Compiling a Monolingual Dictionary as an Active Dictionary-Focusing on the Procedure of Yonsei Contemporary Korean Dictionary Compiling Project -

Ik-Hwan LEE, Kil-Im NAM, Chongdok KIM,Eui-jeong AHN, Jong-Hee LEE,
Institute of Language and Information Studies
Yonsei University,
Seoul, 120-749, Korea.
Tel: (02)-2123-4198
e-mail : nki@lex.yonsei.ac.kr

Abstract
This paper has two major purposes: introducing the procedure of the project of the compilation of *Yonsei Contemporary Korean Dictionary*(*YCKD*), which has been in progress since 2002 as a 7-year-project, and introducing its characteristics as an active dictionary. This paper presents the project from two points of view. First of all, this provides the project plan, focusing on constructing large corpus of contemporary Korean and on developing lexicographer's electronic workbench. Then, this paper explains the characteristics of the future dictionary as an active one. From users' points of view, we pay attention not only to offering users the meaning of a word, but also to making them understand and use their actual language.

The *YCKD* compiling project is going on in three phases. The first phase, a basic-work-phase(Sep. 2002 ~ Aug. 2003), is accomplished and the second phase, draft-composing-phase(Sep. 2003 ~ Aug. 2007) is now under way.

This paper will discuss the following: construction of Korean corpus for compiling *YCKD*, development of aiding tools for editing dictionaries, organization of headwords, and characteristics of *YCKD* as an active dictionary.

1. Introduction

Thanks to the recent development of corpus linguistics, computational linguistics, and lexicography, many changes and developments have been achieved in lexicographical fields. The Institute of Language and Information Studies of Yonsei University published *Yonsei Korean Dictionary* in 1998. Its headwords and examples were obtained from the corpus constructed by Yonsei University for the first time in Korea. The institute also published *Yonsei Elementary Korean Dictionary* in 2001. Sangsup Lee presented the compiling procedure of *Yonsei Elementary Korean Dictionary* at the Euralex'98, which was the result of an analysis of our educational corpus.

The present project which succeeds to these two preceding dictionaries aims to describe two hundred thousand contemporary Korean words based on the corpus from the year of liberation, 1945, to the present[1].

YCKD intends to be an active dictionary[2] and it is a Korean native speaker-oriented dictionary. We define the native speakers of Korean as people who use dictionaries to

choose good and proper expressions when they write or speak. Basically, main users of *YCKD* will be high school students and college students in composition classes and the general public who intend to write good sentences. Like this, *YCKD* characterized as an active dictionary will be a more advanced one than other existing Koreans dictionaries, which are mainly used to look up a word that users do not understand. We developed devices to embody these characteristics as an active dictionary in every step such as constructing corpus, selecting headwords, making-up information items and presenting appendix.

In this paper we introduce the plan of our 7 years dictionary project started in 2002, and present the characteristics of our dictionary as an active one. Our dictionary *YCKD* has several particular goals, which distinguish it from other dictionaries.

First, *YCKD* is a dictionary not only for a better understanding of the words in question, but also for their meanings and their actual usages with adequate expressions. That is to say, from the users' viewpoints *YCKD* is an active dictionary that helps users comprehend and express words. To meet these needs, we select headwords according to the frequency in use, describe meaning of a word and its usage, and develop various patterns of the reference information headed by pragmatic information.

Second, *YCKD* heads for a dictionary preparing for the era of reunified Korea and facilitating communication between North and South Koreas. For this purpose, we use the frequency of words in the North Korean corpus and we include North Korean words in our entries.

Third, *YCKD* is going to be the first dictionary in Korea that includes spoken words and explains spoken usages of the words. Therefore, it is better than any other existing dictionaries which mainly consist of written words. We analyze and treat the spoken language corpus that has already been constructed with various typical spoken data such as actual conversations, many kinds of conferences or meetings, radio forums, TV debates and conversations in TV soap operas.

Fourth, *YCKD* will make the best use of appendix and help high school students, college students and the general public to understand Korean better. The appendix will mostly consist of words and expressions for writing, especially for the composition of logical writing. Besides, the appendix will present everyday composition skills such as resumes and cover letters with good examples.

In section 2 we will present the plan of our project, and in section 3 we will introduce the method of describing the information items for our active dictionary.

2. The Plan of <YCKD> Project

2.1 The Compilation of a Large Corpus for Contemporary Korean

Yonsei Contemporary Korean Dictionary deals with Korean words from the year of liberation, 1945, to the present. Therefore, the corpus as a basic source of dictionary must be constructed according to the time periods. Considering change of Korean, and the kinds and quantities of publications, a large corpus for contemporary Korean has been compiled and divided into three periods: from 1945 to 1965(the first period), from 1966 to 1994(the second period), and from 1995 to the present (the third period).

Now we supplement the first period corpus because the publications of this period are not abundant. This corpus includes sino-Korean and education materials, which are of great value. The volume of this corpus is 10 million. The second and third period corpora will be added to the existing Yonsei Korean Corpus 1-9 composed of 43 million words.

The corpus for *YCKD* will include 100 million words. Corpus compilation and research on construction of a balanced corpus with representativeness are carried on at the same time. The reason is that the corpus will be used for headwords composition, concordance source, and for some frequency information. To compile the balanced corpus composed of 100 million words, first we try to compile the base corpus composed of 10 million words. After testing this 10-million-word-corpus with some statistical analyses, we will enlarge the base corpus to 100-million-word-corpus[3].

Beside the general language corpus, there are some specialized subcorpora such as the spoken language corpus, the North Korean corpus, the corpus of Korean used in Yanbian, Russia, etc, the corpus including sino-Korean and the corpus for classified terminology.

2.2 The Development of Lexicographer's Electronic Workbench

We have many sorts of lexicographer's electronic workbenches, but this paper deals with a concordance program and an editing one.

The major function of a concordance program is to extract a list of all the examples of the target by using a large corpus. YDCONC based on the function of pattern matching was designed and tested in 2002, but there are several limitations of this program. Therefore, a new concordance that can be looked up by the theme and date of the corpus has been developed since 2003.

To compile *YCKD*, we also designed WPacker, a workbench that manages the data files and lexical entries. It is very important to structure lexical entries, especially for developing a CD-ROM dictionary. The WPacker consists of two panes, concordance lists and edit window for the dictionary draft. This is helpful in that the selected examples are easy to move from the pane of concordance list to the pane to edit window. The edit window for dictionary draft was designed on the base of XML. This edit window is also helpful in that the structure of a word is easy to change by being used[4].

2.3 Analysing Corpus and Composing Headwords

We plan to have 200,000 headwords, namely 150,000 general headwords and 50,000 special ones. To extract headwords, we analyze a large Korean corpus (which contains 100 million words) and make a word-frequency list. However, we do not have the word-frequency list now. Thus we use temporarily the headword list constructed as described in Table 1.

Group	Data	Size
I	The headwords of YKD (Yonsei Korean Dictionary)	50,000 words
II	The tokens which appear more than 3 times in the Yonsei Korean Corpus1-9 (excluding group I)	40,000 words
III	The additional headwords extracted from the database of headwords of main dictionaries (excluding group II)	3,000 words
IV	The headwords complemented from the first and third period corpus	6,000 words
V	The headwords complemented from the textbook published after the year of 2000	1,000 words
VI	The selected tokens which appear 1 or 2 times in the Yonsei Corpus1-9	40,000 words
VII	The homonyms omitted in YKD	10,000 words
	Total	150,000 words

Table 1. The Structure of 150,000 general headwords of *YCKD*

3. The Characteristics of *YCKD*

Our dictionary *YCKD* is an "active dictionary for comprehension and expressions". By an "active dictionary" we mean that it actively helps the users to produce texts and express their thoughts and feelings in speaking and writing. *YCKD* aims to provide the users with tools of expressions, whereas the other dictionaries published so far have aimed for comprehension of texts only.

3.1 The Characteristics of Headwords

YCKD provides 200,000 Korean words used from 1945 through 2005. The headwords are listed on the basis of the 100 million words corpus of written Korean and the 1 million words corpus of spoken Korean. Into headwords we put not only written forms but also **spoken** forms like *du* (also, too) or *dwege* (very much).

YCKD lists many new words made from new systems like *bimilbeonho* (password), *mutongjang* (without an account book of bank) and introduces words from technical inventions and foreign origin words like *syopingmol* (shopping mall), *sidirom* (CD-rom). We also provide some dialects if they are used all over the country.

(1) narak (rice-plant) dialect of *byeo*
(2) eolleong (quickly) dialect of *eolleun*

We put some North Korean words into headwords in order to facilitate the communication and cultural exchange between North and South Koreans. We think we should prepare for the unified Korea. Here are some exemples:

- Different spellings for same words; *yeoja*(in S.K.[5]) // *nyeoja*(in N.K.[6])(woman)
- Different forms are used for the same meaning
  ; *byeorakbooja* (S.K.) // *gapjakbooja* (N.K.) (overnight millionaire)
- New words only found in North Korean; *bapgongjang* (factory for cooked-rice)

We also put proper nouns into headwords. Notice that we didn't provide proper nouns in *Yonsei Korean Dictionary* because it is a linguistic dictonary.[7] *YCDK* is an advanced learners' dictionary. So we think that advanced learners will need some information of proper nouns. We are going to choose proper nouns listed in junior high and high school textbooks, and also those found in the corpus. We attach registers to them such as ((biographical)), ((geographical)), and ((title of book)).

(3)Sejong noun ((biographical)) The king of Joseon Dynasty. [1397~1450]

(4)Seoul noun ((geographical)) The capital of Korea. [605.02㎢/ 10,280,523] (in 2002)

(5)Ginesbuk noun ((title of book)) ...

3.2 The Characteristics of Information

Special **usage notes** are provided to show various usages of a headword for the production of texts: for example, the similarities and differences of two or more words, a word and its subordinate words and so on.

For example, we use two special **usage notes** for the headword, *jookda*(to die). One is 'Comparison of several words which have similar meanings with *jookda*(to die). They are *doragasida*(to die used for honorable people) /*byeolsehada*(to die used for honorable people especially in written Korean)/*samanghada*(to die only used for humanbeing)'. These differences are given in a usage note under the headword, *jookda*. The other is 'Several idioms for *jookda*(to die). There are many idioms for 'to die', because the notion of 'death' is very important in human life and deeply related to our everyday life. We try to collect and classify them into groups according to the expressions.

<Special usage note 1: comparison of *jookda*(to die) and a few words >

- *samanghada* ; to die only used for the human being
- *unmeynghada* ; to die used usually in official papers
- *byeolsehada* ; to die used for honorable people especially in written Korean
- *doragasida* ; to die used for honorable people

<Special usage note 2: several idioms for *jookda*(to die) >

- Analogy from a dead body : to close eyes, to lie down, etc.
- There is a different world for the dead : to leave this world, to go to the other world, etc.
- Related to religions : to go to heaven, to go to hell, etc.
- Euphemistic forms : to sleep, to close one's life, etc.

- How to die : to become a fish food, etc.
- Slangs : *duejida*, etc.

We provide pragmatic information of each word. Several registers, for example, spoken/written, archaic, honorific and slang/vulgar are presented in front of the definition.

(6) dwege adverb ***spoken*** very much.
(7) ppyamttagwi noun ***slang*** chic.

The real frequency of a word is provided: how often they are used in written Korean and spoken Korean, and which part of speech is used more often if a word has two parts of speech (see the three figures below).

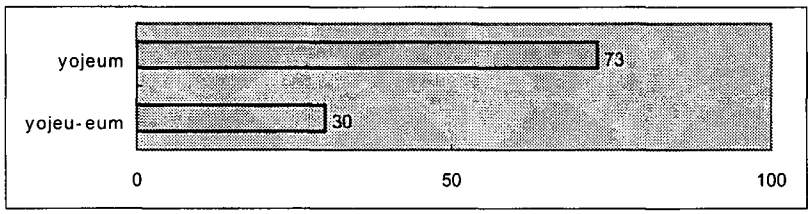

Figure 1. Real frequency of *yojeum* and *yojeu-eum* in written Korean (times)

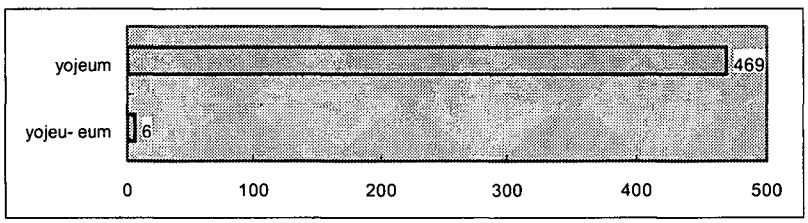

Figure 2. Real frequency of *yojeum* and *yojeu-eum* in spoken Korean (times)

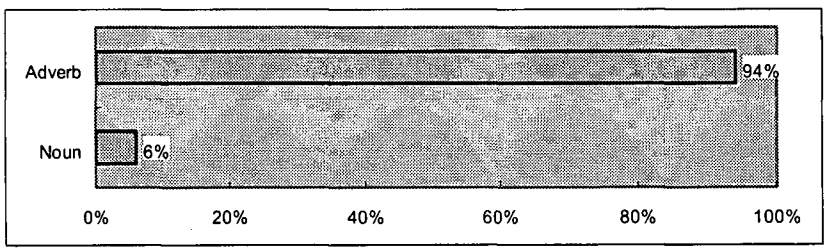

Figure 3. Real frequency of *jinjja* in spoken Korean (%)

3.3 The Differences of meaning descriptions

We describe information related to the usage of a word in grammatical and practical uses in addition to definitions. According to the contents, we use 〔 〕 for morphological/ grammatical information and ( ) for pragmatic/discourse use, as shown in (3).

(8) gyesi-da verb 〔used as an adjective〕 (for honorable people) to exist.

(9) mat-da verb 〔used in the front of a sentence〕 (to agree with a speaker) to be right.

3.4 Various Examples

YCKD provides not only phrasal examples but also many example sentences so that dictionary users understand what a word means easily. Some example sentences are given the source book titles and names of authors if they are suitable for composition or a headword itself has a useful meaning.

3.5. Korean Verb Conjugation Table

At present we cannot find a table of the Korean verb conjugation in any Korean Dictionary. In our dictionary we prepare a large table to show the whole inflectional forms of all the Korean verbs.

First of all, we have selected the 61 typical inflectional endings in the order of frequency in consulting the "Korean Inflectional Endings and Particles Dictionary" (Jong-hee Lee and Hi-ja Lee). This table shows the 61 inflectional endings arranged in a row as in Figure 4. Second, at least 64 verbs are needed to show all the inflectional forms of verbs. The Figure 4 shows them. Since the detailed criteria for selecting the verbs are too complicated to explain here, we just want to make sure that our principle is clear: we gather all the possible inflectional forms of Korean verbs for the conjugation table.

The second column shows part of speech for each word, the third one shows some verbs and adjectives that are chosen for the conjugation table, the forth one shows the meaning of each word, and from the fifth column to the last one show 61 inflectional endings. This table works like this: the example verb of the second row, *sa-da* (to buy) and the inflectional ending of the fifth column, *-neurago* (as a result of) become to *sa-neurago* (as a result of buying), and so on. We can see that the adjective *chup-da* (be cold) (number 30 of the first column) becomes *chuu-n* (which is cold) with an ending *-nun* (relative present).

Since the whole conjugation table is too large, it is impossible to show all the items. The asterisk in the figure below means that the expected verb stem and ending are impossible.

번호	품사	단어	의미	1	2		32	33		61
				이유	현재	...	목적	의도	...	대등
				연결	관형사형		연결	연결		연결
				-느라고	-는		-러	-려고		-지만
1	동사	사다	책을 사다	사느라고	사는	...	사러	사려고	...	사지만
2	동사	보다	영화를 보다	보느라고	보는		보러	보려고		보지만
3	동사	서다	기차가 서다	서느라고	서는		서러	서려고		서지만
		...		...						
21	동사	잡다	손을 잡다	잡느라고	잡는		잡으러	잡으려고		잡지만
22	동사	접다	종이를 접다	접느라고	접는		접으러	접으려고		접지만
		...		...						
30	형용사	춥다	날씨가 춥다	*	추운	...	*	*	...	춥지만
31	형용사	밉다	hateful	*	미운		*	*		밉지만
33	동사	낳다	아기를 낳다	낳느라고	낳는		낳으러	낳으려고		낳지만
34	동사	넣다	책을 가방에 넣다	넣느라고	넣는		넣으러	넣으려고		넣지만
		...		...						
63	동사	부르다	이름을 부르다	부르느라고	부르는		부르러	부르려고		부르지만
64	형용사	빠르다	지하철이 빠르다	*	빠른		*	*		빠르지만

Figure 4. The Korean Verb Conjugation Table

4. Conclusion

In this paper we have presented the construction of a large Korean corpus for our dictionary *YCKD*, the development of the aiding tools for editing dictionaries, the structure of the headwords, and the characteristics of *YCKD* as an active dictionary.

YCKD will be published in 2009. It will have two versions: paper dictionary and electronic one. Among the whole three phases, this project accomplished the first one, basic-work-phase (Sep. 2002 ~ Aug. 2003) last year and this year is the first year of the second phase, draft-composing-phase (Sep. 2003 ~ Aug. 2007). At this phase, each researcher composes drafts by word types such as part of speech, technical terminology, North Korean, and dialects. They also keep developing many devices for the CD electronic dictionary like information items.

Thanks to the Internet facilities, the rapid development of electronic mediums, and the advancement of domestic and foreign lexicography by developing various corpora, we expect a variety of positive changes on the dictionary market in Korea.

In conclusion, *YCKD* will be a modern Korean dictionary which covers modern Korean in the latter half of the 20th century and includes word usage patterns of Korean in the first half of the 21st century. *YCKD* will also be a dictionary for expressions and an active dictionary which will be of help to writing good sentences.

Endnotes

1. Like this, it is very meaningful to select headwords used from 1945, the year of liberation. Since 1945 people have learned Korean at schools and they are the first *Han-geul* 'Korean' generation. Besides, publications in Korean have been increased since that time. For these reasons, including words used from 1945 means *YCKD* covers all the modern Korean spoken by living Korean.

2. Aktives wörterbuch : Ladidlav, Zgusta (1971), *Manual of Lexicography*, Mouton: The Hague-Paris, (recited from Hee-ja, Lee(2003), "A theory of lexicography", *the 4th Korean Language*

Information-Oriented Project Academy, Institute of Language and Information Studies of Yonsei University

3. About some statistical analyses, see Yong-jin Kwak (to appear, 2004).
4. About Wpacker, see Choon-ho Choi(2002)
5. South Korea
6. North Korea
7. Usually 'grand dictionaries' have proper nouns in Korea, except <Grand Korean Dictionary>. It is said, "we do not provide proper nouns like person's name, the name of places, and so on, ... but we put some words into entries like historical and related to Korean culure and literature and so on." (<Grand Korean Dictionary>(1992: p. 9))

References

Choi, Choon-ho. 2002. Design and Implementation of Workbench for Korean Dictionary. Yonsei Univ.

Institute of Language and Information Studies of Yonsei University. 1998. Yonsei Korean Dictionary, DooSan Press, Seoul, Korea.

Institute of Language and Information Studies of Yonsei University. 1998. Yonsei Elementary Dictionary, DooSan Press, Seoul, Korea.

Korean Language Society. 1992. Grand Korean Dictionary, Eomumgak

Kwak, Yong-jin. to appear. 2004. "Automatic Construction of Purposive corpus", Language, Information and Lexicography I, The institute of language culture.

Lee, Hee-ja and Lee, Jong-Hee. 2001. Korean Inflectional Ending and Particle Dictionary, Hankook Munhwasa. Seoul, Korea.

Lee, Hee-ja. 2003. "A theory of lexicography", the 4th Korean Language Information – Oriented Project Academy, Institute of Language and Information Studies of Yonsei University

Lee, Sangsup. 1998. "Compiling a Monolingual Learner's Dictionary on Corpus Linguistics", Actes Euralex'98 Proceedings, Université de Liège, Belgique.