Sangsup LEE, Yonsei University

Compiling a Monolingual Learner's Dictionary on Corpus Linguistic Principles: the Case of YLDCK

Abstract

The Yonsei Learner's Dictionary of Contemporary Korean, to be published at the end of this year, has been compiled on the basis of a forty-three million running words corpus. Its features may suggest something of help to similar projects: 1) All the 52,000 headwords appear at least fourteen times in the corpus. 2) All the examples are taken from the corpus, with least editorial tampering. 3) All possible users, whether native or foreign, are treated as learners who will learn something new, even in the case of "easy words". 4) Great endeavors have been made to integrate semantic and syntactic information.

Keywords: learner's dictionary, corpus, frequency, examples, agglutinative language

Early in 1989, I made a round of visiting British lexicographers and lexicographical institutions consulting them on modern lexicographical methods and also confirming my then newly acquired knowledge of lexicography and corpus linguistics. Following their advice and practices, I and my colleagues, none of whom were lexicographers and had never heard of anything about corpus or corpus linguistics, began building a corpus of written and spoken modern Korean on principles we adopted from Western practices. We also studied books and articles on corpus linguistics and lexicography and attended lexicographical conferences such as the EURALEX. At last some of us ended up as lexicographers and corpus linguists. Our corpus has eventually grown to count forty-three million running words, representing the synchronic state of the language at the present as well as its diachronic changes from the 1960s through the 1990s. The vocabulary of Korean has undergone drastic changes during the last forty years to an unprecedented extent keeping pace with the phenomenal growth of Korea's economy and globalization of its political status.

Four years later we were fortunate enough to be funded by a leading publishing firm to compile a 1800-page monolingual learner's dictionary of contemporary Korean. The firm has published various dictionaries, but never a "learner's" dictionary. As a matter of fact monolingual learner's dictionary is an entirely new concept in our country. There are dictionaries that carry the word "for the learner's" in their title. But such dictionaries are usually the shortened versions of the big dictionaries minus the encyclopaedic items. All our monolingual dictionaries assume to be the repositories of all the possible and probable words in Korean, including archaisms and dialects, most of them unattested.

In such a situation we had to find our lexicographical model among foreign dictionaries. Our choice was the sensational Collins COBUILD Dictionary of the English Language, whose editor John Sinclair had kindly seen to my consultation with his staff on my 1989 visit. However, Korean is an agglutinative language as is well-known, making it necessary for us to make many vital decisions on our own. Our electronic lemmatizer is not wholly satisfactory, mainly because of the highly complicated system of agglutination in which verbs and nouns can possibly have hundreds of different agglutinative particles attached to them, generating

completely unsuspected homophones among themselves, which no computer program we have devised so far can cope with completely. Our lexicographic ancestors happened to adopt the Western convention of entering the infinitive form of the verb. They had to invent an "infinitive" form for each verb to make it look like its Western counterpart. Being completely artificial, such a form is never used in any natural discourse except in grammars and dictionaries. However, we could not change this problem-ridden convention.[1]

We were hardy enough to pitch in the "drudgery", which proved to be quite harmful on some occasions. I, for one, had to forgo my overdue sabbatical leave and prolong my work on a book on English critical history, which, incidentally, is the subject I specialize in as a professor of English. My linguistic colleagues had to convince themselves of the validity and even the unavoidability of corpus linguistic way of thinking about our language. In the process a few of them dropped out of our ranks to keep to their long-cherished view of language as "mainstream" linguists. But most of the young scholars we succeeded in recruiting became enthusiastic participants in our project, willingly converting themselves to corpus linguists, lexicographers and computer programmers and operators.

The YLDCK was going to be a virtual invention because not a single one of us had ever had any real experience in dictionary making. For that matter, there are very few, if any, professional lexicographers in our country, where dictionaries are usually compiled by editing personnel in the publishing houses which are said to provide them generously with high-quality glue and scissors. The first full-size original dictionary of Korean was the six-volume Kun Sajon compiled by the Hangul Society during the 1930s and published only in 1956 after many years of delay due to adverse circumstances of the nation. It became the one great source of all the later commercial dictionaries, making all of them look alike to one another. Their mutual difference is secured mainly through competing among themselves to contain the greatest number of headwords, and recruiting the name of an eminent scholar as the supervisor to appear on the spine of the thick volume.

Now, after more than five years of concerted endeavor of some thirty people, our dictionary is finally in the stage of proof-reading. It is in many ways unique even in the light of advanced lexicographical practices of the West, reflecting the characteristics of the language and special needs of the users. Following is an account of some of its most conspicuous features.

1. The headwords, numbering about 52,000 in all, were selected on the basis of their rank in the word-frequency list of our large corpus. This was the very first application to lexicography of the most important corpus linguistic principle in Korea. The frequency count itself was long overdue. It was the first one in more than forty years since a special team commissioned by the Ministry of Education had collected a two million-word corpus and added up the words using their fingers and abacuses. But strangely enough, no editors of dictionaries of that time and later did make any significant use of the results. This pioneering essential work was all but buried, and was ignored even by elementary textbook writers. Our frequency list shows very interesting facts about the language, including the typically Ziphian distribution of the tokens. The most frequent word in Korean turns out to be *kot,* a dependent noun meaning 'thing,' or 'it'. We wonder what is the most frequent word in Turkish, for example. In English it is the definite article *the.*

There may be some people who will question how we came to put the number of headwords at 52,000. As stated above, YLDCK was designed as a one-volume 1,800-page desk

dictionary, the usual size of similar dictionaries. Our repeated test editing proved that, given the amount of information we decided to include in the treatment of each lexical entry, some 50,000 headwords would fill up the space.[2] In the course of actual compilation about 2,000 additional words had to be admitted. The number 52,000 means that all the headwords in YLDCK occur at least fourteen times in our forty-three million token corpus. That is to say, only words showing the frequency of fourteen or above are given a place in the dictionary. We found that in most cases fourteen examples of a word sufficed on which to base the definition of its meaning and the explanation of its grammatical relations.

2. It turned out that about a tenth of all the headwords, excluding proper nouns, usually entered in our existing single-volume dictionaries are not found in our extensive corpus of contemporary Korean. Many of them are long obsolete, about which our dictionaries often give no notification of their being so, and quite a lot of them are spurious. On the other hand, about 5% of our headwords are not found in other dictionaries. It appears that, in the all-out competition for the greater number of headwords, the compilers have not been above inventing new words on the slightest pretext, little extending their effort to find new words in actual use. Coining words is a notorious lexicographical malpractice. Especially in Korea, coining words is quite temptingly easy, because cultured Koreans have long been used to making up nonce-words by compounding Chinese characters. As a matter of fact Korean words of Chinese origin comprise more than seventy percent of the whole vocabulary, like the words of Latin origin in English. That such made-up words are possible but are highly improbable to occur in the language is attested by large corpora such as ours.[3] Hence the widely deplored discrepancy between the actual state of the language and what is pretended to be its reflection in the existing dictionaries.

3. Practically all the examples of the use of a headword were taken from the corpus. Each of the definitions or explanations of a lexical entry is followed by up to three examples from our corpus. Although naturally occurring examples are often longer than the usual made-up ones found in existing dictionaries, we tried to keep the natural examples intact as far as possible. However, parts considered to be inessential were deleted to save the precious space and certain sensitive proper names were changed. Some examples are left in the form of incomplete sentences, unlike the tidy phrases and neat sentences of made-up examples. "Difficult" words in the examples were left as they were found in the corpus in the belief that the naturalness of their occurrence should be an additional information for the user. Some of them may even be outside the range of the 52,000 words entered in our dictionary. We have not yet cross-checked all the words appearing in the examples at the time of writing. When we have finished all proof-reading we will make a full assessment of such matters.

4. We did not adopt the policy of limited defining vocabulary—for this edition at least. We considered the matter very hard but the idea of devising a language for use in the dictionary did not seem to be in keeping with our corpus linguistic position. The most simplistic idea of a limited defining vocabulary is that the first, say, 2,000 words on the frequency list will be appropriate for such a job. But this is of course naïve to say the least. A dictionary is a special type of text that sometimes needs words of considerably low frequency. However, we tried very hard to avoid using "dictionary language" and sound as natural as we could. We were fully aware that defining vocabulary, limited or otherwise, tends to be a sort of special, that is, artificial, language. So we followed the COBUILD policy.[4] We are not sure that the usual 2,000 word defining vocabulary is really so helpful for the learner, any more than the English

Bible in Basic English substantially increased the number of Bible readers among new proselytes in the English speaking societies.

5. YLDCK is a monolingual "learner's" dictionary aiming to be of practical help for all learners of the language. By learners we mean all people who use the dictionary, whether native or foreign or whether grown-up or growing up, because we believe that anyone who looks up a word in the dictionary or merely thumbs through it is necessarily a learner or a possible learner. We also believe that they should be helped as such, including those whose primary purpose is to find fault with our dictionary. Besides, I do not see any real rationality in the traditional division of the "ordinary" and the "learner's" dictionaries. Because of the regretted lack of good dictionaries, the ordinary Korean relies far less on a dictionary than, say, an ordinary Englishman, for learning the meaning and use of new words. So ours is primarily for domestic learners, unlike British ones which are frankly for the foreign language market.[5]

All users should be made to acquire some new knowledge about the language, which is ultimately based on the extensive corpus. On this principle we tried to offer as detailed semantic and syntactic information of a word as a single-volume dictionary can afford the space for. This claim is borne out, among others, by the fact that such "easy" words as *ka*, to go, and *o*, to come, are given several pages of treatment in our dictionary. This is in direct contrast to the larger existing dictionaries which allow a quarter of a page only very sparingly to a complex lexical item.

6. Detailed syntactic and collocational information is another strong feature. As an agglutinative language, Korean is especially rich in morphemic particles that attach to practically any word-stem to show its grammatical-lexical relationship to the other elements of a text. The particular system of Korean grammar developed over a long period by linguists at Yonsei University, widely acknowledged as the center of Korean language studies, was adapted to suit our lexicographical purposes. Our dictionary amply testifies to the fact that the character of a word is known by the company it keeps. Its company includes the cluster of particles it is surrounded with. We believe that our dictionary is pointing to the (ultimate) integration of the word-book and the grammar-book, the ideal state of the dictionary, which we confess we are short of attaining in this our first lexicographical project.[6] At any rate, we give detailed information about the case marking particles and offer a list of typical words that may fill up the case, with pertinent examples.

7. Adopting the right margin column system initiated by the COBUILD, we extended it to include etymology, irregular conjugation, register, usage notes, case markers with typical case fillers and miscellaneous others, which did not seem to belong with the main definition and example text. Homonyms and antonyms find their place here, too.

Among the margin matters, our etymology seems to deserve a special mention. Since ours is not a full size dictionary, its etymological information is limited to showing the foreign scripts of words of foreign origin which sometimes appear in written Korean, usually in parentheses, i.e., Chinese, alphabet, and Japanese characters. As for the many "false friends," we warn the user very carefully. This is an entirely new dictionary service. For example, *aedubalun*, meaning "a large balloon flown in the air with advertisement written on it or a strip of advertisement attached to it" is usually entered as a English word and is given the spelling *ad-balloon* in Korean dictionaries. Most Koreans believe that it is a good English word, not knowing that it does not exist in the English language. We notify the user that it is a

made-up word using the English forms *ad* and *balloon*. It is intended that the user should not use the Korean invention in his English.

These are some of the features of YLDCK. We believe our experience has something to suggest to those who will apply corpus linguistic principles to a dictionary compilation, especially of a non-I.E. language.

Notes

[1] I (1992) summarized how we built our corpora, and discussed special lexicographical problems arising from Korean being a uniquely agglutinative language. Nam (1992) also gives an in-depth discussion the problem of treating a typical Korean verb as a headword.

[2] Is it merely a coincidence that a typical British monolingual learner's dictionary has around 50,000 headwords? The 1,701 page Cambridge International Dictionary of English (1995), for example, is advertised as containing 50,000 headwords. It might be that the core vocabulary of a language which reflects the advanced state of global culture tends to contain roughly the same number of lexical items. This is very speculative, indeed.

[3] I am of course referring to the famous contrast between the rationalist linguistic "possible" versus the corpus linguistic "probable".

[4] The editor in chief of Collins COBUILD Essential Dictionary says: "The dictionary compilers were asked to keep their explanations simple, but they were left free to choose any words they need. During revision of the dictionary, this aspect was carefully checked". We can say the same.

[5] There are about five million ethnic Koreans living abroad and a growing number of foreigners who are learning Korean. But a mere few millions are not enough to make lexicographers pitch in the drudgery of dictionary making. There are seventy million native Koreans and they are the real target of our endeavors.

[6] This is what I understand to be Sinclair's basic position (1991:134).

References

Hangul Society. 1956. *Kun Sajon*. Seoul: Eul Yoo Munhwa-sa.

Lee, Sangsup. 1992. "The Yonsei Corpora of Korea and lexicographical projects", *Euralex 92 Proceedings*.

Nam, Ki-shim. 1992. "Lexicographical treatment of the Korean verb *mandul-* 'to make' ", *Euralex 92 Proceedings*.

Proctor, Paul (Ed.). 1995. *Cambridge International Dictionary of English*. Cambridge University Press.

Sinclair, John (ed.). 1988. *Collins COBUILD Essential English Dictionary*. London and Glasgow: Collins.

Sinclair, John (ed.). 1991. *Corpus, Concordance, and Collocation*. Oxford University Press.