The KOLON System: Tools for Ontological Natural Language Processing in Korean

Juliano Paiva Junho, Yumi Jo, and Hyopil Shin

Department of Linguistics, Seoul National University,
San 56-1, Sillim-dong, Gwanak-gu, Seoul, 151-742, S. Korea
jjunho@gmail.com, jmocy@snu.ac.kr, hpshin@snu.ac.kr

Abstract. This paper presents and analyzes the KOLON System, created to facilitate Korean Natural Language Processing (KNLP) and to improve experimental results through the KOLON Ontology. It is currently under development at our Computational Linguistics Laboratory, and is based on previous works, namely the Mikrokosmos Ontology and 21st Century Sejong Project. The KOLON System is also extended with software tools to simplify the handling and visualization of the data, as well as the creation of new programs. The mapping of words onto ontological concepts was performed automatically, with faulty information being corrected manually. In order to examine the effectiveness of using KOLON's data, we have rerun a previous sentiment analysis (SA) experiment, changing the approach to include data from the ontology. This new experiment obtained improved results, which is a strong indication that the project will be of use after its completion.

Keywords: KOLON, ontology, Mikrokosmos Ontology, WordNet, FrameNet

1 Introduction

In various fields related to Information Science, such as Artificial Intelligence and Natural Language Processing (NLP), ontologies, also called Knowledge Bases (KB), are formal representations of knowledge and the relations among its concepts, usually composed by individuals, classes, attributes, relations, events, rules and axioms (Nirenburg, 2004). Even though researchers had been building several types of ontologies for decades and have been using them as a formal representation of knowledge within certain domains, ontologies gained momentum with the creation and propagation of the World Wide Web, and, more recently, with the idea of the Semantic Web, introduced by Tim Berners-Lee's homonymous article (Berners-Lee, 2000a; Nirenburg, 2004; Hirst, 2004). However, the ontological structuralization of knowledge alone is not enough to make complete use of information in NLP, since ontologies basically hold semantic atoms of information and interconnect them through different kinds of named relations. For a thorough analysis of linguistic data, we need not only the semantic data from a simple ontology – information about the lexicon, the morphology, the syntax are needed as well. With this in mind, an work-group at our Computational Linguistics Laboratory devised the KOLON Ontology (an acronym for "**KO**rean **L**exicon mapped onto an **ON**tology"), based on ontological and lexical data, through the adaptation of two large works: the Mikrokosmos Ontology and the 21st Century Sejong Project. The KOLON System is the data combined with related software to facilitate its use both for the final user, who may simply want to inspect the data, and for the programmer, who may need to create programs to utilize the data in various ways.

This paper starts with a general presentation of works related to the current project, namely Mikrokosmos Ontology, 21st Century Sejong Project, WordNet and FrameNet, highlighting the

relevant elements of each. KOLON Ontology's structure is then thoroughly discussed and details about the project are presented, together with a comparison of this work with WordNet and FrameNet. The following section mentions the main components of the software library, focusing on `pyKOLON` and `webKOLON`. Finally, this paper describes an application of the KOLON Ontology's data to a previously performed SA experiment showing an improvement in the results, which shows that, even though the KOLON Project is still under development, it can already be used for Korean Natural Language Processing experimentation with improved results. This also implies that upon its completion, KOLON will enable a significant increase in accuracy in Korean Natural Language Processing experiments.

2 Related Work

2.1 The Mikrokosmos Ontology

The Mikrokosmos Ontology (μK) was built by the Computing Research Laboratory at the New Mexico State University (NMSU/CRL), and is based on an older ontology developed at Carnegie Mellon University, whose original 2,000 concepts expanded by NMSU/CRL to around 4,300, covering almost all concepts needed to represent the meaning within the Spanish corpus on company merger and acquisition (Mahesh, 1996). The ontology was written with English as a meta-language, which does not imply that it related concepts with English words: relations were first built between general concepts and, later, the lexemes[1] (English, Chinese and Spanish) were mapped to the resulting ontological structure (Shin, 2007; Mahesh, 1996). Despite Mahesh (1996)'s statement that μK is a world model which can be used by NLP systems that need to represent and manipulate the meanings of texts and whose "central goal is to develop a system that produces a comprehensive Text Meaning Representation (TMR) for an input text in any of a set of source languages", the result was not exactly a complete world model but a world model for the Spanish corpus on company merger and acquisition that they had by that time. Since this is the version of μK that we are currently in the process of extending, we had to modify this to some extent to successfully map Korean lexical items onto it.

2.2 The 21st Century Sejong Project

The 21st Century Sejong Project[2] was a project carried out by the South Korean Government who divided the research among many important universities around the country for a period of around 10 years. It resulted in a large-scale digital linguistic body of dictionaries and corpora, presenting a huge amount of information related to the Korean language (21 C. Sejong Proj., 2007). The lexicon mapped onto the KOLON Ontology consists of 25,459 nouns, 15,180 verbs and 4,398 adjectives[3], combining into a total of 45,037 words; however, due to the fact that Korean is a very polysemous language, the total count comes to 65,326 lexical units with different meanings, i.e. 45% of all the words in the lexicon are polysemous.

[1] Throughout this paper, we have used the term *lexeme* in the sense of being an abstract unit of the lexicon, corresponding to a collection of forms that a certain word can take, such as the lexeme BE for the English words *be*, *being*, *been*, *am*, *is*, *are*, *was*, *were*". However, in addition to this, lexemes are thought to have been systematically disambiguated with additional markers for word sense and part of speech throughout our work on KOLON, e.g.: BASS_1_N: "the lowest part of the musical range"; BASS_2_N: "the lowest part in polyphonic music"; BASS_3_N: "an adult male singer with the lowest voice"; BASS_4_N: "the member with the lowest range of a family of musical instruments"; BASS_5_N: "non-technical name for any of numerous edible marine and freshwater spiny-finned fishes"; BASS_1_A: "having or denoting a low vocal or instrumental range. (English examples extracted from WordNet.)

[2] More information in English about this project can be found at `http://www.sejong.or.kr/eindex.php`.

[3] Nouns, verbs and adjectives are encoded in Sejong data as `NN`, `VV` and `VA`, respectively.

2.3 WordNet

In the mid-eighties, researchers from Princeton University's Cognitive Science Laboratory started a project called WordNet, a dictionary that interconnected words through several semantic relations, such as synonymy, antonymy, hypernymy, hyponymy, holonymy, meronymy, creating the first and most well-known network of English words (Fellbaum, 1998a). After that, hundreds of researchers around the world built other works based on WordNet, e.g. EuroWordNet (originally covering Dutch, Italian, Spanish, German, French, Czech, and Estonian), and, in Korea, KorLex, a project carried out by Pusan National University (Fellbaum, 1998a). Despite the large amount of information provided by the numerous relations in word networks, one weakness of this type of work is that there usually are as many concepts as there are words registered in the lexicon (Shin, 2007).

2.4 FrameNet

FrameNet is an ongoing project at Berkeley University to create an on-line lexical resource of English based on frame semantics and supported by corpus evidence (Ruppenhofer *et al.*, 2006). Its aim is to create a combination of all possible semantic and syntactic elements related to each event-denoting word in all of its senses; however, since words related to artifacts and natural kinds are not annotated, FrameNet cannot be readily used as an ontology, a function which is left for WordNet (Ruppenhofer *et al.*, 2006). Below we present two frames from FrameNet, namely "Apply_heat" and "Ingestion", where the frame elements can be observed as they are annotated in the FrameNet data.

- Apply_heat frame:

  [$_{Cook}$Matilde] FRIED [$_{Food}$the catfish] [$_{Heating_instrument}$in a heavy iron skillet].

- Ingestion frame:

  [$_{Ingestor}$The locals] EAT [$_{Ingestibles}$mainly fish and vegetables].

3 The KOLON Ontology

The KOLON Ontology is a project currently under development by our Computational Linguistics Laboratory. It was born from the need for NLP tools especially created with the Korean language in mind, so that we could improve our experimental results. Our work, which has come to be known as KOLON, is based on two important projects, namely the Mikrokosmos Ontology and the 21st Century Sejong Project.

The mapping process was divided in three phases. In 2007, Seoul National University Computational Linguistics Laboratory performed an automatic mapping of lexemes from the Sejong Electronic Dictionary onto concepts from the Mikrokosmos Ontology. Following this, a team of students from our Department of Linguistics started the work of checking the automatic mapping and complementing the data. The first two mapping results are currently being checked and refined manually by a student who excelled in the second phase, and the some few concepts that are not present in the original μK, but are needed for the mapping of Korean words, are being added. So far, 40,861 senses have been remapped and checked, 63% of the total. In acquiring new Korean lexemes from Sejong, we decided to maintain the original format of the ontology, with only very few adaptations. Special care was taken to avoid creating new concepts to adapt the ontology to the new meanings presented by the Korean words. This was done by keeping a policy of fine-grained mapping, making use of multiple mapping, which causes a great number of Korean lexemes to have a one-to-many mapping between the lexeme and concepts. This happens, especially, with Sino-Korean words, which usually consist of two or more basic units of meaning accommodated

within a single lexical item, e.g.: 출입하다(出入하다) CHWULIP-HATA[4] (Sino-Korean compound containing three morphemes, namely, two Sino-Korean morphemes "enter" and "exit"; and the native Korean verb *hata* ("to do") that creates verbs from Sino-Korean morphemes). Even with the inclusion of 65,326 words compared with the original 16,468 English words in μK, i.e. almost four times more lexical units than the largest set previously existent in the ontology[5], we achieved a growth of only 0.44% from the original μK's 5,425 concepts with only 24 new concepts created.

As we have already mentioned in section 2.2, Korean words are highly polysemous. An example of this high level of polysemy is the word 배 PAY that, according to Sejong, can have 5 different meanings: 1) "*pear*", 2) "*ship*", 3) "*abdomen, stomach, womb*", 4) "*embryo*", 5) "*double*" (since the last two words are Sino-Korean, they can also be written as 胚 PAY and 倍 PAY respectively, however this is rarely done); the word 은 UN means 1) "*silver, silver medal*" (銀 UN) at the lexical level, but can also be 2) a "*topic-contrast particle*" (Sohn (2001), p. 214), 3) a "*verbal past relativizer suffix*" (Sohn (2001), p. 240) or 4) an "*adjectival non-past relativizer suffix*" (Sohn (2001), p. 240), at the morphological level.[6]

3.1 KOLON's structure

The KOLON Ontology is composed of two main databases: the ontological frame-system and the lexical frame-system. The ontological frame-system is based on μK's format of frames, slots, facets and fillers. The ontology is a collection of 5,449 concepts organized in the form of frames. Each frame relates to one single concept and contains all the slots (properties) and fillers (values) related to that concept. The links between concepts are multi-dimensional due to the fact that they can have more than one facet. In short, we have that one frame (concept) can have several slots (properties), which in turn can be composed by several facets (constraints), which can be filled by several fillers (values). Deeper explanations about the structure and the slots and facets' meanings can be found in Mahesh (1996) and Nirenburg (2004). In order to record the mapping of the 65,326 Korean words onto μK, achieved by means of the slot `MAP-LEX`, based on the already existent `LEXE` (English), `LEXC` (Chinese) and `LEXS` (Spanish), we created a new facet, namely, `LEXK` for Korean lexical items. All the mapped lexemes are linked to their corresponding concepts through this. The lexicon structure was created based on μK's frame-system with the lexical information contained in Sejong. We have parsed the entirety of Sejong's `NN`, `VV` and `VA` XML-encoded files (45,037 files in total), each one corresponding to one orthographic word form, extracting the most important information and constructing a lexical frame system, trying to keep its format as similar as possible to the original μK conceptual frame. As in μK, the lexical frames are composed by slots, facets and fillers alike. However, the frame has as its main component a "lex", as it is called in KOLON. A "lex" is a lexeme written in Hangul followed by two integer two-digit indices (entry number and sense number respectively) with its part of speech (NN, VV, VA) appended to the end. By examining the different lexes related to the orthographic form 달다 TALTA in Table 1, we can see that it has several different entries (homonymy), corresponding to the first index, and

[4] All Korean examples are written in Hangul, followed by the transliteration and the translation in English, in the following fashion: 한글 HANKUL ("Hangul"). If applicable, Sino-Korean characters are presented in parentheses following its correspondent Hangul form, such as 한자(漢字). The transliteration system used in this paper is the most internationally known Yale Romanization System. However, data in KOLON is transliterated via the "한글의 로마자표기 방법 HANKUL-UY LOMACA PHYOKI PANGPEP ("Rules of Writing Hangul in Roman Letters")", also called "(Ministry of Culture) Revised Romanization (2000)", the most wide-spread system in South Korea nowadays.

[5] 2,668 Chinese and 719 Spanish lexical units were also mapped in μK.

[6] Nowadays, the Korean language makes use of the Hangul writing system almost exclusively. However, if necessary, it is also possible to use Chinese characters in order to disambiguate the meaning of Sino-Korean words, such as in the case of the word 수도 SWUTO which has four different meanings in Sejong: when written in Chinese characters, all these meanings are clear and unambiguous, each one being written with different characters: 水道 SWUTO ("waterworks, tap"), 首都 SWUTO ("capital city"), 修道 SWUTO ("spiritual practice"), 手刀 SWUTO ("hand side edge").

each entry can still have different related senses (polysemy), corresponding to the second index. A lex performs then the role of a dictionary headword with various sub-meanings. The different meanings can be easily grasped by examining the related concepts.

Table 1: Lexes related to the orthographic form 달다 TALTA.

Lex	Concept
달다_01_01_VA	SWEET
달다_01_01_VV	(YET TO BE MAPPED)
달다_01_02_VA	QUALITY
달다_01_02_VV	ATTACH
달다_01_03_VV	ESCORT
달다_01_04_VV	ADD-TO
달다_01_05_VV	CHARGE
달다_02_01_VV	WEIGH
달다_03_01_VV	WAVE-ENERGY-EVENT
달다_03_02_VV	CHANGE-COLOR
달다_03_03_VV	EFFECT; PATIENCE-ATTRIBUTE
달다_04_01_VV	ENTREAT

3.1.1 Lexical Frames Table 1 presents one lexical frame related to the lex 달 다_01_01_VA (*talta_01_01_VA*). The basic data is saved as a simple tab-separated UTF-8 encoded text file. The columns are: `lex`, `slot`, `facet`, `filler` and `syn_sem_frame` (since for the same verbal or adjectival lex there may exist more than one type of syntactico-semantic restriction frame).

Table 2: An excerpt from a μK lexical frame: 달 다_01_01_VA (*talta_01_01_VA*).

Lex	Slot	Facet	Filler	Syn_Sem_Frame
달다_01_01_VA	SEMANTIC-CLASS	VALUE	GUSTATORY-ATTRIBUTE	0
달다_01_01_VA	TRANSLATION	LEXE	be sweet	0
달다_01_01_VA	MORPHOLOGICAL-STRUCTURE	VALUE	A	0
달다_01_01_VA	THEME	SEM	FOOD	1
달다_01_01_VA	THEME	JOSA	이 I (“nominative particle”)	1
달다_01_01_VA	THEME	LEXK	과일 KWAIL (“fruit”)	1
달다_01_01_VA	THEME	LEXK	김치 KIMCHI (“Korean pickled vegetable”)	1
달다_01_01_VA	THEME	LEXK	채소 CHAYSO (“vegetable”)	1
달다_01_01_VA	FRAME	VALUE	X=N0-이 A	1
달다_01_01_VA	FRAME-TYPE	VALUE	FA	1
달다_01_01_VA	EXAMPLE	VALUE	이 김치는<THEME> 조미료가 들어가서 너무 달다.[7]	1
달다_01_01_VA	EXAMPLE	VALUE	잘 익은 과일은<THEME> 참 달다.	1

Except for the last column, which we needed to add to the structure as explained above, the format follows exactly the one found in μK (Table 2). However, for the lexical frames to be able to hold the information contained in Sejong, we had to create some new slots (`FRAME`, `FRAME-TYPE`, `MORPHOLOGICAL-STRUCTURE`, `SEMANTIC-CLASS`, `TRANSLATION`, `EXAMPLE`) and a new facet, namely, `JOSA` (from the Korean 조사(助詞) COSA (“particle”)). Also, some values that are used as slot names in the ontological frames, are used as facets in the lexical frames (`LEXK`, `LEXE` and `LEXC`). The `FRAME-TYPE` slot informs if the verb is transitive, intransitive, etc. The `FRAME` slot has the information related to the structure of the syntactico-semantic frame (`syn_sem_frame`), and its constraints are given as case-role slots such as `AGENT`, `THEME`, `LOCATION`, etc. The case roles have at least two facets, namely, `SEM` (with semantic information that always relates to a concept in the ontology) and `JOSA` (with a string in Hangul that represents the particle associated to that specific case). Sometimes there can be also a `LEXK` facet that records some more specific lexemes that are frequently used together with the frame. In Table 2, the `THEME` role for the adjective 달다 TALTA (Entry 01, Sense 01) will be marked by the nominative particle 이 I, and will be filled by a word mapped to the `FOOD` concept, or more precisely by some of the words 과일 KWAIL (“fruit”), 김치 KIMCHI (“Korean pickled cabbage”), 채소 CHAYSO (“vegetable”). The `MORPHOLOGICAL-STRUCTURE` slot indicates whether the word is a compound or a single morpheme, and, in the case it is a compound,

[8] Glosses and translation for the example sentences given at the end of this section.

what each of its components are. The SEMANTIC-CLASS slot indicates to which semantic group the word belongs in the Sejong dictionary, which is also mapped to ontological concept.

An important point related to the processing of Sejong's data is that the particles for the semantic roles are all 격조사(格助詞) KYEK-COSA ("case particles") Sohn (2001, p.213), such as the nominative particle 이 I[8]. However, many example sentences in Sejong are marked with 특수조사 (特殊助詞) TUKSWU-COSA ("special particles"), which Sohn (2001, p.214) denominates delimiter particles, such as the particle 은 UN[9]. Whereas the particle 이/가 I/GA marks the nominative case, the particle 은/는 UN/NUN usually serves the purpose of indicating the topic of the sentence. The topic-contrast particle (Sohn, 2001, p.214) 은/는 UN/NUN substitutes the above mentioned subject and topic particles[10]. Sejong reflects the tendency Korean has to do this. With other particles, however, the topic particle does not completely substitutes the other particle, but rather is used in conjunction with them, such as 에 EY ("dative/locative-static"), 에서 EYSE ("locative-dynamic/source"), (으)로 (U)LO ("allative/instrument/capacity") which become 에는 EYNUN, 에서는 EYSENUN, and (으)로는 (U)LONUN[11].

(1) a. 이 김치는 조미료가 들어가서 너무 달다.[12]
b. i kimchi-nun comilyo-ka tuleka-se nemu tal-ta.
this kimchi-TC seasoning-NM enter-and so very sweet-DC.
c. This *kimchi* is very sweet because it has been seasoned with condiment (MSG).

(2) a. 잘 익은 과일은 참 달다.
b. cal ik-un kwail-un cham talta.
well ripe-RL fruit-TC really sweet-DC.
c. Well ripe fruits are very sweet.

By observing glosses (1) and (2), and comparing them with the data from the lexical frame, we perceive the above mentioned point. This is a very common problem we face in order to automatically process the examples given in Sejong, since the particle assigned for a semantic role is very often substituted by a topic-contrast particle.

3.1.2 Ontological Frames The use of an ontological structure, such as μK's, rather than a lexico-semantic structure, such as WordNet, seems promising. An ontological structuralization of the concepts, and consequently of the words mapped to these, gives us the possibility of analyzing not only the linked concepts with its words, but also the type of link and the type of the linked concept that they have.

As seen in Table 3, the structure of the ontological frame is similar to the lexical frame presented above, except for the absence of the syn_sem_frame column. It follows exactly the μK format with the only exception that we have created, as mentioned above, a new slot called LEXK to map the Korean lexes onto the frames.

3.2 The program library pyKOLON and the on-line viewer webKOLON

The KOLON Ontology is the heart of the KOLON System, a collection of programs built in order to facilitate the use and visualization of the ontological data. One of its components, the webKOLON Ontology Viewer, an interface between the KOLON Ontology and the final user, can be accessed at http://word.snu.ac.kr/kolon. This program allows the user to perform

[8] 이 I is used after consonants and 가 GA is used after vowels.

[9] 은 UN is used after consonants and 는 NUN is used after vowels.

[10] Similarly to what happens in Japanese with the particles が GA ("nominative particle") and を WO ("accusative particle") that are substituted by は WA ("topic particle").

[11] Similarly to what happens in Japanese with particles such as に NI ("dative/locative-static"), で DE ("locative-dynamic/instrument"), へ E ("allative") that can have the particle は WA attached to themselves: には NIWA, では DEWA, へは EWA.

[12] The glosses obey the following pattern: "a." is the original Korean; "b.", the transliteration with a second line glossed word by word; and "c.", the translation in English. The labels are the same ones used in Sohn (2001). DC: declarative sentence ending; NM: nominative; RL: relativizer; TC: topic.

Table 3: An excerpt from a μK conceptual frame (FOOD).

Concept	Slot	Facet	Filler
FOOD	DEFINITION	VALUE	a substance taken in and assimilated by an organism to maintain life and growth; nourishment
FOOD	IS-A	VALUE	INGESTIBLE
FOOD	LEXK	MAP-LEX	개암_02_01_NN (KAY-AM) (“cotton put in a falcon’s feed”)
FOOD	LEXK	MAP-LEX	고명_01_01_NN (KOMYENG) (“garnish, trimmings”)
FOOD	LEXK	MAP-LEX	관식_01_01_NN (KWANSIK) (“official food”)
FOOD	LEXK	MAP-LEX	김치_01_01_NN (KIMCHI) (“Korean pickled cabbage”)
FOOD	LEXK	MAP-LEX	날짜_02_01_NN (NALCCA) (“raw food”)
FOOD	LEXK	MAP-LEX	다시_01_01_NN (TASI) (“stock made by stewing”)
FOOD	LEXK	MAP-LEX	두부_01_01_NN (TWUBWU) (“bean curd, tofu”)
FOOD	LEXK	MAP-LEX	식량_01_01_NN (SIKLYANG) (“food, provisions”)
FOOD	LEXK	MAP-LEX	식사_01_02_NN (SIKSA) (“meal”)
FOOD	LEXK	MAP-LEX	식품_01_01_NN (SIKPHUM) (“food, provisions”)
FOOD	MADE-OF	INV	FAT
FOOD	SUBCLASSES	VALUE	FOODSTUFF
FOOD	SUBCLASSES	VALUE	PREPARED-FOOD
FOOD	THEME-OF	INV	ACQUIRE-GROCERY

searches, generating in real time a web page for the easy visualization of KOLON’s data (Fig. 5 and 6). The whole of the webKOLON program is connected to nothing more than the `Kolon` object instantiated by `pyKolon` through which all the KOLON Ontology’s information can be accessed and used, allowing the creation of programs to become an easier task for the researcher, who will be given control to all the information recorded in the data base to perform lexical and conceptual searches.

The webKOLON program was written with two purposes in mind: the first one, and more obvious, as a way to allow end users to browse and inspect the data contained in the KOLON Ontology; and, the second one, as an example of how an application can be built on top of the `pykolon` library.

Figure 1: Screen capture from webKOLON’s lexical frame: 달 다_01_01_VA (*talta_01_01_VA*).

Figure 1 is a screen capture from the actual webKOLON program in action, generating the lexical frame for the lex 달 다_01_01_VA (*talta_01_01_VA*). After the user performs a search for a specific concept or lex, the data stored in the KOLON System’s database is sent to webKOLON, which processes it and generates the page based on templates. All the recursive information contained in the ontological data, in the lexical data are presented with correlation marked as hypertext links that allow the user to navigate through the data and their interconnections, simplifying the visualization of the data.

4 Discussion and Future Work

In this paper, we have introduced the KOLON System, an ongoing project which will facilitate and bring a strong contribution to the future KNLP research at Seoul National University Computational Linguistics Laboratory. As an experiment on how well an ontology like KOLON will

serve us in our KNLP research at the Computational Linguistics Laboratory, we decided to take a previous experiment with acceptable results (Jang and Shin, 2010) and perform it again with data from the ontology. The experiment concerns the application of SA on a corpus of 300 newspaper articles, which present different levels of subjectivity: editorial articles, news articles with opinionated topics and general news articles. The increase in the results that make use of the ontology compared to the previous ones, although not of a very expressive weight in general, between 2.5% to 4.3%, seem promising in the sense that future results may improve even more when our KOLON Project becomes more mature. We have presented the construction of the ontology based on the Mikrokosmos Ontology and the lexicon from the 21st Century Sejong Project, and compared them to related projects, namely WordNet and FrameNet. KOLON, even though not yet fully completed, is on its way to become a body of linguistic data on the Korean language that will allow research on KNLP to grow and become more effective. It has already proven itself of value through the results obtained in the Sentiment Analysis experiment presented. The experiment is a repetition of a previous experiment by Jang and Shin (2010) where we add methods to analyze polysemy in the text by means of concepts from the KOLON Ontology. We will continue to develop the KOLON System through the constant analysis and improvement of its data, by checking the lexical mappings, correcting eventual faulty information, and producing additional pieces of software to facilitate its use.

References

Berners-Lee, T. 2000. The Semantic Web. *Scientific American Magazine*, [Internet] Available from: `http://www.jeckle.de/files/tblSW.pdf` (retrieved 15/Apr/2010).

Fellbaum, C. 1998. Introduction. In C. Fellbaum, ed., *WordNet: An Electronic Lexical Database*, pp.1-19. MIT Press.

Hirst, G. 2004. Ontology and the Lexicon. In S. Staab and R. Studer, ed., *Handbook on Ontologies*, pp.209-229. Springer.

Jang, H.-Y. and H.-P. Shin. 2010. Language-Specific Sentiment Analysis in Morphologically Rich Languages. *Proceedings of COLING 2010.* To appear.

Mahesh, K. 1996. Ontology Development for Machine Translation: Ideology and Methodology. *New Mexico State University CRL report MCCS-96-292*.

Miller, G.A. 1998. Nouns in WordNet. In C. Fellbaum, ed., *WordNet: An Electronic Lexical Database*, pp.23-46. MIT Press.

Miller, K.J. 1998. Modifiers in WordNet. In C. Fellbaum, ed., *WordNet: An Electronic Lexical Database*, pp.47-67. MIT Press.

Nirenburg, S. and V. Raskin. 2004. *Ontological Semantics*. Cambridge, Mass.: The MIT Press.

Ruppenhofer, J., M. Ellsworth, M.R.L. Petruck, C. Johnson and J. Scheffczyk. 2007. *FrameNet II: Extended Theory and Practice*. [Internet]. Available from: `http://framenet.icsi.berkeley.edu/book/book.pdf` (retrieved 13/Jun/2010).

Shin, H.-P. 2007. Mapping Korean Basic Verbs to the Mikrokosmos Ontology. *언어학 (*ENEHAK*)*, 49, 305-324.

Shin, H.-P. 2010. KOLON (the **KO**rean **L**exicon mapped onto the Mikrokosmos **ON**tology): Mapping Korean Words onto the Mikrokosmos Ontology and combining lexical resources. *언어학 (*ENEHAK*)*, 56, 159-196.

Sohn, H.-M. 2001. *The Korean Language*. Cambridge (U.K.): Cambridge University Press.

21st Century Sejong Project. 2007. Official Project Web Site. [Internet]. Available from: `http://sejong.or.kr/eindex.php` (retrieved 15/Apr/2010).