電子化辞書における英語連語の記述・表現法

末松　博、杉浦真弓、有岡昌子、Timothy JONES

(株) 日本電子化辞書研究所

電子化辞書における連語の二段階表現法を英語を例にとって提案する。ＥＤＲの現仕様と比べて、記述力、明瞭度、処理性において優れている。第一段階として、連語を統語機能をノードとする部分統語木で表現する。各ノードをパスにより、形態・統語・意味の全てに渡る制約条件に対応付ける。第二段階として、構成語に対し、独立の見出しを与えることにより、辞書上に部分統語木を分散表現する。見出しに対し共起語識別標識、また、文法情報として下位範疇化情報、パス、等を付与することにより、構成語間の関係を保証する。分解して表現された意味に対しては、単一概念を形成することを示す共起概念識別標識を付与し、単語としての概念と同様に扱う。

DESCRIPTIVE AND REPRESENTATIONAL FRAMEWORKS FOR ENGLISH COLLOCATIONS IN AN ELECTRONIC DICTIONARY

Hiroshi SUEMATSU, Mayumi SUGIURA, Masako ARIOKA, and Timothy JONES

Japan Electronic Dictionary Research Institute, Ltd. (EDR)

Mita Kokusai Building Annex, 4-28, Mita 1-chome, Minato-ku, Tokyo 108, Japan

A two-step framework for representing collocations in an electronic dictionary (ED) is proposed (with English examples) as an alternative to the current EDR specification, with notable advantages of descriptivity, clarity, and processibility. First, collocations are described in a syntactic sub-tree whose nodes are expressed with syntactic functions. Each node is correlated, by node path, with a bundle of constraint features on all levels of morphology, syntax, and semantics. Then, the sub-tree is distributively represented in an ED by assigning independent word entries to constituents. The relationships among constituents are assured by assigning them the same ID's, subcategorization information, and node path.

1 Introduction

EDR is collecting a large number of fixed English expressions composed of more than one word, i.e. compound words, idioms, and collocations, because their treatment is inevitable in constructing large-scale, robust NLP systems. The amount of idioms easily exceeds 15,000 [Cowie (1975)], and that of compounds is theoretically infinite if technical terms in compound form are taken into account.

The descriptive framework for machine use has to be carefully designed from the viewpoints of not only computation but also linguistics. A fixed expression is found in almost any part of speech, shows *syntactic* characteristics among its constituents, and has a close link with meaning.

Previously, the EDR tried a *one-word* approach [EDR (1988)], i.e. registering the expression as one word but allowing declensions and simple insertions. This method focuses on the constituents' characteristic of being located in close positions. However, it has become evident as the research went on that syntactic characteristics of collocations exceed the descriptive power of this approach: it is hardly capable of handling transforms, agreement, and complex insertions.

Our current representation framework [EDR (1990)] is the combinatory approach of *one-word* and *compositional* methods. The leftmost constituent is used as a retrieval entry, and the relationships with the remaining constituents are described by specifying their sequential order and insertable portions, and by specifying the *non-hierarchical* phrase structure. This approach offers more freedom in describing the separability of constituents and agreements. However, it is still problematic in dealing with transforms and in formalizing the insertable portion. Also, it creates information concentration on certain frequently used verb entries, which is a burden for processing syntactic and semantic ambiguities.

There are some compositional approaches in dictionaries specialized for machine translation such as Schenk (1986), Meyer (1990), and Suematsu (1991a). They resemble each other in the sense of having adopted the *syntactic function* (e.g. subject and direct object) as the key feature, and in the sense of allowing *transformation.* Schenk (1986) represents idioms as syntactic sub-trees in the lexicon. This approach is limited to collocations with verbal heads. The rest are treated as one-words, and their constituents' syntactic characteristics are ignored. The latter two share a frame-slot representation. Meyer (1990) registers all collocation information under the syntactic head's entry. This way of registration yields the same 'information concentration' problem as the current EDR approach. Suematsu (1991a) has proposed a distributive representation framework for all types of fixed expressions by assigning independent entries to constituents, which are logically linked by ID's, and by describing their relationships in terms of syntactic functions. This method solves the problem of information concentration, and the other problems encountered by the current EDR approach, and results in most disambiguation being carried out in the dictionary look-up phase.

We have thoroughly reviewed the proposal of Suematsu (1991a), and propose the refined version of descriptive and representation frameworks as an alternative to the current EDR specification. This can be achieved by extending subcategorization representation to the phrase level, by introducing the notion of *path* into these frameworks, and by introducing logical links (ID's) to concept groups in addition to those for word groups. We will show that these frameworks can handle any syntactic characteristic for all types of fixed expressions.

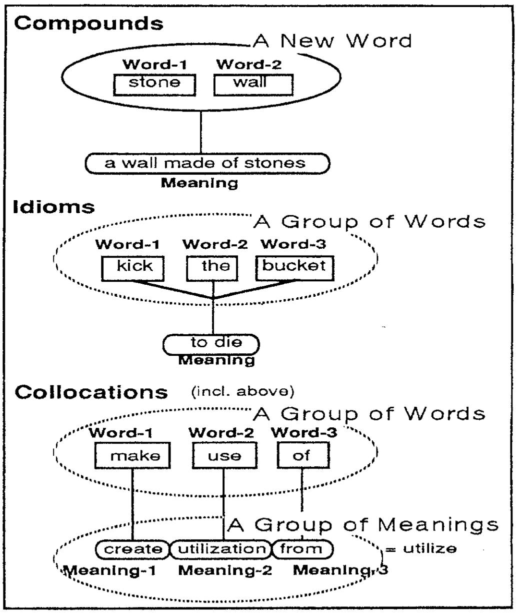

Figure 1. Differences of Fixed Expressions

2 Definition of Collocation and Scope of Discussion

There are three terms relevant to the fixed expression: compound, idiom, and collocation. Their definitions are as shown below [Inoue (1985)], and their intuitive differences are shown in Figure 1.

1. Compound: a combination of words which has become a new word, e.g. a *stone wall*, *part-of-speech*, and *breakfast.*
2. Idiom: an expression composed of more than one word which functions as a unit of meaning, e.g. *kick the bucket*, *take off* (in the sense of *to mimic*), *make an example of*, and *full of beans*.
3. Collocation: any expression the components of which should be learnt together, e.g. all examples described above, plus *make one's mind up*, *make use of*, and *take ~ into consideration/account.*

The focus of a compound's notion is *grammatical unity*, i.e. behavior as a single word. That of an idiom is *semantic unity*, i.e. constituent participation is inevitable in order to express a certain meaning. On the other hand, the meaning of a collocation may be compositional, i.e. the meaning can be constructed from the meaning of each participating constituent. A collocation differs from other ordinary literal expressions from a generation point of view, i.e. in expressing a certain concept in more than a single word, a certain *idiosyncratically fixed* combination of words has to be used. The notion of collocation is the broadest, and includes compounds and idioms.

Compounds which do not involve a space as in *breakfast* are disregarded in this paper since the focus is on those which involve two or more words in a fixed expression. These compounds can be dealt with as one word, or can be treated by applying the framework proposed below.

This paper uses the term *collocation* to represent all these three types of fixed expressions.

3 Issues in Collocation Representation

Issues in representing collocations in the dictionary are as follows.

Linguistic Issue

1. Descriptive power: The framework has to be able to describe all characteristics of collocations on all levels of morphology, syntax, and semantics.

Computational Issues

1. Processibility: The framework has to allow the processing of any variations within a collocation and any combination of collocations within the same phrase and clause.
2. Consistency: Consistency should be maintained throughout the description of all types of collocations. The descriptive framework which deals with phrase level collocations should be shared in describing clause level collocations, and vice versa. For example, *sitting* in a *sitting duck* modifies in a similar fashion as *cold* modifies in *give somebody cold feet.*
3. Avoidance of information concentration: Registering collocations under certain frequently used words results in excessively uneven dictionary loading, which is hardly maintainable. Functional verbs such as *make, take, come, go,* and *get* are involved in many collocations.

4 Distribution and Characteristics of Collocations

Table 1 gives some examples of English collocations. It shows that practically all parts of speech have a collocational expression, and that diverse patterns are found.

Although collocations present a one-word like characteristic, i.e. their constituents are located in close positions, they present the following syntactic characteristics [Cowie (1975)].

1. Separability: a constituent can, in many cases, be modified by A's and Adv's, and thereby be separated from another element. For example:

   make *heavy* use of
   take *good* care of

2. Mobility: a constituent can frequently exchange its position with another constituent through various transforms such as passive, relative, and emphatic transforms and particle movement. For example:

   The governor *made an example of* these prisoners.

   [Passive + Emphatic]
   → *Of* these prisoners *an example* was *made.*

3. Agreement: various types of agreement are involved in some collocations.

   Reflexive pronoun: devote *oneself* to
   bring somebody to *himself*
   Possessive adjective: make *one's* mind up

   In the first example, *oneself* has to agree with the S, and in the second example, *himself* has to agree with the O_d in person, number, and gender. In the third example, *one's* has to agree with the S.

   Also, agreements between a O_d and an A_{pp} are sometimes found.

   make an example of
   → The governor made *an* example of these *prisoners.*
   → The governor made *examples* of these *prisoners.*

Table 1. Distribution of Collocations

Main Function	Part-of-Speech Pattern	Syntactic Function Pattern	Examples
N	A N	TN	mechanical translation
	NN	TN	cannon ball
	N P N	N T_{pp}	part of speech
V	N V	A V	carpet bomb
	V Cnj V		cut and thrust
	V A	V C_s	go berserk
	V Pcl	V A_{pcl}	give up, walk out
	V P	V A_{pp}	look at, rely on
	V Pcl P	V A_{pcl} A_{pp}	catch up with
	V (N) Pcl	V (O_d) A_{pcl}	blow ~ up
	V N P	V O_d A_{pp}	make an example of
	V N Pcl P	V (O_d) A_{pcl} A_{pp}	bring ~ around to
	V N	V O_d	catch sb's attention
	V (N) A	V (O_d) C_o	drive ~ mad
	V (N) D A N	V (O_d) C_o	make ~ a laughing stock
	V (N) D A N	V (O_i) O_d	set ~ a good example
A	A P	C_s A_{pp}	afraid of
Adv	P A P N	A_{pp} A_{pp}	as cool as a cucumber
P	P N P	A_{pp} A_{pp}	in front of
	Pcl P	Apcl App	down to
Cnj	Adv Adv ~ Cnj Adv	A ~ A ~	not only ~ but also ~
Int	A N	TN	Good bye.
Pron	D Pron	DN	a few
S	N V P D N	S V A_{pp}	Time flies like an arrow.

(), ~: unfixed, sb: somebody
First and Second Columns: N: Noun, V: Verb, A: Adjective, Adv: Adverb, P: Preposition, Pcl: Adverbial Particle, Cnj: Conjunction, D: Determiner, Int: Interjection, Pron: Pronoun, S: Sentence
Third Column: D: Determinative, T: Attributive, T_{pp}: Attributive PP (Prepositional Phrase), N: Nominal, Ar: Rightward Adverbial, V: Verbal, S: Subject, O_i: Indirect Object, O_d: Direct Object, A_{pcl}: Adverbial Particle, A_{pp}: Adverbial PP, A_l: Leftward Adverbial, C_s: Subject Complement, C_o: Object Complement

   → The governor made *an example* of this *prisoner.*

   When the *prisoner* is plural, the *example* may be either singular or plural. However, when singular, the *example* has to be singular.

4. Semi-idioms: some expressions, so-called *semi-idioms*, have both idiomatic and literal characteristics. The deletability of *out* in *draw out* shows a literal characteristic. The existence of one-word counterpart *withdraw* shows an idiomatic characteristic.

In addition, there are other syntactic characteristics to which attention has not been paid in Cowie (1985):

5. Dual characteristics: some compound terms present *dual* characteristics of an A and a N. For example, *mechanical translation* can be modified by *perfect* and *perfectly*. In the former case, *translation* is modified, meaning the result of translation is excellent. In the latter case, *mechanical* is modified, meaning the translation process is carried out completely by a machine.
6. Nesting and overlapping possibilities: there are collocations which can take a clause or a phrase in the middle of the lexically specific constituents such as *take ~ into consideration.* The insertable portion can contain again other collocations. And, there are ones which share their elements. For example:

   I *take* that sitting duck *into consideration.*
   I *made use* and *fun of* him.

Therefore, we see that the behavior of collocations should be regarded as continuous with that of less idiosyncratic expressions, which have constraints on various levels of morphology, syntax, and semantics.

5 Problems in Preexisting Frameworks

5.1 ONE-WORD APPROACH

When we tried to represent collocations in the one-word registration approach [EDR (1988)], we encountered the following problems.

Linguistic problems: It is hard to:

1. cope with positional exchange of collocation constituents. (Mobility)
2. cope with conjugations and declensions.
3. capture the dual aspect of parts of speech in some compounds.
4. formalize the insertable (i.e. modifying) portion. (Separability)
5. specify fully the form of agreement.

Computational problems: It is hard to:

6. keep the description consistent in all types of collocations. (Consistency)
7. extract literal meanings in the processing. (Processibility)
8. cope with nesting and overlapping possibilities. (Processibility)

The foremost difficulty is number 1, which accelerates the number of variations to be registered. For example, the idiom *make an example of somebody* shows 50 variations due to conjugation, declension, and transforms.

The existence of semi-idioms indicates that the representation framework which deals with normal expressions has to be combined with collocation representation.

Therefore, we see that some better syntactic framework has to be involved in expressing collocations.

5.2 CURRENT EDR APPROACH

Our current representation framework [EDR (1990)] is the combinatory approach of *one-word* and *compositional* methods. The first (leftmost) constituent is used as a retrieval entry, and the relationships with the remaining constituents are described by specifying their sequential order and insertable portions, and by specifying the *non-hierarchical* phrase structure.

This approach offers more freedom, compared with the one-word approach, in describing the separability of constituents and agreements. However, it is still problematic in dealing with transforms, in formalizing the insertable portions, and in describing complex, hierarchical collocations such as *kill the goose which lays the golden eggs*. Also, it creates the problem of information concentration on certain frequently used verb entries.

5.3 COMPOSITIONAL APPROACHES

There are some compositional approaches in dictionaries specialized for machine translation, such as Schenk (1986), Meyer (1990), and Suematsu (1991a). They resemble each other in the sense of having adopted the *syntactic function* (roughly equivalent to the grammatical case) as the key feature, and in the sense of allowing *transformation.*

Schenk (1986) represents idioms as syntactic sub-trees (called *basic S-trees*) in the lexicon. This approach is limited to idioms with verbal heads since transforms are observed mainly in this type of idiom. The rest are treated as one-words, and their constituents' syntactic characteristics are ignored. Also, the lexicon depends on the system design, and is hardly neutral to applications. The matching of the leaf of the basic S-tree becomes difficult as the depth of the tree grows.

The latter two share a frame-slot representation. Meyer (1990) registers all collocation information under the syntactic head's entry. The application domain is very limited because it is an experimental system. This way of registration yields the same problem as the current EDR approach, i.e. information concentration on certain frequently used lexical entries. The matching of the leaf of the syntactic tree is also problematic as in Schenk (1986).

Suematsu (1991a) has proposed a distributive representation framework for all types of fixed expressions, under independently assigned constituent entries, which are logically linked by ID's, and the relationships of which are described in terms of syntactic functions. This method solves the problem of information concentration, and provides most disambiguation to be carried out in the *dictionary look-up* phase. This way of representation seems promising in NLP, because it avoids all problems in the one-word and compositional approaches.

6 Refinement of Syntactic Approach

In dealing with collocations syntactically, the framework must be able to deal with syntactic relationships among constituents (i.e. syntactic structure), and their behavioral constraints on all levels of morphology, syntax, and semantics. This chapter discusses requirements for a descriptive framework.

The syntactic sub-tree or its equivalent is necessary, because there has to be an intermediate node, in order to group constituents together. In this way we can capture their behavior as phrase or clause in modification relations and transforms, as discussed in Chapter 4. There is the benefit that it can capture any combination of the syntactic relationships of constituents, as long as they obey the condition of one-head.

There is another benefit of adopting a syntactic subtree: the constraints can be correlated with any node in the sub-tree by providing *node path* information, on condition that the sisters in the tree are differentiated. This way of node identification is superior to any other method, such as assigning sequential identification numbers to each node as has been adopted in Meyer (1990), because it shows relationships with other constituents, and can be utilized also for processing as will be shown in Chapter 9.

Subcategorization patterns, classified according to *syntactic functions* (e.g. O_d and C_o), should be included. This is because part-of-speech information is not sufficient to capture constituent behavior; word dependencies can be captured in the subcategorization framework; and *obliqueness hierarchy* among sisters can be utilized in identifying the linear positions of collocation constituents. The obliqueness hierarchy is the relative order observed among syntactic functions as shown below.

Obliqueness hierarchy for a sentence [Suematsu (1991a)]:

$S \ll O_i \ll O_d \ll A_{pcl} \ll A_{pp} \ll C_s \ll C_o \ll A_l \ll A_{cnj}$

Obliqueness hierarchy for a phrase:

$D \ll T \ll T_{pp}$

In Table 1, the contrast of *make somebody a laughing stock* and *set somebody a good example* shows the necessity of syntactic functions since they present an identical part-of-speech pattern, but present different functions. The first example is of the O_d C_o pattern. The second one is of the O_i O_d pattern. Also the table shows that colloca-

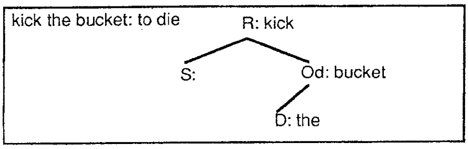

Figure 2. Syntactic Sub-tree

tions can be captured by subcategorization patterns which follow the obliqueness hierarchies.

Inhibited elements have to be specified, since a syntactic tree itself describes only argument information for each constituent, and unlike the one-word approach, it does not specify any inhibited adjunct. This is a trade-off for having a flexible system. For example, *bucket* in *kick the bucket* does not allow any modification such as follows.

* He kicked the *peaceful* bucket.

7 Descriptive Framework

7.1 BASIC FRAMEWORK

From the above discussion, it is evident that the syntactic sub-tree whose nodes are expressed in terms of syntactic functions is sufficient as a basic descriptive framework. The BNF representation for the syntactic sub-tree is as follows.

```
[sub_tree] ::=  [synt_func]
             |  [synt_func [sub_tree]]
             |  [sub_tree] [sub_tree]
```

where, synt_func is a syntactic function and sub_tree is a syntactic sub-tree.

The graphical representation of the sub-tree for *kick the bucket* is shown in Figure 2, and the exhaustive list of syntactic functions and their definitions are shown in Table 2.

7.2 DIFFERENTIATION OF SISTERS

When a certain syntactic function appears more than once as sisters (e.g. A_{pcl} and A_{pp}), they are differentiated according to objective word-order around their head; for the purpose of path identification in correlating the constraints with the node; and for the purpose of linear position identification of a collocation constituent. Repetitive syntactic functions are marked by numerical suffixes which are assigned from the closer positions to their head as follows:

... 3 « 2 « 1 « HEAD « 1 « 2 « 3 ...

Examples are:
After the head verb: A_{pp1} A_{pp2}
Before the head noun: T_2 T_1

This marking method is better than assigning numerical suffixes just from left-to-right, because constituents are fixed in relation with their head.

7.3 EXAMPLES OF SYNTACTIC SUB-TREES

Some examples of syntactic sub-trees from various types of collocations are shown below.

Predicate
make use of → $[_{R}[_{S}]$ mak(e) $[_{Od}$use] $[_{App}$of $[_{Op}]]]$

Predicative
(be) afraid of → $[_{Cs}$afraid $[_{App}$of $[_{Op}]]]$

Nominal
a sitting duck → $[_{R}[_{D}$a] $[_{T}$sit(ting)] duck]
part of speech → $[_{R}$part $[_{Tpp}$of $[_{Op}$speech]]]

Parallel
not only ~ but also ~
→ $[_{R}[_{L1}[_{Ar}[_{Ar}$not] only]] but $[_{L2}[_{Ar}$also]]]
It rains dogs and cats.
→ $[_{R}[_{S}$It] rain(s) $[_{Al}[_{L1}$dog(s)] and $[_{L2}$cat(s)]]]

Sister
from A to Z → $[_{R}[_{App1}$from $[_{Op}$A]] $[_{App2}$to $[_{Op}$Z]]]
not ~ at all → $[_{R}[_{Ar}$not] $[_{Al}$at $[_{Op}$all]]]

Comparison
The pen is mightier than the sword.→ $[_{R}[_{S}[_{D}$The] pen] is $[_{Cs}$might(ier) $[_{Acnj}$than $[_{S}[_{D}$the] sword]]]]

R is assigned to the top of the sub-tree as a default value when the syntactic function is not fixed. When the root is a V, it can be S, O_d, O_p, C_s depending on the usages of the V. When the function of the root is fixed, the corresponding value is assigned as in the example of *be afraid of.*

7.4 PATH IDENTIFICATION AND CONSTRAINTS

Path information to identify a node is given in terms of syntactic functions as follows.

PATH <(Own) (Mother) (Grandmother)...(Root)>

In identifying *laughing* in *make somebody a laughing stock*, the path is described as follows.

PATH <T C_o R> (i.e. the T directly under C_o which is directly under R)

Using the path, any constraint can be correlated with the constituent on all levels of morphology, syntax, and semantics. A morphological constraint is, for example, a headword (HEADWORD). Syntactic constraints are: e.g. part of speech (POS), number (NUM), person (PER), gender (GEN), countability (COUNT), verb form (VFORM), and agreement with the S, the O_d, and others (AGR). Semantic constraints are: e.g. the semantics of the constituent (SEM), typical semantics (TYPSEM), and a conceptual relation with the concepts of other constituents, if the collocation is semantically decomposable.

Constraints which can be regarded as applying to the whole expression, such as transforms and non-decomposable semantics, are assigned to the sub-tree as a whole.

Note there is no necessity to specify on the intermediate node phrase and clause categories such as a noun phrase, verb phrase, and that-clause, as is currently adopted in the EDR Word Dictionary specification, because all intermediate nodes correspond to terminal symbols, i.e. constituent words, in this framework.

Note also that this way of correlating constraints with the node gives, better descriptive clarity than the current EDR method, which describes various constraints on the nodes of the sub-tree.

7.5 INHIBITION

We can inhibit the cooccurrence of words by specifying inhibitional syntactic function subcategorization (ISUBCAT) under the head node. Adjuncts such as *peaceful* and other prepositional phrases modifying *bucket* in *kick the bucket* are prohibited by specifying as follows.

ISUBCAT <T, T_{pp}>

The mechanism of inhibition must be considered carefully, because the correspondence between the syntactic functions and the immediate dominants is not always one-to-one. Syntactic functions S, O_i, O_d, O_p, C_s, C_o, and D are exclusive in that each corresponds to one and only one immediate dominant, and can be viewed as a slot having the capacity of only one room for an immediate dominant. Hence, specifying them in ISUBCAT affects only one cooccurrence of a certain type of immediate dominant.

However, other types of adjunctive syntactic functions such as A_{pp} and T_{pp} are inclusive, as they can be shared by more than one sister, either obligatorily or optionally. Hence specifying them in ISUBCAT affects all adjuncts of the types specified. This is sometimes inadequate as a behavioral description of a collocation. It is wrong when the head has a lexically specific adjunct, but inhibits a second adjunct of the same type.

To cope with this situation, numerical suffixes are used to show the inhibition of a cooccurrence of the same syntactic function. ISUBCAT <T_2> means the first T is described in the tree and the second T and so on are prohibited to cooccur.

O_d can sometimes be assigned to two immediate dominants as in:

Forgive me my sins.

There is no need to specify ISUBCAT <O_{d2}> for ordinary mono-transitive verbs, because the second direct object can be regarded as exceptional, and because O_d can be regarded as belonging to the exclusive syntactic function category. When O_d without the numerical suffix is assigned to a

Table 2. The Syntactic Functions Used in the Syntactic Sub-tree Description

Root (R)		Definition	Root of the syntactic sub-tree
Subject (S)		Definition	A nominal argument which controls the inflection of the predicate verb in the form of NOM + V.
		Characteristics	1. Becomes the first candidate for the antecedent of a reflexive pronoun. 2. Takes nominative case in the declension.
Object (O)			
	Direct Object (O_d)	Definition	A nominal argument which receives the performance expressed by the predicate verb in the form of V + NOM.
		Characteristics	1. Usually expresses inanimate things. 2. Takes objective (accusative) case in pronominal expression. 3. Located after a verb and an indirect object.
	Indirect Object (O_i)	Definition	A nominal argument which is related through a O_d in the form of V + NOM.
		Characteristics	1. Usually expresses a human being. 2. Takes objective (dative) case in pronominal declension. 3. Located before a direct object.
	Preposition Object (Op)	Definition	The object of a preposition
Complement (C)			
	Subject Complement (C_s)	Definition	An argument with nexus relationship with the subject
		Characteristics	1. Impassivizable 2. Takes nominative case in pronominal declension.
	Object Complement (C_o)	Definition	An argument with nexus relationship with the object
		Characteristics	1. Impassivizable 2. Takes objective case in pronominal declension.
Adverbial (A)			
	A_{pcl} A_{pp} A_{cnj}	Definition	An adverbial argument distinguished according to the head's part-of-speech category
		Characteristics	1. Typically a space adverbial 2. Relatively fixed compared with other optional adverbials
	Al	Definition	An adverbial argument which modifies leftward, other than the above
	Ar	Definition	An adverbial element which modifies rightward
Determinative (D)		Definition	Determiners and possessive expressions
		Characteristics	1. Takes genitive case in pronominal declension.
Attributive (T)			
	T	Definition	An attributive which modifies rightward
		Characteristics	1. Typically a qualitative adjective
	T_l	Definition	An attributive which modifies leftward, with non PP form
	T_{pp}	Definition	An attributive which modifies leftward, with PP form
Parallel (L_1, L_2, ...)		Definition	Parallel relation, i.e. *and, but, or,* ','. Assigned according to the objective word-order.

NOM: An argument with nominal function. Based on Inoue (1985), Quirk (1972), and Cowie (1975).

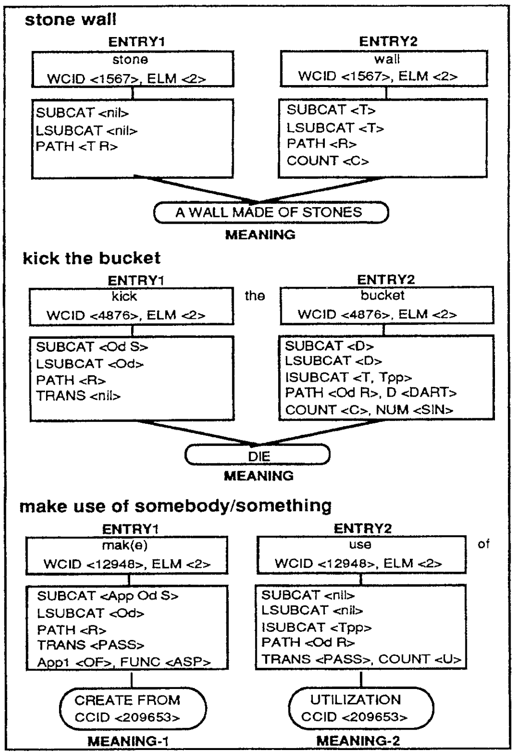

Figure 3. Dictionary Structure for Collocations

certain immediate dominant, it occupies the O_d position and there is no room for the others.

7.6 DESCRIPTION EXAMPLE

The description for *kick the bucket* is as follows.

Syntactic Tree

$[_R[_S]$ kick $[_{Od}[_D$the$]$ bucket$]]$

┌ TRANS <nil> ┐ # No transforms allowed #
│ COLTYPE <N> │ # Non compositional #
└ SEM <to die> ┘ # Meaning of the collocation #

Constituents Information

┌ PATH <R> ┐ # Refers to the root node. #
│ HEADWORD<kick> │ # Fixed portion is described. #
│ CONJUG <kick kicks kicked kicked kicking>
└ POS <V> ┘ # Part of speech is verb. #

┌ PATH <S R> ┐ # Refers to the subject node.#
└ TYPSEM <human> ┘ # Typical semantics: human. #

┌ PATH <D O_d R> ┐ # Refers to the D of O_d #
│ HEADWORD<the> │
└ POS <DART> ┘ # DART: Definite Article #

┌ PATH <O_d R> ┐ # Refers to O_d #
│ HEADWORD<bucket>
│ DECL <bucket nil> │ # Plural form prohibited. #
│ POS <N> │
│ COUNT <C> │ # Countable #
│ NUM <SNG> │ # Singular #
│ PER <3RD> │ # 3rd person #
└ ISUBCAT <T, T_{pp}> ┘ # No modification allowed. #

8 Dictionary Representation Framework

The syntactic sub-tree is distributively represented in the dictionary in order to avoid excessive information concentration, and to manage any spatial and temporal combinations of constituents in parsing as follows.

1. Lexically specified content words (i.e. V, N, A, and Adv) are given independent word entries.
2. Functional word information is basically described in the content word entry. The D is described in the head N. The Pcl is described in the head V. The Cnj and P are described in their head and/or dependent, according to which item is specified in the collocation.
3. Each constituent of a collocation is assigned the same word-cooccurrent ID (WCID) to show that the constituents belong to a single expression.
4. When the semantics of the collocation is decomposable (i.e. the meanings of its constituents contribute to the meaning of the fixed expression), meanings are linked with the same concept-cooccurrent ID (CCID) to show that they form a single concept. A concept group, denoted by the same CCID, is treated as a single concept, and assigned relationships with other concepts. The relationships are: equivalent, super/sub, and similar.
5. When the semantics of the collocation is not decomposable, each constituent entry is linked to the same meaning represented by the collocation.
6. In each entry, the following fields are described in terms of syntactic functions to assure the correct relationships among constituents.
   a) A subcategorization pattern (SUBCAT)
   b) Lexically specified subcategorized-for elements (LSUBCAT)
   c) Inhibitional elements (ISUBCAT)
   d) The entry's *path* in the syntactic sub-tree (PATH)
7. Relevant constraints on all levels of linguistics are described under these entries.
8. Lexical variants (e.g. take into *consideration/account*) are assigned the same word-cooccurrent ID because their *exclusivity* is guaranteed in this framework.

Figure 3 shows examples: *stone wall*, *kick the bucket*, and *make use of ~*.

In the case of collocations with decompositional meaning, independent word entries with exactly the same behavior can be collapsed into one entry by listing their WCID's.

9 Processing Collocations

According to this representation framework, collocations are disambiguated (or selected in generation) using constraints on all levels of logical link, morphology, syntax, and semantics. The most notable differences from other compositional approaches are: using *logical link* in the phase of dictionary look-up, and using *path* in the syntactic processing phase.

9.1 DICTIONARY LOOK-UP

In analysis, most disambiguations can be conducted by implementing a simplified parsing mechanism in the dictionary look-up phase, i.e. the phase to load information relevant to the literal senses.

Upon detecting the possibility of a collocation constituent, only the WCID is loaded into the buffer. And, upon detecting another constituent whose WCID coincides with a previously extracted WCID, information on the collocation is loaded under these constituents. The number of constituents of content words (denoted by the value in ELM in Figure 3) can be combined to detect more precisely the collocation possibility.

In the morphological analysis, the candidates are checked as to whether they conform with morphological constraints, such as declension and conjugation forms.

9.2 IDENTIFICATION METHOD

In the syntactic analysis, the possibility of a collocation is confirmed by examining the constraints on the local relationships among the constituents with the same WCID.

1. Path constraint: the value of PATH subtracting the constituent-candidate's own syntactic function (i.e. the first value in the field) has to coincide with the PATH of the head candidate.
2. Lexical specificity constraint: the constituent-candidate's own syntactic function has to be described in the LSUBCAT of the head candidate.
3. Saturation constraint: both SUBCAT and LSUBCAT have to be saturated, because the participation of all lexically specific and non-specific constituents are necessary.

Other information such as ISUBCAT, TRANS (transform), and AGR are utilized to assure the correct relationships.

9.3 GENERATION METHOD

Collocation generation can be achieved in a similar way to the analysis. When a (group of) meaning(s) which corresponds to a collocation has been selected because it is inevitable or preferred for the sake of clarity of meaning, syntactic and morphological information under the entries with identical WCID's are utilized. The constituents are correlated with each other using path and subcategorization information, and are, in principle, ordered in the sequence of the obliqueness hierarchy. The corresponding word is generated in a way that satisfies syntactic and morphological constraints required for the selected collocation.

10 Applications of the Framework

We have dealt with a fixed expression's many-to-one, and many-to-many correspondences between words and concepts respectively, from the viewpoint of cooccurrence, by introducing logical links in both the word level (WCID) and the concept level (CCID). We can apply this cooccurrence framework to one-to-many correspondence between them, i.e. the definition of a word using defining senses, as is currently carried out by Longman. The method is the application of the method discussed above: describe a syntactic sub-tree for a definition, and convert it distributively into the dictionary space. A WCID and CCID are used to express their unity on the word level and the concept level. And the 'equivalent' relation is used between the CCID and the concept ID for the word defined.

This cooccurrence framework is useful for rephrasing, paraphrasing, and summarizing through the links of equivalent, super/sub, and similar concept relationships within a language. This is because the framework provides the way to control the number of constituents in an expression.

Also, this framework can correlate a single word in one language with more than one word in another language. For example, the Japanese word *ani* corresponds to *elder brother* in English. In this framework, the word *ani* is linked to a single concept that defines *ani*, which is linked by the equivalent relation to a combination of concepts which share the same CCID composed of the concept *older* and the concept *brother*. These concepts are in turn linked to the English words *elder* and *brother*.

Furthermore, the framework can contribute to the efficient examination of the *granularity* of the concepts under the same headword. This is because the concept of a constituent which participates in the semantically decomposable collocation can be easily compared with the other literal concepts, since they are registered under the same headword. The concept of *use* in *make use of ~* is the ordinary sense of *utilization*, and hence, there is no need to provide an extra concept, i.e. one that differs from the literal sense of the word use.

11 Conclusion

Based on the proposal of Suematsu (1991a), the revised powerful version of descriptive and representational frameworks for compounds, idioms, and collocations has been proposed as an alternative to the current EDR specification. The descriptive framework is a syntactic sub-tree which heavily utilizes syntactic functions and their obliqueness hierarchies. In order to identify any constituent of any type of collocation, the notion of *path* has been introduced. Constraints on all levels of morphology, syntax, and semantics are correlated with constituents, using the path.

The sub-tree is *distributively* represented in the dictionary by assigning an independent word entry to each lexically specified node of content words. Word-cooccurrent ID's (logical links) are described, to show they form a single expression. A semantically decomposable collocation is assigned meanings corresponding to the constituents just as in expressing literal meanings, and connected by concept-cooccurrent ID's to show that they form a single concept. Constituents of a collocation, whose semantics is not decomposable, are pointed to the same meaning. Subcategorization patterns, and the entry's path in a syntactic sub-tree, are described in order to assure the correct relationships among collocation constituents.

Acknowledgement

The authors are grateful to Mr. Toshio Yokoi, Mr. Seibi Chiba, Mr. Shin'ya Amano, Mr. Chiaki Harada, and the other members of the Word Dictionary Group of the Japan Electronic Dictionary Research Institute, for opportunities to discuss the subjects mentioned in this paper, and for providing them with valuable comments.

Also, the authors are thankful to Dr. Takenori Makino in Toho University for advising them in various ways.

References

Cowie, A. P.; Mackin, R.; and McCaig, I. R. 1975 *Oxford Dictionary of Current Idiomatic English.* Oxford University Press.

EDR 1988 *Word Dictionary, TR-008* . EDR Technical Report, November.

EDR 1990 *English Word Dictionary, TR-026* . EDR Technical Report, April.

Inoue, Yoshimasa 1985 *A Comprehensive Dictionary of English Grammar.* Kaitakusha, Tokyo.

Meyer, Ingrid; Onyshkevych, Boyan; and Carlson, Lynn. 1990 Lexicographic Principles and Design for Knowledge-Based Machine Translation. CMU-CMT-90-118. Carnegie Mellon University.

Quirk, Randolph; Greenbaum, Sidney; Leech, Geoffrey and Svartvik, Jan 1972 *A Grammar of Contemporary English.* Longman Group UK Ltd.

Schenk, André 1986 Idioms in the ROSETTA Machine Translation System. COLING '86. Pages 319-324.

Suematsu, Hiroshi 1991a A Simultaneous Selection Algorithm for English Sentence Patterns. *Proceedings of the International Workshop on Electronic Dictionaries*, TR-031. Japan Electronic Dictionary Research Institute, Tokyo. Pages 190-203. February.

Suematsu, Hiroshi; Sugiura, Mayumi; Arioka, Masako; and Jones, Timothy 1991b Dictionary Representation Frameworks for English Compounds and Idioms. *Proceedings of the 43rd Convention IPSJ.* Fall.