Korean grammatical collocation of predicates and arguments

Byungsun Park, Beom-mo Kang
Korea University

1. Introduction

This paper investigates the grammatical collocation of predicates and arguments in Korean in an objective way through statistical methods. We investigate the statistical meaning of the co-occurrence of the two words and the semantic relationship between them. [1]

In the light of Korean grammatical structure, we study the collocation of the argument and the predicate using statistical measures. The definition of the collocation in Korean is clear; however, on the basis of previous studies, we can say that it is words with adjacency and high co-occurrence relations.

Collocation has been an active field of study in English, and the results of these studies have been used to do part-of-speech tagging and analyse the grammatical structure of large corpora.

Collocation studies of western languages generally focus on words close to each other. However, in studying Korean, on the other hand, more meaningful results are produced by examining the co-occurrence of words in a grammatical relationship within a sentence, such as subject and object.

We first investigate in how far the correlation between the occurrences of two words is statistically meaningful, and then analyse the semantic relationships between them. The Korean language has many auxiliary words used for case marking. For example, the subjective case is marked by the auxiliary words '-i', '-ka'. Korean has developed its case mark system in this way. Therefore, Korean language order is relatively flexible. In this view, rather than the relationship between keyword and adjacency word, the relation between keywords and words that are grammatically related to the keyword is more meaningful for collocation. This paper focuses especially on predicates as keywords.

In this paper, we focus on a methodology for analysing data, so we just introduce Korean grammatical collocation by one word 'boda *–see* ' as an example. The verb 'boda' is extracted from the Korea-1 corpus (H. Kim & B. Kang, 1996) that was balanced on a 10 million word scale.

2. Data analysing by statistics

2.1 Data processing

At first, the data in this paper was selected from the 100 most frequent verbs in the Korea-1 corpus. Then, concordances of these verbs were extract by a concordance package, and the argument of every sentence was marked by hand. An example is given below.

S *seongyosa*-neun amu maleobsi i O chobeol-eul [bo-goman] isseul ppun …
S *missionary-sub without saying anything this* ***O*** *punishment-obj [see(stem)–only(ending)] stand just …*
" The missionary, without saying anything, stood there, seeing this punishment…"
(**S**: subject marker, **O**: object marker)

In this way, the grammatical markers for every concordance sentence were marked, and the frequency of the marked word for each argument extracted.

2.2 Statistical approach

Up to now, a large part of linguistic research has focused on determining whether the expressions in a language are appropriate or not through intuition. This type of research is usually based on the intuition of specifically native-speakers, rather than on objective data. In our research, however, we are interested in finding out what are meaningful collocations not through intuition, but through actual language use. Since the amount of language data used in corpus linguistics is usually large, a statistical

---

[1] This paper is part of the research project in which the authors collaborated with Jong-sun Hong, and Ho-cheol Choe of Korea University.

approach is essential. Moreover, a statistical approach reveals important information for natural language processing and specific data for use in a general theoretical approach.

When two words that have a grammatical relation to each other co-occur and also when the frequency of co-occurrence is higher than we can usually expect, we can say that the two words have a collocation. From this point of view, statistical methods are used in this research. How exactly the statistical measures are used will be described in the following section.

2.2.1 t-score

The t-score shows the difference between expected frequency of co-occurrence in the population and observed frequency of those in the sample. In this paper, the population is the Korea-1 corpus. The bigger the difference, the higher the degree of collocation. If a word occurs x times in the population, and the number of observed words in the text is y, we can calculate the probability of x/y. For example, a word, A, occurs 5000 times in the 10 million corpus, and the size of the observed text is 1000 words, we can expect that a word, A, can occurs 0.5 times in the observed text. If the observed frequency of A is larger than 0.5, it can be said that A is in collocation to the node. However, we consider not only the difference between observed frequency and expected frequency but also the statistical significance. In the character of t-score, the larger the observed frequency compared to the expected frequency, the more exaggerated the t-score is expressed. Consequently, the t-score is not simply comparing the difference between A and B. For example, if the t-score of A is 6.22 and the t-score of B is 3.11, it does not mean that A is two times larger than B in collocation. In this view, statistical significance is important.

The t-score formula is shown below:[2]

$$t = \frac{O - E}{\sqrt{O}}$$

Where:

O = the observed frequency of occurrence of the word within the span

E = the expected frequency of occurrence of the same word

And also, E is calculated like this:

$$E = \text{the frequency of occurrence of the population } X \quad \frac{span\_size}{population\_size}$$

2.2.2 MI-score (Mutual Information score)

The MI-score is different from the t-score when expressing collocation. The t-score is the measured collocation by fixed node and co-occurring word. This means that the MI-score does not express the difference between observed and expected frequencies in terms of standard deviation, but represents the amount of information to which each of the two words, the node and its collocate, are related. So, the MI-score provides information about the words in relation to each other by comparing the observed probability of their co-occurrence with the expected probability, assuming that they are distributed randomly. In the MI-score, it is assumed that the amount of information that each of the two words contains is same. However, in this research, the MI-score measure shows that Korean is very different from English.

Especially collocation studies of western languages generally focus on the words close to each other. we believe, on the other hand, that more meaningful results are produced by examining the co-occurrence of words in a grammatical relationship within a sentence. we therefore attempted to test the MI-score in Korean collocation research, but the result was not appropriate to Korean collocation research.

The MI-score formula is shown below:

$$I = \log_2 \frac{O}{E}$$

3. The collocation of the main argument

In this paper, the collocation is calculated by the statistical measure as previously stated. The population is 10 million words of the Korea-1 corpus. The sample, the span, is calculated based on the

[2] This formula and the explanation were described in Barnbrook (1996).

fact that one sentence in Korean usually consists of 12 words on average (B. Kang 1999, B. Park1997), so the span is 12 * the frequency of concordance.

Even if the score is a negative number, it is also meaningful because negative numbers mean the two words usually do not co-occur. Especially when a negative number also has statistical significance, it gives a meaningful result.

3.1 The collocation of 'boda (see)' arguments

The verb 'boda' needs a subject and an object as arguments. The meanings of the verb 'boda' are 'to see' and 'to induce or deduct'. The verb 'boda,' which means to see, only needs a subject and an object, but when it means to 'induce or deduct,' it needs a subject (nominative), an object, and also an adverbial expression. The result of the statistical measure of this is shown in the table below. Statistical significance depends on the individual research propositions, so, in this paper, we assume the level of significance of the t-score and MI-score to be over 1.96.

3.1.1 The main arguments

Table 1: t-score of the verb 'boda' [3]

Argument Mark	Word form	Frequency	t-score
O	moseub('shape')-eul(Obj)	39	5.187
O	geol('thing')	28	4.462
O	na('I')-leul(Obj)	29	4.045
Oko	issda('be-past')-ko(Q)	47	3.914
O	yeonghoa('film')-leul(Obj)	20	3.838
O	eolgul('face')-eul(Obj)	18	3.655
S	Taisu(P)	13	3.602
O	nunchi('sense')-leul(Obj)	12	3.288
S	na('I')-neun(Sbj)	71	3.076
O	sigieoi('watch')-leul(Obj)	8	2.702
S	jeo('Ip')-neun(Sbj)	13	2.688
O	Taisu(P)-leul(Obj)	7	2.644
O	ccog('side')-eul(Obj)	8	2.570
O	bakk('outside')-eul(Obj)	7	2.519
O	sonhai('loss')-leul(Obj)	7	2.444
O	kkol('side')-eul(Obj)	6	2.315
O	pihai('damage')-leul(Obj)	8	2.194
S	jeoi(Ip)-ga(Sbj)	9	2.049
O	byeol('star')-eul(Obj)	5	2.037
Oko	eobda('nothing')-go(Q)	14	2.036
S	nu('who')-ga(Sbj)	12	2.032
O	Soohai(P)	4	1.996
O	duismoseub('back')-eul(Obj)	4	1.956

As table1 shows, the objectives are mostly over 1.96, which is statistically significant, and some subjectives (nominatives) are also included in the table. But the other argument has not appeared. Some proper nouns in subject places are not significant because this depends on the specific text. Therefore they just appear without any collocational meaning.

Table 2: MI-score of the verb 'boda'[4]

Argument Mark	Word form	Frequency	MI-score
O	Taisu(P)-leul(Obj)	7	10.219
S	Taisu(P)	13	10.113
S	Byeonghoa(P)-neun(T)	2	9.412
S	Yeongjin(P)	2	9.412

[3] Suj: subject case word, Obj: object case word, P: proper noun, Q: quotation mark word.
[4] C: adverbial case word.

S	juinajeossi(‘owner’)-neun(T)	2	9.412
O	Suhai(P)-leul(Obj)	4	8.827
Cr	silyeon(‘loss of love’)-eulo(C)	1	8.412
Cr2	yongdo(‘usaage’)-lona (C)	1	8.412
Cx	bangjeung(‘corroboration’)-eulo(C)	1	8.412
Cx	sayong(‘using’)-ilagoman(C)	1	8.412
Cx	junggandangieoi(‘middle step’)-lo(C)	1	8.412

The MI-score, unlike the t-score, shows that the co-occurring words with low frequencies are usually beyond the statistically significant score, 1.96. So the words with high MI-score are usually low frequency items. This is very different from high t-score word list. But in English collocation research, the high score t-score words and the high score MI-score words have a similar word list. In this aspect, the collocation in Korean is very different from collocation in English. This paper just represents the sample of the statistically significant words. We can conclude from this result that the degree of dependency of the two words is very different in Korean. The assumption behind the MI-score measure is that the dependency between two words is same. But Korean is somehow different in this respect, so that we need other statistical measures for the mutually dependent approach. Since this analysis of the MI-score is not needed, it will not be mentioned here.

3.1.2 The argument in subject

Table 3: the t-score of the subject place

Argument Mark	Word form	Frequency	t-score
S	Taisu(P)	13	3.602
S	na(‘I’)-neun(Sbj)	71	3.076
S	jeo(IP’)-neun(Sbj)	13	2.688
S	jeo(IP)-ga(Sbj)	9	2.049
S	nu(‘who’)-ga(Sbj)	12	2.032
S	namdeul(‘others’)-i(Sbj)	4	1.818
S	sonyeon(‘boy’)-eun(T)	4	1.792
S	nai(‘I’)-ga(Sbj)	35	1.734
S	Byeonghoa(P)-neun(T)	2	1.412

There are 5 words that have statistical significance in the subject place of ‘boda (see)’. The word with the highest t-score is a proper noun, but, as previously stated, this just depends on the specific text of the corpus. Due to its dependency on the text, this is not so meaningful from a semantic point of view. Except for this, the word with the highest t-score word is the first person pronoun. The words, ‘na (I)’, ‘jeo (I-modest expression)’, and ‘jieo (I-modest expression), are Korean first person pronouns. From this perspective, the grammatical collocation of the verb ‘boda (see)’ is related to the cognition of visual sense and the induction or deduction. We can also conclude from this that the first person pronoun is the main subject of a cognition and an induction or deduction. The interrogative word ‘nuga (who)’ is also related to a person, so we can conclude in the same way.

Table 4: the t-score of the object place

Argument Mark	Word form	Frequency	t-score
O	moseub(‘shape’)-eul(Obj)	39	5.187
O	geol(‘thing’)	28	4.462
O	na(‘I’)-leul(Obj)	29	4.045
Oko	issda(‘be-past’)-go(Q)	47	3.914
O	yeonghoa(‘film’)-leul(Obj)	20	3.838
O	eolgul(‘face’)-eul(Obj)	18	3.655
O	nunchi(‘sense’)-leul(Obj)	12	3.288
O	sigieoi(‘watch’)-leul(Obj)	8	2.702
O	Taisu(P)-leul(Obj)	7	2.644
O	ccog(‘side’)-eul(Obj)	8	2.570
O	bakk(‘outside’)-eul(Obj)	7	2.519
O	sonhai(‘loss’)-leul(Obj)	7	2.444
O	kkol(‘side’)-eul(Obj)	6	2.315
O	pihai(‘damage’)-leul(Obj)	8	2.194

O	byeol(‘star’)-eul(Obj)	5	2.037
Oko	eobda(‘nothing’)-go(Q)	14	2.036
O	Suhai(P)	4	1.996
O	duissmoseub(‘back’)-eul(Obj)	4	1.956

In this table, the mark ‘Oko’ is an object quotation clause. This is an argument of the verb ‘boda (see)’ as an opinion and an induction or deduction. It is 19 word forms that have statistical significance in the object argument. Among them, it is 2 word forms that appear in the ‘Oko’ place. The other word forms are usually related to the main meaning of ‘boda’ that is the cognition of sight. The words, ‘nunchi (sign)’, ‘pihai (damage)’, and ‘habui (consent),’ are the object arguments of ‘boda (see)’, and it means that the meaning of ‘boda’ is expanded to the cognition of abstraction. In the dictionary, this expression is written as an idiom. This is an important piece of information for translation and Korean teaching as a second language.

Table 5: the t-score of the other argument place

Argument Mark	Word form	Frequency	t-score
C	nun(‘eye’)-eulo(C)	5	1.568
C	bigoanjeog(‘pessimistic’)-eulo(C)	2	1.379
C	eolgul(‘face’)-lo(C)	2	1.049
C	geungjeongjeog(‘optimistic’)-eulo(C)	2	1.047
C	silyeon(‘loss of love’)-eulo(C)	1	0.997
C	yongdo(‘usaage’)-lona (C)	1	0.997
C	bangjeung(‘corroboration’)-eulo(C)	1	0.997
C	sayong(‘using’)-ilagoman(C)	1	0.997
C	junggandangieo(‘middle step’)-lo(C)	1	0.997

The other argument does not appear with statistical significance. We can suppose that this other argument helps the meaning of the object. These meanings are usually an instrument of the verb ‘boda (see)’.

4. The negative score

As previously described, if the t-score is negative but has statistical significance, then we should consider the word form as statistically significant. Because a negative t-score means that the two words have a tendency not to co-occur. There are 90 word forms that have statistical significance with negative scores. Even if we take a stricter approach to conformity with statistical significance, which is 3.96, then 71 word forms are in this standard. So we can predict that the argument of the verb ‘boda’ appears in various word forms. The table below shows a sample:

Table 6: the t-score of the statistical significant negative[5]

Argument Mark	Word form	Frequency	t-score
Sp	geos(‘thing’)-eun(T)	1	-86.080
Op	geos(‘thing’)-i(C)	1	-81.200
O	uli(‘we’)	1	-64.038
Op	geos(‘thing’)-eun(T)	2	-60.160
Op	geos(‘thing’)-eul(Obj)	1	-52.169
Oko	geos(‘thing’)-elo(C)	1	-46.019
S	uri(‘we’)	2	-44.575
Sp	na(‘I’)-neun(T)	1	-44.084
O	geo(‘he’)-neun(T)	1	-33.096
Sp	nai(‘I’)-ga(Sbj)	1	-23.739
S	nai(‘I’)	1	-23.032
O	geos(‘thing’)-do(T)	1	-20.903
Op	geos(‘thing’)-do(T)	1	-20.903
O	deung(‘etc’)-eul(Obj)	1	-18.857
Op	Hangug(‘Korea’)	1	-17.727

[5] Sp: a subject that follows a predicate, Ob: a object that follows a predicate.

O	geos('thing')	4	-16.872
O	geogeos('that')-eun(T)	2	-16.586
Sp	geu('he')-nuen(T)	4	-15.048

As this table shows, the frequencies of most word forms are low, but the absolute value of t-score is very high. In this table, however, it is very noticeable that the first person pronoun 'Na (I)' has a negative t-score. This is contrary to the previous description that a first person pronoun is the main subject of cognition and induction or deduction. The reason for this difference is the location in the sentence. This means that the collocation depends on the location whether is in front predicate or not.

5. Conclusion

To sum up, this research focused on a grammatical collocation approach using the verb 'boda (see)' as an example. As a result, the verb 'boda (see)' has many statistically significant collocations in the object argument. In the subject argument, there were 5 word forms that had statistical significance. These word forms are usually first person pronouns, and this is related to the meaning of the verb 'boda (see)'. This type of research is expected to be very useful for natural language processing and Korean teaching as a second language.

References

(K) - Korean

B. Kang. 1999 Frequencies and Language descriptions *Research for language information –1*. Seoul: Yonsei University, Center for Language and Information Development.(K)

B. Kang 1999 *The text genre and language character of Korean.* Seoul, Korea University Press. (K)

B. Park. 1997 *The character of word use in Korean spoken language*. MA thesis, Korea University (K)

D. Lee. 1998 *The semantic research of Korean collocation.* Seoul, Korea University MA thesis, Korea University (K)

G. Barnbrook. 1996 *Language and Computers*. Edinburgh University press

H. Kim & B. Kang 1996 Korea-1 Corpus: the design and the organization *Korean Linguistics 3.* Seoul, The association for Korean linguistics. (K)

J. Hong et al. 1998 *The dictionary of modern Korean verb structure*. Seoul, Doosandonga Press. (K)

J. Hong, B. Kang, & H. Choe 2000 The research of applied analyzing of Korean collocation information, *Korean Linguistics 11*. Seoul, The association for Korean linguistics. (K)

J. Kim. 2000 *Korean collocation research*. PhD thesis, Kyunghee University (K)

J. Sinclair. 1991 *Corpus, Concordance, Collocation.* Oxford and New York, Oxford University Press

J. Yun. 1997 *Korean structure analyzing by co-occurrence based word relation.*, PhD thesis, Yonsei University. (K)

U. Paik. 1996 *The introduction of statistics*. Seoul, Jayuacademi Press. (K)