Development of the Integrated Concordancer for the Corpus of the 17th to 19th Century Culinary Manuscripts

Doo-hyun Paek
Kil-im Nam
Mi-hyang Lee
Eui-jeong Ahn
Hyeon-ju Song
Kyungbook National University

The aim of this project is to develop the Integrated Concordancer for food-related terms used in Korean culinary manuscripts from the XVIIth to the XIXth century. The Integrated Concordancer may be utilized by Korean linguists who wish to make use of culinary manuscripts as research materials for the history of the Korean language. Additionally, it might be useful for culinary scholars of traditional foods, and also for the general public. The tasks of the current project are twofold. The task is, firstly, to construct a corpus by collecting hand-written culinary manuscripts written between the XVIIth century and the XIXth, and develop a web-based search engine. Secondly, to extract headwords of everyday words from the corpus of the XVIIth to XIXth century manuscripts and compile a source book for traditional culinary terms by making and utilizing concordance data by frequency, part of speech, and semantic pattern. The current project is a two-year project-starting in August 2007, ending in July 2009-which will eventually become available to the public.

1. Introduction

In the last ten years, there have been a lot of corpus building projects in Korea including "The twenty-first century Sejong Project", which has been commissioned by the government.[1] These projects have contributed to the development of electronic dictionaries and the Korean wordnet, the establishment of ontology, and much more. Furthermore, they have been utilized to make various commercial products. However, the corpus building projects that have been undertaken in Korea so far have only focused on a general purpose corpus. Accordingly, there is a need for studies on the establishment of a diachronic and specialized corpus and the methodology of its application, which the current project aims to achieve.

This project is based on the construction of a historical corpus. It is of great significance the fact that for this historical corpus of a specialist area will be constructed and annotated in a way easy ways for the general public to use.

The purpose of the current project is to develop the Integrated Concordancer for food-related terms used in 17th to 19th century Korean culinary manuscripts. The Integrated Concordancer may be utilized by not only culinary scholars of traditional foods and the general public who are interested in traditional Korean foods but also by Korean linguists who wish to make use of culinary manuscripts as research materials for the history of the Korean language. The tasks of the current project are twofold: (1) to construct a corpus by collecting hand-written culinary manuscripts written between the XVIIth and XIXth century and develop a web-based search engine, and (2) to extract headwords of everyday words from the corpus of the XVIIth to XIXth century manuscripts and compile a source book for traditional culinary terms by making and utilizing concordance data such as frequency, part of speech, and semantic pattern.

Unlike printed books, the manuscripts written in Han-geul from the XVIIth to XIXth century were mostly written by women, and accordingly, the text is characterized by life- and culture-related content and realistic descriptions. To date, the main materials for the history of the

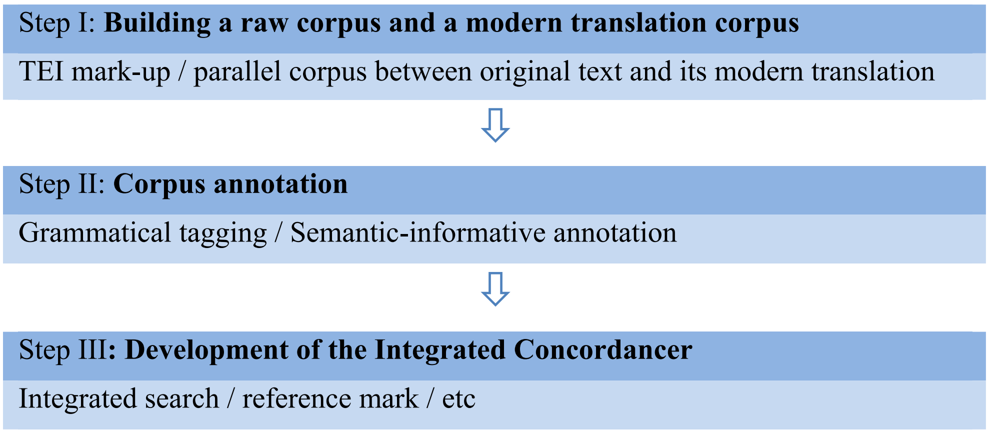
[1] * King Sejong is the one who invented Han-geul, the Korean alphabet. *The twenty-first century Sejong Project* is a Korean corpus building project in 10 years (1998-2007), that is a leading national project.

Korean language are XVth century literature and printed books that were printed after the invention of Han-geul (1443). Since most of the printed books were Buddhist books, Confucian books, translated books, medical books, wordbooks, and educational books, they failed to reflect everyday language and various social aspects. On the contrary, the cooking manuscripts written in Han-geul contain more varied contents than the printed books, and have a rich expression, which makes the manuscripts valuable. The current project will contribute to offering Korean linguists and the general public detailed information about the recipes of traditional foods that were available to the XVIIth to XIXth century nobility in Korea. The current convention for storing historical data of the Korean language is that the data is published sorted by an ascending or descending order of keywords. While this concordance data has academic value, it is difficult for the general public to use this data. The Intergrated Concordancer of this project overcomes this problem as it can search in all the areas of culinary terminology, including old Korean and its modern translated language.

This project is funded as part of a Basic Research Project by the Korea Research Foundation, which is a government-run organization in Korea that supports academic research and domestic/international academic exchange or collaboration.

2. Plan of project progress

The current project consists of three major phases: (1) building a raw corpus, (2) tagging the corpus, and (3) developing the Integrated Concordancer. The three phases are outlined in Figure 1 below.



Step I: **Building a raw corpus and a modern translation corpus**

TEI mark-up / parallel corpus between original text and its modern translation

⇩

Step II: **Corpus annotation**

Grammatical tagging / Semantic-informative annotation

⇩

Step III**: Development of the Integrated Concordancer**

Integrated search / reference mark / etc

Figure 1. Major tasks in each of three research phases

The steps I and II for corpus building will be introduced in section 2.1. and 2.2. and the development of the concordance engine and final results will be introduced in section 3.

2.1. *Building a raw corpus and a modern translation corpus*

Several steps are needed to use old manuscripts written in Han-geul. The first stage is to decipher the manuscripts accurately. The decipherment of the original text is confirmed through discussion among researchers of the text reading, and then passed on be reviewed by expert. Also, the data input is reviewed meticulously to ensure there are no typographical errors.

At this stage the data is a modern translation, which will be paralleled with the original text in the long run, is to be inputted. In addition, a separate file lists the notes for important but infrequent terms including names of foods, ingredients, cookers, and measuring units. Figure 2 below illustrates a part of *Jusikshiui*, which introduces 99 recipes written in the 19th century by a noble family.

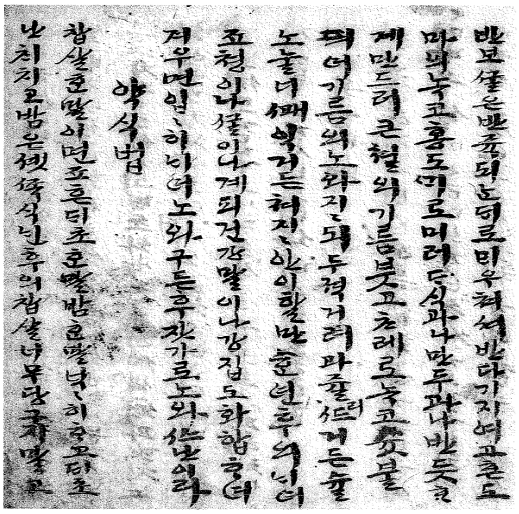


Figure 2. An example of the original text of *Jusikshiui* (written in 19th century)

The next requirement after deciphering the texts and inputting in the data is to record text-external information so that it can be used in addition to the electronic documents as well. To do this, DTD study concerning Old Korean has been conducted, based on TEI. The bibliographic information of the data is placed in the header of the file; the bibliographic information includes the information about the original text, the decipherment, the input, pages, researchers, and so on.

The DTD study about Old Korean requires writing a mark-up manual. The standardization of mark-up helps researchers to exchange their data and expand their corpus more conveniently.

This study only used relevant elements among TEI P3 tagsets, and annotation tagging is based on XML.

The target data of this study is a parallel corpus that maps the original text to its modern-translation. Therefore, work is needed on how to match two parallel sets of data.

In the case of a parallel corpus, it is not desirable to have three separate files, which are an original text file, translation file and indexing file, in terms of programming. Hence, a single file needs to contain all the information about the original text, translation, and index, which is what the authors of this paper are working on.

2.2. *Corpus annotation*

It is impossible to annotate the large corpus completely manually, so this study attempts to develop a tagging program for the historical corpus, which is a semi-automatic tagging program. The first step to make this program is to annotate a sample text manually and make a preliminary dictionary based on the annotation. The next step is to build a model based on the list made in the first step and apply the model to the program, as a so-called example-based analyzer.

This tagging program carries out the analyses of word classes. The example-based analyzer makes it possible for annotators to find and tag unknown or unanalyzed words after the first analysis, which will lead to perfect analysis.

Next, morphologically-annotated culinary books need to be semantically annotated. It calls for the creation of a tagset of semantic-informative annotation, which in turn requires a study about WordNet, Thesauruses and Ontology. In particular, a study about the terminology of culinary science is very important. These studies enable the establishment of a tagset for semantic-informative annotation.

The major target of the current study is culinary terminology. Therefore, most words are hyponyms of *FOOD*, and they can be classified into the following: food name, ingredients, the five senses, cooking method, degree, tableware, cooking instruments, measuring unit, serving manner, storage (method, place and duration), and the best season for cooking.

Table 1 below presents some examples of semantic tagset.

food name	dumpling (*mandubeop*), a kind of rice cake mixed with chestnut (*bamseolgitteok*), a kind of light nutrious liquor made of a little malt (sogukju), plum vinegar (*maesilcho*)
ingredient	rice (*ssal*), flour (*milgaru*), ferment (*sulmit*), soy sauce mixed with vinegar (*choganjang*)
5 senses	bitterish (*ssabssalhada*), sweet (*dalda*)
cooking method	to steam (*jjida*), to parch/roast (*bokda*), to boil (*samda*), to slice (*jeomida*)
degree	to be soft (*mureuda*), to get lumpy (*manguri saenggida*), to congeal (*eonggida*)
tableware	spoon (*sutgarak*), plate (*jeopsi*), vessel (*geureut*)
cooker	strainer (*che*), earthenware steamer (*siru*)
measuring unit	18 liter (*mal*), 1.8 liter (*doe*)
serving manner	to serve cold (*siwonhage naenda*)
storage	to tie and hang (*maeeoduda*)
best season for cooking	January (*jeongwol*), December (*seotdal*)

Table 1. Examples of semantic tagset

2.3. *Development of the integrated concordancer*

The search engine that the authors are working on is designed for both linguists and the general public who are interested in traditional Korean food and culture. However, since the data is from 17th to 19th century manuscripts, it is hard for people who are not Korean linguists to be able to understand them only by looking at the original text. Therefore, the search engine should carry out an integrated search rather than separate searches of grammatically-tagged corpus, semantically-annotated data, examples, and wordlist.

Furthermore, this program will be developed as a web application so that it can be used by anyone who has Internet access, without installing special software.

3. Characteristics of the integrated concordancer

As mentioned in Section 2, a raw corpus turns into a parallel corpus between the original text and its modern translation, which is broken into arbitrarily determined sentences. Whether a user types in an Old Korean word or a Modern Korean word, he or she will get the same result. In order to make this possible, a wordlist for Old Korean and its modern translation will be created along with a major explanation, and this dataset will be operated in the system internally.

As described in Section 2.4, the major culinary terms are tagged with semantic annotation, which can be retrieved with exemplar sentences.

The examples from the corpus have citation information like other concordancers. In the current study, the citation information includes recipe, text title, and page number.

Since non-experts would not be familiar with technical terms, they may want to search only food names which they are interested in. This would be possible by selecting a list of food names under the program menu.

This search engine has the function of selecting the corpus type. When a raw corpus is selected, it is possible to search by word form. When a grammatically tagged corpus is selected, it can be searched by word class or word form as well as morpheme.

The exemplar sentences that are retrieved by keyword search include tags that offer textual or extra information such as TEI tags. However, this information is unnecessary when one is only interested in exemplar sentences. Therefore, the suggested program will be designed so that this unnecessary information will not appear through filtering.

4. Conclusion

The current study should contribute to understanding the life style of Koreans and be applicable to modern lives. In addition, it will provide various traditional recipes that have been forgotten in modern Korean society. The examples extracted from the Concordancer will be used for the compilation of a glossary of culinary terminology.

References

Barlow, M. (2002). "ParaConc: Concordance Software for Multilingual Parallel Corpora". Language Resources for Translation Work and Research.

Gale, W. A.; Church, K. W. (1993). "A program for aligning sentences in bilingual corpora". *Computational Linguistics* 19.

Kennedy, G. (1998). An *introduction to corpus linguistics*. London: Longman.

Kil-yong, S. (2006). "About the Tagging Tool for the Concordances to Doklipsinmun Corpus". *Journal of Korealex* Volume 7. Korealex.

Kytö, M.; Rissanen, M.; Wright, S. (eds.) (1994). Corpora across the Centuries-Proceedings of the First International Colloquium on English Diachronic Corpora. Amsterdam: Rodopi.

Martínez Magaz, J. (2006). "TRADI IMT (xx-xxi): Recent proposals for the alignment of a diachronic parallel corpus". *ICAME Journal* 30.

McEnery, T.; Wilson, A. (1996). *Corpus Linguistics*. Edinburgh: Edinburgh University Press.

Rissanen, M. (1989). "Three problems connected with the use of diachronic corpora". *ICAME Journal* 13.

Sinclair, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Svensén, B. (1987). *Practical Lexicography*. Oxford: Oxford University Press.

Zgusta, L. (1971). *Manual of Lexicography*. The Hague: Mouton.