The Present Conditions and Tasks in Constructing the Database of Korean Literary Materials Centering on the Korean Poetry Corpus

Kim Byong-sun

The significance of the compiling project of the database of Korean literary materials lies in supplying well-organized research materials and theory and principle of scientific literary study for related researchers, in enhancing the research capabilities of Korean literature, and in developing the Korean language.

First of all, I will generally introduce how the database of Korean literary materials has been compiled and how the related study has been carried out. And then I will introduce in detail the compiling process and the principle underlying the database of modern Korean poetry, the project of modern poetry corpus that I am trying to construct, and also show the present contents of compilation and a plan for their future use.

The database on modern poetry is as follows and only a part of it is served on the Internet: 1) The modern poetry corpus (10,300 poems) containing text information of modern poetry together with the metadata; 2) the dictionary of usage of modern poetic diction (610,000 items) showing the usage and information of the sources of modern poetic diction; 3) the explanatory dictionary of modern poetic diction (44,000 items) containing the whole word list of the dictionary of usage and explanations on important words; 4) the whole list of modern poetry anthologies (2,430 books) published in 1908-1970; 5) and the index of articles on modern poetry (17,500 articles) contained in the collections of poems published in 1923-1950.

I hope that scientific study on modern poetry based on such a database will flourish. At present, the quantitative way of study on modern poetry is being made on a basic level, and one of the ultimate purposes will be to find the stylistic fingerprint of each individual poet. This database will be supplied to related researchers on condition that joint research be carried out.

Keywords: Korean Poetry Corpus, statistic research, humanities computing, corpus linguistics

The Review of Korean Studies Vol. 8, No. 4 (105-140)

Introduction

First of all, this article deals with the significance and necessity of compiling a database of literary materials. It also introduces how the database of Korean literary materials is being compiled and how the related study is being carried out on the whole. And next, I will introduce in detail the compiling process and principle of the database on modern poetry, and the project of building a modern poetry corpus (abbr. KoPoCo)[1] that I am trying to construct, and also discuss the present contents of compilation and a plan for their future use. Finally, I will suggest the development tasks in compiling the project of the database of Korean literary materials and a plan for the common use of the modern poetry corpus.

1. The Significance of Compiling the Database of Literary Materials

Compiling the database of literary materials primarily means to turn literary works published in paper books into computerized electronic media. It can be itemized as follows: 1) To input the original text itself according to a fixed form; 2) to input literary works and their lists; 3) to input other literary information into the computer. These input materials are processed in due course in the next stage and the processed information is supplied to ordinary people or literary researchers through the media of the Internet or CD-ROM.

There are two ways to input the original text of a literary work: One is to input the text in letters, and the other is to scan the images of the text and input them in picture files. In case of classics, it is basically more important to input the text literally than to input even the images of the text. Thus, when it can be stored in digital form, the text of a literary work will have a new means of preservation. Even though the durable period of preservation by electronic media has not been tested, it is true that the conditions of preservation and duplication of digital materials are much better than those of paper books. Also, a computerized literary work can be easily adaptable to any kind of new media afterwards. The e-book of the present is a good example.

Literary works and their lists are used as primary materials of the literary information service. The information of the lists is also used as metadata on the original text materials of a literary work as mentioned above. In the meantime,

1. KoPoCo is short for Korean Poetry Corpus. It is composed of the first two letters of each word.

most of the works, i.e., the materials of bibliography, have been entered through an information-oriented project of the library or a construction project of the digital library. But all data have not yet been fully entered for all literary work in the process, i.e., the index of articles.

Finally, other literary information contains the personal information on writers, the history of literature, literary language, and others. These pieces of information are partly treated on an internet site specializing on literature and are actually compiled in the higher-level database. For example, the personal information on writers can be treated in the information on prominent figures; the information on the history of literature in the historical or chronological information; and the information on a literary language in the general language information. At present, an exclusive service of information on literature cannot be found.

The compiling project of the database of literary materials will result in the more scientific and precise study of literature. First of all, an accurate study on materials will be carried out through a variety of information areas organized in lists. Apart from the previous practice of subjective and intuitive study, a more objective and scientific study will be carried out by means of more abundant and inclusive materials being available through the computer. In the past, a study on materials was despised as a low-level activity and, in fact, a study on Korean literary materials themselves has not yet been carried out seriously. When the appropriate and elementary materials can be entered into the computer, researchers will be able to carry out much higher-level studies with the assistance of computer performing simple and laborious chores.

The significance of the compiling project of the database on Korean literary materials is to supply well-organized research materials and a theory and principle of scientific literary study for related researchers in order to enhance the research capabilities of Korean literature and to develop Korean language.

2. Examples of Literary Informatization

The present compilation project of the database of literary materials deals with the published materials of Korean literature, i.e., the printed and published literary works, and covers not only Korean classical literature but contemporary literary works as well. However, the practical compilation of the database is limited to classics, modern works, and contemporary works published before 1980.

When searching the internet, we can find that many individuals and institutions have entered and released the texts and lists of literary works. Some are acceptable

to a certain degree but most are composed of fragmentary and incomplete contents. Those which can be satisfactorily available as a well-organized database of materials are scarce. Accordingly, this article will look into some institutions and individuals that have expert knowledge and interest in the projects of the database of literary materials and have been persistently pursuing them.

The research activities of the Korean Computing Society established in 1988 made possible the computer processing of Korean linguistic and literary materials. This society, composed of linguistic and literary scholars, concentrated its efforts on processing old *hangeul* (the Korean language). Members majoring in Korean linguistics were interested in the input and process of organizing the materials of the history of the Korean language. On the other hand, members majoring in Korean literature were interested in those of classical literature. The activities of this society motivated its members and many Korean linguistic and literary scholars to computerize materials. Consequently, research on the analysis of input materials as well as the input of the materials such as original texts and lists has been carried out, making various remarkable achievements and resulting in publishing.

The first entered and processed research materials are as follows: The database of research materials on modern Korean literature (Lee Sun-young 1990),[2] the concordance of Kim So-wol's poetic diction (Kim Byong-sun 1993a),[3] the concordance of Songgang's poetry (Kim Hung-gyu 1993), an electronic anthology of modern Korean poetry (The Institute of Korean Culture ed. 1996),[4] overall information on modern Korean literature (The Institute of Humanities Information ed. 1999),[5] and the database of materials of the original texts of classical Korean literature (Kim Jin-young 2002).[6]

---

2. In the preface of the book edited by Lee Sun-young (1990), there are remarks on computer processing. It seems that materials were stored in the database.
3. It was a new experiment as a concordance of Sowol's poetic diction. Refer also to another book by Yoon Ju-eun (1991).
4. It contained 11,000 poems of 316 poets in 1910-1990. There are many ways of searching for the works and a lot of information on the explanation of poetic diction and the introduction of poetry is available.
5. This CD-ROM contains a total of 120,000 pieces of literary works between 1895 and 1994; 4,000 writers, 7,000 pictures, 1,000 poems, and the entire texts of 200 novels. The service of *Hanguk hyeondaemunhak daesajeon* (Grand Dictionary of Modern Korean Literature) of KRPIA is based on this CD-ROM.
6. According to this thesis, Prof. Kim compiled a database of the original and variant texts of *pansori* materials and is planning to publish a commentary dictionary of usage in *pansori* literature.

The research on literary informatization itself is as follows: A study on computer processing of literary materials (Kim Byong-sun 1992a), the compilation and usage of the old *sijo* database (Kim Hung-gyu, Kim Byong-sun, and Wu Eung-sun 1992), and a study on the construction and service of a digital library of modern literature (Kim Byong-sun 2002a; 2002b).

Meanwhile, the ten-year 21st Century Sejong Project (1997-2007) contains a compiling project on the corpus of literary works, the genre of which has been generally limited to novels. An additional project on the corpus of literary works to be inputted under this project will be of use in literary studies.

Structure of the Database on Modern Poetry

As a concrete example, I will discuss the corpus of modern Korean poetry, the so-called KoPoCo, compilation project, supported by The Academy of Korean

Table 1. Structure of Database on Modern Poetry

Classification	Name of Table	Scale	Description
Information on the List and Original Texts	List of Modern Poetry	20,078 types	Collection of the name of poet, title, genre, code of the anthology, publisher, publication date, and version information.
	Original Text of Modern Poetry	10,099 types	Collection of original text of the corpus of compilation object and metadata information.
	List of Anthology of Modern Poetry	2,437 volumes	Collection of the name of poetry book, name of poet (editor), code of the anthology, publisher, publication date, and classification of work form.
Processed Information	Concordance of Modern Poems	610,000 items	KWOC list of Korean modern poetry.
	Statistical Information		Collection of statistical information on poetic words, parts of speech, form of poetry, poet, and period.
Related Information	Dictionary of Modern Poetic Diction	44,000 items	Collection of diverse information in dictionary with the keywords of the concordance as head words.
	Index of Articles in Literary Journals	2,606 volumes	Collection of index of literary works in *Gaebyeok* and *Jogwang*.
	Corpus of Individual Poets		Collection of works by Kim So-wol, Kim Hyeon-seung, Sim Yeon-su, and Kim Hak-cheol.

Studies and another compiling project of the database on modern poetry I am trying to carry out on a personal level. This database has been constructed for nearly twenty years since personal computers appeared and that its usefulness for literary research has been fully verified in view of its structure and informational technique. Other databases have not opened their processes or contents concretely. We cannot grasp the facts of them. Discussion on the basis of this database will be a good example for the later compilation project of the database of Korean literature. Focusing on the following table that shows the database on modern poetry, I will briefly introduce its contents, scope, future plans, and refer to each of the conspicuous aspects of the table.

1. Lists and Information on Original Texts

a. Table of the List of Modern Poetic Works

When compiling the database on modern poetry, I took primarily as targets of compilation *Hanguk hyeondaesi jaryo jipseong* (A Compilation of Materials of Modern Poetry, abbr. *Compilation*), and *Hanguk hyeondaesi jaryo daege* (An Outline of Materials of Korean Modern Poetry, abbr. *Outline*). When publishing a concordance which does not support the full text in order not to violate the copyright, researchers will have to approach the original or copy texts easily. This table contains a total of 20,078 items of article index information including 18,064 pieces of information on original poems, translated poems, old or modern *sijo*, and Chinese poems, etc. contained in *Compilation* and *Outline* and other information on the titles both of the poetry and chapters therein.[7] Basically, it contains information on poets, titles, genre, code of poetry, sources in the *Compilation* and *Outline*, and year of publication. Every field was entered according to the writing of the original text, and a field for searching *hangeul* was added to information on poets and titles.

Considering the relations with the adjacent tables in all the databases, the titles of poems were encoded in digits: four digits indicating the year of publication and three digits, which was optionally given to the poems published in the same year and connected with hyphen.[8] All the databases share a unique key value for each work and will be expanded by giving the digits of each work in

7. The information on the scope will be available as of October 2005.
8. Considering that thousands of poems will be published in the 2000s, this system will be revised:

the code of the anthology. For example, the code of the poem "1925-007" indicates a poem published in 1925,[9] that is, *Jindallaekkot* (Azalea) by Kim So-wol, and that of "1925-007-005" represents his fifth poem, which would be, *Yet niyagi* (An Old Tale).[10]

In addition to the information on all the original, final, and variant texts, a copy version of each work was included and some works that changed their title in the re-compiling process was marked. In this index, the first ten words of each work are also contained.[11]

b. Table of the Original Text of Modern Poetry

This table contains total 10,099 poems as an index to original poems in *Compilation* and *Outline*. Twenty-two poems were not inputted because of missing pages. This table also contains the results of the KoPoCo project inputting the original texts of modern poetry and the constitution and contents of KoPoCo will be dealt with later. The original text of each work is stored in the memo field of this table and the mark for lines was substituted with a single slant (/) and the mark for stanzas with double slants (//). Information on the frequency of syllables, words, lines, and stanzas was also contained and will be easily calculated through my programming modules and query languages.

When searching for information on frequency in this table, an epic titled *Gukgyeongui bam* (The Night of the Border) by Kim Dong-hwan is the longest work with 4,716 words, 956 lines, and 112 stanzas while *Paengi* (A Top) by Hwang Sun-won is the shortest with only two words, one line, and one stanza.

c. Table of the List of Modern Poetry

Centering on original, edited, and translated poetry anthologies, this table con-

---

a front number of four digits and a back number of four digits. For example, 2004-0193 would indicate the 193rd poem published in 2004.

9. The numbers do not represent the order of publication of the poem. They are optionally given for convenience's sake.
10. As the numbers of the lines are given, this code system will be conveniently used as a standard search key for all the databases of modern poetry. Furthermore, when the numbers of the phrases and syllables are given in the future, the encoding of all the literary and linguistic constituents of modern poetry will be completed.
11. This can be used as an index to *Compilation* and *Outline*. This information is being served through my internet site "Hangugui hyeondaesi" (Korean Modern Poetry): www.koreanpoetry.net.

tains information on the titles of poems, the names of poets, and editors (in the case of translated poetry, the original poets and translators), the codes and titles of the poems, publishing companies, and the dates of publication. It also contains distinguishable information on original, edited, and translated editions. The titles of the poems were not inputted according to the writing of the original text but followed modern rules of spelling and spacing. For the convenience of searching, the name codes of poets in hangeul were added to the names of poets, most of which were written in Chinese.[12] The date of publication contains only the year and month of publication. When the month of publication is unidentified, it is marked with "0." Accordingly, when the poetry in the lists is contained in *Compilation* and *Outline*, the source is also shown.

This table contains a total of 2,437 kinds of poetry from *Gyeongbucheoldo norae* (Songs about the Seoul-Busan Railroad) by Choe Nam-seon in 1908, to *Gim hyeonseung jeonjip* 1 *si* (A Complete Collection of Kim Hyeonseung Poetry 1, Siinsa) in 1985.[13] Meanwhile, a unique key value is given to a poem and is shared by all the databases. The compilation project of the modern poetry corpus is based on the information of this table and contains information on present conditions, the status of the texts, and the kind of edition. The pictures and other information of the poetry anthologies will also be contained.

2. Processed Information

a. Table of the Concordance of Modern Poetry

The concordance of modern poetry was the first processed information based on the KoPoCo. Through an extra programming, the text of KoPoCo was processed to KWIC index, of which the unit of word simply became a keyword. But after five years of manual analysis of morphemes, basic form or original form was extracted from each word. And then, after two years of processed morpho-semantic analysis, homonyms and part of polysemy were extracted and serve as search headwords.

To show the usage of each keyword, eight to ten words before and after the

12. In most of the retrieval systems, the search for the Korean sounds of the Chinese vocabulary is impossible. For the convenience of search, the *hangeul* field for searching was added to major fields apart from the fields according to the writing of the original text.
13. This database was constructed with reference to many available lists and needs to be confirmed by the original text. In particular, materials on poetry after 1950 are being secured.

keyword were presented, and in information on sources, the name of the poet, the code of the poetry anthology, the title of the work, the order of the line, and the genre of the work were added. Extra codes were added for the connection with *The Dictionary of Modern Poetic Diction*, and the codes of poems connected to information on sources were connected to *The Table of the Whole Lists of Modern Poetry.* Almost 610,000 items of poetic diction are contained therein as of September 2005.

The compilation of the concordance is basically meaningful for looking into the usage of the same poetic diction with a context. On the other hand, it is also meaningful for looking into the poetic diction according to the paradigmatic relation as aside from its syntagmatic relation. When the poetic diction in the word form is rearranged through sorting and the cross reference is available, it will be of great help to the fixing of input errors.

In concordance compiling process, there is a way to apply the function of Arae Hangeul itself (Kim Byong-sun 1992b) and another way to use some automatic indexing program.[14] As this corpus was too enormous and unique, I myself created the computing program called Ttokttoksae. The result was *Hyeondaesieo yongryesajeon* I (The Dictionary of Usages of Modern Poetic Diction I) using practical headwords as keywords, after entering the MS Access database of Microsoft Corp. and several stages of proofreading. This dictionary was published on CD-Rom by Nuri Media (Ltd.) under the name of *Hanguk hyeondaesieo yongryesajeon* (The Dictionary of Usage of Korean Modern Poetic Diction) together with some related tables and statistical information. Any institutions or universities contracted with this company can search for it online (http://www.krpia.co.kr).

b. Table of Information on Statistics

A variety of statistical analyses of the corpus of modern poetry such as frequency research can be easily processed by the very query of MS Access. MS Access is not a special program for statistical analyses but is easy to handle and browse. It is not restricted in the number of processible records and can process literal information without any serious difficulty. The usability of this program is

14. Hgrep by Hangeul and Computer, Ltd., and Kkamjjaksae, developed by Prof. So Kang-chun et al. of Jeonju University, belong to this category. For further information, refer to the site of "Hangukhak jaryoui jeongbohwa" (Informatization of Materials on Korean Studies, (www.koreanpoetry.net/computing/materials4.htm).

excellent in statistical analyses. Furthermore, the use of this query will make possible high-level statistical analyses. The results of this project were stored as different fields in major tables of the database or as different frequency tables.

The frequency of basic headwords and representative synonym among the statistical materials related to the frequency of poetic diction are stored together with the table of the dictionary. The statistical materials related to poetic forms contain the frequency of works of each genre, the poetic diction of each genre, and the lines and stanzas of poetry. The frequency of poetic diction and the lines and stanzas of each work are added to the table of the original text. Information on poets and poetry contains the number of poems, poetic diction, and the kinds of poetic diction and the average of them among a total of 345 poets. The frequency of usage of the poetic genre of each poet is also contained.

The frequency data on the basis of the year contain information on how often 41,973 poetic words were used every year or every five years from 1923 to 1950. By means of these data, it can be affirmed when and which poetic words began to be used and which were preponderantly used.

Table 2 below contains information on the parts of speech containing not only the frequency and the number of types of headwords of each part of speech but also the number of types of the representative synonym. The related information on the number of types of headwords and representative synonym is also contained.[15]

3. Related Information

a. Table of Dictionary of Modern Poetic Diction

This table contains all the headwords and representative synonym in the concordance of modern poetry as basic items and presents lexical information on them. A total of 44,000 items are contained. Headwords, superscripts, and parts of speech are indicated as basic information. In case the headword is Chinese, a foreign language, or an adapted foreign language, the original language is indicated. The explanation of meaning about parts of important poetic words is served. Headwords follow the way of inscribing "head words" in *Pyojun gugeo-daesajeon* (Grand Dictionary of Standard Korean, abbr. *Standard Dictionary*), and in case of homonyms, superscript is used according to the dictionary. This

15. Information in parenthesis denotes abbreviations used in the database.

Table 2. The Table of Frequency of Parts of Speech

Part of Speech	Frequency	Number of Types of Headwords	Frequency of Headwords	Number of Types of Representative Synonym	Frequency of Representative Synonym	Types of Headwords & Representative Synonym
Noun (n)	250,888	23,580	10.64	19,690	12.74	1.20
Incomplete Noun (nd)	16,859	256	65.86	213	79.15	1.20
Pronoun (m)	32,955	133	247.78	57	578.16	2.33
Proper Noun (p)	7,143	2,669	2.68	2,416	2.96	1.10
Verb (v)	149,046	7,082	21.05	5,744	25.95	1.23
Auxiliary Verb (vd)	15,612	41	380.78	29	538.34	1.41
Adjective (j)	59,255	3,246	18.25	2,809	21.09	1.16
Auxiliary Adjective (jd)	2,260	25	90.40	13	173.85	1.92
Adverb (a)	45,443	3,767	12.06	3,029	15.00	1.24
Exclamation (i)	4,424	250	17.70	195	22.69	1.28
Numeral (b)[16]	1,743	130	13.41	121	14.40	1.07
Pre-Noun (e)	23,571	411	57.35	371	63.53	1.11
Obscurity (u)[17]	492	94	5.23	94	5.23	1.00
Suffix (hb)[18]	46	21	2.19	21	2.19	1.00
Prefix (hf)[19]	5	5	1.00	5	1.00	1.00
Stem of a word (ws)[20]	404	236	1.71	231	1.02	1.02
The Ending (wt)[21]	4	4	1.00	4	1.00	1.00
Postposition (wx)[22]	1,864	23	81.04	21	88.76	1.10
Total	612,014	41,973	14.58	35,063	17.45	1.20

16. Those that are used as pre-nouns in fact are classified as pre-nouns.
17. They are what cannot be ascertained because of missing or obscure letters.
18. Those suffixes of the words that have not been recorded in the dictionary are solely designated. *-jjae, -tuseongi, -ryu, -naegi, -jabi,* etc.
19. Those affixes of the words that have not been recorded in the dictionary are solely designated. *heot-, doe-, hol-, cheo-,* etc.
20. Those are mostly the roots of declinable words but functionally have the characters of nouns and adverbs. They are treated according to the *Standard Dictionary*.
21. They are obscure and independent ones such as *-keoniwa, -deoni, -guryeo,* etc.
22. Besides *-gatchi,* several independent postpositions are also included.

table contains not only the vocabulary of the *Standard Dictionary* but also what are not contained in the dictionary.

As the items of headwords primarily treat the written form of poetry, standard written forms are indicated as being falsely written, dialectical, or other written forms. This field for standard writing of related words is used as basic materials for statistical projects showing semantic phenomenon. That is, like the table of the dictionary of usage of poetic diction, the representative synonym equivalent to each headword was indicated and the relation between headword and its synonym was also revealed. The typical types and frequency are as follows:

- Words of fallacy (4,651 types): written headwords deviant from orthography.
- Synonym (2,750 types): headwords and representative synonym are heteronyms mutually.
- Abbreviated words (267 types): headwords are abbreviated forms of the synonym.
- Ancient words (244 types): headwords are types of ancient representation of the synonym.
- Dialects (164 types): headwords are dialects of the synonym.

b. Other Related Matters to Literature

In "Munyeji surok hyeondae munhak gisasaegin" (The Index of Articles of Modern Literature Contained in Literary Journals) are contained the index of articles of modern literature in general literary journals such as *Gaebyeok* (A New World) and *Jogwang* (The Morning Sunlight). A total 2,606 pieces of information and some original full texts of the literary works are contained, which will be expanded to the indexes of articles of all literary journals in the future. Search services are available on my site "Hangugui hyeondaesi" (Korean Modern Poetry).

Hyeondae Siin Jeongbo Corpus (The Corpus of Information on Modern Poets) that I have personally secured contains the corpora of Kim So-wol and Kim Hyeon-seung.[23] Accordingly, there are also the corpora of Sim Yeon-su (poetry) and Kim Hak-cheol (essays) that some students have compiled under

23. Besides this server, there is a lot of information on modern and contemporary literary writers, literary works, research data, the original texts, and lists. They are always available if you visit my office.

my supervision during their course work at graduate school. These are a little different from the forms of KoPoCo because they were compiled for a immediate purpose to write a dissertation.

The Corpus and Concordance of Modern Poetry

1. The KoPoCo Project

a. The Performing Process of the Project

The corpus of Korean modern poetry (KoPoCo) was compiled at the end of the 1980s aiming for a certain degree of the possibility of processing old *hangeul*. At first, to realize the writing of the original texts written in old hangeul correctly was the main task of input. In spite of a craving for the use of the database in the processing, there was no appropriate function for old *hangeul* yet. After the word processor program like Arae Hangeul has allowed the input of old *hangeul* or expanded Chinese characters, the database system does not support it for a long time. Accordingly, the theoretical study on what system had to be applied to input data and to process it made little progress, and compiling the concordance of input works could be the prime tasks of humanities scholars inevitably.

The first model text of my choice to compile the concordance was the poetic works of Kim So-wol. Fortunately, I obtained a revised edition of textual criticism on the poetic works of Kim So-wol by Prof. Chon Chong-gu. So, I dared to compile the concordance of Kim So-wol's poetic works not by manually. I made the utmost use of the macro function of Arae Hangeul used as an input tool and made an index in the form of KWOC (Keyword out of Context). I also made a variety arrangement of lists such as normal and reversed ones through the various functions of processing information regards Arae Hangeul such as sorting. To treat statistical information, MS Excel was used. The results were the publication of a total of three books on *Sowolui sieowa geu sseuimsae* (The Poetic Diction and Usage of Kim So-wol, Hangukmunhwasa).[24] This was the first concordance of modern poetry compiled by diligent computer processing in Korea.

Since I started my position at The Academy of Korean Studies in 1993, a large-scale project for constructing the database on modern poetry began on the

24. Refer to Kim Byong-sun (1992a) for compiling the concordance on a word processor.

basis of a short article *Hyeondaesi deitabeiseu guchuke gwanhan yeongu* (A Study on Compilation of the Database on Modern Poetry, Kim Byong-sun 1993b). For almost four years after 1994, the inputting of the original texts of modern poetry and the project of concordance was accomplished under the name *Hanguk hyeondae sieosajeon pyeonchan* (The Compilation of Dictionary of Poetic Diction of Korean Modern Poetry). The results have been available on the internet under the name "*Hanguk hyeondae sieo yongrye sajeon*" (A Dictionary of Usage of Modern Korean Poetic Diction) by Nuri Media Ltd. For almost four years since 2001, a statistical study on the corpus of modern poetry has been carried out and will be available in books and on the internet when the project is completed this year.

b. The Range of KoPoCo

The periodical range of the database on modern poetry is the same as that of Korean modern poetry. Even though there is argument about the starting point of modern poetry anthology, *Haepariui norae* (Songs of a Jellyfish) of Kim Eok published by Joseondoseo Ltd. in 1923 was viewed as the first modern poetry anthology and the anthology up to the present since then has become the subject of study. Of course, poetic works are generally published through journals or literary coterie magazines before they are published in an anthology. In this case, the period of publication is very important and the poetic works will be valuable when they are published in an anthology. The information on the original texts of poetic works in the database on modern poetry is based on the poetic works contained in an anthology.

It is not easy to access the early works of modern literature that become targets for compiling the database. It is true that anthologies before the 1960s are regarded as valuable editions at libraries. It is absolutely necessary to seek confirmation of the printed original texts when researchers want to use a dictionary of usage or other processed information. Fortunately, as most modern poetry was published through copying or reprinting using the brisk photographic reproduction business of the 1980s, it became easy for general researchers to access these original materials. Accordingly, when compiling this database, the materials of photographic reproduction for researchers to access, that is, *Compilation* and *Outline* were treated as the first targets of compilation.

Information on the original texts set a limit to the year, taking 1950 as the base line. While literary works were rather plentifully contained in *Compilation* and *Outline* up to 1950, those after that time contained only part of the materials.

Table 3. The Present Conditions of Compiling the Corpus of Materials

Materials	Volumes	Contained Poetry	Original Poetry	Edited Poetry	Translated Poetry	Contained Corpus
46	228	193	23	12	192	Outline
31	150	134	16	N/A	34	Total
77	331	287	32	12	194	

It was very difficult to deal with the huge texts of that time at once.[25] This corpus contains only original creative poems. It means to exclude translated and adapted poems. Another corpus of translated poems will be compiled later. The corpus of original poems contains not only modern free-style poetry, prose poetry, epics, dramatic poetry, folk poetry, and *changga* but also the genre of modern *sijo* as well.

There are many poetry anthologies in *Compilation* and *Outline*, and a total of 77 volumes contains a total of 331 poetry anthologies. Among them, 47 poetry anthologies (40 original and 7 edited) are contained altogether in both *Compilation* and *Outline*. The types of poetry and corpus contained in them are as follows.[26]

2. The Production of the Concordance of Modern Poetry

a. The Lists of Poetic Works and Setting of Definitive Editions

KoPoCo contains 9,893 poems of modern original poetry including the period 1923-1950 to compile the concordance of modern poetry.[27] In the case of the titles of poems, the original titles were investigated and fixed as standard titles. For example, when the same poem was published many times, the title itself was changed in some cases, and the writing system in other cases. In these cases, through investigating variant texts, the titles worthy of the original or of the

25. The latest poetry contained in *Compilation* and *Outline* was *Gongcho o sang-sun siseon* (Selected Poetry of Gongcho Oh Sang-sun) published by Jayumunhaksa in 1963.
26. In addition to *Compilation* and *Outline*, Kim Myeong-sun's *Saengmyeongui gwasil* (The Fruit of Life) published by Hanseongdoseo Ltd. in 1925 is also contained. The original text of this poetry belongs to The Academy of Korean Studies.
27. The original texts of 47 poems were not inputted because the whole or parts of them were lost for various reasons.

highest popularity were fixed as standard titles.[28] When a certain poet published several different works with the same title, where possible they were distinguished by adding the numerals ① ② ③, etc. to the back of the original titles in the order of publication.

When the same work or similar work was published repeatedly, analytical information on the original and variant works was included. The principle of the original text decision gave priority to the final work published in individual poetry during the lifetime of a poet. As a result, 8,300 of 9,893 works were classified as original and the other 1,500 works as variants.[29] Besides this, information on the original texts and genres, the location contained in *Compilation* and *Outline*, and a standard key to search, etc., are essentially contained. The date of creation, and the titles and subtitles of series of works, etc. are contained as optional information.[30]

When the same work was published in various poetry anthologies, it was proper to contain all information on publication in the list of works. However, in order to compile the database of the original texts it was also proper to contain only one typical representative work from among the same works. Otherwise, various statistics on them will not become reliable. The principles for decision of representative works are as follows:

Firstly, the frequency of publication of each work was confirmed. In the table of the lists of works, the same titles were primarily compared. When the titles were published differently, the beginnings of works were compared to confirm it was the same work.[31] When the title of a work was changed or the contents were completely changed, it was treated as a totally different work. When the same work was published repeatedly, it was named as the first edition or the second

28. In this case, it was difficult to find out the "sameness" of the works, but such was confirmed by comparing the beginning or middle of the works. When fixing the headwords of the original text, not a few of the same works could be found.
29. In fact, it was not so easy to fix a single work. For example, a work of a certain origin was divided into more than two works, and vice versa. In case of a series of poems under same title, the whole had to be taken as a work sometimes, and at other times had to be divided into a respectively independent work.
30. When stored in the table of the MS Access database, it is very convenient to extract statistical information. Accordingly, the database on modern poetry will be operated according to this structure of file for the time being. If the usage of XML for literature is developed still more in the future, this database will be changed into an XML database.
31. Sometimes, the titles and beginnings of a part of certain works were changed. In this case, the concordance of the original text made confirmation possible by cross reference.

edition according to the order of publication.[32]

Secondly, when fixing the objects of compilation, the last poem in the anthology published during the lifetime of a poet was given priority. This was named as "the final version."[33] When a poem was collected in the anthology published together with the poems of various poets, it was excluded in the final version. However, when it was not published in the individual collection, it was recognized as the final version.[34] When the titles or writings were slightly different, standard titles were fixed through this endeavor.

Finally, the part that needed proofreading in the final version was treated by textual criticism referring to the works published previously. Thus, the completed work by proofreading was named as "the revised edition."

Through this process, the final revised editions, except the variant versions (the works slightly different from the final version), and the copy versions (the works identical with the final version but different in its period of publication) contained in the corpus of modern poetry were used as basic materials for further various statistics and processing. Thus, 8,509 of 10,099 poems were recognized as the final versions and the other 1,590 poems as the variant versions.

The poetic works of Kim So-wol had the highest frequency of republication. It means they had a great number of variant and copy versions. A total of 248 poems of his works were contained in the collections of poems up until 1950. Among them, 95 poems were recognized as variant versions, 8 poems as copy versions, and 145 poems as final versions. Among them, the works reproduced more than four times are as follows: *Ganeun gil* (The Way to Go) 4 times, *Geumjandi* (Golden Turf) 4 times, *Meon huil* (Far Future) 6 times, *Sakjuguseong* (Tortoise Castle at Sakju) 4 times, *San* (Mountain) 4 times, *Sanyuhwa* (Blossoming Mountain) 4 times, *Wangsimni* 5 times, *Imui norae* (A Lover's Song) 5 times, *and Jindalnaekkot* (Azalea) 6 times, etc.

b. The Principle of Input of the Original Text

When inputting the text of modern poetry, I tried to grasp the underlying structure

32. Except for some particular instances, the first published works seem to be contained mostly in various journals, literary coterie magazines, and literary pages of newspapers rather than in a collection of poetry.
33. There is one task left to fix the final version of the same works contained in the poetry published after 1950 through another project in the future.
34. A common collection of poems of two or three poets was regarded as an individual one.

of poetic works rather than the outward appearance from of the original texts or the reproduction of the types of printing. The underlying structure means not the outward form, that is, the surface structure in the process of editing and publication but the form of the work laid out in the poet's mind in his final writing. In other words, the underlying structure of poetic forms such as lines and stanzas varies according to the size of printing paper, format, and a printing device. Feeding data into a computer followed the distinction between lines and stanzas regardless of the outward aspects deriving from the characteristics of these devices. Even in this case, when the end of a line coincides with that of a page, there comes a problem whether the line continued on the next page can be regarded as the beginning of a new line or as that of a new stanza or sometimes as same line. This case was treated by means of referring to other published editions of the same work or considering the context of the work if variants are not available.

Thus, the surface structure of a work such as printing type, the size and the margin of space, etc. was not taken into consideration and the marginal mark of a page was not set up.[35] Also, in accordance with the modern computer processing environment, all the works were inputted in the form of horizontal typesetting in spite of their original typesetting. As a system of input, TEI Lite, a simplified system of SGML, was used and as an implement of input, and the Arae Hangeul word processor program was used.[36] When inputting works fixed as final versions, faultless documents, which do not have superfluous, lines and spaces are favored. It was made a rule to transcribe a letter into a computer code.[37]

In the case of compound words, according to the principle of the *Standard Dictionary*, those words that are not contained as compound words are divided into more than two independent words and are treated. In the case of idiomatic phrases all composed of Chinese characters, those contained in the dictionary are treated according to the dictionary. But those that are not contained in the dictionary, if possible, are also divided into more than two independent words

35. In the corpus of modern Korean poetry, the goal for reproducing the printed original text was not set up. How to set up this goal and treat it will be studied later.
36. The 3.0 edition of Arae Hangeul (HWP word processor) began full-fledged service of old letters of Hangeul and expanded Chinese characters. The 2002 edition of Arae Hangeul serves Unicode and cuts down much inconvenience as a literal system for inputting modern literary works.
37. The long vowel mark (-) and inserted vowels are excluded. Inserted siot is maintained. In case of contracted forms, a syllable may be added in the process of revealing the original. For example, in the case of *-yeo*, it is reconstructed and treated as *ieo* as *meokyeo-> meokida*.

and are treated. When divided, each word that is not contained in the dictionary is tied together as the original. In the case of spaced words in the original text and contained as one word in the *Standard Dictionary*, they are optionally tied together and recorded by the compiler in a word.[38] When the errors of the original text seemed to be those of proofreading, they were amended at the stage of input. But unprinted letters in the original text and unintelligible letters because of censorship, so-called missing letters, were not amended.

c. The Procession and Conversion of the Original Text into the Database

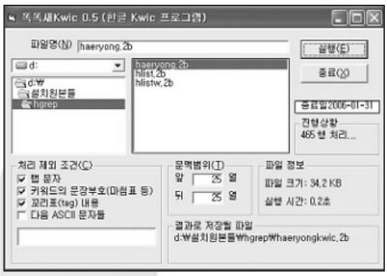
The first task to process the original text was to compile the concordance. The concordance becomes the bases of the investigation into used vocabulary and the treatment of statistics.[39] Also, the concordance has an enormous influence on re-proofreading of the original text. When producing the concordance of the text of Kim So-wol in the early 1990s, the text processing was done by the help of Arae Hangeul. As the corpus of modern poetry contained more than 600,000 words, the compilation of the corpus was beyond the capacity of Arae Hangeul. Accordingly, another program had to be used, and arrival of the program Hgrep was primarily awaited. This program was developed by the National Institute for the Korean Language to analyze the corpus of the *Standard Dictionary*. It has a function of the concordance generating at the level of the phonemes of *hangeul*.

---

38. In the case of words that can be tied and spaced out, their meanings are taken into consideration and are treated. In general, when two words are tied into a word, the meaning becomes special or changes slightly. Such a task of fixing the boundary of poetic diction was one of the most difficult ones. Even though this task had been carried out continuously from the stage of input to that of proofreading, not a few words were newly spaced out or tied together after making the concordance. At the early stage of input, it was thought desirable to record each spaced word as each headword as a principle, but in the end, they were treated according to the *Standard Dictionary*.
39. When proofreading, this concordance was fully used. At first, graduate students majoring in modern literature joined in inputting, but they were not accustomed to the orthography principle and language system at that time. A proper principle was not prepared at the early stage of input. But computer culture and information processing techniques have been continuously developed, and the principle has also been changed, resulting in not a few errors. Accordingly, even after the primary proofreading on the computer screen, and the next output proofreading, this proofreading continued in the compilation of the concordance. Then, not individual works, but the whole corpus was batched. It was very efficient to proofread through confirming the text when compiling the concordance. Also, it was possible to review the original text through statistics when mistaking words of low frequency for errors.

This program was not finally available because it used only the two-byte file of the Arae Hangeul 97 edition. Finally, by using Visual Basic language, I have developed the Ttokttoksae (2bkwic.exe) program that automatically produces the concordance of KWIC form from the 2 byte text file of Arae Hangeul's own according to given conditions.[40]

This concordance was imported into MS Access. As the recorded items of the concordance numbered more than 610,000, an exclusive program for the database had to be used for the purpose of browsing these voluminous materials.[41] The MS Access program is the most typical database program contained in MS Office package and makes it possible to browse data and considerable programming with query language or Visual Basic applications.[42] After all, related information on the concordance, various lists of the poetry and the poetic works, and the dictionary of poetic diction were arranged according to the table of MS Access and were laid out in terms of interrelation.[43]



Picture 1. Pop-Up Window of Ttokttoksae 2bkwic.exe.

d. The Normalization and Standardization of the Data

① The Meaning and Necessity of Normalization

Before the orthography of Hangeul was established in the 1930s, the object works of KoPoCo compilation began. Then, the poets and editors did not understand the orthography fully. For example, a certain poet did not adhere to a certain writing rule consistently, and different writing appeared even in the same work. Accordingly, it was necessary to input consistently and apply the principle of proofreading relating to the textual criticism in the input of this text. At the

40. "2bkwic.exe" is the program producing the concordance for two bytes file of Arae Hangeul 97. Several options will be available. At present, it has been developed even to treating even the Unicode text file.
41. MS Excel program was once used, but was unsuitable to treat enormous materials because of its small capacity of 65,000.
42. MS Access is superior in function to the dBase program and Clipper of early DOS period.
43. When XML is used full scale in the future, this table will be easily changed into the form of XML materials.

stage of input, the correction of errors was mainly done, and full-fledged proofreading was done later after producing the first dictionary of usage.

When sorting the headwords in the dictionary of usage, some were falsely written in the diversity of writing, and so-called literal obscurity in expressing a word in various letters could be found. In spite of proofreading them according to a unified principle, the personal and characteristic writing of a poet had to be paid attention to because the objects of compilation were poetic works. However, when compiling, sometimes it was very difficult to distinguish whether certain erroneous writing derived from the poet's intention, inattention, errors of editing and proofreading, or those of printing. Therefore, the following principles were generally applied and the project of normalization was accomplished.

First of all, the poet's writing was given priority, instead of writing according to modern orthography, and was standardized to a certain degree. In actual processing, the original or basic form was reconstructed in the writing form, inflected or conjugated, of actual works according to the following principles. The *Standard Dictionary* became a standard of fixing vocabulary, and the headwords in this dictionary were used untouched.[44] Dialectic, archaic, and erroneous words as well as standard words were contained as headwords in the dictionary. When the writing of poetic diction was applicable to such a word, they were also recognized as untouched.

Next, the words that were not contained in the *Standard Dictionary* were standardized according to the poet's writing.[45] In this case, the compiler optionally processed them. When poetic words written by a poet are pronounced and the sounds are the same or analogous to those of the headwords in the *Standard Dictionary*, they are substituted with the headwords of the dictionary. They are mostly related to problems of syllabication and anti-syllabication (*yeoncheol*). When pronounced differently, the substitution of positive and negative vowels is partly allowed. In the case of the writing of lowest frequency, it is integrated into the similar writing of the highest frequency.

In the case of the low frequency of loan words and foreign words, they are

44. Postpositions were not taken as independent headwords. But only *-gatti* was separately treated. In fact, as a field of postpositions and suffixes through the analysis of morphemes in the dictionary of usage of poetic diction was processed, the frequency of these postpositions and suffixes could be searched for through the task of standardization.

45. Some errors can be found in the *Standard Dictionary* and are optionally corrected and processed.

disregarded and changed to the present standard writing. In case of their high frequency, they are processed according to the unvaried writing of the works. Meanwhile, the writing of Chinese place and personal names are processed according to the present orthography of loan words. However, those that were introduced long ago and have merged with our cultural heritage are written according to their present pronunciation.[46]

② The Standardization of the Writing of Chinese Characters

The writing of Chinese characters followed a consistent process of proofreading. There were diverse Chinese characters expressing the same word and also not a few erroneous ones among them. When looking into the deviant writing when compared to the standard, it seemed difficult for a poet to adhere to the writing of certain Chinese characters in the process of the actual writing of poetic works, editing, and printing. In modern times, the writing of Chinese characters was not optional, and the types of letters varied according to an environment of a print shop. Moreover, changes in Chinese characters did not lead to a particular new meaning. Accordingly, the writing of Chinese characters began to follow the process of standardization.

At the stage of computerization, this problem faced a new phase. In computer processing, printed letters change into the codes of the computer and the problem of ambiguity arises. In particular, the writing of Chinese characters could be proofread by means of a textual criticism of the writing itself, but a problem of writing the same Chinese characters variously in abbreviated, simplified, and the popular and regular forms of Chinese characters was beyond textual criticism. In poetic works at this time, abbreviated, popular form and correct form of Chinese characters were used without any particular rule, and sometimes, Japanese forms and popular forms of Chinese characters were also used.[47] In recent Unicode, the writing of abbreviated, simplified, and regular form of Chinese characters is registered entirely and it is not difficult to realize these characters in computer inputting. But it will only be the input of the writing of poetic works that will cause a lot of problems in the process of search and use in the future.

Accordingly, in the database on modern poetry, all the writing of Chinese

46. For example, "泰山" is written not as *taisan* but as *taesan* and "李太白" not as *litaibai* but as *itaebaek*.
47. For example, "勞働" and "身軆" are misused words-of "勞動" and "身體."

characters was changed into the correct forms of characters and then inputted. Also, even the Chinese characters of daily use were changed into the correct ones, according to the agreement of usage of grammar.[48] The Chinese characters registered as different sounds and various sounds according to the rule of initial pronunciation were inputted fit to the context of the word.

③ The Lemmatization and Fixation of Headwords

The analysis of morphemes, or so-called lemmatization, is often a basic task on the corpus based linguistic study, and this project could not be an exception in order to fix the headwords of the dictionary of usage. This project covering a total of 610,000 words was accomplished and demanded lots of time and effort.

Though automatic morphological analysis programs for modern Korean language were developed and have been used in some fields, it was unfortunate not to be able to use such analytical machines because there are many old *hangeul* and non-standard expressions in the modern poetry corpus. Of course, in the case of the pure modern language, the morphological analysis is not so easy. Accordingly, a certain degree of automation for sorting and analyzing postpositions and suffixes of the same form through compiling the reversed lists of vocabulary was used, but most of the task was manually accomplished.

In the analysis, a word was divided into two constituents, that is, semanteme and morpheme, the meaning and form part, and treated according to the *Standard Dictionary*. They were divided into basic forms of headwords, postpositions, or suffixes and recorded in the concordance. But form parts were not analyzed minutely.

④ Morpho-Semantic Analysis

In the concordance of modern poetry, the analyses of homographs were made. The words of the same written form but different meaning were sorted while attaching a superscript to each word according to the *Standard Dictionary*.[49]

---

48. For example, Jongno is often written as "鍾路" but the correct writing is "鐘路." Thus, the correct form as "鐘路" was adopted.
49. In the database on modern poetry, the field of every superscript was realized as the number of two figures. So, the number 1-9 was shown as 01-09. No superscript was attached to the unique headword without a synonym in the *Standard Dictionary* but "00" was attached to it in the database on modern poetry. Hereafter, the two-digit numbers immediately following the headwords in this paper should be understood in this context.

Those words unregistered as headwords in the *Standard Dictionary* were given the numbers of 90-99. They amounted to 10,781 types in all and covered 25% of all the headwords of 41,973 types.

Homographs could be found in 2,674 types. Morpho-semantic analyses were manually made considering the context of these words using a lot of time and effort. The words with more than seven types of homographs are as follows (the number in the bracket shows the types of homographs): *chida* (15), *su* (15), *jang* (14), *sang* (12), *yeong* (12), *I* (12), *dae* (11), *yeon* (11), *yang* (10), *dong* (9), *bun* (9), *seong* (9), *jeong* (9), *je* (9), *gyeong* (8), *weon* (8), *ja* (8), *tada* (8), *hae* (8), *gong* (7), *do* (7), *tteuda* (7), *ri* (7), *mi* (7), *bal* (7), *bae* (7), *sa* (7), *so* (7), *si* (7), *yeok* (7), *il* (7), *jeo* (7), *jeok* (7), *jeon* (7), *jong* (7), *ju* (7), *ji* (7), *jida* (7), *jin* (7), *jjak* (7), *cheong* (7), and *chae* (7).

Meanwhile, the polysemic analyses of some high frequency words were made in the database on modern poetry. Let us give an instance of polysemic analysis using, "*hae* 01." This word as a noun was divided into three and lemmatized symbols to the superscript: ① a daily expression of "*taeyang* 02," ② a cycle of the earth's revolution (year), and ③ the duration of daylight (brightness). As an incomplete noun, it was classified as the unit of a cycle of the earth's revolution.

⑤ The Fixation of Representative Synonym

In the meantime, there are words that have the same meaning with different spelling. Speaking of abbreviated and original words, they have to be taken as the same meaning though they are recorded as two separate words. "*Gal* 02" is an abbreviated word of "*gaeul* 01." In spite of the different spelling, they have the same meaning. "*Gaeul* 01" is given as synonymous with "*gal* 02."

Besides, these cases can equally be shown in the relations between words written in Chinese characters and native Korean words, adapted foreign languages and the Korean language, archaic language and modern language, abbreviated words and original words, erroneous writing and correct writing, peculiar writing and normal writing, etc. In particular, in case of erroneous writing, there are dialectical writing and deviant writing from orthography. In case of peculiar writing, there are humble words, honorific words, and pen names or abbreviated names shown in the personal names. They can facilitate the researcher's retrieval and furthermore, can be used as materials for the semantic research of the works.

There were 6,445 types of headwords with representative synonym and they

Table 4. The Lemmatization of Kim So-wol's *Jindallaekkot*

Key Word	Headword	Suffix / Postposition	Relation	Representative Synonym	Super script	Part of Speech	Meaning
Na	Na	0		Na 03 m	03	M	
Bogiga	Boda	Giga		Boda 01¤–v	01㉠	V	
Yeokgyeoweo/	Yeokgeopda	Eo		Yeokgeopda 00 j	00	J	
Gasil	Gada	Sil		Gada 01¤–v	01㉠	V	
Ttaeeneun/	Ttae	Eneun		Ttae 01 n	01	N	
Maleopsi	Maleopsi	0		Maleopsi 00 a	00	A	
Gohi	Goi	0		Goi 01 a	01	A	
Bonae	Bonaeda	0		Bonaeda 00 v	00	V	
Deuriurida//	Deurida	Orida		Deurida 01 vd	01	Vd	
Yeongbyeone	Yeongbyeon	E		Yeongbyeon 02 p	02	P	寧邊
Yaksan/	Yaksan	0		Yaksan 04 p	04	P	藥山
Jindallaekkot/	Jindallaekkot	0	Synonym	Jindallae 00 n	01	N	
Areum	Areum	0		Areum 01 n	01	N	
Ttada	Ttada	Da		Ttada 01 v	01	V	
Gasil	Gada	Sil		Gada 01¤–v	01㉠	V	
Gile	Gil	E		Gil 01 n	01	N	
Ppeuriurida//	Ppeuriuda	Rida	Error	Ppeurida 00 v	90	V	
Gasineun	Gada	Sineun		Gada 01¤–v	01㉠	V	
Georeumgeoreum/	Georeumgeoreum	0		Georeumgeoreum 00 n	00	N	
Nohin	Noida	Nieun		Noida 00 v	00	V	
Geu	Geu	0		Geu 01¤†e	01㉡	E	
Kkocheul/	Kkot	Eul		Kkot 01 n	01	N	
Sappeuhi	Sappeuhi	0	Synonym	Sappeun 01 a	02	A	
Jeuryeobalkko	Jeuryeobaltta	Go	error	Jireubaltta 00 v	90	V	
Gasiopsoseo//	Gada	Siopsoso		Gada 01¤–v	01㉠	V	
Na	Na	0		Na 03 m	03	M	
Bogiga	Boda	Giga		Boda 01¤–v	01㉠	V	
Yeokgeoweo/	Yeokgeoptta	Eo		Yeokgeoptta 00 j	00	J	
Gasil	Gada	Sil		Gada 01¤–v	01㉠	V	
Ttaeeneun	Ttae	Eneun		Ttae 01 n	01	N	
Jugeodo	Jukda	Eodo		Jukda 01 v	01	V	
Ani	Ani	0		Ani 01 a	01	A	
Nunmul	Nunmul	0		Nunmul 01 n	01	N	
Heulniurida	heulnida	Orida		Heulnida 00 v	00	v	

were analyzed one by one with reference to the *Standard Dictionary*. Those words with more than ten spellings related one representative synonym are as follows (the number in the bracket shows the types): *eomeoni* (17), *dugyeoni* (12), *abeoji* (12), *jogeumahada* (12), *sanbonguri* (11), *bbeokkugi* (11), *ipsul* (10), *eodi* (10), *gejibai* (10), *haneul* (10), *harabeoji* (10), and *chuseok* (10). For reference, for those words with synonyms, the highest frequency word *eomeoni* appeared a total of 1,163 times and the spellings of the headwords are as follows (the number in the bracket shows the frequency): *mo* (1), *mochin* (4), *ame* (1), *eomani* (2), *eomassi* (2), *eomae* (33), *eomeo* (3), *eomeoni* (804), *eomeonim* (77), *eomeoi* (8), *eome* (38), *eomuni* (3), *eomunim* (1), *eomma* (150), *eommae* (8), *omae* (28), and *omma* (1).

For example, the following shows the contents of the concordance of *Jindallaekkot* (Azalea) written by Kim So-wol through the analysis of morphemes, the analysis of homonyms, and the fixation of the representative synonym.

3. Statistical Analysis of Poetic Diction

After compiling the above database on Korean modern poetry, I have approached several studies in a quantitative way. First of all, I tried to study on the quantitative phenomenon of formal constituents such as the stanzas, lines, words and syllables of poems through *Hyeondaesiui gyeryangjeok munche yeongu siron* (A Tentative Study on Quantitative Literary Styles of Modern Poetry) in 2000 (Kim Byong-sun 2000). Through a humble and average survey, it was revealed that a modern poem is composed of four stanzas, four lines per stanza, and about twelve syllables per line. It was also revealed that modern Korean poetry has something to do with the tradition of the styles of songs coming from the styles of *chang-ga*.

Since 2002, after this study, I have tried to research *Hanguk hyeondaesie daehan gyeryangjeok muncheron yeongu* (A Study on Quantitative Literary Styles of Korean Modern Poetry) for four years. This project has paid attention to the quantitative phenomenon of poetic words themselves and the forms of modern poetry. In the process, it has produced some fruitful results. In the following, the poetic words of high frequency in modern Korean poetry (up to 30 places) will be analyzed in brief.

Table 5. List of Poetic Words of High Frequency in Korean Modern Poetry

Ranking	Headword	Frequency	Percentage (%)	Accumulated Frequency (%)
1	Na 03 m	11,309	1.8551	1.8551
2	Gada 01 v	5,126	0.8408	2.6959
3	I 05 e	4,635	0.7603	3.4562
4	Hada 01 v	4,525	0.7423	4.1958
5	Eopda 01 j	4,401	0.7219	4.9204
6	Geot 01 nd	4,154	0.6814	5.6018
7	Geu 01 e	4.081	0.6694	6.2713
8	Neo 01 m	3,939	0.6461	6.9174
9	Oda 01 v	3.551	0.5825	7.4999
10	Itta 00 j	3,217	0.5277	8.0276
11	Han 01 e	3.136	0.5144	8.5420
12	Bam 01 n	3.102	0.5088	9.0508
13	Ulda 01 v	2,721	0.4463	9.4972
14	Sori 01 n	2,639	0.4329	909301
15	Sok 01 n	2.607	0.4276	10.3577
16	Ttae 01 n	2,583	0.4237	10.7814
17	Maeum 01 n	2,481	0.4070	11.1884
18	Doeda 01 v	2,341	0.3840	11.5724
19	Uri 03 m	2,339	0.3837	11.9561
20	Haneul 01 n	2,225	0.3650	12.3211
21	Boda 01 v	2,196	0.3602	12.6813
22	U 05 n	2,167	0.3555	13.0367
23	Itta 00 vd	2,100	0.3445	13.3812
24	Geu 01 m	2,081	0.3414	13.7226
25	Saram 00 n	2,015	0.3305	14.0531
26	Gatta 00 j	2,014	0.3304	14.3835
27	Gil 01 n	2,002	0.3284	14.7119
28	Geudae 00 m	1,998	0.3277	15.0396
29	Boda 01 vd	1,932	0.3169	15.3565
30	Gaseum 01 n	1,904	0.3123	15.6689

The types of poetic words were 41,632 in all. The highest frequency word as shown in the Table 4 is the pronoun *na* (I or me). This word appeared two-and-

a-half times as many times as the second place word. If a related word *nae* (1,492 frequency) is added, the gap becomes enormous. In common sentences of everyday language, the pronoun *na* takes nothing more than eighth place and only 40% of the lowest frequency restricted noun *geot*. Considering this phenomenon (The National Institute for the Korean Language 2002), it shows that it is one of the characteristics of poetic works as compared with common sentences. We can come to a tentative conclusion that modern poems are poems about *na*, the speaker of the poem or the poet itself, and it shows well the lyrical characteristics of modern poetry. Besides this, among the words above the high 30%, one-syllabic Korean native words such as *hae* (the sun), *dal* (the moon), and *byeol* (a star) take higher places. Also, those words that make pairs semantically, such as *gada-oda* (go-come), *itta-eoptta* (exist-not exist), *ulda-utta* (cry-laugh), and *gibbeuda-seulpeuda* (be pleased-be sad) belong to higher frequency words. This is revealed as another characteristic.

Future Tasks and Prospects

1. The Expansion of the Lists and Original Texts

The present database was constructed from the lists of collections photo-printed in the *Compilation* and *Outline*. From now on, the lists of collections and poems that are not contained in these materials will be expanded to include the 1980s.[50] Moreover, the information on the location of the materials in the concordance will be included. Also, the concordance of magazines related to literature (literary coterie magazines, literary magazines, journals, and literary pages of newspapers) will be constructed. At present, the concordance of literary articles of *Gaebyeok* (A New World) and *Jogwang* (The Morning Sunlight) has been constructed and will be expanded upon in the future. To do this, it is necessary to search the possessions of libraries and personal collections. In fact, it will be a

50. It is also necessary to construct a database of information on the publication of poetry in *Munyeyeongam* (The Literary Yearbook), published once a year by the Korean Culture & Arts Foundation. At present, it is partially available on the internet. But the system is a rough one. The format becomes different every year. The search system is not served well. It can be searched for with only the title of *A Handbook* and is of little value. Moreover, materials before 1975 have not been computerized yet.

project demanding a heavy budget.

The corpus of modern poetry will also be expanded. The first corpus of 1923-1950 will be followed by the second one of 1950-1980. The poetic works of 1951-1960 will be treated as the next step. Afterwards, the corpus of modern poetry will be expanded every five years.[51] An extra corpus compiling only translated poems will also be constructed.

2. Qualitative Improvement of the Corpus

When compiling poetic works in the corpus of modern Korean poetry, textual criticism was followed without any satisfactory results. As several principles on fixing definitive editions were stated previously, the literary materials should be expanded until after 1950 together with a full investigation into the lists and original texts. In particular, definitive editions should be fixed through common efforts rather than personal and fragmentary efforts. The fixation of definitive editions of poetic works should be incessantly continued and the association for individual poets also be established. When scholars majoring in a certain poet establish an association and a working group, operate a website and fix standard lists and original texts, the database of modern Korean poetry will be complete.

A variety of materials, information, and supplementary comments related to fixing the revised edition resulting from the activities of such a working group should be opened. At the same time, the literary works contained in magazines and coterie magazines should be referred to. In the corpus of modern poetry, these works were not referred to.

The percentage of errors of the present corpus is assumed to be about 0.03%. These errors can be divided into those of input of the original texts, lemmatization, and morpho-semantic analysis. Therefore, 0.03% of all 600,000 words total 180 errors. It also means that this corpus is an excellent one of good quality. Incessant efforts to decrease the rate of errors will be continued. To do this, a feedback system to adapt the existing analysis to a new one will be developed and a mutual reference system applied to the nGram technique will also be developed. In the future, I hope that even the analysis of polysemy will be made satisfactorily.

---

51. As a time factor was added to basic word items when constructing the database, a synchronic and diachronic study will be possible. At first, the 1923-1950 period will be set and studied. That period will also be subdivided. According to the output of the database, the quantity of literary works increased every five years in 1923-1950, except for the period 1923-1930.

3. The Information Processing and the Development of a Service System

The database of modern Korean poetry contains all the materials in the Korean language. It is necessary to expand this to serve as useful information worldwide. It is not efficient to translate the whole database into foreign languages. Only keywords and headwords will be translated into English. By means of this database, a comparative study on the poetic works of other culture areas will be possible.

Also, it is necessary to develop a system that will enable us to drive effectively forward the projects of corpus, indexes, and statistics. At present, the projects from the input of the works to the indexes and databases of the concordance are not correlated together. It is necessary to prepare a plan for automating the input of the works, normalization, and lemmatization (sampling of basic forms, etc). To do this, it is also necessary to use positively an XML document form and prepare a feedback system to adapt the existing results of analysis to a new analysis of the works. At the same time, it is necessary to standardize various analytical projects. It is also necessary to prepare a plan for co-working through the internet.

I will try to promote the connecting of information mutually by means of studying the compatibility of information on the explanation of the dictionary with the *Standard Dictionary* of the National Institute for the Korean Language. A plan should be prepared for serving the database on modern poetry to another database through a website. Accordingly, a system should be established to input and output immediately varied statistical information on the materials of the database on modern poetry. And this should be served through the internet at the same time. The present system of the MS Access database will be hugely expanded in its scope, and it is advisable to use another related database based on XML. At present, the web server to use MS SQL is under study and the development of an exclusive search engine for modern poetry is being considered.

4. The Quantitative Phenomena of Works and Their Interpretation

It seems that my database on modern poetry is well laid out in its own way and has varied usefulness. Moreover, the poetic words of this database have the information on the period of publication,[52] and a part also provides spatial infor-

52. They have basic information on the period of publication of the poetry with only confirmed information on the period of creation.

mation, that is, the place of publication. If, in the abstraction of statistical information, the information on time can be used as a variable, the analysis of diachronic phenomenon will be possible on an axis of time-series.

Though the database on modern poetry was constructed after due consideration quantitatively and qualitatively, a study on how to make good use of it has not yet been carried out. How and in what aspect can quantitative phenomenon be measured? Can literary explanation be possible by means of the materials of revealed quantitative phenomenon? What relation has the verbal phenomenon of poetry to that of everyday common language? Can it be possible to analyze themes by quantitative phenomenon? Theoretical backgrounds to answer these questions should be prepared. Together with the considerable development of computer linguistics, a linguistic analysis on the corpus of modern poetry will be possible. Accordingly, a deep analysis on poetic diction and syntactical and semantic structure will be possible. Thus, the day will possibly come when a poet's secret of poetic creation can be ultimately revealed according to mathematical models.

In spite of various possibilities, a statistical and linguistic analysis on literary works has not been actively carried out anywhere in the world. Because of the difference of languages in various culture areas, it is very difficult to adapt recklessly a statistical method of a culture area to the literary works of another culture area. Accordingly, we are entrusted with a mission to develop delicate methods of analysis and explanatory abilities through consistent study.

At present, I put emphasis on a quantitative analysis of the results of the construction of the database.[53] I will study into what meaning it will be interpreted later on. Furthermore, I will investigate the stylistic fingerprints of a poet. Just as everyone has his own unique fingerprints, so does a poem. I will find out through statistical phenomenon how they are constituted. The analysis of the characteristics of the use of words will help discover the fingerprints. In particular, the analysis of particular poetic words used by a poet will play an important role. In a word, I aim at finding out an idiolect of a poet.

This does not satisfy fully the explanation of individual stylistic fingerprints. The analysis will be expanded into the analyses of the characteristics of sounds

53. I have abstracted information on statistical materials and frequency of the number of literary works, words and syllables according to each genre, poet and period and information on interrelations, average and distribution.

used in poems at the level of phoneme and pronunciation, the characteristics of the construction of lines and stanzas in a poem, and the syntactical characteristics of poetic sentences. It will be necessary to study a plan for showing a poetic world by drawing a meaning map of poetic words and showing them as coordinates on the map. Also, the theory and instrument to show various linguistic and semantic phenomena visually will have to be developed.

Furthermore, I aim to abstract the stylistic DNA of a poet. Through explaining the literary styles of poets objectively relating to the phenomenon of their use of words, I shall investigate into the unique inherence and relationships with other poets. I shall also research a plan for visualizing these results.[54]

Finishing this thesis, I will explain the usage method of the database on modern poetry. This database on modern poetry is conditionally opened. *Hyeondaesieo yongryesajeon* (Dictionary of Usage of Modern Poetic Diction) is being served on the internet for a fee, and the information on the lists of modern poetry is for free. I would like to serve as many materials as possible, but I am sorry I cannot because of an inadequacy of various given conditions. In particular, the service of the information on the original texts will be a violation of the Copyright Act. The database can not be reconstructed according to each researcher's demand. For researchers who want to make intellectual and complex searches on this database, the materials will be served according to the following conditions.

1. Materials will be served free on demand for scientific research. However, your contribution will be expected to a certain degree to expand and improve the database on modern poetry. For example, the input of materials that have not been inputted, donations of materials, the proofreading of already inputted materials, etc.
2. When quoting the processed results of your study, it is cordially requested to disclose the source (The Database of Korean Modern Poetry constructed by Kim Byong-sun) of your quotation.

54. My task is limited mainly to the fixation and input of the works and their basic constitution and analysis. This task tends to depend on information technology. Afterward, together with the future development of the technique of processing such linguistic information, this database will become better. First of all, the general development of the techniques for processing natural languages is expected. The techniques of analyzing language structure, constructing the thesaurus, and compiling an electronic dictionary will have to be secured.

3. In case of request for a special search, it may be desirable to conduct collaborative research with me.

References

Bae Hee-suk. 1999. "Munhakjakpumui yangjeok bunseokgwa computer ui hwalyong" (A Quantitative Analysis of Literary Works and Use of Computer). *'99 chugye haksulbalpyohoe jeongujaryojip* (A Collection of Research Materials in an Academic Meeting, Autumn 1999). French Association of Culture and Arts.

Butler, Christopher. 1985. *Statistics in Linguistics.* Oxford: Basil Blackwell.

Kim Byong-sun. 1992a. *Gugeowa computer* (The Korean Language and Computer). Seoul: Hansil Publications.

_____. 1992b. "Chaepteo 8: Munheon jaryo saeginui weolli" (Chapter 8: The Principle of the Indexing Literary Materials). *Gugeowa computer* (The Korean Language and Computer). Seoul: Hansil Publications.

_____. 1993a. *Sowolsiui eohwiwa geu sseuimsae III* (Poetic Words and Usage of Kim So-wol III). Seoul: Hangukmunhwasa.

_____. 1993b. "Hyeondaesi database guchuke gwanhan yeongu" (A Study on Constructing the Database on Modern Poetry). *Hangugeomun Vol. III* (The Journal of Language and Literature of Korea Vol. III). Seongnam: The Academy of Korean Studies.

_____. 2000. "Hyeondaesiui gyeryangjeok munche yeongu siron" (A Tentative Study on Quantitative Literary Styles of Modern Poetry). *Haksuldaehoe balpyononmunjip* (A Collection of Papers in Conference). The Society of the Korean Language and Literature).

_____. 2001. *Hanguk hyeondaesieo yongryedaesajeon* (CD-ROM) (Grand Dictionary of Usages of Korean Modern Poetic Diction (CD-ROM)). Seoul: Nuri Media, Ltd.

_____. 2002a. "Eomunhak jeonjadoseogwan network guseong" (The Construction of a Network of a Digital Library of Linguistics and Literature). *Gugeo yeongu jaryo guchuk 1* (The Compilation of Research Materials of the Korean Language 1). Seoul: The National Institute for the Korean Language.

_____. 2002b. "Hyeondaemunhak yeonguui jeongbohwa - hyeondaesi yeongu jeonjadoseogwan guchuk bangan" (Informatization of Research on Modern

Literature - A Plan for Constructing a Digital Library for Studying Modern Poetry). *Hangugeowa jeongbohwa* (The Korean Language and Informatization). Seoul: Taehaksa.

_____. 2004. "Hanguk hyeondaesi database ui guseonggwa geu hwalyong bangan" (The Construction and Plan of the Database of Korean Modern Poetry). *Hanguk eoneomunhak* (The Journal of the Korean Language and Literature) Vol. 53. Seoul: The Korean Language and Literature Society.

Kim Byong-sun and Jeon Jeong-gu. 1993. *Sowolsiui eohwiwa geu sseuimsae* (Poetic Diction and Usages of Kim So-wol) I-II. Seoul: Hangukmunhwasa.

Kim Hung-gyu. 1993. *Songgangsiui eoneo-- Computer cheorie uihan siga yonglye saegin yeongu* (Poetic Diction of Songgang-- A Study on the Concordance of Usage of Poetry Processed by Computer). Seoul: Korea University Press.

Kim Hung-gyu, Kim Byong-sun, and Wu Eung-sun. 1992. *Hangeul yetgeuljaui computer cheori hyoyulhwa bangangwa ie uihan gosijo database system gaebal yeongu* (An Effective Plan for Computer Processing of Old Hangeul and a Study on Developing a Database System for Old *Sijo*). Seoul: The Korean Computing Society.

Kim Hi-chan. 2000. *Hangugeo malmungchiui gyeryangjeok cheori jeolcha yeongu* (A Study on Quantitative Processing of the Corpus of the Korean Language). M.A. Dissertation, Seoul National University.

Kim Jin-young. 2002. "Pansori jaryoui database guchuk hyeonhwanggwa yeongu jeonmang" (The Present Condition of Constructing the Database of Pansori Materials and Research Prospects). *Hangugeowa jeongbohwa* (The Korean Language and Informatization). Seoul: Taehaksa.

Kwon Young-min, ed. 1990. *Hangukgeundaemunindaesajeon* (Grand Dictionary of Modern Literary Figures in Korea). Seoul: Asea Munhwasa.

Lee Sun-young, ed. 1990. *Hangukmunhak nonjeo yuhyeongbyeol chongmong-nok* (The Total List of Types of Writings in Korean Literature) I-III. Seoul: Hangukmunhwasa.

Muller, Charles. 1992. *Initiation aux Mēthodes de la Statistique Linguistique.* Translated by Bae Hee-suk, 2000. Seoul: Taehaksa.

Seo Sang-kyu and Han Young-gyun. 1999. *Gugeo jeongbohak ipmun* (An Introduction to a Study of Korean Language Information). Seoul: Taehaksa.

The Institute of Humanities Information, Seoul National Univ., ed. 1999. *Hanguk hyeondaemunhak 100 nyeon* (100 Years of Korean Modern Literature). Seoul: Munhak Sasangsa Inc.

The Institute for Korean Culture, Korea University, ed. 1996. *Jeonja sijip, hangugui hyeondaesi (CR-ROM)* (Electronic Anthology, Korean Modern Poetry (CD-ROM). Seoul: Daehan Printing and Publication Co., Ltd.

The National Institute for the Korean Language, ed. 2002. *Hyeondae gugeo sayong bindo josa* (Research on Frequency of Use of the Modern Korean Language). Seoul: The National Institute for the Korean Language.

Yoon Ju-eun. 1991. *Kim Sowol siui eohwiwa geu hwalyong gujo* (The Poetic Diction of Kim So-wol and Its Structure of Use). Seoul: Hakmunsa.

Internet

www.koreanpoetry.net
www.korean.go.kr
www.krpia.co.kr

Kim Byong-sun is Professor of Korean Literature at The Academy of Korean Studies. As a scholar who is interested in modern poetry of Korea as well as digitalization of Korean studies materials, he is an active member of the Council of Evaluation of Information Processing. He was a visiting scholar at the Digital Library Research Laboratory at Virginia Tech in the U.S. www.aks.ac.kr/~kimbs