Sangsup Lee, Lexicographical Center, Yonsei University

The Yonsei Corpora of Korean and Lexicographical Projects

ABSTRACT: Koreans need a dictionary that will do full justice to the phenomenal changes in the language in recent decades. The Yonsei Lexicographical Center has structured several corpora of Korean, totalling over 15 million tokens to date, collected on the basis of a survey of the reading habits of adults, library lendings, and lists of important books made by Publishers Cooperative, and has developed computational tools specially keyed to process the agglutinative character of Korean. The corpora are the basis on which a dictionary worthy of the electronic information age is being compiled.

1. Some problems in Korean lexicography

The Korean language, spoken by almost seventy million people, is one of Northeastern Asian languages, whose kinship with other languages remains still very speculative. The prevailing mode of its grammatical representation by modern scholars has been based on Western, more specifically, Anglo-American, linguistic concepts and categories, and has always been rather strained. For example, most Korean adjectives behave very much like verbs, a fact that makes linguists hesitate to explain them in terms of Western definitions of the two categories.

Another case in point is the agglutinative characteristics of the Korean language. These indeed pose the greatest difficulty to linguists whose orientation lies in the grammar of inflections and well-defined case frames of Indo-European languages. There are legions of other linguistic peculiarities in Korean which are not readily amenable to the adopted framework.

Likewise all Korean lexicographical practices so far seem to have failed to do full justice to the language largely because they have followed Western lexicographical conventions. In Western dictionaries, the infinitive form of a verb is entered as the headword. This practice seems to have been understood by pioneering Korean lexicographers as a prescription to invent "infinitive" forms for Korean verbs. The result is the purely artificial constructs purporting to be "infinitive" forms of verbs which, of course, can never occur in natural language situations. Although they are used only in theoretical and meta-linguistic discourses, they have become an indelible part of every educated Korean's linguistic conscience. Every child is taught this "new speak" in order to follow instructions in grammar and to use dictionaries. The made-up forms are referred to as "representative," not "infinitive," but it is never clear what those artificial constructs are representative of.

Unlike verbs, adjectives are entered in natural forms, that is, the stems of the adjective plus declarative predicate end marker are entered which native Koreans feel represent ordinary uses of adjectives. As mentioned above, Korean verbs and adjectives behave so very much alike that the polar difference in the treatment of the two parts of speech cannot help embarrassing learners of Korean using a dictionary.[1]

Since the agglutinative characteristics of Korean cannot find counterparts in any Indo-European languages, their lexicographical treatment has been the thorniest problem, though some theorists have proposed that agglutination be treated like somewhat uncomfortably complex inflection. So far, the extremely high semantic as well as syntactic productivity of the agglutinative elements of the language remains largely unexplained in dictionaries.

According to ordinary lexicographical practices, a word for entry is a discrete string of characters separated from other strings by a blank space on either side when written or printed. But a string of Korean characters with a space on either side is usually the stem of an entry word followed by agglutinative particles, which is referred to as an *ojol* or "word-phrase."[2] Thus the sentence, *naendul ochi hagessumnika?* (roughly, "What can even I do?") is composed of three *ojol*. *Naendul* is an *ojol* where *na* means "I" and *-endul* is the subjective agglutinative particle meaning "even". *Ochi* is an adverb, which, in this case, is without any agglutinative. *Hagessumnika* is made up of the verb stem *ha*-(meaning "do") and the cluster of predicate particles *-gess-*, *-umni-*, and *-ka*. Korean dictionaries usually enter *na* ("I") and *ha*-("do") as may be expected, but as for the particles, such as *-endul*, *-gess-*, *-umni-*, and *-ka*, they are so numerous and their possible combinations so varied and unexpected that many of them and their combinations fail to appear even in the larger dictionaries now available.[3]

However, tracking down all of them may cause far less difficulty than defining each of them. They are incomparably more abstract and elusive than, say, English inflectional morphemes, prepositions and conjunctions, and they continue to turn up in unexpected combinations. Quite an impressive deal of these unexpected combinations, however, strikes the hearer or the reader as most appropriate for the particular occasions. Meanings seem to grow subtler every day, leaving the lexicographer way behind in his drudge, caught in the web of airy particles.

It is no wonder that all learners of Korean suffer the acutest headache over its particles. Dictionaries do not always provide good relief for the obvious reasons. Moreover, the ordinary learner cannot easily separate the stem and the particles: he or she cannot easily look up *ha*-("do") and then each of the particles, *-gess-*, *-umni-*, *-ka*, and combine them into a word-phrase meaning "can (I) do?" addressed, in this case, to a person or persons higher in social position than the addresser. He or she may have learned some Korean word-phrases but much less of what they may encounter in actual situations. This is one of the main reasons that, while there are many fine Target Language-Korean dictionaries, there are so few, if any, good Korean-Target Language dictionaries. Only the most frequently used particles are entered in Korean-Target Language dictionaries, and their counterparts in the target language are suggested in the vaguest way. Except perhaps for Japanese, there seem to be no counterparts to the Korean particles in any other important languages of the world.

Like Japanese, Korean has a great infusion of vocabulary of Chinese origin. The situation is in some respects similar to that of English and many other European languages where words from Latin abound. As is very well-known, practically every Chinese syllable or character is a bundle of meanings, so that it is quite easy for a person with some Chinese to make up nonce-words by combining characters in certain ways. This linguistic productivity is all right. But outsiders will be surprised and dismayed when they learn that this productivity is a major cause of lexicographical chicanery in Korea. Since the sheer volume of vocabulary entered in a dictionary is popularly recognized as the most important single measure of its worth, dictionary makers everywhere are ever so keen to include as many words as possible. So are Korean dictionary makers, with a vengeance. A recent one-volume Korean dictionary claims that it treats 450,000 words. An instantaneous check confirms the high visibility of spurious words made up of Chinese characters that serve to aggrandize the mere number of headwords in the dictionary. The majority of headwords of Chinese origin which are given a single, simple definition are such nonce-words, neither actually used, nor palatable enough to induce people to use.

2. The Yonsei Corpora

The Lexicographical Center at Yonsei University was organized in 1986 for the express purpose of compiling a series of dictionaries of Korean utilizing contemporary language engineering techniques and availing itself of advances in Korean linguistics. Especially, new trends in corpus-based research have helped define the kind of work a fresh lexicographical project faces these days. The university environment has brought the participation of linguists, computer scientists, information scientists, and psychologists. More than three hundred dues-paying members of the faculty have endorsed the project. Funded mainly by the University and several public research organizations, the Center is structuring a database of large scale corpora of contemporary Korean and developing tools specially keyed for processing the agglutinative characteristics of the language. We at the Center have diligently studied the theory and practice of corpus linguistics and computational lexicography and their applications in producing highly innovative dictionaries in the West.

It is clear that any serious lexicographical project for the Korean language should start with a thorough-going review of such problems as mentioned above. The Lexicographical Center concluded that such a review required large scale data of natural Korean and that the data should be so collected as to reasonably represent the actual language use. The Center launched out on its work with conducting a survey of the average amount of time adult Koreans spend on reading different kinds of printed texts in their daily life. Planned and conducted by a psychologist on the Center staff, the survey was to determine if tentatively the relative value of each kind of printed matter in forming and reinforcing their vocabulary. The Center was provided with the addresses of one thousand randomly chosen adults by Gallup-Korea, Inc., and we sent out specially trained students all over Korea to interview them with questionnaires, which had been prepared through discussions with a group of educationists, editors, writers, and publishers.

The results were very interesting, to us at any rate. Following are the figures the survey turned up showing the relative value of each of the main categories of printed texts. (The value of each subdivision of the main categories is not given here.)

Daily newspapers	36.6%
Magazines	22.8%
Books of fiction	20.4%
Books on general culture and information	11.2%
Biographies	9.0% (100%)

We decided to collect a corpus of contemporary Korean from these text categories keeping the suggested proportion between them. The first batch of about one million *ojol* or word-phrases was from materials published from 1980 through 1985. The next two batches of a million each were from 1986 and 1987 respectively. The three batches were keyed into the computer system and were subsequently structured into a textbase. This corpus of three million contemporary Korean *ojol* is called Yonsei Corpus I, the first of its kind in the country. Frequency lists of such elementary features as characters, syllables and *ojol* were almost instantly produced, and, with the development of computational tools for analyzing the *ojol* into the stem and the particles, a tentative frequency list of words has now come to be available. Our tool is not perfect yet; there are so many homographs and unexpected exceptions, which all need human intervention. KWIC indices of words and of particles are, of course, available.

After a preliminary examination of Yonsei Corpus I, we decided to implement a different principle for collecting another corpus. The information scientist on our staff suggested that we collect material from books most frequently checked out by students in our Central University Library. The books should be evenly distributed among the ten categories of the Dewey Decimal System. The result is Yonsei Corpus II of another million *ojol*. This tidy corpus is our favorite, though its basic configuration is much alike to its predecessor.

We adopted still another principle for a larger corpus. Early in 1990, the Publishers Cooperative of Korea issued a list of several hundred books selected by a group of scholars and critics as representing the achievement of Korean writing during the previous decade (the 1980s). A cultural critic on our staff made a selection of about one hundred books from the list, and these were keyed in in 1991, bringing the total to ten million tokens. As of the end of March this year, we have already keyed in about five million running word-forms from texts published in the 1970s. We foresee that around the end of August of this year, we will have collected ten million tokens from the 1970s material. This will be called Yonsei Corpus V. After that we will build up Yonsei Corpus VI also of ten million tokens from the 1960s material. A great deal of recent publications are available in machine readable forms, saving us lots of time and labor. These and other keyed-in material will make up the 1990s corpus.

It may be noticed that no mention of Yonsei Corpus IV has yet been made. It will be a corpus of spoken Korean, promising a very hard work not to be avoided. We are planning on it. It will be well on its way in August. We will temporarily decrease the rate of collecting corpora at the end of 1993, for, thenceforward, we will concentrate on writing as well as electronically structuring a dictionary of contemporary Korean which will reflect the reality of the language as closely as possible. Relevant language engineering tools and database techniques will have advanced to meet our immediate needs by that time. This is Phase I of our total plan.

In all, the Phase I Yonsei Corpora will comprise a total of about thirty-five million word-forms collected from material of the 1960s through the 1990s. We hope that the exuberant growth of our language during the present generation will be represented in adequate lexicographical terms for world-wide use.

The Yonsei Corpora as they stand now are already attracting commercial as well as research-oriented language engineering organizations and individuals.[4] We may be able to proceed on our projects in the near future largely with funds from interested people who will pay for the kind of information and technology we can provide. However, practical language engineering is a very new concept in our country, even newer than our concept and implementation of corpus linguistics and computational lexicography.

3. Lexicographical projects

Our Phase I objective is to complete a dictionary of contemporary Korean which will be a real aid to foreign as well as native learners of the language. By a learner we mean anybody who wants any bit of information about the language. In this view, an educated adult Korean becomes also a learner when he or she looks up some word or grammatical point in the dictionary. Accordingly our dictionary will have such features of the learner's dictionary as limited defining vocabulary, detailed grammatical information, usage notes, etc. We are now engaged in treating words to be used in defining and explaining the entry words. The treatment of agglutinative particles still poses big problems. We are devising user-oriented ways to represent and explain them.

Its electronic version will be a co-product, developed together with the book version. We look forward to its being incorporated into ordinary Korean word processors. We believe that our work will contribute not a little to the development of Korean machine translation systems. It will also open a new era of bilingual dictionaries in the country, especially the combinatory-explanatory variety.

And our long-term enterprise will be the compilation of an unabridged dictionary of the Korean language on historical principles, a global system which will contain all valuable information about the language and continue to record all the significant changes that may happen in the language. We will expand and consolidate our corpus until it can fully represent the Korean language since the last decade of the nineteenth century when the vernacular was officially decreed to be used in documents, national textbooks, and newspapers.

Endnotes

[1] For example, the verb *ha*-("do") and the adjective *cha*-("cold") are entered respectively as *hata* and *chata*. *Hata* is the "infinitive" form which is never used, whereas *chata* is an ordinary adjective plus predicate declarative end marker meaning "am, are, is cold". Although both words are given the same agglutinative particle *-ta*, its function in the respective forms is entirely different. *Ha-* and *cha-* are never used separately but with particles, so that they are treated as stems.

2 *Ojol*, or "word-phrase," came to be represented graphically only at the end of the nineteenth century. Before that, characters were written and printed without any sort of spacing between them, though punctuation was used rather loosely. It was thanks to learning written Western languages that a few bright pioneers hit upon the idea of the use of spacing between units of words in order to help people catch the meaning more easily and clearly. Chinese and Japanese never adopted this convention of spacing which was a great invention of the Latin Middle Ages. Spacing is not gratuitous at all: for most people it is the chief reminder that their speech is composed of discrete words. Hence its basic importance for lexicography!

3 According to my count of the *ojol* with *o*-("come") as the stem there are 292 different types of them distributed among 5,320 tokens occurring in our monitor corpus of 2 million wordforms. Computational lexicographers will instantly see the great possible difficulty of devising tools for automatic lemmatization of Korean *ojol*. Our computer experts have developed a tool which is not yet fool-proof, but promises reasonably fine performance when further trained.

4 In the near future a monitor corpus of Korean of one million tokens will be contributed to centers of natural language research anywhere in the world. We would like to promote the study of Korean as a major language of the world.

KEYWORDS: the Korean language, agglutinative, Yonsei Corpora, Yonsei University, computational lexicography