Constructing a Syntactic-Semantic Information Dictionary of Predicates Oriented to Korean Information Processing

BI Yude
PLA University of Foreign Languages, Luoyang,China
E-mail: biyude@yahoo.com.cn

Abstract

In any NLP systems, including that of MT, syntactic and semantic information dictionary is an essential component. Based on the achievements in semantic project studies both at home and abroad, the present paper provides an integrative description of the syntax and semantics of Korean predicates, with an aim to construct an information-processing-oriented dictionary of syntactic and semantic information. The semantic framework is drawn from theta structure theory and semantic field theory. We begin with a semantic classification of Korean predicates, which is followed by a detailed description of the semantic properties of these predicates. And in the construction of the dictionary, we integrate syntactic and semantic properties in a structural way.

0. Introduction

Generally speaking, there is always an E-dictionary with syntactic and semantic information in any NLP systems. It is of vital importance to develop such a dictionary and to provide a description of its properties, because it determines the performance and quality of the whole system. This has been paid due attention to in the field of computational linguistics. As a result, much manpower and material resources have been put into the development of such dictionaries.

At present, theoretical studies in semantics and the development of information processing both exert direct influence on the construction of semantic knowledge projects. The former is reflected in the expression of semantic knowledge, and the latter in the scale of the semantic knowledge database already constructed and in its specific knowledge content.

As a branch of linguistics, computational linguistics should focus on language itself. This point has been brought to more and more scholars' mind when computational linguistics has come through several decades' development (Liu Yongquan 1997). It must be pointed out that, however, focusing on the linguistic and human aspect of computational linguistics does not mean overlooking methodology and empirical efforts. On the basis of linguistics and the semantic syntagmatic relations between predicates and nominal elements, this study aims to construct a syntactic-semantic information dictionary of modern Korean predicates oriented to information processing.

1. The state of affairs

1.1. Studies in Europe and America

These large scale studies on natural language semantics are centered on English and other major European languages. Among the most influential ones can be mentioned *The WordNet, The MindNet* and *The FrameNet*. The WordNet, which was started by scholars from Princeton University in the 1980s, is a dictionary based on psycholinguistic principles. To the extent that lexical information is organized in accordance with word meaning rather than word forms, The WordNet is basically a semantic dictionary (Miller 1990). The MindNet is a semantic knowledge database, consisting of a word-searching sub-program, which first tries to identify the meaning of each word, and then the semantics of the whole sentence (Richardson 1998). The FrameNet (Bake 1998) is also based on English.

1.2. Studies in China

In recent years, many achievements have been made in Chinese semantic projects. Mr. Dong Zhendong developed *The HowNet* through years of hard work on his own. The Semantic Group of Project No. 905, led by Mr. Lu Chuan and Mr. Lin Xinguang, improved Ch. J. Fillmore's Case Grammar, studied the case frame and semantic syntagmatic relations of Chinese predicates in a dynamic semantic framework, and compiled *A Dictionary of Modern Chinese Verbs for Personal and MT Use*. This dictionary has been published. Mr. Huang Zengyang, drawing insights from his original HNC theory, established *The Local Association Network at the Lexical Level*. This Network provides a complete set of means to describe the semantics of Chinese words and has been realized in computer program. His idea, which combines the semantic classification and the identification of semantic roles of verbs, creates an explanatory model for the distinction, description and representation of the semantic structure of verbal sentences. Moreover, this model is both computable and psychologically realistic.

1.3. Studies in Korea

Following the design principles of The WordNet and The EuroWordNet, the Century Plan of South Korea began to provide detailed descriptions of the syntactic information of predicate verbs in 1998. In addition, The Annual Report on E-Dictionary of the Language Branch of the Century Plan 2000 has set as its major objective to "hold symposia on semantic classification and semantic relations", and to "found a semantic classification system from multi-angles, which is very important for the semantic description of predicates" (South Korean Culture Ministry 1999; 2000). Besides, Piao Deyu (1998) gave a list of seven categories of the basic semantic classification of verbs, including state, mentality (perception, cognition and sensation), action, change, accomplishment, instantaneity and movement.

2. The basis, objective and significance of this study

2.1. Ferdinand de Saussure, the founding father of modern linguistics, took language as a system of symbols, and classified the relations among such symbols into paradigmatic and syntagmatic relations. This distinction opened up a new era for linguistic studies. These two relations are well presented at the phonological, lexical, syntactic and semantic levels. Inspired by the syntagmatic relation, we study the syntagmatic relations holding between verbs and nominal elements at the semantic level, for the specific purpose of language information processing. As is generally accepted, predicate verbs constitute the core of syntactic and semantic structures. Therefore, the syntagmatic properties of predicate verbs regarding syntax and semantics directly determine the properties of syntactic configuration.

2.2. Increasing attention is being paid to semantics in modern linguistics. Meaning establishes itself as the core of communication. Both phonology and syntax serve for the smooth communication of meanings. A solution to the problems facing semantics is necessary not only for social communication, but for NLP as well. Natural language information processing, having gone through the word level and the phrase level, has reached its present stage, namely, the sentence level. Intelligence systems for knowledge projects have developed into the substantial level of knowledge acquisition and intelligent modeling. At this stage, more urgent than ever before are the basic theoretical studies on syntax, semantics and pragmatics. Such studies, which present themselves as highly thorny leading topics in the field of language information processing, have attracted extensive attention of many experts around the world. The focus of research is currently put on the acquisition of syntactic and semantic knowledge. And, our ultimate goal is to initiate a linguistic knowledge representation system oriented to Korean information processing. In our view, this system is an integration of the conceptual systems of materials and motion. The present study is just one component of that overall system.

2.3. In terms of linguistics *per se*, this study is of both theoretical and empirical significance. On the one hand, syntactic-semantic studies have become a hot topic for modern linguistic researches. It is easily perceived that the same language form may have a rich variety of semantic elements. For example, in the configuration "N1-il +V", the semantics of N1 varies with the meaning the verb: it may be agent, result, purpose and others. The property of verbs determines the number of arguments and their categories (Bi Yude 2000). Conforming to the current tide in modern linguistics, the present research gives prominence to semantics. It is in this sense that our study forms an important part of the

overall studies on the Korean language.

On the other hand, human society has stepped from the industrial age into an era of information. That initiates the need for the computerization of mental labor. In other words, the computer should be able to deal with natural languages, including complex semantic issues, one of which is the semantic syntagmatic relations of predicate verbs and nominal elements. Studies on such relations, which are of great significance to language information processing, will offer a linguistic guarantee for Korean information processing and Korean-Chinese MT.

3. Working principles

We adopted a pragmatic methodology for two reasons. One is to avoid difficulties, into which Chomsky an pure rationalism has run. The other is to shun the weak points of empiricism that ignores rational reasoning and, therefore, lacks logical rigor. Pragmatism is basically a reconciliation of the disputes between empiricism and rationalism.

Empiricists advocate that all concepts are learned through experience, and that only linguistic knowledge thus proved is of significance. This methodological principle has attracted the attention of the computational linguists at home and abroad. That is manifested in the ever-fining granulitization of linguistic knowledge, in the increasingly precise auto-tagging of unbounded large-scale real texts in language corpus, and in the ever-increasing importance of the lexicon in NLP.

By contrast, rationalists adopt a rule-based auto-syntactic-semantic way of analysis, while putting sort of constraints on natural language, thus the concept of “bounded language”. It is because “The computerization of natural languages needs the support of knowledge of various kinds, including both fine knowledge obtained by way of empirical practices and crude knowledge acquired through rationalist means” (Feng Zhiwei 1996).

We know that linguistics is an empirical science and researches on language theories must take their empirical value into account. Therefore, the choice of linguistic research methodology should also consider this point, which is of special importance in the computerization of natural languages. “What linguistics discusses is the relevant facts from experience, namely, explanation for language-internal and -external facts. The motive is to provide an easy-to-operate system for applied researchers.” (Zheng Ding’ou 1999)

4. Semantic classification of predicates

Classification is both one way by which human beings understand the world, and more important, it is also the result. Taxology, the science of classification, has developed from qualitative classification to that by mathematical means. As a basic science, taxology is applied in all branches of science. Its basic criteria include mutual-exclusiveness, exhaustiveness and universalness.

With the development of semantics, the theory of semantic field has been attracting increasing attention of researchers in the fields of linguistics and computer science. This theory was put forward by a group of linguists in Germany and Switzerland in the 1930s. Jost Trier was the most prominent one. He pointed out that, in every language, “each word is located within its relative concepts. These words and the one word that they all refer to constitute a self-contained whole. Such structures may be referred to as word field.”

Semantic field can be grouped into paradigmatic field and syntagmatic field. The former refers to all the words of the same category; and the latter is composed of by words that can be combined with one another but are of different categories. The elements of each semantic field are interrelated in meaning. And each semantic field is made up of by various sub-semantic-fields according to a common semantic element. It is verb semantic field and its sub-semantic-fields that we intend to study. Differently put, we pay primary attention to the semantic classification system of verbs and the semantic combinational relations (syntagmatic field) of verbs with nominal elements.

We classify modern Korean verbs into four hierarchical semantic groups: semantic field, sub-semantic-field, word semantic groups and semantic classes.

1) Semantic field: state, relation, action;

2) Sub-semantic-field: for example, the semantic field of action includes the sub-semantic-fields

of location, placement, physical action, creative activity, intelligent activity, linguistic activity, social activity, physiological activity, etc.;

3) Word semantic groups: the semantic field of location movement, for instance, includes spatial movement, vertical movement, origin-based movement, terminal-based movement, etc.;
4) Semantic classes: each word semantic group contains verbs of specific semantic classes, among which verbs referring to the whole are standard verbs, while those collective ones are called the enumerative verbs.

5. Property description of the dictionary structure

Semantic classification is, up to now, far from satisfactory for information processing. Words of the same semantic class may have rather different internal and collocation properties. Semantic classification may meet the demand of explanatory adequacy but not that of descriptive adequacy (Yu Shiwen 2000). Then, when developing semantic dictionaries, we should base them on semantic classification, and depict the semantic syntagmatic relations of each predicate (even each semantic element) with nominal elements from the perspective of property description. As a result, we may provide a more efficient and accurate linguistic support for NLP systems.

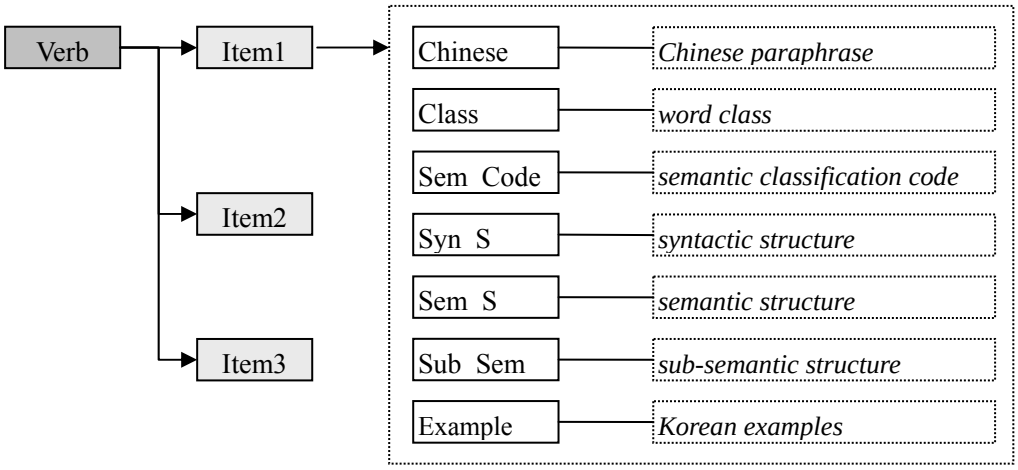
Predication is a kind of cognitive operation, resulting in the decomposition of propositional structures into predicates and arguments. The number and category of arguments in any given argument structure are determined by the properties of verbs. The other way round, we can identify the property and category of verbs in different structures on the basis of the category of the nouns in the structures concerned. L.Tesnier's Dependency Grammar and G. Helbig's Valency Grammar offer a quantitative analyzing and classifying mechanism for verbs. With further help from Fillmore's semantic case role analysis, we can provide a qualitative and quantitative description of the semantics for verbs and verbal sentences in natural languages. In order to design NLP systems, a semantic descriptive system must be first established that is both explanatorily adequate and psychologically realistic. The key step is to find a semantic binding mechanism for the various concepts in natural languages.

According to our study combined with properties of the Korean language, it is of great benefit to the acquisition of syntactic and semantic knowledge to integrate the syntactic and semantic information of Korean predicates, and to give a semantic classification of these predicates. Further, a predicate forms the core of a sentence, the basic structure of which may be viewed as the mapping of the syntactic properties of a predicate. Such properties include argument property, theta role property, the categorical properties of arguments and the syntactic properties of theta roles, just to name a few. (Tang Tingchi 1996)

With respect to the design of dictionary structure, we integrate dictionary and grammar into one, namely the grammaticalization of lexical items and the lexicalization of grammar. Each lexical element of each predicate is taken as a construct, which embodies information of semantic element forms, Chinese paraphrase, word class, syntactic structure, semantic structure, semantic properties and exemplary sentences. Each record in the dictionary is a construct, equaling to a lexical entry as specified in Chomsky's Principles and Parameters framework. In terms of argument structure theory, such a construct is the integration of argument properties, theta role properties and syntactic features.

ITEM ————semantic element (form)
[Chinese] ————Chinese paraphrase
[Class] ————word class
[Sem_Code] ————semantic classification code
[Syn_S] ————syntactic structure
[Sem_S] ————semantic structure
[Sub_Sem] ————sub-semantic structure
[Example] ————Korean examples

If a verb have three semantic elements, there are three constructs in its hypogyny.



Notes:

1) The dictionary has a vocabulary of 3200, or more accurate, 3500 with semantic elements included. These words are mostly taken from the list of basic predicates of the South Korea Century Plan, with some verbs and adjectives added from other sources.
2) As for the different meanings of the same word, subscript is used for distinction. In the case of homophones we apply superscript. E.g., $sata_1$ (to hire), $sata_2$ (to recognize), $sata_3$ (to buy), $sata_4$ (to feast), $sata_5$ (to provoke) and $musta^1$ (to ask), $musta_2$ (to bury). Homophones are in nature different words of the same form, different from the semantic elements of a single word. Semantic elements are derived from the original meaning of a given word.
3) The dictionary is realized electrically in the form of Tree View characterizing directness and a clear hierarchy. In the dictionary, each word is presented as a one-bar node and the number of semantic elements decides the number of sub-nodes. As a construct, each semantic element contains sub-nodes indicating such information as Chinese paraphrase, word class, semantic classification code, syntactic structure, semantic structure and semantic properties.

6. Empirical analysis

Taking the verb ***/sata/***(originally meaning buy) as an example, we give a description of the syntactic and semantic information of modern Korean predicates as follows (Note: the examples are translated into Wide International Phonetic Symbols)

sata

1[Chinese] 雇佣/guyong/(to hire)
[CLASS] V
[Syn_S] N0 N1-eul V
[Sub_Sem] N0=person N1=thing
[Sem_S] AGT+THM+V
[Example]uri iuseun ilga chincheogi eopsseoseo sarameul saseo jangnyereul chireotda.(As our neighbor has no relatives, they hired some people for the funeral.)

2[Chinese] 认可/renke/(to recognize)
[CLASS] V
[Syn_S] N0 N1-eul V
[Sub_Sem] N0= persoṇ | collective N1=abstract (contribution | labour force)
[Sem_S] AGT+THM+V
[Example] sajangeun gim bujangui jigeumkkajiui gongnoreul nopi saseo miguk jibujangeuro seungjinsikyeotda.(The head of the company highly appreciated Minister King's contribution till today, and promoted him to general manager of the American branch.)

3[Chinese] 买/mai/(to buy)
[CLASS] V
[Syn_S] N0 N2-eseo | egeseo N1- eul V
[Sub_Sem] N0=person | collective N1=thing (including abstract ones) N2=person | collective
[Sem_S] AGT+THM+SRC+V
[Example] byeongjuneun han goldongpum sujipgaeseo i dojagireul baengman wone satda. (He bought this

porcelain from an antique collector with 1 million dollars.)

minhoneun sinaee nagan gime daehyeong seojeomeseo chaegeul han gwon satda. (Our company bought the copyright from the author.)

4[Chinese] 请客/qingke/(to feast)
[CLASS] V
[Syn_S] N0 N2-ege N1- eul V
[Sub_Sem] N0=person N2=person N1=food|supper|wine
[Sem_S] AGT+THM+PAT+V
[Example] oneureun nega naege sulhanjaneul saya doeji ankenni? (Shouldn't you invite me for a drink today?)

5[Chinese] 惹，讨/tao, re/(to offend, to provoke)
[CLASS] V
[Syn_S] N0 N2-egeseo | lobuteo Npr1- eul V
[Sub_Sem] Npr1=trust | doubt | compain N0=person N2=person
[Sem_S] THM+SRC+V
[Example] gim gyejangeun geu illo gwajangegeseo miumeul satda.(Mr. King, our group head, was criticized by the sect leader for that.)

sseuldeeomneun de doneul sseoseo gwadaepyoin myeongsuneun haksaengdeullobuteo wonseongeul satda. (The course representative was criticized by all classmates for squandering the money.)

7. Work to be done

At present, we have taken from dictionaries more than 3000 pieces of the semantic information of predicates, including Chinese paraphrase. For some of them, we provide a description of their syntactic structure, semantic structure and semantic properties, with examples. In the next step, we are going to do the same to others. All examples are taken from dictionaries and the South Korea Century Plan Corpus.

The knowledge database of the concepts of the material category has been left for anther research project. After finishing this work, we will combine these two databases and conclude the construction of the knowledge database for modern Korean.

Reference

[1] Miller,G., et al. 1990. Introduction to WordNet: an on-line lexical database. In *International Journal of Lexicography* 3, No.4.

[2] Richardson,S.D.,William B.Dolan, and Lucy Vanderwende. 1998. MindNet: acquiring and structuring semantic information from text. In *Proceedings of COLING'98*.

[3] Bake, C.F., C.J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet Project. In *Proceedings of COLING'98*.

[4] Korean Culture Ministry. 1999. The Annual Report on E-Dictionary of the Language Branch of the Century Plan 1999.

[5] Korean Culture Ministry. 2000. The Annual Report on E-Dictionary of the Language Branch of the Century Plan 2000.

[6] Bi Yude. 2000. The Patternization of Syntactic Semantic Structures. *PLA University of Foreign Languages Journal*, No.1.

[7] Feng Zhiwei. 1996. *The Computerization of Natural Languages*. Shanghai: Shanghai Foreign Languages Education Press.

[8] Liu Yongquan. 1998. *Applied Linguistics*. Shanghai: Shanghai Foreign Languages Education Press.

[9] Tang Tingchi 1996. Argument Net, P&P Grammar and Machine Translation. *The Chinese Language* 4.

[10] Yi Mianzhu. 1999. *The Grammar of Location: Theory and Practice*. Harbin: Heilongjiang Education Press.

[11] Yu Shiwen. 2000. *A Collection of Papers on Computational Linguistics* 4. Beijing: The Computational Linguistics Institute of Beijing University.

[12] Zheng Ding'ou. 1999. *Studies on Lexical Grammar and Chinese Syntax*. Beijing: Beijing Language and Culture University Press.

[13] Park dekyu. 1998. *Studies on the Valency of Korean Verbs*. Seoul: Extensiveness and Quietness Press.

[14] Lin Xinguang. 1999. *Lexical Semantics and Computational Linguistics*. Beijing: The Chinese Language Press.