\) on XP says that the node cannot be expanded internally with any modifiers. In addition, all the lexical categories are connected with \(\{\ldots\}\) which has only a single member. The flexibility will increase as more lexical items are capitalized, as more lexical items appear in \(\{\ldots\}\), and as \(<...>\) or \(\ldots/\) disappears, eventually to become regular/non-idiomatic phrases.

One of the most innovative properties of our approach is that we can formalize precisely the properties of potentially numerous different types of idioms. This is possible because we are using regular phrases built from phrase structure rules to represent the peculiarities of idiomatic expressions. The regularity and/or the flexibility of them are captured on the basis of the regular phrases involved, and the irregularity and/or the rigidity of them on the basis of the four restrictive mechanisms applying to the phrases. The resulting framework can naturally capture not only the irregularities/peculiarities but also the regularities of idioms. In addition, the interactions between the different types of restrictive mechanisms enable the framework to analyze a variety of different types of idioms.

### 4.3 Comparison with Approaches under the LinGO Project

There are some traditions of developing mechanisms and resources for dealing with “multiword expressions,” which include idioms and other expressions with idiosyncratic properties. Especially notable is the one under the “LinGO Project,” which is based on Head-Driven Phrase Structure Grammar (HPSG) frameworks and the LKB system (Sag et al. 2002, Copestake et al. 2002, and so on). In this section, we will compare the approaches under this project with our approach briefly.

In Sag et al. (2002: 3-8), multiword expressions are classified into “lexicalized phrases” and “institutionalized phrases.” The former are described as having “at least

---

15 Sag et al. (2002) defines multiword expressions as “idiosyncratic interpretations that cross word boundaries (or spaces)” (p. 2) and lists the following ongoing projects on multiword expressions: the ParGram Project at Xerox PARC, the XTAG Project at the University of Pennsylvania, the LinGO Project and the FrameNet Project (p. 3).

16 We can get information about the LinGO Project and related frameworks/systems from the site [http://lingo.stanford.edu] and about the Multiword Expression Project from the site [http://mwe.stanford.edu]. The LKB system is “a grammar and lexicon development environment for use with unification-based linguistic formalisms.”
A representational system of idiomatic constructions ...  507

partially idiosyncratic syntax or semantics, or contain ‘words’ which do not occur in isolation,” and the latter as being “syntactically and semantically compositional, but occur with markedly high frequency (in a given context).” The former are further classified into “fixed expressions,” “semi-fixed expressions” and “syntactically-flexible expressions.” Non-decomposable idioms (i.e. IPs) like kick the bucket belong to semi-fixed expressions, and decomposable idioms (i.e. ICEs) like spill the beans belong to syntactically-flexible expressions. That is, as in Nunberg et al. (1994) and Jackendoff (1997), they also assume that idioms can be divided into distinct classes. However, as Espinal & Mateu (2010) shows with reference to V one’s head off idioms, the distinction is not clear-cut and uniform (cf. sec. 2).  

In addition, Sag et al. (2002: 1) assume that “the various kinds of multiword expressions should be analyzed in distinct ways.” As for non-decomposable idioms, “the whole entry (a list of words)” is treated as “a single listeme that is associated with a single semantic relation (Sag et al. 2002: 9) or as “a word with spaces” (Copestake et al. 2002: 1945). It does not make a big sense, however, to assume that a part of “a single listeme” or “a word with spaces” can have its own inflectional endings. In our analysis, the inflecting word itself is a regular independent word (cf. the analysis in (12)). As for decomposable idioms, on the other hand, the meaning of the whole phrase, e.g., of let the cat out of the bag, comes from the idiomatic senses of cat and bag, ‘secret’ and ‘hiding place,’ respectively (Sag et al. 2002: 13, Copestake et al. 2002: 1945-6). They need to posit idiomatic senses of the component words for their mechanisms to work in accounting for the syntactic flexibility of decomposable idioms. First of all, it is not clear how we can get the idiomatic senses of the words inside an idiom. For example, they need to provide a theory to derive the meaning of ‘secret’ from the literal/regular meanings of cat. It would be easier and more reasonable to assume that the whole idioms have their meanings directly, regardless of their classes (cf. the analysis in (13)). Our framework accounts for the syntactic flexibility with mechanisms involving regular phrases, without assuming words with idiomatic meanings and other unmotivated mechanisms.

17 As we can see in (18) in chapter 5, Korean also has some idioms which cannot be classified into either decomposable or non-decomposable idioms.

18 Copestake et al. (2002: 1945) agree with Sag et al. (2002) in making “a strong distinction between decomposable and non-decomposable idioms” and treat them in distinct ways.
From a meta-theoretical point of view, it does not seem to be very reasonable to treat different types of idioms, if such distinctions exist, in distinct ways. Putting aside the issue whether idioms can be assigned to different classes or not, it is not easy to agree with the idea that non-decomposable and decomposable idioms are so different from each other that they should be treated with such different mechanisms as they propose. For example, *kick the bucket* is treated as a single “listeme/word,” but *spill the beans* as a phrasal unit. It is not clear what makes them belong to the two distinct grammatical categories. Note that both of them contain an inflecting verb as a component word. In addition, we need to decide whether the *V one’s head off* idioms belong to a lexical category, to a phrasal category, or to a third category.

In addition to these theoretical differences between the LinGO approaches and our approach, the two approaches have different objectives. Firstly, they focus on building resources which can be implemented into the LKB system. However, we are not interested in implementing the resources built from our framework into a specific system. Our objective is to provide a guide for building theory/system-neutral resources for dealing with expressions containing idioms. Secondly, we hope that our guide will be useful in building good resources for Korean multiword expressions, regardless of whether it would be useful for other languages or not.19

5. An Analysis of Korean (Idiomatic) Constructions

In the previous chapter, we have laid out the framework for the analysis of idiomatic expressions and have attested its validity with reference to English idioms. In this chapter, we will provide an analysis of some examples of Korean idioms. There are not only numerous research papers and books (KH Moon 1996, HS Min 2003, J-h Park 2003, M-G Park 2003, DH Lee 2007, HS Kim 2011, etc.) but also some dictionaries on Korean idioms (YJ Park & KB Choi 1996, Y No 2002, KB Choi 2014).

19 Although there are many works dealing with Korean idioms from a theoretical point of view (cf. chapter 5), there are not many works, except C Chung & J-B Kim (2007), dealing with them from a computational point of view. For example, we could not find any papers focusing on Korean idioms from the previous proceedings of “SIGLEX-MWE: Workshops on Multiword Expressions” (http://multiword.sourceforge.net/PHITE.php?sitesig=CONF).
As in C Chung & J-B Kim (2007), we would be able to distinguish between non-decomposable idioms and decomposable idioms in Korean, as well.\footnote{C Chung & J-B Kim (2007) distinguish two different types of Korean idioms and provide an HPSG analysis of them. Idioms like \textit{kkuueng kwu-e mek-un sosik} ‘no news (at all),’ \textit{mwul cha-n ceypi} ‘very neat/swift/slim looking,’ and \textit{kkuu-eta noh-un poli calwu} ‘dumb like a wallflower’ are classified as non-decomposable idioms. Idioms like \textit{kkoli-lul chi} ‘to seduce’ and \textit{kani klu} ‘be very brave’ are classified as decomposable idioms. They capture differences in the degree of idiomaticity of decomposable idioms with reference to various syntactic processes. According to them (p. 422), most of them allow scrambling, case marking or adverbial modification of their components. Many of them, however, do not allow adnominal modification, topicalization, pronominalization or passivization, and none of them allow cleft or afterthought constructions.}

(16) A Non-decomposable Idiom
a. so kwi-ey kyeng ilk-
cow ear-at Buddhist scripture read
[literally] ‘to read Buddhist scriptures to a cow’s ears’
b. so kwi-ey kyeng cacwu ilk-
often
c. [so kwi-ey e ilk-un] kyeng
-Relativizer

(17) A Decomposable Idiom
a. nwun-ey khongkkakci-ka ssui-
eye-at bean pod-Nom be put
[literally] ‘bean pods are put on the eyes’
b. nwun-ey twukkewun khongkkakci-ka ssui-
thick
c. [nwun-ey e ssui-n] khongkkakci
-Relativizer

The idiom in (16a) is used to describe a situation where someone has done something vainly: ‘to sing psalms to a dead horse, to beat a drum to a cow’s ear, to bang one’s head against a brick wall.’ The fact that the expressions in (16b-c) do not have this idiomatic meaning shows that the idiom cannot be expanded internally and that its parts cannot be extracted. The idiom in (17a) has the idiomatic meaning of ‘to be blinded by love, one cannot love and be wise.’ Unlike (16a), however, it keeps its idiomatic meaning in (17b-c).
As in English, there are some idioms which cannot be categorized either as a non-decomposable one or as a decomposable one:

(18) a. euy(-ka) eps-
     euy-Nom there is not any  [literally] ‘there is not any euy’
     b. euy(-ka) cengmal eps-
         really               ‘to be really preposterous’
     c. *[e eps-nun]       euy
         -Relativizer

The meaning of the idiom [euy(-ka) eps-] in (a) is ‘to be preposterous.’ However, the putative noun euy does not exist as a separate word and does not have any meaning at all in itself (at least in current Korean). As a component part of the idiom does not have any (literal or idiomatic) meaning and, hence, the semantics of the whole idiom is opaque, the idiom can only be regarded as non-decomposable. Please note that, as is argued by Sag et al. (2002) and Copestake et al. (2002), the meaning of a decomposable idiom is supposed to come from the idiomatic senses of its component parts (cf. section 4.3). However, although the idiom does not allow extraction of its components as is shown in example (c), it allows internal expansion as is shown in (b), which is a property of decomposable idioms. According to Sag et al. (2002: 4), “non-decomposable idioms are not subject to syntactic variability.” Henceforth, we will not assume that the distinction between the two classes is real.

Y No (2002) classifies Korean idioms according to the number of fixed inflectional affixes they have. We will provide an analysis of the idiom in (18) euy(-ka) eps- ‘to be preposterous,’ as an example of those which do not have any such affixes.
A representational system of idiomatic constructions ... 511

(19)  
euy(-ka) eps- ‘to be preposterous’

\[
\begin{align*}
<S> & \\
/\text{NP}/ & \text{VP} & \\
\{\text{euy(-ka)}\} & \text{AdvP}_a & V' \\
\{\text{EPS-}\} & \\
\end{align*}
\]

S: ‘[x is α preposterous]’

Here we have a clausal unit idiom. There is only one verb, i.e. \(\text{EPS-}\), and it can have all the inflectional endings of verbs. The whole clause is enclosed with \(<\ldots>\) because the subject cannot be extracted by such processes as relativization (cf. (18c)). However, the VP is not enclosed with \(/\ldots/\) because the verb \(\text{EPS-}\) can be modified by adverbs like \(\text{cengmal} \) ‘really’ (cf. (18b)). Only the subject NP is enclosed with \(/\ldots/\) because the (putative) head noun \(\text{euy}\), \(^{22}\) although it can be marked with the subject marker \(-\text{ka}\), cannot be modified by any adjectives (‘[[\text{khu-n ‘big’/ cak-un ‘small’}] euy]’).

The idiom in (19) is peculiar in the sense that the VP inside the idiom is allowed to be expanded with an adverb. Although the adverb phrase in the VP is in a position to modify the verb \(\text{EPS-}\), it actually modifies the whole idiom. We can implement this unusual property of the idiom into our framework by integrating the meaning of the adverb phrase, which is represented as α in (19), directly into the clausal level meaning.

We will consider two more examples from Y No (2002), which contain one or more fixed inflectional affixes: [\(\text{yatcap-a po-}\) ‘to despise’ and [...] \(\text{V}_1\)-\(\text{tunci} \ldots \text{V}_2\)-\(\text{tunci} \text{kan-ey}\) ‘regardless of whether … or …’ Only the affix \(-a\) is allowed after

\(^{21}\) We assume that the set of adverbs allowed in the construction (19) is the same as the set of those allowed before the verb \(\text{eps}\) in regular phrases. If the two sets are not the same and only a few adverbs are allowed in (19), these adverbs would have to be listed in the construction and the VP would have to be enclosed by \(/\ldots/\) (cf. footnote 12).

\(^{22}\) As the noun \(\text{euy}\) does not mean anything in itself, it cannot be listed as a member of the lexical entries component in (9). It can appear only as a part of an idiom/construction.
yatcap- in the former, and only the affix -tunci is allowed after the two verbs involved in the latter.

(20) yatcap-a po- ‘to despise’

\[
<V'>
\]

\[
V' \quad V
\]

\[
\{yatcap-a\} \quad \{PO-\}
\]

\[
V': '\text{despise (y) (x)}
\]

(21) \ldots V1-tunci \ldots V2-tunci kan-ey ‘regardless of whether \ldots or \ldots’

\[
<PP'>
\]

\[
\quad NP \quad P
\]

\[
\quad S_1 \quad S_2 \quad N \quad \{ey\}
\]

\[
\ldots V_1-tunci \quad \ldots V_2-tunci \quad \{kan\}
\]

PP: ‘[regardless of whether S_1 or S_2]

As no parts of these idioms can be extracted, they are enclosed by <j>. However, the former is not enclosed by /j/ because delimiters like -mun, which indicates a meaning of contrast, can be inserted in between (e.g., yatcap-a-mun po-ass-ciman ‘although \ldots has despised \ldots’).24

We assume that the idiom in (20) has a structure of “serial verb constructions.”

\[\text{The structure in (20) is taken from C-H Roh (2013: 78).}\]

\[\text{It is true, however, that yatcap-a po- does not allow regular adverbs to be inserted in between (\text{yatcap-a cengmal po-ass-}). As there are many environments where only delimiters are allowed in Korean, we may have to introduce a new notation, say !...!, to be placed on the upper V' in (20). Then, the node <!V'> would indicate that it is allowed to be expanded only with delimiters, excluding regular modifiers.}\]
In general, the string \{V-e/a po\} is analyzed as a sequence of a main verb and an auxiliary verb. However, as an idiomatic expression, [yatcap-a po] does not seem to contain po- as an auxiliary verb, which means ‘to try’ and takes a VP complement (cf. C-H Roh 2013: 76). It would make more sense assuming that the idiomatic meaning ‘to despise’ comes from the combination of two main verbs po- ‘to see/regard’ and yatcap- ‘to make a low estimate of’ than from the combination of the auxiliary verb po- ‘to try’ and its VP complement. Please note that the verb yatcap- is inside the VP in the latter assumption. Without positive evidence for the latter assumption, we will assume that the two verbs involved are both main verbs constituting a serial verb construction. The idiom in (21) has syntactic anomalies as well as semantic anomalies. It has an abnormal NP structure. No regular NPs in Korean contain two clauses as the dependents of its head noun. In addition, the head noun kan is a “clitic,” which requires a modifier as its host. However, it is not clear whether the V-tunci can be a legitimate host or not, because it is not a (regular) modifier.25

Lastly, we will examine idioms like kkoli(-lul) chi- and pihayngki(-lul) thaywu-, which require their own complements (cf. C Chung & J-B Kim 2007: 437, HS Kim 2013: 60). The former means ‘to seduce’ and the latter ‘to overpraise, to throw bouquets, to flatter (jokingly).’ The latter is mainly used in an informal situation.

\[(22)  \]  
\[a. \text{kkoli(-lul) chi-} \] 
\[\text{tail-Acc swing [literally] ‘to swing a tail’} \] 
\[b. \text{kkoli(-lul) mence chi-} \] 
\[\text{first ‘to seduce first.’} \] 
\[c. \text{[meyli-ka chi-n] kkoli Mary-Nom -Relativizer [no idiomatic meaning]} \] 

\[(23) \]  
\[a. \text{pihayngki(-lul) thaywu-} \] 
\[\text{plane-Acc let take [literally] ‘to let (someone) take a plane’} \] 
\[b. \text{pihayngki(-lul) caewu thaywu-} \] 
\[\text{often ‘to overpraise (someone) often’} \] 

25 As for the structure in (21), we have not decided whether V-tunci as a whole or only -tunci will be put into \{…\}. That depends on whether we treat -tunci as a clitic or not. If it is a clitic, it will be put into \{…\} alone. On the other hand, if it is an inflectional affix, V-tunci will be put into \{…\}.
c. [apeci-ka] thaywu-n [pihayngki]
   father-Nom -Relativizer [no idiomatic meaning]

As we can see from the examples in (b), the idioms can be internally expanded, but, as we can see from the examples in (c), their component parts cannot be extracted maintaining their idiomatic meanings.\(^{26}\)

One special property of the idioms in (22) and (23) is that they require NPs as their own complements, although they consist of a verb and its object NP:\(^{27}\)

\[(24)\]

\[\text{a. meyli-ka cyon-hanthey kkoli-lul chi-ess-ta.} \]
   Mary-Nom John-to -Past-Decl
   ‘Mary seduced John.’

\[\text{b. na-nun ttayttaylo tongsayng-ul pihayngki-lul thaywu-ess-ta.} \]
   I-Top sometimes younger brother-Acc -Past-Decl
   ‘I overpraised my younger brother jokingly sometimes.’

In sentence (a), the idiom \(kkoli(-lul)\) \(chi\)- requires its own complement PP \(cyon-hanthe\) and, in sentence (b), the idiom \(pihayngki(-lul)\) \(thaywu\)- takes its own object NP \(tongsayng-ul\).

For the analysis of sentence (24a), C Chung \& J-B Kim (2007: 438) assume that the verb \(chi\)- itself has two complements, \(kkoli-lul\) and \(cyon-hanthey\). Under this approach, they have to assume that \(chi\)- is ambiguous between a literal meaning ‘to swing’ and an idiomatic meaning ‘to seduce.’ It seems to be more natural to assume that the meaning ‘to seduce’ comes from the whole idiom \(kkoli(-lul)\) \(chi\)- rather than from the verb \(chi\)- itself. In addition, the complement \(cyon-hanthey\) does not seem to be required by the verb \(chi\)- but by the whole idiom.

\(^{26}\) According to C Chung \& J-B Kim (2007: 429), the idiom \(kkoli(-lul)\) \(chi\)- allows “topicalization”:
   \(kkoli-nun meyli-ka cence chi-ess-ta\) ‘Mary made seduction first.’ If this topicalization is a subtype of extraction, we would have to classify sub-classes of extraction (cf. footnote 24). However, it is possible that the expression \(kkoli-nun\) is a scrambled element rather than a topologized element.

\(^{27}\) Of course, causative verbs like \(thaywu\)- in Korean can have two accusative NPs as their complements. However, the “second object” \(tongsayng-ul\) in (24b) is different from one in regular causative constructions. In regular causatives, the second object is more naturally realized with an \(-eykey\) PP, i.e. \([NP-eykey]\). However, if we replace \(tongsayng-ul\) with \(tongsayng-eykey\), the sentence can only have a literal meaning. Thus, we assume that \(tongsayng-ul\) can better be analyzed as a direct object of the idiom \(pihayngki-lul\) \(thaywu\).
A representational system of idiomatic constructions ...  515

HS Kim (2013: 60) assumes that there are two direct objects involved in (24b). However, although he posits a restriction which says that *pihayngki* is “dependent on” *thaywu-,* it does not make sense to assume that the verb *thaywu-* requires two direct object NPs. Please note that (24b) is not an example of “Double/Multiple Accusative Constructions.” In these constructions, “the semantic relationships between the accusative NPs are strictly restricted to a part-whole and type-subtype relations” (Chae & Kim 2008: 888). The relationship between *tongsayng* and *pihayngki* is far from being either part-whole or type-subtype. We can analyze the sentence as follows (cf. Chae & Kim 2008: 890).

(25)

\[
\begin{array}{c}
\text{VP}_2 \\
\text{NP} \\
\text{tongsayng-ul} \\
\text{pihayngki-lul} \\
\text{taywu-ess-ta}
\end{array}
\]

As an (idiomatic) construction, VP$_1$ is enclosed by <...> (cf. (23c)), and carries the meaning of ‘to overpraise jokingly,’ together with the subcategorization frame [OBJECT NP] (cf. footnote 10). In this analysis, the two NPs do not have any direct relationship. The NP *pihayngki* is inside the construction as the object of the verb *taywu-* and the NP *tongsayng* is the object of the idiomatic VP *pihayngki-lul taywu-.* As idiomatic phrases are listed in the component of constructions rather than in the component of lexical entries in our approach (cf. (9)), we are not obliged to consider the verb *thaywu-* as a verb which takes two (direct) object NPs. The analysis here shows that idiomatic parts and non-idiomatic parts of a sentence are integrated into each other very naturally and are treated uniformly and systematically in our framework (cf. the analysis of the *V one’s head off* idioms in section 4.2.).
6. Conclusion

In this paper, we have introduced a framework for representing idiomatic expressions, i.e. constructions as we define them. In general, previous analyses on idioms are not flexible enough to deal with the gradational nature of syntactic and semantic anomalies of idioms, and/or are not explicit enough to be useful in such areas as natural language processing. We believe, however, that the present framework has necessary flexibility and explicitness for treating them. First of all, it turned out to be very useful in formalizing various types of idioms precisely, because it can capture not only the irregularities but also the regularities of idioms naturally. In addition, it has been shown that idiomatic parts and non-idiomatic parts of a sentence can be analyzed uniformly and systematically.

We have focused on providing analyses of idioms. However, the present framework can be easily adapted for treating other multiword expressions like not only clichés and a variety of collocational expressions but also “compound nominals,” proper names and so on (cf. Jackendoff 2013: 75, Sag et al. 2002). When we analyze these multiword expressions of a language under our framework, the resulting resources will become very valuable for the processing of the language. As they will provide pre-analyzed parts, i.e. constructions, to the sentences concerned directly, the accuracy and speed of parsing will be enhanced significantly. In addition, they will play an important role in lexical and structural disambiguation, which is one of the most problematic areas in natural language processing.

References


Copestake, Ann, Farbre Lambeau, Aline Villavicencio, Francis Bond, Timothy Baldwin,
A representational system of idiomatic constructions ...


