Informatization and Use of Korean Language Data

Seo Sang Kyu

This article aims at making a general survey of the results of informatization of the Korean language data and of the important results and characteristics of research using the corpora of the Korean language from the mid-1980s to the present according to several major points. Meanwhile, through a variety of studies and projects concerning the computerization of Korean language, Korean language data useful to various research and development of language processing techniques have increased more diversely in quantity and type than ever before. We can positively ascertain that practical and useful methods of research based on corpora and analyses of usage are becoming one of the major methods in studying the Korean language.

Keywords: Informatization, Linguistic Informatics, Korean Language Data, corpus, corpora

Introduction

Nowadays, in the middle of the 2000s, the word "informatization" has become common in every field of studies. Setting limits to the studies of the Korean language, almost twenty years have passed since discussion on how to use computers in the studies of the Korean language began in the middle of the 1980s.[1] Meanwhile, through a variety of studies and projects concerning the computerization of the Korean language, Korean language data useful to various research

[1] As indicated in Seo Sang Kyu (2002), emphasis was put on implements to accurately and rapidly supply useful data for scholars of the Korean language, or the necessity of mechanization and computerization of existing data in the early days.

The Review of Korean Studies Vol. 8, No. 4 (71-104)

and the development of language processing techniques have increased more diversely and enormously in quantity and type than ever before. In addition to the accumulation of basic materials and developing the stage of techniques to deal with them in the early days, research has been expanded to groping for using various linguistic data for a practical linguistic study, the diversification of data, figuring out the aspects of understanding the Korean language practically through spoken language, and even to the stage of developing processing techniques to cover syntax and semantics.

This article aims at making a general survey of the results of the informatization of Korean language data and of the important results and characteristics of research using the corpora of the Korean language. In fact, the informatization of the Korean language has already been recollected many times in the academic world. In particular, most os the recollections and arrangements as to informatization were made in *Hangugeowa jeongbohwa* (The Korean Language and Informatization) by Hong Yun-pyo, et al. (2002).[2] Therefore, this study will reflect recent new information but inevitably has to depend on some previous studies.

The Present Situation of Informatization of Korean Language Data

The primary aim for the informatization of Korean language data is the construction of the "corpus."[3] This is a primary source of information among the various stages of informatization mentioned in Hong Yun-pyo (2002: 26).

In 1987, the Yonsei Korean Corpora (abbr. YSC hereinafter) was constructed for the first time. When many institutions joined in collecting and constructing the enormous corpora earnestly, the construction of corpora generally meant the computerization of documentary data, that is, the written language. Though this phenomenon has generally continued in almost the same way up to the present, it is a big change to begin the earnest construction of a spontaneous spoken corpus through the 21st Century Sejong Project. At the turn of the 21st century, a variety of corpora such as an enormous speech corpus by means of the development of speech processing

[2] I have already published such theses on the present situation and the results of the informatization of the Korean language.

[3] Refer to Seo Sang Kyu (2002: 255-257) or Chapter 2 of Seo Sang Kyu and Han Young-gyun (1999) for basic understanding of the term, 'corpus', concepts, and types. Various types of corpus are easily classified in the table of Hong Yun-pyo (2002: 31).

technology, a learner's corpus collecting the writing and speech of learners of the Korean language, and a regional dialect corpus, etc., have evolved.

Table 1. The Present Situation of Developing Corpora of Some Major Institutions as of the End of 2004

Corpus (Constitution)	Period of Construction	Scope (Unit: Phrase) & Objects of Collection	Purpose
Yonsei Korean Corpora (YSC) (Center for Linguistic Informatics of Yonsei Univ.)	1987 to present	▲ Written corpus of modern Korean (YSC1-3,YSC5-9, about 41.85 million) ▲ Spoken/quasi-spoken corpus (YSC4, about 770 thousand) [*] ▲ Special corpus (about 25 million)[†] ▲ Morphological annotated corpus (about 1.8 million) ▲ Semantic annotated corpus (about 1 million)[‡] ▲ Korean learner's corpus (total 9 types, about 660 thousand)	Lexicography (*Yonsei Korean Dictionary*, etc.), studies on the Korean Language, developing a balanced corpus, Korean language teaching.
Corpora of the National Institute of the Korean Language [§]	1992-1999	▲ Spoken corpus of modern Korean (about 34 million) ▲ Quasi-spoken corpus[**] (about 2.02 million) ▲ Corpus of oral traditions (about 2.2 million) ▲ Corpus of historical documents (about 6.04 million)	Lexicography (Standard Korean Dictionary), studies on the Korean language.
KOREA-CORPUS 1 (Institute of Korean Culture, Korea Univ.)	1995	▲ Corpus of Modern Korean (About 12% of quasi-spoken corpus among about 10 million of written corpus)	Developing a balanced corpus, studies on the Korean language, lexicography.
Corpora of Constructing the Basis of Korean Language Processing (KAIST)	1994-1997	▲ Written corpus of modern Korean (About 71.58 million, or 58.63 million)[††]	Information processing of the Korean language.
Korean Informatics Base Ⅱ(KAIST)	1998	▲ Written corpus of modern Korean (About 10 million in the raw corpus, about 200,000 morphological annotations)	Information processing of the Korean language, developing a balanced corpus.

[*] They are composed of spoken language and quasi-spoken language (drama, script, scenario, etc.) recorded in actually used speech. Conversation (26%), lectures (24%), counseling (14%), dramas and scripts (13%), DJ broadcasting (13%), discussions (8%), meetings (2%)

[†] Refer to Seo Sang Kyu and Han Young-gyun (1999: 259) for the concrete contents of this corpus.

[‡] This semantic annotation means the distinctive polysemy of each word. Refer to Seo Sang Kyu (2001).

[§] Referred to Kim Hung-kyu, et al. (1998: 86) and to Seo Sang Kyu and Han Young-gyun (1999: 264).

** Scenarios, dramas, the scripts of drama broadcasts, etc.

†† The number in Seo Sang Kyu and Han Young-gyun (1999: 263) was based on the original report of KAIST and the number in Kim Hung-gyu, et al. (1998: 75) on the number revealed in the later process of the 21st Century Sejong Project. There is a considerable gap in the quantity of data.

1. The Present Situation and Characteristics of Major Korean Language Corpora

The Korean language major corpora since 1987 are as in Table 1 below.[4]

Table 2 is a survey of the 21st Century Sejong Corpora.

Table 2. The Present Situation of the 21st Century Sejong Corpora as of the End of 2004

Corpus (Constitution)	Period of Construction	Scope (Unit: Phrase) & Objects of Collection	Purpose
The 21st Century Sejong Corpora *	1998 - the present	**<Basic Section>** ▲ Written corpus of modern Korean (about 84.58 million) (Raw corpus about 63 million, morphological annotations about 12 million, semantic annotations about 8 million, syntactic analyses about 360,000, etc.) ▲ Corpus of oral traditions (about 2.36 million) **<Special Section>** ▲ Corpus of spoken transcription (Raw corpus about 3.37 million, morphological annotations about 690,000) ▲ Korean-English parallel corpus (Raw corpus about 4.07 million, morphological annotations about 710,000) ▲ Korean-Japanese parallel corpus (Raw corpus about 830,000, morphological annotations about 170,000) ▲ North Korea's corpus of overseas Korean language (Raw corpus about 8.27 million, morphological annotations about 1.19 million) ▲ Corpus of historical data (Raw corpus about 4.78 million, annotations about 580,000) ▲ Korean-Chinese-French-Russian parallel corpus (Raw corpus about 150,000) ▲ Corpus of technical terms (Raw corpus about 1 million)	Development of basic language corpus (studies on the Korean language, information processing of the Korean language), construction of a unified national corpus

* Refer to Bak Yeong-sun (2004: 32). But be careful about the erroneous calculation in the table of the report. The number of the special sections was reconstructed on the basis of the report.

[4] Table 1 was made on the basis of Seo Sang Kyu (2002a: 276) and revised according to various new facts from related reports, etc.

In the above tables, we can understand some of the characteristics of the Korean language corpora that have been constructed up to now.

Together with the process of informatization of the Korean language, the purposes of constructing the corpora have also been diversified. The primary purpose of constructing the early corpora of the Korean language, e.g. the Yonsei Korean Corpora and the Corpora of the National Institute of the Korean Language, etc. focused on the lexicography of the Korean language, and has been expanded naturally to become a study on Korean language informatics and then to Korean language processing. In the process of studying the lexicography of the Korean language, investigation into word frequency, selection of headwords, analysis of meaning and usage of words, and the search for examples to be contained in the lexicon are all based in the corpora.

Through the 21st Century Sejong Project, unified corpora became possible and the types of corpora became diversified. Through the 21st Century Sejong Project that began in 1998, previous corpora became unified and the development of various corpora has been attempted. This unifying project means to construct enormous corpora according to a unified type and system. Almost simultaneously, the Yonsei learner's (error) corpus of the Korean language began to be developed,[5] and the efficiency of the corpus began to be boosted in Korean language teaching and in the field of Korean language education.

Most of the corpora constructed up to now are synchronic corpora centering on the modern Korean language. On the contrary, the corpora of the National Institute of the Korean Language and the 21st Century Sejong Project contain diachronic corpora that contain a wide range of data throughout the times.

Recently, spoken corpora have been actively constructed and have begun to be studied. It is absolutely necessary to construct spoken corpora for balanced studies on the Korean language. However, written corpora were taken as the primary focus of collection. If there were any spoken corpora, most of them were quasi-spoken ones practically realized in written corpora. Also, for the purpose of balancing all the corpora, only a portion of the spoken corpora (approximately around 10%) were collected in general. For example, the Yonsei Korean Corpora 4 that began to construct spoken corpora relatively earlier was com-

[5] This was the corpus I have collected with the help of teachers of various institutions and individual researchers centering on the Center for Linguistic Informatics Development of Yonsei University since 1999. Refer to Seo Sang Kyu (2003: 227-231) for further information.

posed of almost pure spoken corpora (conversation, lectures, counseling, discussion, meetings, DJ broadcasting, etc.), but it is difficult to say whether all of them are pure spoken data because a portion of quasi-spoken corpora (dramas and scripts) were included in part.

There are largely two ways of materializing spoken language data according to purpose. One of them is to construct speech corpora on the basis of speech processing technology and the other is to construct spoken corpora on the basis of figuring out linguistic characteristics of spoken language.[6] The corpus of the spoken transcription of the 21st Century Sejong Project can be a typical example of the latter. We can also refer to Yi Yong-ju (2002) and Sin Ji-yeong (2004) for research and results relating to the former.

How to define spoken corpora and which material to contain can be changed according to the researcher's purpose. The 21st Century Sejong Project defines the corpora of spoken transcription as follows: "The corpora of large-scale recorded speech, either the data of spontaneous or controlled discourse, the corpora of strengthened orthographic or phonological/ phonetic transcription, and the corpora that can be processed mechanically by means of putting in writing according to a standardized recording system."[7] Up to now, the corpora of spoken transcription of the 21st Century Sejong Project are the only full-fledged spoken corpora that satisfy this definition that have been constructed consistently.

At the early stage of the 21st Century Sejong Corpora in 1998-1999, spoken and quasi-spoken corpora consisting of about one million phrases were constructed but quasi-spoken data were given a great deal of weight. In this respect, daily spoken data of the Korean language began to be earnestly constructed in 2000 when the task constructing the corpora of spoken transcription was carried out by the team for constructing special data of the 21st Century Sejong Project.

These corpora of spoken transcriptions are fewer in quantity than written corpora constructed simultaneously. The main reasons fort his are the time and cost necessary for collecting data (tape and video recording), and transcribing it, etc.[8]

In the 21st Century Sejong Project, at first a balanced corpus unifying written and spoken languages was planned. But because of many differences of written characters and the degree of difficulty in the process of construction between

[6] Refer to Gu Hyeon-jeong and Jeon Yeong-ok (2002: 20-23) and Kim Hyeong-jeong (2004) for the characteristics and differences of these two corpora.

[7] Refer to Kim Hyeong-jeong (2004: 166) for this definition.

Table 3. Comparison of the Present Situation of Constructing Written/ Spoken Corpora of the 21^{st} Century Sejong Project as of the End of 2004 (Unit: Phrase)

	Written Language	**Spoken Language**	**Ratio**
Raw Corpora	About 63 million	About 3,37 million	5.3%
Corpora of Morphological Annotation	About 12 million	About 690,000	5.8%

these two types of linguistic data, it is now thought that the written and spoken formats need their own respective balance. Afterward, in order to reveal all the aspects of our daily linguistic lives, it will be necessary to research more earnestly into the type of utterances of spoken language and the actual ratio to vitalize the construction of consequent corpora.

2. The Present Situation of Annotation of Korean Corpora

As shown in the Tables 1 and 2, we can see that the annotation of Korean corpora has been vitalized considerably. The morphological annotation analyzing the phrases of raw corpora according to grammatical structure has accumulated in quantity to more than 10 million phrases. These data contribute to extracting information on morphology, collocation, word frequency, and are also useful in searching for usage. The morphological annotated corpora of the 21st Century Sejong Project have been constructed in the whole fields of written language, spoken language, including North Korean and overseas historical data, and are expected to play an important role in the future.

However, many problems still remain. One is that the annotation system developed up to now or the analytical results of annotating implements are beyond the strata and knowledge of Korean linguistic analysis in many cases. Another is that the annotation system cannot reflect the hierarchical structure of language. In particular, when annotating an enormous corpus of more than 10 million phrases, we cannot help but depend on programmed automatic annota-

[8] Kim Hyeong-jeong (2004: 170-171) said that "However, the more basic reason for the quantitative problem is that the process constructing the spoken corpora can be a difficult task that requires a lot of time and effort. On the basis of raw corpora in 2002, it took about 35 hours for one-hour of recorded data (about 7,443 phrases) to be put into writing as textual data involving 'recording, primary and secondary transcriptions, markup, and header tasks'"

Table 4. An Example of the Structural Annotated Corpus of the 21st Century Sejong Project

```
; 일할 사람이 없다:2
(S   (NP_SBJ (VP_MOD 일/NNG + 하/XSV + ㄹ/ETM)
       [J8](NP_SBJ 사람/NNG + 이/JKS))
     (VP 없/VA + 다/EF + :/SP + 2/SN))
```

tion. Meanwhile, it can be indicated that modifications to basic errors in the program itself have not been sufficiently carried out.

The syntactic annotation in the annotated corpora of the 21st Century Sejong Project follows the stage of establishment of its basic methodology and the development of tools.

Recently, the annotation of the corpus has tried to add more precise information and semantic annotations to distinguish homonyms in the 21st Century Sejong Corpora.

It has also become known that the National Institute for the Korean Language is planning to construct a "lexical annotated corpus" of 3 million phrases for the purpose of investigating modern Korean lexicology. These two corpora contain annotations of information on distinguishing homonyms in common on the basis of *Pyojun gugeodaesajeon* (Standard Korean Dictionary, 1999). Accordingly, it will become possible to distinguish homonyms, which has been a critical point in investigating Korean word frequency based on enor-

Table 5. An Example of the Semantic Annotated Corpus of the 21st Century Sejong Project

```
4BS_0140002750 일부 일부___02/NNG
4BS_0140002760 대기업들은 대/XPN + 기업___01/NNG + 들/XSN + 은/JX
4BS_0140002770 기술 기술___01/NNG
4BS_0140002780 개발보다는 개발/NNG + 보다/JKB + 는/JX
4BS_0140002790 손쉬운 손쉽/VA + ㄴ/ETM
4BS_0140002800 외국 외국___02/NNG
4BS_0140002810 기술 기술___01/NNG
4BS_0140002820 도입에 도입/NNG + 에/JKB
4BS_0140002830 급급하고 급급/XR + 하/XSA + 고/EC
4BS_0140002840 있다고 있/VV + 다고/EC
4BS_0140002850 분석했다. 분석___02/NNG + 하/XSV + 았/EP + 다/EF + ./SF
```

Table 6. An Example of the Semantic Annotated Corpus of the Korean Language

```
60032704 아줌마 1 아줌마 nng -
60032705 하나가 1 하나 nr 2-
60032705 하나가 2 가 jks
60032706 반찬들을 1 반찬 nng -
60032706 반찬들을 2 들 xsn 5-①
60032706 반찬들을 3 을 jko
60032707 들고 1 들 vv 2-①
60032707 들고 2 고 ec
60032708 나와 1 나오 vv - I ①㉠
60032708 나와 2 아# ec
60032709 아무렇게나 1 아무렇 vj -①
60032709 아무렇게나 2 게나 ec
60032710 놓고 1 놓 vv 1- I ①
60032710 놓고 2 고 ec
60032711 돌아가 1 돌아가 vv - II②
60032711 돌아가 2 아# ec
60032712 버린다 1 버리 vx 2-
60032712 버린다 2 ㄴ다 ef
60032713 . 1 . sp
```

mous corpora.

The following is a part of semantic annotated corpora of the Korean language (1 million phrases) that I have established since 1999 and it suggests another method of semantic annotation. These corpora minutely distinguished and annotated the polysemy of 10 thousand basic words, not to mention distinguishing homonyms. This dictionary was established on the basis of the semantic annotation of *Yonsei hangugeo sajeon* (Yonsei Korean Dictionary, 1998).[9]

3. The Present Situation of Balanced Corpora of the Korean Language

In Table 1, we can see that we have made an effort from the early stage of constructing the corpus to develop a balanced corpus of the Korean language.

[9] Refer to Seo Sang Kyu (2001b) for the concrete contents of these corpora.

The major and unique balanced corpora with more than 1 million phrases that have been widely known up to now[10] are as follows: YSC 1 (1990), YSC 2

▼ The Structural Analysis of Korean Balanced Corpora (1)

	YSC 1	YSC 2
Year	1990	1990-1991
Scale	3 million (target)/2,881,175 phrases	1 million/1,107,363 phrases
Institution	Center for Linguistic Informatics Development by Yonsei University	Center for Linguistic Informatics Development by Yonsei University
Purpose	Lexical research for selecting headwords of dictionary	Understanding of statistical characteristics of Korean lexicon
Origin	Published prints (1975-1985)	Published prints (1980-1989)
Classification	Unique classification	Dewey decimal system
Construction	Newspaper(33%), magazines(20%), novels & essays(18%), taste & culture(10%), memoirs, biography, nonfiction(9%), textbooks(5%), broadcasting scripts(5%)	Series(7.8%), philosophy(9.9%), religion(10.7%), social science(12.8%), language(5.7%), pure science(11%), applied science(11.7%), arts(8.1%), literature(11.2%), history(11.3%)
Sample	5,000 phrases and below per text (minimum 120 phrases-maximum about 37,000 phrases) /749 samples	About the first 20 pages of text (minimum 1,040 phrases -maximum 8,110 phrases)/ 234 samples
Method of Sampling	· Constructed on the basis of the results of the quantity of reading according to the types of reading materials by a questionnaire on “Reading Situation of Modern Koreans.” · Half-reduction sampling per year.	· To include as many words as possible in every subject, selected samples of the books of high frequency of borrowing centering on general books 1987-1988.
Characteristics	· Excluded translated books · Unique markup system (the beginning of a sample, the end, information on bibliography, classification of subjects, area of sampling, etc.)	· Excluded theses for degree, periodicals, reference books · Corpus without newspapers or magazines · Included parts of translated books · Independent markup system

[10] The corpora mentioned here were made mainly by institutions. The contents have become known through the Internet, CD-ROM, reports, thesss, etc. Refer to Seo Sang Kyu and Han Young-gyun (1999: 255-266) for further corpus information.

▼ The Structural Analysis of Korean Balanced Corpora (2)

	Corpora of Dictionary of Korean Word Frequency ***(Gwahakbaekgwasajeonjonghapchuppansa)***	**KOREA-CORPUS OF Korea Univ. 1 (KOREA-1 Corpus)**
Scale	1 million/1,047,376 words	1 million/10,082,000 phrases
Year	1993	1995
Constitution	Mun Yeong-ho, et al.	Institute of Korean Culture, Korea University
Purpose	· Precise analysis of language, scientific application of speech and writing · Machine translation, automatic information processing, computer linguistics, automatic lexicography, etc. · To found a basis for linguistic data essential to modern applied linguistics and information processing	· Studies of linguistics and computer linguistics · Lexicography of the Korean language
Origin	Published prints (1960-1980, materials from immediately before and after the Liberation)	Published prints (1970s-1990s), quasi-spoken/spoken language
Method of Classification	Unique classification	Unique classification
Construction	Literature and arts (44%), socio-politics (25%), press reports (20%), scientific technology (11%)	Spoken/quasi-spoken language (11.7%), newspaperS (20.7%), magazines (9.8%), books and information (33.5%), books and suppositions (21%), others (3.3%)
Scale of Sample		Around 50,000 phrases of a single text, maximum 100,000 of more than two texts
Method of Sampling	Fixed about 180 sections and subjects and genre of each section item by item, classified them into literature and arts, socio-politics, press reports, scientific technology, and applied multi-grade methods of sampling.	Constructed according to the media of text (spoken language, newspapers, magazines, books, etc.) and contents (humanities, arts, society, nature, etc.). Referred home and foreign corpora like Brown, COBUILD, and Yonsei Korean Corpora.
Characteristics	· Support of the scale and credibility of corpora by means of statistical method · Excluded books written both in Hangeul and in Chinese characters, foreign books, and dictionaries · Included translated novels.	· Included part of the data from 1920-1960 · TEI markup system (mainly header information)

▼ The Structural Analysis of Korean Balanced Corpora (3)

	Base of Korean Language Information II	**21st Century Sejong Balanced Corpora (1998)**
Year	1994-1998	1998 (2000)
Scale	10,000 thousand phrases/real quantity unidentified	10,000 thousand /10,090 thousand
Institution	KAIST	Ministry of Culture and Tourism /National Institute of the Korean Language
Purpose	Construction of basis of Korean information processing	Study on the Korean language, construction of a basis of language information
Origin	Published prints	Published prints (90%. 1920-1998), spoken/ quasi-spoken language (10%)
Method of Classification	Dewey decimal system and others	Unique classification
Construction	※Unidentified ratio by Dewey decimal system, recorded literature (30%), non-literature (40%), newspapers, news(30%).	· Written language (90%): newspapers (20%), magazines (10%), book-information (35%), book-imaging (20%) · Spoken language (5%) · Quasi-spoken language (5%)
Scale of Sample	Around maximum 10 thousand phrases / 700 samples	Minimum of 419 and maximum of 146,000 phrases (average 19,842 phrases)/509 samples
Method of Sampling	Considered a balance of each field to the utmost.	Constructed according to the ratio of each genre
Characteristics	Use of SGML tags (title, author, year of publication, place of publication, classification (Korean decimal classification), mark of text information)	· Sejong 98 Corpora 80% (8 million)/Corpora of the National Institute for the Korean Language 15% (1.5 million)/ KAIST Corpora 5% (500,000) · Included translated books

(1991), the corpora of *Joseoneobindosusajeon* (Dictionary of Korean Word Frequency, 1993), the KOREA-CORPUS of Korea University 1 (1995), the Base of Korean Language Informatics II (1998), the 21st Century Sejong Balanced Corpora (1998), the Corpora of Korean Language Teaching (1998), etc.[11]

The structural characteristics of these corpora are as follows:[12]

As shown above, a variety of attempts using diverse conditions such as purpose and scale, method of construction, characters of original data and method of classification, scale of samples, method of sampling, etc. were made by YSC 1 in 1990 and were earnestly continued. But no advanced discussion has been stimulated since the 21st Century Sejong Balanced Corpora in 1998[13].

4. The Present Situation of Constructing the Database of Image and Speech

One of the most important results of the informatization of the Korean language in 2000 was to explore and input various data and to construct digital image scanning through Hangeul yetmunheon gichojaryojosa (A Research on Basic

Year	Quantity (Kind)	Input Pages	Contents (Unit: Kind)
2001	183	15,700	Hangeul old documents (100), Korean traditional novels (55), arts, sciences, education (12), data of life and arts (16), etc.
2002	280	15,734	Hangeul old documents (85), journals (192), documents (1), block books (2)
2003	1,550	20,104	Literature (300), data of life and arts (50), data of mechanization and informatization of Hangeul (1,000), data of North Korea and overseas (200)
2004	187	30,517	Old literature and old books
Total	2,200	82,055	

[11] These corpora are excluded here because they were constructed for Korean language teaching and learning as a foreign language. Refer to Seo Sang Kyu, Yu Hyeon-gyeong and Nam Yunjin (2002: 134).

[12] These tables are quoted from Seo Sang Kyu (2002b, 2002d: 122-124). The former (written in the scale of the corpora) shows the target quantity when constructed and the latter the actual constructed quantity.

[13] Refer to Seo Sang Kyu (2002d) for various problems relating to Korean balanced corpora.

Data of Hangeul Old Literature) by Professor Hong Yun-pyo, chief researcher, as a part of constructing the Digital Hangeul Museum.

This project was to construct a database of the image data of “important data on Hangeul literature dated from the creation of Hunmin jeongeum (Hangeul) to the early part of the 20th century.” The image data were constructed as a database primarily on the basis of image scanning and, if not possible, as slides or pictures taken by digital camera. The results are as follows[14]:

This project has three characteristics. Firstly, the data are very widely composed of “important data on Hangeul literature dated from the beginning of the creation of *Hunmin jeongeum* (Hangeul) to the early part of the 20th century.” Secondly, while the existing corpora focused on securing a text database through inputting ordinary documents, this project tried to construct the data as types of image data. This enabled researchers to overcome the limits to depending on black and white photographic editions. It could solve the problem at a stroke where it had been very difficult or almost impossible for users to verify directly the original corpora among the corpora already known up ‘til then. Thirdly, this database itself has the intention of collecting and recording *hangeul* culture since the creation of *Hunmin jeongeum*.

Together with this project, another important project in the field of informatization of the Korean language is the collection of speech data “to investigate the regional distribution of the Korean language” by the National Institute of the Korean Language. This project aims at “investigating and transcribing dialects unique to major locations according to each region for ten years after 2004, constructing a web database, preserving and making use of it” (http://www.Korean.go.kr/000_new/50_saup_data.htm). In 2004, basic investigation, such as the method and items of investigation, and since 2005, questionnaires and regional investigation were initiated.

Also, another case is the project to digitalize speech data of the Seoul dialect. The National Institute of the Korean Language has been recording much of data such as spoken data of recitations, interviews, lectures and free talking speech according to sex, age, and education level of the speakers of the Seoul dialect since 2002. In 2003, “the corpora of spoken data of recitations of Seoul dialect” containing recited speech

[14] This table was made on the basis of reports presented over four years. Refer to pages 143-159 of the concerned report for lists of concrete data for the year 2001, pages 233-239 for the year 2002, pages 179-186 for the year 2003, and pages 31-38 for the year 2004.

from 120 persons x 930 sentences were published (on 5 volumes of DVD). In 2005, the database of intonation of standard Korean will be constructed.

The Present Situation of Research Using the Corpora of the Korean Language[15]

Among the theses in academic journals contained in *Gugeo yeongu nonjeo mongnok* (The List of Theses on Studies of the Korean Language) 3 (1971-2003) distributed recently by the National Institute of the Korean Language, the distribution of such words as "*malmungchi (malmodum)*, corpus, and informatization" contained in the titles can be analyzed by year as follows[16]:

	Corpus	Informatization	Computerization	Computation by Computer	Total
1971-79	0	0	0	0	0
1980-89	1	2	3	4	10
1990-99	33	16	11	25	85
2000-03	77	31	4	8	120
Total	111	49	18	37	

As shown in the above, the period in which technical terms like *malmungchi* as related to informatization began to appear in the titles of theses was during the 1980s, which coincided with the history of constructing corpora[17].

The studies from the end of the 1980s when the corpora of the Korean language began to be constructed can be subdivided into three short periods. According to Seo Sang Kyu (2001a), they are the period of establishing the basis of informatization of the Korean language (from the end of the 1980s-the begin-

[15] In this respect, Seo Sang Kyu (2001a, 2002c) rearranged the research of the former part of the 2000s. Therefore, this chapter will complement and describe new results after that time on the basis of Seo Sang Kyu (2002c).

[16] This is nothing more than a rough sketch of the developing stage of informatization in our academic world. When analyzing the contents of theses practically, a lot of theses have been treating corpora since 1990.

[17] Noma Hideki (2005: 17) indicates that one of the characteristics of studies on the Korean language in Korea was considerable development of corpus linguistics and lively studies on corpora since the 1990s, especially the latter part of the 1990s.

ning of the 1990s), the establishment of lexicography and corpus linguistics in the 1990s, and the full-fledged establishment of Korean language informatics after the middle of the 1990s.

In the middle of the 1980s, as shown in the previous section, not only the construction of corpora as basic data for lexicography but also searches for methods for Korean data processing by computer had been vigorously carried out. *Hangugeo jeonsanhak* (A Study on Korean Language Computerization) published by the Computerization Society of the Korean Language in 1991 dealt with Hangeul old letters, data of dialects, and the processing and concordance of literary data. *Gugeowa Computer* (The Korean Language and Computer) (Kim Byong-Sun. 1992. Hansil) was the first introductory book using technical terms such as *Gugeo jeonsanhak* (Korean Language Computerization) and *Gugeo jeongbohak* (Korean Language Informatics) at that time.

At the turn of the 1990s, the research area of the Korean language using corpora was enlarged and research methods of lexicography and corpus linguistics were introduced. As mentioned many times, the arousing of interest due to *Malmungchieoneohak* (Corpus Linguistics) was related to the projects of lexicography vigorously carried out by various institutions during the 1990s. In particular, the following research activities compiling dictionaries became motives: *Yonsei hangugeo sajeon* (Yonsei Dictionary of the Korean Language) by the Committee for Lexicography of the Korean Language, Yonsei University (1987-1998), *Pyojun gugeodaesajeon* (Standard Korean Dictionary) by the National Institute of the Korean Language (1992-1999), and another dictionary by the Institute for Korean Culture at Korea University (1995-).

The following theses and books have exerted considerable influence on the introduction and settlement of the concepts of corpus and methodology of corpus linguistics in Korea.

* Yi Sang-seop (1988). "Mungch eoneohakeuro bon sajeon pyeonchanui silje munje" (Real Problems of Lexicography from the Viewpoint of Corpus Linguistics). *Sajeonpyeonchanhak yeongu Vol. II* (A Study on Lexicography Vol. II). The Committee for Lexicography of the Korean Language, Yonsei University.
* Yi Sang-seop (1990b). "Mungchieoneohak: sajeonpyeonchanui pilsujeok gaehyeom" (Corpus Linguistics: An Essential Concept of Lexicography). *Sajeonpyeonchanhak yeongu Vol. III* (A Study on Lexicography Vol. III). The Committee for Lexicography of the Korean Language, Yonsei

University.

* Yi Sang-seop (1995a). “Malmungchi geu gaenyeomgwa guhyeon” (Corpus: Its Concept and Realization). *Sajeonpyeonchanhak yeongu Vol. V & VI* (Combined Edition) (A Study on Lexicography Vol. V & VI Combined Edition). The Committee for Lexicography of the Korean Language, Yonsei University.
* Yi Sang-seop (1995b). “Mungchi eoneohakui gibon jeonje” (A Basic Prerequisite to Corpus Linguistics). *Sajeonpyeonchanhak yeongu Vol. V & VI* (Combined Edition) (A Study on Lexicography Vol. V & VI Combined Edition). The Committee for Lexicography of the Korean Language, Yonsei University.

I shall make a survey of research carried out on the Korean language using corpora centering on various characteristic facts.

1. Publication of Specialized Academic Journals and Introductions

It was at the turn of the 1990s that corpora began to be used in studies on the Korean language. The most remarkable characteristics were the publication of academic journals relating to the informatization of the Korean language.

* *Sajeonpyeonchanhak yeongu Vols. 1-11* (A Study on Compiling Dictionary Vols. 1-11). 1998-2002. The Committee for Lexicography of the Korean Language, Yonsei University, Institute for Linguistic Informatics.
* Seo Sang Kyu, ed. *Eoneo jeongboui tamgu Vol. 1* (An Inquiry into Linguistic Informatics Vol. 1). 1999. The Institute for Linguistic Informatics Development, Yonsei University.
* *Eoneojeongbowa sajeonpyeonchan Vols. 12-13.* (Linguistic Information and Lexicography Vols. 12-13). 2003. The Institute for Linguistic Informatics, Yonsei University[18].
* *Hangugeo jeonsanhak Vols. 1-2* (A Study on Computerization of the Korean Language Vols. 1-2). 1991 and 1998. The Society for

[18] “Studies on Lexicography” and “An Inquiry into Linguistic Information” were united into “Linguistic Information and Lexicography” in 2003. Vols. 12-13 were published in a combined edition.

Computerization of the Korean Language.
* *Eoneojeongbo Vols. 1-4* (Linguistic Informatics Vols. 1-4). 1997-2000. Institute for Linguistic Informatics, Korea University.
* Bak Yi-jeong. 2001-2002. *Gyeryangeoneohak Vols. 1-2* (Quantitative Linguistics Vols. 1-2).
* *Hanguk sajeonhak issues 1-5* (Korean Lexicography Issues 1-5). 2003-2005. The Society for Korean Lexicography.

Introductory books published after the 1990s are as follows:

* Seo Sang Kyu and Han Young-gyun. 1999. *Gugeojeongbohak ipmun* (An Introduction to Korean Informatics). Taehaksa.
* Yu Seok-hun, tr. 1999. *Eoneowa Keompyuteo* (Language and the Computer). Korea University Press.
* Han Young-gyun. 1999. *Jeonjamalmungchireul iyonghan sajeon pyeonchanron* (A Theory of Lexicography Using Computerized Corpora). A Report from the Ministry of Culture and Tourism.
* Bae Hi-suk, tr. 2000. *Tonggyeeoneohak* (Statistical Linguistics by Charles Müller). Taehaksa.
* Kang Beom-mo. 2003. *Eoneo, Keompyuteo, Corpus eoneohak: Keompyuteo reul iyonghan gugeo bunseokui gichowa iron* (Language, the Computer, and Corpus Linguistics: The Elements and Theory of Analysis of the Korean Language Using Computers). Korea University Press.

Books published in North Korea are as follows:
* Mun Yeong-ho. 1990. *Gesangieoneohakgaeron* (An Introduction to Computer Linguistics). Sahoegwahakchulpansa.
* Gwon Jong-seong. 1994. *Joseoneo jeongbo cheori* (Korean Language Processing). Gwahakbaegwasajeonjonghapchulpansa.

2. Studies on Corpus Construction and Processing

a. The layout and construction of corpus

The use of a corpus in studies of the Korean language has been vigorously discussed. In particular, the discussion was centered on the scale of the corpus, the practical method of collecting data and its construction, and the method of creating the necessary balanced for it to be considered in the construction of larger corpora.

Corpora and Database Construction in General

* Hong Yun-pyo. 1990. “The Method of Korean Data Processing Using a Computer,” A Special Lecture at the Society of Korean Language Research.
* Han Young-gyun. 1993. *Keompyuteo reul iyonghan gugeo jaryo bunseoke daehan gichojeok yeongu* (A Basic Study on Analyzing Korean Data Using a Computer). The National Institute of the Korean Language.
* Han Young-gyun and Yu Dong-seok. 1993. *Keompyuteo e uihan gugeo-jaryo cheorireul wihan gichojeok yeongu* (A Basic Study on Korean Data Processing by Computer). The National Institute of the Korean Language.
* Jeong Gwang, Yi Gi-yong, Kim Heung-gyu, Im Hae-chang, and Kang Beom-mo. 1995. *Hangugeo Deitabeiseu ui seolgye mit eungyongeul wihan gicho yeongu* (A Study on Layout and Application of Databases of the Korean Language). Mineumsa.
* Jeong Gwang. 1996. *Gugeo eohwi Deitabeiseu guchuke daehan yeongu* (A Study on Constructing the Database of Korean Vocabulary). The National Institute of the Korean Language.
* Kang Beom-mo, ed. tr. 1997. *Jeonja text buhohwa gaeseol: TEI Light* (Construction of Encoding Electronic Text: TEI Light). The Institute of Korean Culture, Korea University.

Methodology of Constructing Corpora and Types of Text

* Jeong Chan-seop, Yi Sang-seop, Nam Gi-sim, Han Jong-cheol, and Choe Yeong-ju. 1990. “Wurimal nanmal bindo josa pyobonui seonjeong gijun” (The Criteria of Sampling of Korean Word Frequency). *Sajeonpyeonchanhak yeongu Vol. III* (A Study on Lexicography Vol. III). The Committee for Lexicography of the Korean Language, Yonsei University.
* Yi Sang-seop. 1990. “Nanmal bindoreul chujeonghagi wihan malmungchi jaryo sujipui silje” (Collecting Corpus Data for Estimating Word Frequency). *Sajeonpyeonchanhak yeongu Vol. III* (A Study on Lexicography Vol. III). The Committee for Lexicography of the Korean Language, Yonsei University.
* Han Young-gyun, Nam Yun-jin, Ryu Bin, and Kim Hyeon-jeong. 1993.

Gyunhyeong Corpus ui guchukeul wihan pyojun bangbeopnon mit sibeom Package guchuk, yeongubogoseo (A Research Report on Standard Methodology for Constructing a Balanced Corpus and the Construction of a Model Package). KAIST.

* Jeong Yeong-mi. 1995. “Gugeo eohwiui tonggyejeok teukseonggwa iui eungyong” (Statistical Characteristics of Korean Vocabulary and their Application). *Sajeonpyeonchanhakyeongu Vol. V & VI (Combined Edition)* (A Study on Lexicography Vol. V & VI Combined Edition). The Committee for Lexicography of the Korean Language, Yonsei University.

* Kim Heung-gyu and Kang Beom-mo. 1996. “Goryeodaehakgyo hangugeo malmodum 1: seolgye mit guseong” (The KOREA-CORPUS of Korea University 1: Layout and Construction). The National Institute of the Korean Language.

* Kim Hung-gyu, Seong Gwang-su and Hong Jong-seon. 1998. “Daegyumo hangugeo Deitabeiseu ui dawonjeok tonggye bunseok yeongu” (A Study on Pluralistic and Statistical Analysis of a Variety of Korean Databases). *Hangugeo jeonsanhak Vol. II* (A Study on Computerization of the Korean Language Vol.II). Taehaksa.

* Kang Beom-mo. 1999. *Hangugeoui text genre wa eoneo teukseong* (The Text Genre of the Korean Language and Linguistic Characteristics). Korea University Press.

* Nam Yun-jin. 1999. “Gyunhyeong malmungchi guchukeul wihan silheom-jeok yeongu (1)” (A Tentative Study on Constructing a Balanced Corpus 1). *Eoneojeongboui tamgu 1* (Research on Linguistic Informatics 1). The Center for Linguistic Informatics Development, Yonsei University.

* Kim Han-saem and Seo Sang Kyu. 1999. “Malmungchiui guchukgwa hwalyong” (The Construction and Use of a Corpus). *Eoneojeongboui tamgu 1* (Research on Linguistic Informatics 1). The Center for Linguistic Informatics Development, Yonsei University.

* Kang Beom-mo, Kim Hung-gyu and Heo Myeong-hoe. 2000. *Hangugeoui text genre, munche, yuhyeong: Keompyuteo wa tonggyejeok gibeopui iyong* (Korean Text Genre, Literary Style and Types: Computers and the Use of Statistical Technique). Taehaksa.

* Kang Beom-mo, Yi Yu-seon and Cha Jae-eun. 2002. *Dagugeo eohwi Database guchuk bangbeopnon yeongu mit mohyeong gaebal (1)* (A Study on the Methodology Constructing a Database of Multilingual Vocabulary and Model Development 1). The Institute for Korean Culture, Korea

University.

* Gwak Yong-jin. 2003. “Hapmokjeokjeok malmungch jadong guchuk” (Automatic Construction of a Purpose-Built Corpus). *Eoneojeongbowa sajeonpyeonchan Vols. XII-XIII* (Linguistic Informatics and Lexicography Vols. XII-XIII). The Center for Linguistic Informatics, Yonsei University.
* Kang Beom-mo, Cha Jae-eun and Yi Yu-seon. 2005. *Dagugeo eohwi Database guchuk bangbeopnon yeongu mit mohyeong gaebal (2)* (A Study on the Methodology Constructing a Database of Multilingual Vocabulary and Model Development 2). The Institute of Korean Culture, Korea University.

b. Annotation and processing of corpus

The methodology of annotation to extract varied and precise information from the constructed corpora and practical studies has been vigorously carried out.

* Yim Hong-bin. 1998. “Hangugeo jeongbocheorireul wihan eojeol bunseok pyojiui pyojunhwa yeongu” (A Study on the Standardization of Marking Phrasal Analysis for Information Processing of the Korean Language). *21segi sejonggyehoek gugeo gichojaryo guchuk* (The Construction of Basic Data of the Korean Language in the 21st Century Sejong Project) The Final Report. The Ministry of Culture and Tourism.
* Han Young-gyun. 1998. “Muneo Corpus ui hyeongtae jeongbo juseokeseo seonhyeoldoeeoyahal myeot munje” Some Problems to be Solved in the Annotation of Morphological Information of Written Corpus). *Hangugeo jeonsanhak* (Computerization of the Korean Language) Vol. II. Taehaksa.
* Han Young-gyun. 1999. “Hangugeo muneo Text ui hyeongtae, tongsajeok juseok sangui gibbon munje” (Morphology of Written Text of the Korean Language and Basic Problems of Syntactic Annotation). *Eoneojeongboui tamgu 1* (Research on Linguistic Informatics 1). The Center for Linguistic Informatics Development, Yonsei University.
* Seo Sang Kyu. 2001. “Malmungchiui juseokgwa hangugeo gibbon eohwi uimi bindo sajeon” (Annotation of the Corpus and the Sense Frequency Dictionary of Korean Basic Vocabulary). *Gyeryangeoneohak Vol. I* (Quantitative Linguistics Vol. I). Bak Yi-jeong.
* Sang-kyu Seo, Jinung Kim & Hansaem Kim. 2001. Yonsei Sense Frequency Dictionary Based on Sense-Tagged Corpus.

Sajeonpyeonchanhakyeongu Vol. XI No. 2 (A Study on Lexicography Vol. XI No. 2). The Center for Linguistic Informatics, Yonsei University.

* Seo Sang Kyu, Seo Eun-a and Yi Byeong-gyu. 2002. “Tongsa jeongbo juseok malmungchiui guchuk” (Construction of the Annotated Corpus of Syntactic Information). *Hangugeo gueo yeongu (1)-gueo jeonsa malmungchiwa geu hwalyong* (A Study on Spoken Korean 1-Corpus of Written Transcription and Its Use). Hangukmunhwasa.
* Min Gyeong-mo. 2003. “Hyeondaegugeo hyeongtae jeongbo juseok muneo malmungchiui juseok pyogi bangane daehayeo” (A Study on the Marking Method of Written Corpus Annotated with the Morphological Information of Modern Korean). *Eoneojeongbowa sajeonpyeonchan Vols. XII-XIII* (Linguistic Informatics and Lexicography Vols. XII-XIII). The Center for Linguistic Informatics, Yonsei University.
* Kang Beom-mo and Kim Ei-su. 2004. “Sejong gumunbunseok malmungchireul wihan gumun bunseok bangbeop” (A Syntactic Analysis Method for the Corpus of Syntactic Analysis of the Sejong Project). *Corpus wa eohwi Database* (Corpus and Lexical Database), edited by Kang Beom-mo, Bak Byeong-seon, Yi Bong-won and Jo Jin-hyeon. Seoul: Weolin.
* Kim Jong-bok, et al. 2004. *Hangugeo jeongbohwawa gumunbunseok* (Informatization of the Korean Language and Syntactic Analysis), (Minyeon Series Eomun and Minsok 6). Seoul: Weolin.

c. Studies on basic linguistic informatics based on corpora

The most vigorous studies in basic linguistic informatics using corpora investigate word frequency.

Investigation into Word Frequency Using Corpora

* Mun Yeong-ho, et al. 1993. *Joseoneobindosusajeon* (Dictionary of Korean Word Frequency). Pyeongyang: Gwahakbaekgwasajeonjonghapchulpansa.
* Kim Jeong-su, Kim Hi-rak, Jeong In-sang, Jo Nam-ho & Yi Jun-hi. 1994. *Yet hangeului eumjeol josa yeongu* (Research on the Syllables of Old Hangeul). Seoul: Hangukchulpanyeonguso.
* Kim Heung-gyu & Kang Beom-mo. 1997. *Hangeul sayongbindoui bunseok* (The Analysis of the Frequency of Use of Hangeul). The Institute of Korean Culture, Korea University.
* Kim Heung-gyu, Seong Gwang-su & Hong Jong-seong. 1998. “Daegyu-

mo hangugeo deitabeiseu ui dawonjeok tonggye bunseok yeongu" (A Study on Plural Statistical Analysis of a Large-Scale Korean Database). *Hangugeo jeonsanhak* (Computerization of the Korean Language) Vol. II. Taehaksa.

* Seo Sang Kyu. 1998. "Malmungchi bunseoke gibaneul dun nanmal bindoui josawa geu eungyong, Yonsei malmungchirel jungsimeuro" (Investigation into Word Frequency Based on the Analysis of Corpora and Their Application Centering on the Yonsei Corpus). *Hangeul* (Hangeul) Issue 242.
* Seo Sang Kyu. 1998. "Hyeondae hangugeoui eohwi bindo (sang, ha)" (Word Frequency of Modern Korean Vols. I-II) (Unpublished). The Center for Linguistic Informatics, Yonsei University.
* Jang Seok-bae. 1999. "Malmungchi gyumowa eojeol yuhyeong jeunggaganui sangwanseonge daehan yeongu" (A Study on the Relativity between the Scale of a Corpus and the Increase of Phrasal Types). *Eoneojeongboui tamgu 1* (Research on Linguistic Informatics 1). The Center for Linguistic Informatics Development, Yonsei University.
* Kang Beom-mo & Kim Heung-gyu. 2004. *Hangugeo hyeongtaeso mit eohwi sayong bindoui bunseok 2* (Analysis of Frequency of Use of Korean Morphemes and Lexicon 2). The Institute of Korean Culture, Korea University.
* Kim Han-saem. 2003. *Hanguk hyeondae soseolui eohwi josa yeongu* (An Investigation into Word Frequency in Korean Modern Novels). The National Institute of the Korean Language.
* Jo Nam-ho. 2002. *Hyeondae gujeo sayong bindo josa-hangugeo hakseupyong eohwi seonjeongeul wihan gicho josa* (Investigation into Frequency of Use of Modern Korean-Basic Research on Selecting a Lexicon for Learning the Korean Language). The National Institute of the Korean Language.

d. Studies on Korean grammar based on corpora

Recent studies on Korean grammar based on corpora are divided into two categories. One describes lexical and grammatical characteristics wholly shown in the corpora through quantitative analysis, and the other uses the practical usage of corpora as supplementary demonstration data according to the deductive hypothesis of a researcher in terms of lexical and grammatical description.

At the beginning of the 1990s, a certain amount of usage was secured and the analyses and descriptions of lexical grammar were tried on the basis of the results of observation. Most of the articles contained in the following books have such trends in common:

"GentaiChosengo no TeidoHukusi ni chuite--Hukusi 아주 no *Teido* to *Yotai* no Imi o Chusin ni" (On Gradable Adverbs of Modern Korean Centering on the Senses of "Grade" and "Manner" of the Adverb *aju*) by Seo Sang Kyu (1991),[19] *Gugeo josaui yongbeop* (Usage of Korean Postposition) edited by Nam Gi-sim (1993), *Gugeo yeongyeoleomiui sseuim* (Usage of Korean Linking Suffixes) edited by Nam Gi-sim (1994) and *Gugeomunbeopui tamgu III* (Inquiry of Korean Grammar III) edited by Nam Gi-sim (1996).

Afterward, as shown below, the results of analyzing corpora came from most of these theses.

Studies on the Korean Language Using the Corpora

* Seo Sang Kyu. 1992. *Gendai kankokugo no giseigitaigo no yogenkechugobunseki oyobi yorei deta besu.*

* A Report of Joint Research Presented to the Japanese Ministry of Culture in 1991. Tokyo University of Foreign Studies.

* Seo Sang Kyu. 1993. "Hyeondae hangugeoui sinyungmalui munbeopjeok gineunge daehan yeongu-purimalgwaui gyeolhapgwangyerel jungsimeuro" (A Study on the Grammatical Function of Modern Korean Mimesis and Onomatopoeia Centering on Connection Related to the Predicate). *Joseon Hakbo Vol. 149* (Korea Academic Journal Vol. 149).

* Seo Sang Kyu. 1996. "Umjikssiui sinyungmal chihagi-daneogyeolhabui tonggyebunseok" (On the Usage of Mimesis and Onomatopoeia of Verbs-Statistical Analysis of Word Combination). *Daedongmunhwayeongu Vol 30* (Daedong Culture Research Vol. 30).

[19] In advance of this, "Siganbusaui siganpyosigineunge daehayeo" (On the Function Indicating the Time Sequence of Temporal Adverbs) (Joseon Academic Journal Vol. 133) presented by Seo Sang Kyu (1989) in Japan did not use the word 'corpus' but could be the early fruit of a study analyzed on the basis of usage extracted from mainly literary works. This tendency was spontaneous, apart from the early studies on corpora carried out in Korea at that time. Noma Hideki (1993, 2002: 375) called the research methodology in Japan that shared the linguistic methodology of corpora and the basic attitude in Korea as 'linguistic realism.' Refer to Noma Hideki (2002: 365-385) for his research results in Japan according to this methodology. Recently, Noma Hideki (2005) explains very well the trends and problems of realistic studies using corpora both in Korea and Japan.

* Yim Chil-seong, Mizuno Shunpei, and Kitayama Kazuo. 1997. *Hangugeo gyeryangyeongu* (Quantitative Study on the Korean Language). Jeonnam University Press.

* Nam Yun-jin. 1997. *Hyeondaegugeoui josae daehan gyeryangeoneohakjeok yeongu* (A Quantitative & Linguistic Study on Postpositions of Modern Korean). A Doctoral Dissertation, Seoul National University. (2000. Gugeohakhoe gugeohakchongseo 36. Taehaksa.)

* Kang Hyeon-hwa. 1998. *Gugeoui dongsayeongyeol guseonge daehan yeongu* (A Study on Construction of Korean Verb Connections) (Malmungchigibangugeoyeonguchongseo 2). Hangukmunhwasa.

* Yu Hyeon-gyeong. 1998. *Gugeo hyeongyongsa yeongu* (A Study on Korean Adjectives) (Malmungchigibangugeoyeonguchongseo 3). Hangukmunhwasa.

* Yi Hi-ja and Yi Jong-hi. 1998. *Sajeonsik Text bunseokjeok gugeo josaui yeongu* (A Study on Korean Postpositions by Analyzing Lexical Texts). (Malmungchigibangugeoyeonguchongseo 1). Hangukmunhwasa.

* Yi Hi-ja and Yi Jong-hi. 1999. *Sajeonsik text bunseokjeok gugeo eomiui yeongu* (A Study on Korean Suffixes by Analyzing Lexical Texts). (Malmungchigiban gugeoyeonguchongseo 5). Hangukmunhwasa.

* Kim Han-saem. 1999. *Hyeondae gugeo gwanyongguui gyeryangeoneohakjeok yeongu* (A Quantitative & Linguistic Study on Modern Korean Idioms). An M.A. Dissertation, Yonsei University.

* Jeong Hi-jeong. 2000. *Gugeo myeongsaui yeongu* (A Study on Korean Nouns). (Malmungchigibangugeoyeonguchongseo 6). Hangukmunhwasa.

* Han Song-hwa. 2000. *Gugeo jadongsa yeongu* (A Study on Korean Intransitive Verbs). (Malmungchigibangugeoyeonguchongseo 7). Hangukmunhwasa.

* Kim Jin-hae. 2000. *Yeoneo yeongu* (A Study on Collocations). Hangukmunhwasa.

* Hong Jong-seon, Kang Beom-mo & Choe Ho-cheol. 2001. *Hangugeo yeoneo gwangye yeongu* (A Study on Relations between Korean Collocations). Weolin.

* Yi Sang-eok. 2001. *Gyeryanggujeohakyeongu* (A Study on Quantitative Korean Linguistics). Seoul National University Press.

* Nam Gil-im. 2003. “Malmungchie gibanhan busaui hyeongtae, tongsa jeongbo cheori yeongu” (A Study on the Morphology of Adverbs Based using Corpora and Syntactic Information Processing). *Eoneojeongbowa*

sajeonpyeonchan Vols. XII-XIII (Linguistic Informatics and Lexicography Vols. XII-XIII). The Center for Linguistic Informatics, Yonsei University.

* Nam Gil-im. 2004. *Hyeondae gugeo "ida" gumun yeongu* (A Study on Syntax ida in Modern Korean). (Malmungchigibangugeoyeonguchongseo 12). Hangukmunhwasa.

* Kim Han-saem. 2005. *Gugeo sugeoui gugeojeongbohakjeok yeongu* (A Linguistic and Informational Study on Korean Idioms). A Doctoral Dissertation, Yonsei University.

* Choe Un-ho. 2005. *Hangugeo cheorieseo gumukkeum'eul wihan myeongsaui teukseong yeongu* (A Study on Characteristics of Nouns for "Chunking" in Korean Language Processing). A Doctoral Dissertation, Seoul National University.

One of the recent characteristics of studies on Korean grammar is the deepening of studies into spoken language.

* Jeon Yeong-ok. 1998. *Hangugeo damhwae natanan banbokpyohyeon yeongu: yuhyeong, bunpo, gineung* (A Study on Repeated Expression Shown in the Discourse of the Korean Language: Types, Distribution & Function). A Doctoral Dissertation, Sangmyeong University.

* Sin Ji-yeon. 1998. *Gugeo jisiyongeon yeongu* (A Study on Indicative Verbs / Adjectives of the Korean Language). (Gugeohakheo gugeohakchongsea 28). Seoul: Taehaksa.

* An Ui-jeong. 1998. *Hangugeo ipmalmungchi jeonsa bangbeop yeongu* (A Study on the Method of Transcription of Korean Spoken Corpora). An M.A. Dissertation, Yonsei University.

* Yang Yeong-ha. 2000. *Bangsong sangdam daehwaui gujowa chegye bunseok* (An Analysis of the Structure and System of Conversation during Broadcasting Consultation). An M.A. Dissertatiion, Sangmyeong University.

* Kim Jeong-seon. 2001. *Sanggeorae daehwaui jinhaeng gujowa seoldeuk chaekryak* (The Progressing Structure of Business Talks and Tactics of Persuasion). A Doctoral Dissertation, Hanyang University.

* Seo Sang Kyu & Gu Hyeon-jeong, co-ed. 2002. *Hangugeo gueo yeongu (1)-gueo jeonsa malmungchiwa geu hwalyong* (A Study on Korean Spoken Language 1-The Corpus of Spoken Transcription and Its Use). Hangukmunhwasa.

* Kim Hyeong-jeong. 2002. *Hangugeo ipmal damhwaui gyeolsokseong yeongu* (A Study on Cohesion in Korean Spoken Discourse). An M.A. Dissertation, Yonsei University.

* Gwon Jae-il. 2004. *Gueo hangugeoui uihyangbeop silhyeonbangbeop* (Methods of Realizing Intentions in Spoken Korean Language). Seoul National University Press.

* Seo Sang Kyu & Gu Hyeon-jeong, co-ed. 2005. *Hangugeo gueo yeongu (2)-daehaksaeng daehwa malmungchireul jungsimeuro* (A Study on Korean Spoken Language 2-Centering on the Spoken Corpus of Collegians). Hangukmunhwasa.

e. Applied linguistic study based on corpora

The fields of studies actively applying the basic technology of the corpus are lexicography and studies on the patterns of areas using language, and fields are expanding to studies on language acquisition, linguistic pathology, and language teaching based on the a learner error corpus. In particular, the field of language teaching has been conspicuously developed. An evaluation of the lexicon essential to teaching lexis and grammar, statistical research on the lexicon and sentence patterns, methodology of construction to use the learner error corpus, a study on error analysis, development of a learner dictionary, and the construction of an informatization system of education have been vividly discussed. Studies on the so-called "learner error corpus (or learner corpus)" have been actively carried out.

* Yu Seok-hun. 2001. "Oegujeoroseoui hangugeo hakseupja malmungchi guchukui pilyoseonggwa jaryo bunseok" (The Necessity of Constructing the Learner Error Corpus of the Korean Language as a Foreign Language and Analysis of the Data). *Hangugeo gyoyuk* (Korean Language Teaching) Vol. 12 No. 1.

* Seo Sang Kyu. 2002. "Hangugeo gibbon eohwiwa malmungchi bunseok" (An Analysis of the Basic Vocabulary and Corpus). *21segi hangugeo gyoyukhakui hyeonhwanggwa gwaje* (The Present Situation of 21st Century Korean Language Teaching and Its Task) edited by Bak Yeong-sun. Seoul: Hangukmunhwasa.

* Seo Sang Kyu, Yu Hyeon-gyeong & Nam Yun-jin. 2002. "Hangugeo hakseupja malmungchiwa hangugeo gyoyuk" (The Learner Error Corpus of

the Korean Language and Korean Language Teaching). *Hangugeo gyoyuk* (Korean Language Teaching) Vol. 13 No. 1. Seoul: The International Society of Korean Language Teaching.

* Yu Hyeon-gyeong & Seo Sang Kyu. 2002. “Hangugeo hakseupja malmungchie natanan busa sayonge daehan yeongu” (A Study on the Use of Adverbs Shown in the Learner Error Corpus of the Korean Language). *Ijungeoneohak* (Bilingualism) Vol. 20. Seoul: the Society of Bilingualism.
* Yi Ik-hwan & Seo Sang Kyu. 2002. *Gibon eohwi seonjeong mit sayong siltae josareul wihan gicho yeongu* (A Preliminary Study on Selection of a Basic Lexicon and the Situation for Its Use). The National Institute of the Korean Language.
* Yi Jeong-hi. 2003. *Hangugeo hakseupjaui oryu yeongu* (A Study on the Learner Errors of the Korean Language). Bakijeong Publication Co.
* Seo Sang Kyu. 2003. “Hangugeo hakseupja malmungchi guchukui siljejeok munje” (Practical Problems of Constructing the Learner Error Corpus of the Korean Language). *Hangugeo gyoyukgwa hakseupsajeon* (Korean Language Teaching and the Learner Dictionary) edited by Seo Sang Kyu. Hangukmunhwasa.
* Go Seok-ju, et al. 2004. *Hangugeo hakseupja malmungchiwa oryu bunseok* (The Learner Error Corpus of the Korean Language and Error Analysis). Hangukmunhwasa.

Meanwhile, Yi Mi-hye (2005: 351) surveyed the research trend of the field of Korean language teaching, indicating that it has practical data in common, that is, the layout of a syllabus through the analysis of Korean language corpora in a series of articles studying the syllabus of teaching grammar.

3. Development of the Lists and Books of Usage

Lists

* The List of Writings on Korean Linguistics (Policy on the Korean Language): December, 1991. The National Institute of the Korean Language.
* The List of Writings about Korean Linguistics 1 (1991-2001): December, 2002. The National Institute of the Korean Language.
* The List of Writings about Korean Linguistics 2 (1981-2002): December, 2003. The National Institute of the Korean Language.
* The List of Writings about Korean Linguistics 3 (1971-2004): December,

2004. The National Institute of the Korean Language.

* The Construction of Korean Research Data of North Korea 1 (1946-2000): December, 2004. The National Institute of the Korean Language.

Books of Usage by Jeong Ho-seong

* 2001. *Juyo eohwi yongrye sujip mit jeongni (hyeongyongsa pyeon)* (The Collection and Arrangement of the Major Usage of Vocabulary Items Centering on Adjectives). The National Institute of the Korean Language.

* 2002. *Juyo eohwi yongryejip (dongsa pyeon sang/ ha)* (The Book of Usage of Major Vocabulary Items Centering on Verbs Part I-II). The National Institute of the Korean Language.

* 2003. *Juyo eohwi yongryejip (myeongsa pyeon sang/ jung/ ha)* (The Book of Usage of Major Vocabulary Items Centering on Nouns Part I-III). The National Institute of the Korean Language.

Conclusion

Through this article, we have surveyed the present situation of research on the Korean language using corpora as well as that of informatization of the Korean language from the mid 1980s to the present according to several major points.

Presently, the informatization of Korean language data is advancing toward developing a variety of new data that meet various purposes aside from the simple computerization of written language. And the annotation that supplies corpora with a variety of linguistic annotated informatics is expanding to the stages of syntax and semantics as well as morphological informatics.

Finally, we can positively ascertain that practical and useful methods of research based on corpora and analyses of usage are becoming settled as one of the major methods in researching the Korean language.

References

Bak Yeong-sun. 2004. *21segi sejonggyehoek gugeojeongbo gwanri ssenteo unyeong* (The Management of Administration Center for Korean Language Information according to the 21st Century Sejong Project). The Ministry of Culture and Tourism / The National Institute of the Korean Language.

Hong Yun-pyo. 2002. "Gugeohak yeonguwa jeongbohwa" (Research on Korean Linguistics and Informatization). Pp. 15-53 in *Hangugeowa jeongbohwa* (The Korean Language and Informatization). Seoul: Taehaksa.

______. 2005. "Gugeosa yeongurel wihan jeonjajaryo guchukui hyeonhwangg-wa gwaje" (The Present Situation and Tasks for Constructing Electronic Data for the History of the Korean Language). Pp.50-76 in *Gugeosa yeongu eodiggaji wa itneunga* (Where is Research on the History of the Korean Language?) (The Argument Presented at the Academic Meeting for the History of the Korean Language). The Institute for Korean Studies, Yonsei University.

Kim Byong-sun. 1997. "Hanguk jeonsan eoneohakui hyeonhwanggwa gwaje" (The Present Situation and Tasks for Computerized Linguistics in Korea). *Hangugeomun* (The Korean Language and Literature) Vol. 5. The Academy of Korean Studies.

______. 2000. "Bukhanui eoneojeongbocheori yeongu" (A Study on the Language Processing of North Korea). *Gukjegoryeohakhoe seoul jihoe nonmunjip* (Articles of the Seoul Branch of the International Korea Society) Vol. 2. The International Korea Society.

Kim Heung-gyu, et al. 1998. *21segi sejonggyehoek gugeo gichojaryo guchuk* (The Construction of Korean Basic Data of the 21st Century Sejong Project, A Repot of Academic Services). The Ministry of Culture and Tourism.

Ministry of Culture & Tourism / The Foundation for Globalization of the Korean Language. 2001. *Cyber hangeulbakmulgwan guchuk unyeong saeop 2001nyeondo gyeolgwa bogoseo* (The 2001 Report on the Project of Construction and Administration of the Cyber Hangeul Museum).

Ministry of Culture & Tourism / The Foundation for Globalization of the Korean Language. 2002. *Digital (Cyber) hangeulbakmulgwan guchuk unyeong saeop 2002nyeondo gyeolgwa bogoseo* (The 2002 Report on the Project of Construction and Administration of the Cyber Hangeul Museum).

Ministry of Culture & Tourism / The Foundation for Globalization of the Korean Language. 2003. *Digital (Cyber) hangeulbakmulgwan guchuk unyeong saeop 2003nyeondo gyeolgwa bogoseo* (The 2003 Report on the Project of Construction and Administration of the Digital / Cyber Hangeul Museum).

Ministry of Culture & Tourism / The Foundation for Globalization of the Korean Lanoguage. 2004. *Digital hangeulbakmulgwan guchuk unyeong saeop 2004nyeondo gyeolgwa bogoseo* (The 2004 Report on the Project of

Construction and Administration of the Cyber Hangeul Museum).

Noma Hideki. 2002. “A Study on Grammar and Lexicology of Modern Korean in Japan since the 1980s-Development of Linguistic Realism.” *Hangugeo eohwiwa munbeopui sanggwangujo* (Relative Structure of Korean Lexicon and Grammar). Seoul: Taehaksa.

______. 2005. “Kankoku to nihon no kankokugo kenkyu--Gendaikankokugo no bunpokenkyu o chusin ni.” *Nihon Gatkai* Vol. 24. Tokyo: Meiji Syoen.

Seo Sang Kyu. 2001a. “Malmungchirel iyonghan gugeo munbeop yeonguui hyeonhwanggwa banghyang” (The Present Situation and Direction for Studies on Korean Grammar Using Corpora). Pp. 89-126 in *21segi gugeo jeongbohwawa gugeo yeongu* (Pp. 89-126 in the 21st Century Informatization and Studies on the Korean Language). The Research Center for the Korean Language, the Institute of Korean Culture, Korea University (ed.).

______. 2001b. “Malmungchiui juseokgwa hangugeo gibon eohwi eumi bindo sajeon” (The Annotation of Corpora and a Dictionary of Sense Frequency of Korean Basic Vocabulary”. *Gyeryangeoneohak Vol. 1* (Quantitative Linguistics Vol. 1). Bakijeong Publication Co.

______. 2002a. “Hangugeo malmungchiui guchukgwa gwaje” (The Construction of Corpora of the Korean Language and Tasks). Pp. 255-292 in *Hangugeowa jeongbohwa* (Pp. 255-292 in The Korean Language and Informatization). Taehaksa.

______. 2002b. Pp. 301-313 in “Hangugeo gyunhyeongmalmungchiui hyeonhwanggwa gwaje, hangugeohagui oneulgwa naeil (2002 Hanguk Eohakhoe Gukje Haksuldaehoe)” (Pp. 301-313 in The Present Situation and Tasks of Korean Balanced Corpora and the Present and Tomorrow of Korean Linguistics-2002 The International Academic Meeting of the Korean Language Society). The Korean Language Society.

______. 2002c. “Gugeojeongbohak yeonguui hyeonhwanggwa banghyang” (The Present Situation and Direction for Research on Korean Language Informatics). Pp. 431-463 in *Gugeohak yeongu 50 nyeon* (pp. 431-463 in 50 Years of Research on Korean Linguistics). The Institute for Korean Culture, Ewha University, ed. Hyean.

______. 2002d. “Hangugeo gyunhyeong malmungchiui hyeonhwanggwa gwaje” (The Present Situation and Tasks of Korean Balanced Corpora). Pp. 149-173 in *21segi gugeohakui hyeonhwanggwa gwaje* (The Present Situation and Tasks of 21st Century Korean Linguistics) Edited by Bak Yeong-sun. Seoul: Hangukmunhwasa.

______. 2003. "Hangugeo hakseupja malmungchi guchukui siljejeok munje" (The Practical Problems of Constructing the Learner Error Corpus of the Korean Language). Pp. 209-267 in *Hangugeo gyoyukgwa hakseupsajeon* (Korean Language Teaching and the Learner Dictionary) Edited by Seo Sang Kyu. Hangukmunhwasa.

Seo Sang Kyu and Han Young-gyun. 1999. *Gugeojeongbohakipmun* (An Introduction to Korean Language Informatics). Seoul: Taehaksa.

Seo Sang-kyu, Kim Jinung, and Kim Han-saem. 2001. Yonsei Sense Frequency Dictionary Based on a Sense-tagged Corpus. Pp.19-38 in *Sajeonpyeonchanhakyeongu* (A Study on Lexicography) Vol. 11 No. 2. The Center for Linguistic Informatics Development, Yonsei University.

Seo Sang Kyu, Yu Hyeon-gyeong and Nam Yun-jin. 2002. "Hangugeo hakseupja malmungchiwa hangugeo gyoyuk" (The Learner Error Corpus of the Korean Language and Korean Language Teaching). *Hangugeo gyoyuk* (Korean Language Teaching) Vol. 13 No. 1.The International Society of Korean Language Teaching.

Sin Ji-yeong. 2004. "Eumseong Corpus rel hwalyonghan gugeo yeongu" (Research on the Korean Language Using the Speech Corpus). Hangugeohak (Korean Linguistics) Vol. 23. Seoul: The Korean Language Society.

The Supporting Center for Industry of Speech Informatics, Wongwang University. 2004. *Jiyeokeo eumseongjaryoui chegejeokin sujip mit gwanrie gwanhan yeongu* (A Study on the Systematic Collection and Management of Speech Data of Dialects) (A Final Research Report). The National Institute of the Korean Language.

Yi Mi-hye. 2005. "Hangugeo munbeopgyoyuk yeongusa" (A History of Teaching Grammar of the Korean Language). Gukje Hangugeo Gyoyukhakhoe Je15cha Gukje Haksuldaehoe (The 15th International Academic Meeting of the International Society of Korean Language Teaching). The International Society of Korean Language Teaching.

Yi Tae-yeong. 2003. "Gugeo yeonguwa malmungchiui hwalyong" (Research on the Korean Language and the Use of Corpora). *Text eoneohak* (Text Linguistics) Vol. 15. The Society of Korean Text Linguistics.

______. 2005. "Bangeon malmungchiui jeonsanhwawa hwalyong" (The Computerization and Use of Dialectal Corpora). *Hanguk Eohak* (Korean Linguistics) Vol. 21. Seoul: The Korean Language Society.

Yi Yong-ju. 2002. "Eumseong Corpus ui gonghakjeok eungyong mit guknaeoe hyeonhwang" (The Technological Application of the Speech Corpus and the

Present Situation Inside and Outside of the Country). Pp. 323-340 in *Hangugeowa jeongbohwa* (The Korean Language and Informatization). Taehaksa.

Yim Yong-gi, et al. 2004. *21segi sejonggyehyeok gugeoteuksujaryo guchuk* (The Construction of Korean Special Data for the 21st Century Sejong Project) (A Research Report). The Ministry of Culture and Tourism / The National Institute of the Korean Language.

Seo Sang Kyu is a professor in the Department of Korean Language and Literature, Yonsei University. He is interested in studying about Korean grammar, the spoken language, and corpus linguistics. inaka@yonsei.ac.kr