A Study on the Structure of Korean Knowledge Database[1]

Yude Bi, Binhong Wu, Jianguo Xiong

PLA University of Foreign Languages, P.O.Box 036-50, 471003 China
biyude@gmail.com, 3241122lm@sina.com, jianguoxiong@163.com

Abstract. Knowledge is most often and directly expressed with natural language. Therefore, the representation and acquisition of knowledge play a key role in studies on information processing and natural language interpretation. Also, they are the principal issues in establishing a knowledge database. Only highly formalized descriptive systems or knowledge databases can be processed by computers. On the basis of the descriptive mechanisms of lexicalist theory, the present paper tries to provide a detailed description and integration of the syntactic, semantic and other information of Korean, in the form of lexical structures.

Keywords: Korean, knowledge database, lexical structure

1 Introduction

Studies on knowledge projects or intelligent systems have developed to such an extent that exploration into knowledge acquisition and intelligent simulation has become possible. Now, we must pay closer attention to basic theoretical studies, especially in the field of syntax and semantics. These studies present themselves as pioneering research topics in the field of language information processing and have attracted the attention of many experts both at home and abroad. Focus is placed on the description and acquisition of syntactic and semantic knowledge. The present study is intended to construct a language knowledge presentation system for Korean information processing. The core of this system is a knowledge database encompassing both syntactic and semantic information. Doubtlessly, the design of this database plays a crucial role, because it has a direct impact on the operation and finally the overall quality of the system.

2 Theoretical background

As stated above, syntactic and semantic information constitute the core of a knowledge database for information processing, therefore, finding an efficient way for the description of such information determines the final adequacy of the database. This calls for a clarification of the relations between syntax and semantics. On the one hand, semantics puts constraints on the range of available syntactic expressions. And syntax must be loyal to semantic content for the proper communication of the speaker's ideas and intentions. On the other hand, semantics must follow syntactic rules in order to get the ideas properly expressed.

The form of a sentence is subject to both syntactic and semantic constraints. For syntactic constraints between constituents, we have relations such as subject-predicate, verb-object, modifier-head, etc. For semantic ones, we have relations between action and agent, patient, instrument, time, place and others. Such semantic relations are addressed as syntactic-semantic relations. Both syntactic and semantic structures are formed in the derivation or composition of a sentence. In language use, meanings are always expressed with certain syntactic structures or forms. Syntactic meaning is no exception.

[11] This work was supported by the National Social Science Foundation of China (No. 05BYY019) and Brain Korea 21 Project, The school of information technology, KAIST in 2005.

However, semantic and syntactic structures do not have one-to-one correspondence. One syntactic structure may have several semantic relations; one semantic relation may also be expressed by a number of distinctive syntactic structures. Syntax is a dynamic process and must be based on semantics. The constituents of a sentence are confined to the theta-roles permitted by the verb. No doubt, it is not necessary for all theta-roles (case) to be overtly expressed in a sentence. For example, in (1b), the possible agent and patient are not expressed syntactically. Semantics, by contrast, is static and represented by inherent theta-roles and arguments. Semantically, a verb constrains the number and property of the constituents it takes, resulting in different semantic structures. Syntactically, the verb's own features put limit on the possible projection positions of these constituents, resulting in different syntactic structures. The two, semantic and syntactic structures are linked through argument structure. Based on the facts in Korean, we divide syntactic structures into basic structure[2] and surface structure. The former produces abstract sentences, while the latter generates specific ones.
In Korean, this kind of structure is formed by a verb, into which nouns are incorporated, together with auxiliary case markers.

(1) a. 김 선생님이 학생들에게 수학을 강의하신다.
b. 김 선생님은 학생들에게 수학을 강의하신다.
c. 김 선생님께서 학생들에게 수학을 강의하신다.
d. 김 선생님은 학생들에게도 수학을 강의하신다.
e. 김 선생님은 학생들에게 수학도 강의하신다.

(2) a. 그는 사과를 먹었다.
b. 사과는 그가 먹었다..

Sentence (1) may have different forms depending on varying contexts, with its basic syntactic structure "N0 N1-에게 N2-을 V" unchanged. Sentence (2) have two different syntactic forms in accordance with different foci, while its basic structure "N0 N1-을 V" remains constant. In other words, surface structure may be extended to different forms depending on pragmatic needs.
The verb "끼여들다" has two forms, "N0 N1-에 V" and "N0 S-데에 V", as in:

(3) a. 그는 이번 일에 끼여들지 않았다.
b. 한국 전자는 이동 통신 사업에 뒤늦게 끼여들었다.
c. 우리가 일하는 데에 끼여들 생각을 말아라.
d. 명수가 영희와 내가 대화하는 데에 끼여들려고 했다.

Another example is the verb "가까이하다", with its two basic structures: "N0 N1-를/을 V" and "N0 N1-와/과 V".
Given the varying correspondence between semantic and syntactic structures, the basic syntactic structure "N0 N1-를 V(subject + object + predicate)" may take one of the following semantic structures:

a. agent+ patient + action
b. agent + result + action
c. experiencer + exp-theme + action
d. Subject + patient + action

For example:

[2] It is called syntactic argument structure by Hong Jae-Sung (2001). In semantic structure, theta-roles are unmarked, while in the structure "N0 N1-에게 N2-을 V", the constituents are marked.

(4) a. 그는 유리창을 깨드렸다./He broke the window.
b. 그는 소설을 써냈다./He wrote a novel.
c. 그는 어머님을 사랑한다./He likes his mother.
d. 그는 돈을 잃었다./He lost his money.

When semantic structure is converted to syntactic structure, it is first represented as argument structure in logical form, which is then converted into basic syntactic structure via predicate logic calculus. After that, basic structure is turned into different surface structures depending on different linguistic contexts and pragmatic needs. This is shown by the sentences in (4). When basic structure is transformed into surface structure, some constituents may be dropped. In syntax, we have default logical subject and object. Sometimes, we may even drop the predicate. From the matrix of varying surface structures, we abstract the basic structure and get the deep semantic structure. The ultimate aim is to get a clear picture of the compositional relations between the verb and its constituents. In different syntactic transformation matrices, the semantic relation remains unchanged.

Reorientation calculus projection
Semantic structure→argument structure→basic structure→surface structure
(deep syntactic operations)(surface syntactic operations, pragmatic selection)

For instance, an event formed by the verb "읽다/read", and constituents "나/I, me", "교실/classroom" and "책/book" has the following semantic structure (a), argument structure (b), basic structure (c) and surface structure (d):

a. agent + location + patient + time
b. V, agent [_N1, _N2]
location patient
or expressed in function: f(x1,x2,x3)
c. 나-는 교실-에서 책-을 읽-는다./He is reading in the classroom.
d. 나-도 교실-에서 책-을 읽-는다./He is also reading in the classroom.
책-은 내-가 교실-에서 읽-는다./The book, he is reading the classroom.

3. The idea underlying the knowledge database design and its linguistic description

In the study of grammar, the word *form* has two meanings. One refers to the formalization of descriptions, with emphasis on simplicity, precision and formulation. The other refers to something in contrast with *meaning*, or *content*. This latter sense of *form* carries an ontological flavor. Ontologically, syntax, which is different from semantics or pragmatics, is conceived in language forms. In this sense, semantic analyses that cannot be formally tested are meaningless for syntactic studies. Syntactic analysis must be tied to language forms. Only through formal devices can we construct a consistent system of syntax. It is definitely true that formal analysis presents itself as the distinctive feature of syntax as differentiated from semantics and other branches of linguistic studies. That accounts for the prominence of the formal school in the contemporary syntactic circle.

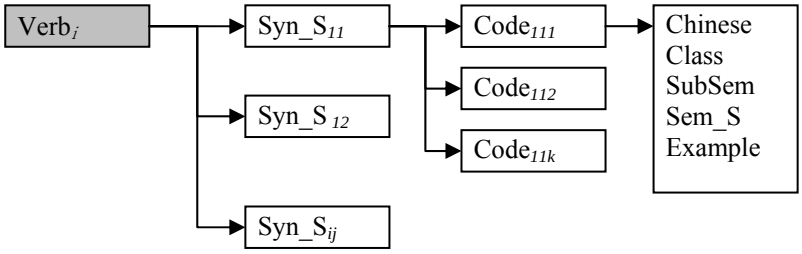
Predicate is the core of a sentence, and its syntactic properties (including the properties of its arguments and theta-roles, the categorial feature of its arguments, and the syntactic features of its theta-roles) are projected onto the basic structure. Based on the facts of Korean, our study attempts to integrate the syntactic and semantic information of Korean verbs, and to classify Korean predicates

based on their syntactic and semantic properties. This work will doubtlessly help in studies on the acquisition of syntactic and semantic knowledge.

In the design of the knowledge database, we adopt an integrational approach, trying to integrate the lexicon and the grammar. In other words, we try to grammaticalize the lexicon and lexicalize the grammar. This knowledge database has the verb as its basic unit, comprising three levels. The first level is basic syntactic structure information; the second level consists of syntactic-semantic classification coding information, and the third level embraces Chinese paraphrase, parts of speech information, semantic structure, semantic features (elements) and exemplifying sentences.

Since this knowledge database encompassing both syntactic and semantic knowledge, we might as well call it Korean Syntax-Semantic Knowledge Database (KSSKD).

According to our definition, the knowledge structure of a verb in the lexicon has the following the form:



Verb----morphological verb
Class----Parts of speech information
Code----syntactic-semantic classification code
Syn_S----basic syntactic structure (syntactic expression)
SubSem----semantic features
Sem_S----semantic structure (semantic expression)
Example----examples
i-----verb number
j---- basic syntactic structure number, range: $1<j<n_i$
k---- semantic structure number, range: $1<k<n_{ij}$

As mentioned before, each verb may have different numbers of basic syntactic structures. For each basic syntactic structure, we have different semantic structures or different syntactic meaning, defined as $Code_{ij}$.

For example:

&가까이하다
*N0 N1-을 V
$Code_{i11}$
亲近/to show intimacy
타
Agent + patient + verb
N0=인물,N1=인물
그 친구를 너무 가까이하지 말아라.
$Code_{i12}$
接近/to come close to
타

Agent + patient + verb
N0=인물,N1=사물(술, 책, 컴퓨터)
그는 요즘 술을 멀리하고 책을 가까이하려고 노력한다.

* N0 N1-와 V

$Code_{i21}$
亲近/ to show intimacy
타
Agent + comitative + verb
N0=인물,N1=인물
명수는 나쁜 친구들과 가까이하더니 젊은 나이에 결국 콩밥을 먹는 신세가 되었다.

$Code_{i22}$
接近/to come close to
타
Agent + comitative + verb
N0=인물, N1=사물(술, 책, 컴퓨터)
기영이는 늘 책과 가까이하는 생활을 해 올 수 있었던 것을 만족해한다.

*N0i N1j-와 (서로) V ↔ N0i N0j-와 (서로) [대칭]

$Code_{i31}$
与……接近/to come close to…
자
Agent + comitative + verb
Ni=인물, Nj=사물
찬우는 미란이와 그 일을 계기로 서로 가까이하게 되었다

*N0i N0j-와 (서로) [대칭]

$Code_{i41}$
与……接近/to come close to…
자
Co-agent + verb
Ni=인물, Nj=사물
찬우와 미란이는 그 일을 계기로 서로 가까이하게 되었다.

In the database, Chinese paraphrase is an abstract description of the concept of each syntactic semantic term, when property description is a more specific and detailed formal coding of this concept. Syntactic semantic code is in fact the code for syntactic semantic relations, reflecting not only the place of each semantic term in the whole semantic field, but its relations with other terms as well.

From either an engineering or linguistic angle, our motive in constructing this knowledge database is to describe whether certain linguistic forms are acceptable or not, and what transformational relations exist between the acceptable ones. In other words, the transformation between linguistic forms can be defined formally by way of computation.

KSSKD is in nature an information lexicon/dictionary revolving around predicate verbs with both syntactic and semantic information. It has 3200 lexical entries, and 15, 000 semantic terms. Over 700 entries have been coded. Most of the chosen words are in the Korean Sejong Project predicate list. We have also chosen words without the list, including both verbs and adjectives.

The three levels of the knowledge database are defined mathematically as:

$$Group = (P, R)$$

$$P = \{ V_1, \ldots, V_i, S_{11}, \ldots, S_{ij}, C_{111}, \ldots, C_{ijk}, L \}\ 1<i<n,\ 1<j<n_i,\ 1<k<n_{ij}$$

$1 \le n \le Q1$(number of predicates)，$1 \le n_i \le Q2$(number of basic syntactic structure), $1 < n_{ij} < Q3$ (number of semantic coding)

$R = \{R_1, R_2, R_3\}$

$R_1 = \{V_i, S_{ij}\}$ //relation between predicates and basic syntactic structure

$R_2 = \{S_{ij}, C_{ijk}\}$ // relation between basic syntactic structure and syntactic semantic terms

$R_3 = \{C_{ijk}, L\}$ // relation between syntactic semantic terms and the lexicon

V —morphological verb

S —basic syntactic structure

C —syntactic semantic code

L – Lexicon

The above mathematical definition represents an abstract model, which enables us to operate on the database through computers.

4. Perspectives

In face of a net age and knowledge economy era, language information processing has two issues to deal with. The first one is the shaping of correct theoretical conception. That is: the nature of linguistics should be clarified in the first place in order to create a unique environment for linguistic studies; the interfaces between linguistics and other branches of sciences should be given primary importance for ultimate breakthroughs, with an eye to provide an operational system for application program developers; linguistic studies should be based on empirical application; a consistent and verifiable hypothesis should be put forward that satisfies descriptive and explanatory conditions, so that a strong linguistic support will be there for computer processing of language at every level. The second one concerns engineering technology. Only implication of language project can information products (such as electronic dictionary, machine translation system) be developed. As a result, computer technicians should devote themselves to the embodiment of certain advanced linguistic ideas, to constructing basic studies platform, application key-technique platform, and system development platform for language information processing. The aim is to provide program support for development of computer software or hardware. Only if linguists and computer scientists work together will a new picture be painted beautifully in the relevant field.

References

1. Yude Bi: Constructing a Korean Syntactic-semantic Information Dictionary for Information Processing. Journal of Foreign Studies Institute, No.4. Luoyang (2002a)
2. Yude Bi: A Study on the Semantic Categories of Korean Sentences. National Language, No.5. Luoyang (2002b)
3. Weidong Zhan: Principles in Defining and Relativity in Semantic Categories. World Chinese Teaching, No.2. Beijing (2001)
4. Tinchi Tang: Theta-net, P&P Grammar and Machine Translation. Chinese Linguistics, No.4. Beijing (1996)
5. Dexi Zhu: Questions and Answers on Grammar, The Commercial Press, Beijing (1985)
6. Hong Jae-Sung: Modern Korean Verb Syntax Dictionary. Donga Press, Seoul (2001)
7. Gwangsu Song: Representation of Case & Semantics of auxiliary word, Worin Press, Seoul (1999)
8. Hyeon-Kwon Kim & Jong-Myeong Kim: Semantic description and clustering of Predicate, In Proceeding of Workshop on 21century Sejong Project , Seoul(2001)
9. Grimshaw, J.: Argument Structure, Cambridge, Mass: MIT Press. (1990)
10. Dowty, D.: Thematic Proto-roles and Argumen Selection, language 67. (1991)
11. Bake, C.F., C.J. Fillmore and John B. Lowe.: The Berkeley FrameNet Project, In Proceedings of COLING'98 (1998)