A Study on a Conceptual Map of Korean Words

Sun-Mee Bae
KORTERM, KAIST
373-1 Guseong-dong,
Yuseong-gu
Taejon 305-701, Korea,
sbae@world.kaist.ac.kr

Chung-Kon Shi
CHSS, KAIST
373-1 Guseong-dong,
Yuseong-gu
Taejon 305-701, Korea,
chungkon@kaist.ac.kr

Key-Sun Choi
KORTERM, KAIST
373-1 Guseong-dong,
Yuseong-gu
Taejon 305-701, Korea,
kschoi@world.kaist.ac.kr

Abstract

A multi-lingual lexical semantic wordnet called CoreNet has been developed by KAIST KORTERM. CoreNet is constructed based on one shared semantic hierarchy oriented from NTT thesaurus. Korean wordnet in CoreNet consists of 2,937 conceptual nodes (semantic categories) with 12 depth levels and of 51,172 senses for nouns, 5,290 for verbs, and 2,081 for adjectives in Korean. As a primary work for constructing a conceptual map of Korean words, this paper aims to show the concept distributions of Korean words in CoreNet based on the depths and semantic categories. The analysis results on concept distributions shows that WORK<ABSTRACT> and HUMAN ACTIVITY are the most broadly distributed concepts in nouns and verbs, while ABSTRACT RELATION, STATE, and ATTRIBUTE are the most ones in adjectives. This study provides the indispensable statistical data in order to construct a conceptual map of Korean words. Moreover, it allows to structurally and totally understand structure of Korean wordnet, to review proper specifications of semantic categories and correct assignment of concepts for Korean words and to prospect the next version of Korean wordnet.is a Word.

Introduction

A multi-lingual lexical semantic wordnet called *CoreNet*[1] has been developed by KAIST KORTERM. This wordnet is constructed based on the shared semantic hierarchy which is originated from NTT thesaurus (Ikehara, S. et al. (1997)). Whereas NTT thesaurus consists of 2,170 hierarchical semantic categories, Korean wordnet in CoreNet has 2,937 semantic categories which reflect the necessary concepts identified in the Korean language. The other characteristics of Korean wordnet is that the same semantic categories are applied to nouns, adjectives and verbs[2]. Every possible meaning of a word in *Urimal* Korean dictionary is mapped onto one or more concepts. Korean wordnet in CoreNet consists of 2,937 conceptual nodes (semantic categories) with 12 depths, and of 51,172 senses for 21,368 nouns, 5,290 for 1,758 verbs, and 2,081 for 813 adjectives.

In Section 2, we introduce semantic hierarchical structure and syntactic case frames of Korean wordnet with concepts and words. In Section 3, in order to construct a conceptual map of Korean words, we analyze concept distributions of Korean words (nouns, adjectives and verbs) based on the depths and semantic

[1] In this paper, the term "wordnet" refers to a set of words. The CoreNet system v 1.0 has Korean, Chinese, and Japanese. English and German will be included in CoreNet v 2.0.

[2] Different concept systems are applied to nouns and predicates in NTT thesaurus.

Petr Sojka, Key-Sun Choi, Christiane Fellbaum, Piek Vossen (Eds.): GWC 2006, Proceedings, pp. 147–156.

categories. Finally, we discuss conclusions and future works.

1 Semantic Hierarchical Structure of Korean Wordnet

1.1 Construction of Korean Wordnet

Korean wordnet in CoreNet is constructed using large corpora, *Urimal* Korean dictionary, and a single concept system. The noun wordnet is firstly constructed, and then secondly wordnet of predicates using lexical net of nouns. We briefly review construction method of Korean wordnet (cf. for details, see Lee, J.-H. et al. (2002) and Key-Sun Choi and Hee-Sook Bae(2004)).

Since it is difficult to deal with all nouns in *Urimal* Korean dictionary, we selecte basic nouns from KAIST POS-tagged corpora (1999, 2003) based on term frequency, and the information for basic nouns is extracted using *Urimal* Korean dictionary. The basic nouns (25,368 nouns with 69,242 senses) cover about 91.1% of all nouns in the corpora. The construction of Korean noun thesaurus can be considered as assigning semantic categories to each sense of the basic Korean nouns. NTT thesaurus has 2,710 hierarchical semantic categories and the relation between senses is Has-a or Is-a one. According to Key-Sun Choi and Hee-Sook Bae(2004), the construction of Korean noun thesaurus is assigning one of 2,710 semantic categories to each 69,242 senses.

For initial semantic category assignment, we use the translated noun list of NTT thesaurus into Korean. Then, we assign a semantic category by matching Korean words with their equivalent list for the semantic category in the NTT thesaurus. If no equivalent can be found in the translated word list, a genus term for the word is extracted from descriptive statements of a machine-readable dictionary. Some of translation errors are removed by manual correction of experts. To correct translation errors and wrong assignment between semantic categories and senses of nouns, word sense disambiguation is manually performed. The most difficult problem arises from the difference in concept division systems between Korean and Japanese. In Japanese, for example, concepts like FURNITURE has subordinate concepts like DESK, CHAIR, and FIREPLACE, while in Korean, FIREPLACE is dealt with as part of KITCHEN. These problems arise from the difference in the way of thinking and culture (Key-Sun Choi and Hee-Sook Bae(2004)).

For second semantic category assignment, assuming that meanings falling under a concept are defined by similar words in the dictionary, the definition of the dictionary is used to expand the thesaurus. For the senses-assigned-semantic categories, the definitions of the word senses are clustered by semantic categories. A cluster of the definitions is made per semantic category. The similarity between the definition and the cluster has to be computed to retrieve relevant clusters for that definition. The similarity between the sense and each cluster is computed and the most relevant cluster is founded, then we can find proper semantic categories of that cluster. Our structured version of the Korean dictionary includes such lexical relation information as synonyms, abbreviations, antonyms, *etc*. It is reasonable that the two senses linked by this lexical relation information (except for antonyms) fall under the same concept. The process of word sense disambiguation was manually performed in order to assign proper semantic categories to every possible meaning of a word, and translation errors were removed.

After construction of noun wordnet, wordnet of predicates is constructed by manual selection of specialists about each argument based on web environment considering usages extracted from large corpora. Syntactic case frames of predicates are also completed by manual selection of each argument and automatic mapping between selected argument and noun wordnet. Predicate wordnet consists of 5,290 senses for 1,758 verbs, 2,081 senses for 813 adjectives, 989 syntactic case frames for verbs, and 1,289 frames for adjectives.

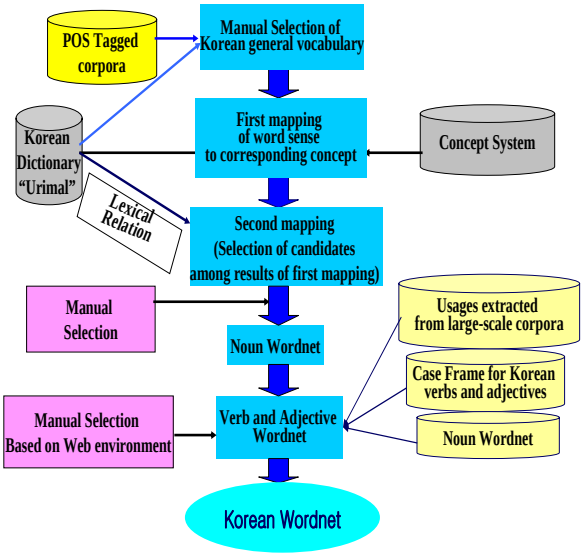

Fig. 1. Construction Process of Korean Wordnet

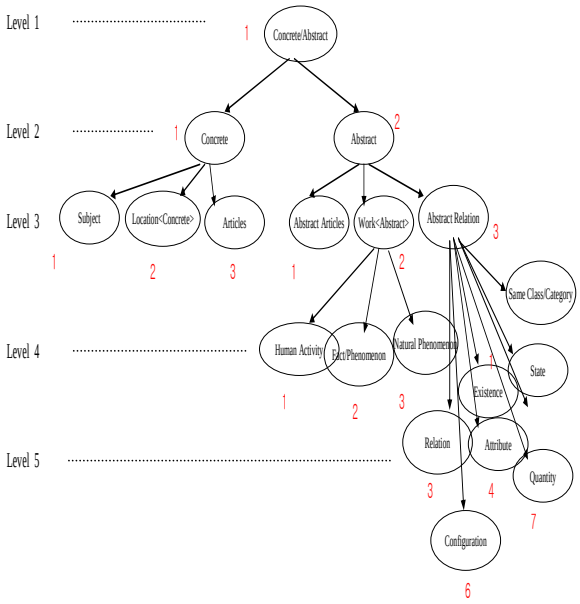

Fig. 2. Conceptual Hierarchy in Korean Wordnet

1.2 Conceptual Structure of Korean Wordnet

Korean Wordnet in CoreNet has 12 depths and 2,937concepts. The top level node starts with CONCRETE/ABSTRACT. The second level has two nodes: CONCRETE and ABSTRACT, and the third level has 6 nodes such as SUBJECT, LOCATION <CONCRETE>, ARTICLES, ABSTRACT ARTICLES, WORK<ABSTRCT>, ABSTRACT RELATION, and the fourth level has HUMAN ACTIVITY, FACT/ PHENOMENON, and so on. Each node has its proper concept identification number such as 1, 11, 12, 111, 112, 113, etc. The number of a chipper means the number of level. For example, the concept SUBJECT has 111 concept identification number and 111 with three digits means that the level of concept is third one and the concept number of its upper node is 11. The following <Fig 2> shows conceptual structure of Korean wordnet:

CoreNet Browser v 1.0 developed by KAIST KORTERM permits easily to navigate the whole conceptual hierarchy, and semantic network of each word (Fig. 3). All concepts are aligned with three languages: Japanese, Korean and Chinese. All words of these three languages (nouns, verbs and adjectives) are categorized into a single concept hierarchy. The same whole hierarchy of 2,937 concepts can be represented by spots (Fig 4).

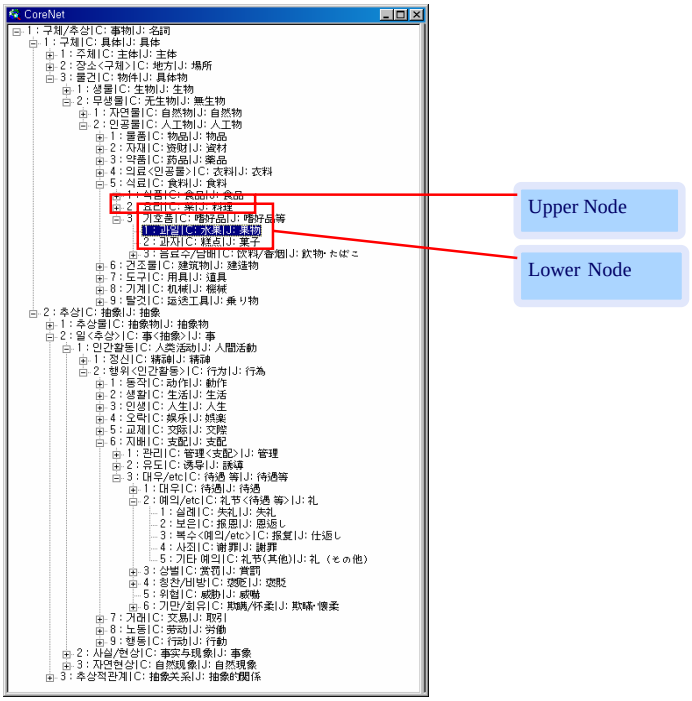

Fig. 3. A Part of Conceptual Tree (C: Chinese, J: Japanese)

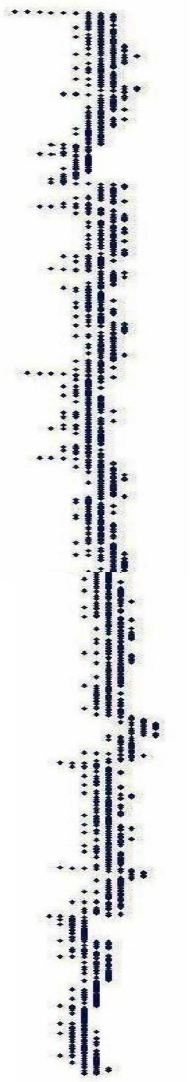

Fig. 4. Whole Conceptual Tree Represented by Spots

1. 3 Multiple mapping between concepts, word senses, and syntactic frames

The first purpose of CoreNet is mainly to remove semantic ambiguities in natural language processing using the two functionalities. The first functionality is mapping between every possible meaning of a word in the dictionary and one or more concepts in CoreNet. For example, eight meaning in the dictionary of the word *sagwa* is mapped onto six concepts; FRUIT, APPOLOGY, SCIENTIFIC DOMAIN, OFFICEAL, and so on. The following window shows this functionality well. Since the word *sagwa* is mapped onto the concept identification number and its concepts such as 11322531 [FRUIT], 111131224 [OFFICEAL], 1211112 [SCIENTIFIC DOMAIN], 122126324 [APPOLOGY], etc. We can confirm each sense of *sagwa* by dictionary search linked by web version of the dictionary. We can also easily navigate lexical semantic network of Korean words by typing concept itself, concept identification number and/or word itself or by clicking graphic window or conceptual tree.

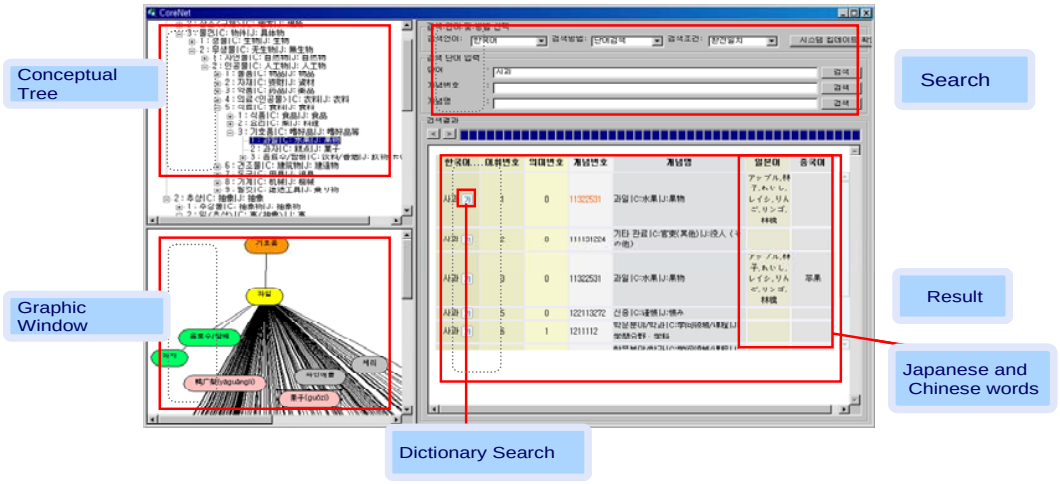

Fig. 5. Screenshot of mapping between concepts and word senses

Secondly, a syntactic-semantic structure is mapped onto the predicate-argument structure. For example, a Korean adjective *ganeulda* has a set of 6 senses in the dictionary; the word senses are mapped onto the 5 concepts such as SLEEM, SOUND<OTHER ASPECT>, SMALL, etc. This set of predicate concepts is identical to nouns' concept. Each predicate has its unique argument structure. For instance, in the syntactic case frames of predicates, *ganeulda* is mapped onto two concepts (e.g., SLEEM and SOUND<OTHER ASPECT>) whose argument structures would be different depending on the concepts. Each argument is represented by a set of possible concept filler (e.g., [VOICE], [WRITING STYLE]) and syntactic role (e.g., subject, dative, and object) in Korean phases. Each syntactic argument structure of each predicate is verified by Korean linguists. The following screenshot shows 20 frames of *ganeulda* according to its possible concept filler of its arguments:

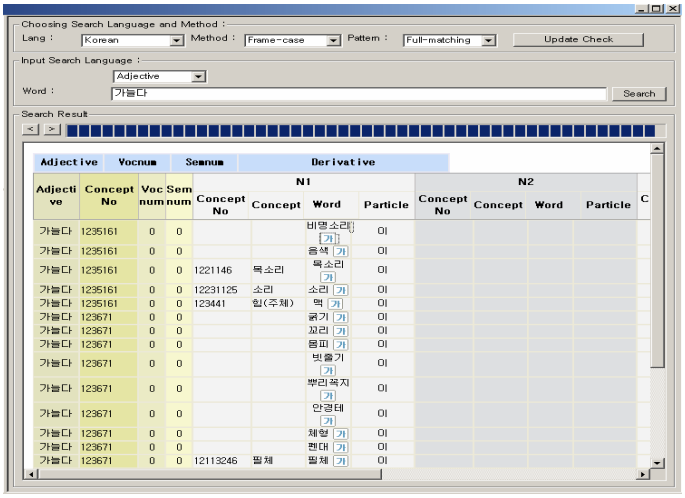

Fig. 6. Screenshot of Mapping Syntactic Frame, concepts and word senses

2 Concept Distributions of Korean Words in CoreNet

Now, we analyze distributions of Korean words in Korean wordnet based on the depths and semantic categories in order to construct their conceptual map. The following table shows the

distribution of 2,937 concepts according to the depth level.

Level	Num.of Concept	Distribution Rate
1	1	0.03
2	2	0.06
3	6	0.2
4	21	0.71
5	106	3.6
6	288	9.8
7	586	19.95
8	912	31.05
9	718	24.44
10	240	8.17
11	41	1.39
12	16	0.54
Total	2937	100

Table 1. Concept Distributions according to the depth level

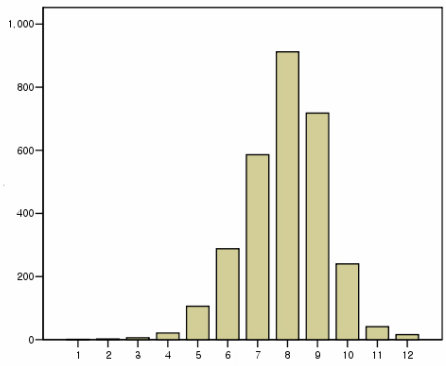

Fig. 7. Concept Distributions

The <Table 1> shows that the semantic categories in 7th, 8th, and 9th levels are well-specified. The Korean words are also broadly distributed in 7th, 8th, and 9th levels. We will consider distributions for nouns, verbs, and adjectives to draw conceptual map of Korean words.

2.1 Concept distribution of Korean Nouns

Korean wordnet in CoreNet has 51,172 senses for 21,368 nouns. 51,172 senses are assigned to 12 levels such as <Table 2> and <Fig 8>. The Korean nouns are broadly distributed in 7th, 8th, and 9th levels.

Level	Num. of Nouns
1	2
2	10
3	11
4	135
5	3321
6	6738
7	11078
8	15456
9	11129
10	2770
11	325
12	197
Total	51172

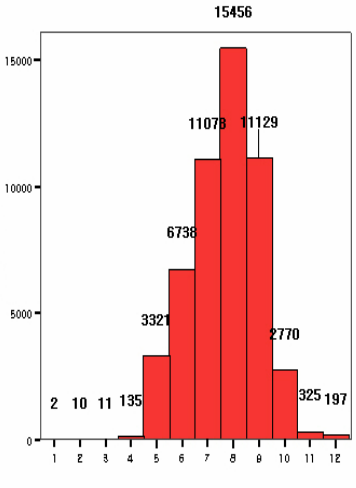

Fig. 8. Distribution of Nouns according to the levels

Table 2. Distribution of Nouns

The best top 10 concepts of noun distribution are [DISEASE] (Lev 9), [PERSON'S NAME] (Lev 8), [DOING MAN] (Lev 7), [DOCUMENT] (Lev 6), [MIND] (Lev 6), [DEGREE] (Lev 5), [HOUSE] (Lev 8), [MAN;OTHER POSITION] (Lev 7), [OTHER OFFICIAL] (Lev 9), and [PERIOD (NATURE/HUMAN ACTIVITY/Etc.] (Lev 7) (cf. See the table 3):

Concept ID Number	Num of Nouns	Concept	Level	Percentage	WORDS
122323213	429	**DISEASE**	9	0.80%	liver cancer, hepatitis , tuberculosis, pneumonia, migraine
12113121	382	**PERSON'S NAME**	8	0.70%	name, surname, alias, title, family name
1111336	330	**DOING MAN**	7	0.60%	conspirator, member, supervisor, connoisseur, surveillant
121142	311	**DOCUMENT**	6	0.60%	housekeeping book, memoranda, estimate sheet, surveying report
122111	284	**MIND**	6	0.50%	heart, competitive spirit, resolution, wariness, warning
12325	264	**DEGREE**	5	0.50%	notables representing various social circles, each class, class, grade
11322611	264	**HOUSE**	8	0.50%	temporary building, shop, building, prison, gallery
111132A	260	**MAN; OTHER POSITION**	7	0.50%	Inspection, review, advisor, Governor, legation
111131224	254	**OTHER OFFICIAL**	9	0.40%	auditor, chief public prosecutor, policeman, schoolmaster
1239221	245	**PERIOD(NATURE /HUMAN ACTIVITY/Etc.**	7	0.40%	the blooming season, prescription, turn of life, transitional period

Table 3. Best Top 10 Concepts in Noun Distributions

Now, we consider concept distributions of nouns according to the depth level. In level 2, the concrete nouns occupy 40% and the abstract nouns 60%. This means that in Korean, the number of abstract nouns is bigger one of concrete nouns. For nouns in level 3, WORK<ABSTRCT> occupies 30.30%, ARTICLES 20.20% ABSTRACT RELATION 17.22%, SUBJECT 14.00%, ABSTRACT ARTICLES 12.47%, LOCATION<CONCRETE> 5.78%. For nouns in level 4, HUMAN ACTIVITY occupies 21%, INANIMATE 15%, MAN 12%, ABSTRACT ARTICLE<MIND> 9%, FACT/PHENOMENON 5%, ANIMATE 5%, NATURAL PHENONMENON 4%, ABSTRACT ARTICLE<ACTION> 3%, TIME 3%, QUANTITY 3%, STATE 3%, ESTABLISHMENT 3%, LOCATION 2%, QUALITY 2%, ORGANISATION 2%, ATTRIBUTE 2%, REGION 2%, RELATION 1%, CONFIGURATION 1%, and SAME CLASS/ CATEGORY 1%. The analysis results on concept distributions of nouns in level 3 and level 4 shows that WORK<ABSTRACT> [122] and HUMAN ACTIVITY [1221][3] are the most broadly distributed concepts in Korean nouns. This method permits one to understand noun distributions in each level at a glance.

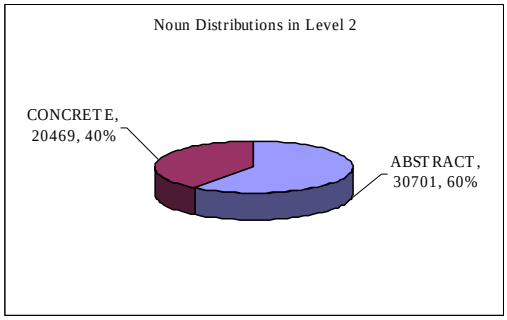

Fig. 9. Noun Distributions in Level 2

[3] HUMAN ACTIVITY [1221] is a subordinate concept of WORK<ABSTRACT> [122].

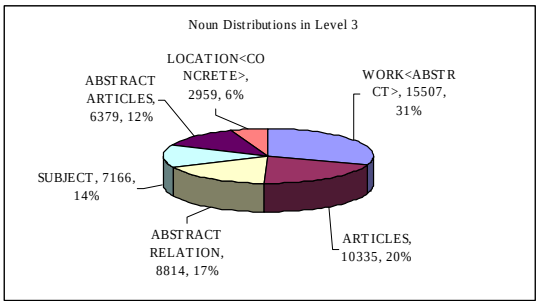

Fig. 10. Noun Distributions in Level 3

2.2 Concept distributions of Korean Adjectives

For adjectives, Korean wordnet in CoreNet has 2,081 senses for 813 adjectives and 1,289 syntactic case frames. 2,081 senses are assigned to 12 depth levels such as Fig 11. The Korean adjectives are broadly distributed in 6th, 7th, and 8th levels.

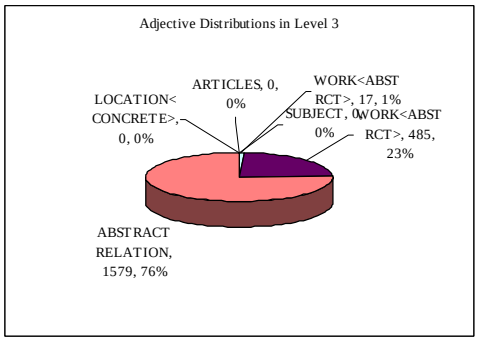

Fig. 11. Adjective Distributions in Level 3

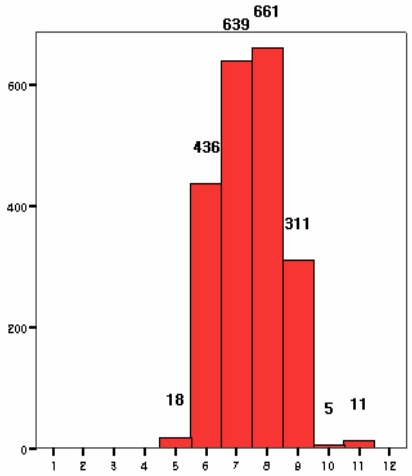

Fig. 12. Distribution of Adjectives

Consider distribution of adjectives according to the levels. For adjectives in level 3, ABSRACT RELATION occupies 76% and WORK<ABSTRACT> 23%. For adjectives in level 4, STATE occupies 41%, ATTRIBUTE 14% HUMAN ACTIVITY 13%, CONFIGURATION 9%, NATURAL PHENOMENON 5%, RELATION 3% QUANTITY 3%, FACT/PHENOMENON 3% EXISTANCE 1%, and SAME CLASS/CATEGORY 1% (cf. Fig 13). The analysis shows that STATE and ATTRIBUTE occupy 55% in level 4, and that ABSRACT RELATION [123], STATE [1235] and ATTRIBUTE [1234] are the most broadly distributed concepts in Korean adjectives. The distribution in level 5 shows more detailed information in adjectives: ASPECT occupies 30%, MINE 11.3%, PROPERTY 10.2%, FEELING 8.9%, SHAPE 8.6%, INANIMATE PHENOMENON 7.5%, POWER 3%, CHANGE 2.5%, CASE 2.1%, DEGREE 2%, SUPERIRITY /INFERIORITY 1.7%, SAME/DIFFERENCE 1.3%, SUITABILITY/INSUITABILITY 1%, and LARGE/ SMALL 1%.

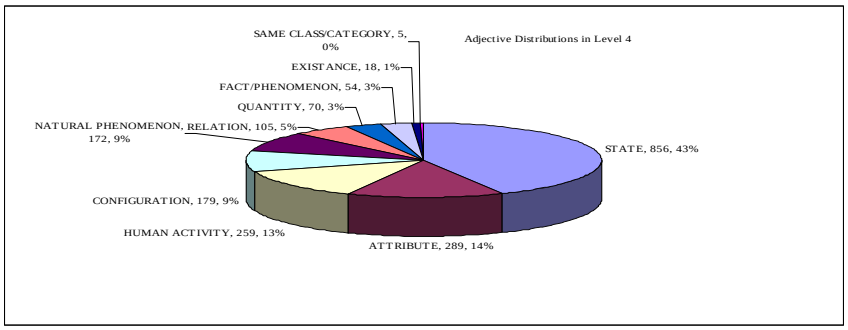

Fig. 13. Adjective Distributions in Level 4

2.3 Concept distribution of Korean verbs

For verbs, Korean wordnet consists of 5,290 for 1,758 verbs and 989 syntactic case frames. The Korean adjectives are broadly distributed in 8th and 9th levels.

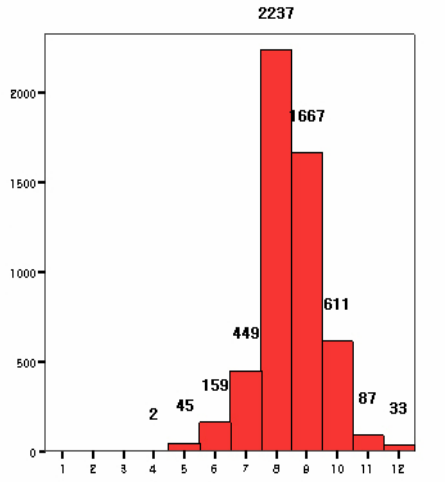

Fig. 14. Distribution of Verbs

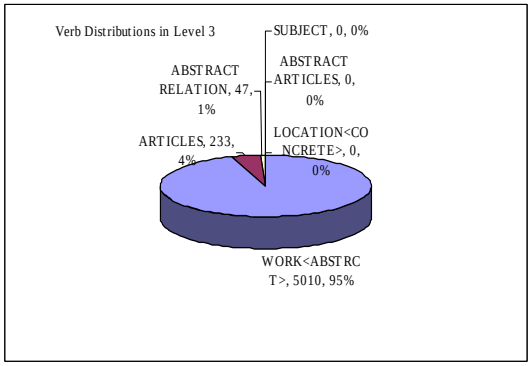

Fig. 15. Verb Distributions in Level 3

For example, for verbs in level 3, WORK<ABSTRACT> occupies 95%, ABSTRACT RELATION 4%, and ABSTRACT ARTICLE 1%. One may see the WORK<ABSTRACT> has an overwhelming distribution rate. For verbs in level 4, HUMAN ACTIVITY occupies 50%, FACT/ PHENOMENON 36%, NATURAL PHENOMENON 8%, and RELATION 2%. Likewise nouns, WORK<ABSTRCT> [122] and its subordinate concept HUMAN ACTIVITY [1221] are the most broadly distributed concepts in Korean verbs. In level 5, CHANGE occupies 36%, ACTION<HUMAN ACTIVITY> 32%, MIND 18%, INANIMATE PHENOMENON 4%, and ANIMATE PHENOMENON 4%. In level 6, LABOR 8.4%, PROCESS 6.8%, APPEARENCE/DISAPPEARENCE 5.6%, MEETING/ PARTING 5.2%, BUSINESS 5.0%, ACTION 4.9%, CONTROL 4.5%, ENTERANCE 4.3%, THOUGHT <MIND> 4.2%, EMOTION 3.9%, INCREASE/

DECREASE 3.7%, SHAPE OF OBJECT 3.7%, and ASSOCIATION 3.6%.

Conclusion

We introduced a semantic hierarchical structure of Korean wordnet in CoreNet which consists of 2,937 conceptual nodes (semantic categories) with 12 depths, and of 51,172 senses for nouns, 5,290 for verbs, and 2,081 for adjectives. We analyzed concept distributions of Korean nouns, adjectives and verbs based on the depths and concepts. For nouns, abstract nouns occupy 60% and concrete 40%. The analysis results on conceptual map showed that WORK<ABSTRACT> [122] and HUMAN ACTIVITY [1221] were the most broadly distributed concepts in nouns and verbs, while ABSRACT RELATION [123], STATE [1235] , and ATTRIBUTE [1234] were the most ones in adjectives.

This study certainly provides the indispensable statistical data in order to construct conceptual map of Korean words. Moreover, it allows one to structurally and completely understand the structure of Korean wordnet, to review proper specifications of semantic categories and correct assignment of concepts for Korean words and to proceed to the next version of Korean wordnet. In future work, we will respectively connect concept distributions of nouns, verbs, and adjectives and study their relations to construct conceptual map of Korean wordnet. We can also apply this method for the other languages in CoreNet.

Acknowledgements

This work was supported by the Brain Korea 21 Project, School of Information Technology, KAIST in 2005.

References

Ikehara, S. et al(1997) *The Semantic System*, volume 1 of Goidaikei . A Japanese Lexicon, Iwanami Shoten, Tokyo.

Hangeul Society, ed.(1997) *Urimal Korean Unabridged Dictionary*, Eomungag .

KAIST Corpus (1999, 2003). (in Korean) http://morph.kaist.ac.kr/kcp/

Lee, J.-H. et al. (2002) *Semi-Automatic Construction of Korean Noun Thesaurus by Utilizing Monolingual MRD and an Existing Thesaurus*, Proceedings of the 16th PACLIC, Jeju

Key-Sun Choi and Hee-Sook Bae(2004) *Procedures and Problems in Korean-Chinese-Japanese Wordnet with Shared Semantic Hierarchy*. 2nd International Wordnet Conference, Mazaryk University, Brno (Czech Republic), January.

Sun-Mee Bae and Chung-Kon Shi (2005) *Study on nouns in KAIST Wordnet*, Proceedings of 2005 Korean Society for Language and Information Summer Conference, pp. 39-54.

KORTERM(2005), KAIST Wordnet 1, 2, 3: KAIST Press