The use of Symbolic Words in Korean Newspapers

Hye-min Jo[1] and Beom-mo kang[2]

[1]graduate school, Korea University, [2]dept. of linguistics, Korea University

Jhmin9471@naver.com, bmkang@korea.ac.kr

Abstract

This paper analyses the meaning and usage of symbolic words in newspapers. The main concern of this paper is to present statistical basis for semantic connections of symbolic words and various aspects of their co-occurrence in order to further advance the existing lexical network research, and to visualize and objectify words through their network. The corpus used for this study is Trends 21 Corpus, a corpus of Korean newspaper texts, comprising around 400 million words (Kim, et al. 2010). We have tagged the texts with a morphological analyzer.

Keywords

corpus, Trends 21 corpus, symbolic word, network

1 Introduction

Research on language is increasingly carried out as interdisciplinary studies together with such disciplines as computer science and statistics, and recently, objective linguistic research, as opposed to intuitive and subjective research, has been receiving more attention. Accordingly, building and analyzing a lexical network mentioned in existing studies on language also need to be tested using more systematic and statistical methods. Words generally are not in a state of disorder but structured very organically and their semantic characteristics are defined through their relationship with other words. Research on lexical network or word co-occurrence, therefore, is essential to word studies.

2 Previous Studies

2.1 Network

In this section we examine previous studies about networks. Most studies of the network have been handled the sociology field. Therefore the main focus of these previous studies will be sociology. Nooy, et al.(2005) investigated social network analysis integrating theory, applications, and professional software for performing network analysis (Pajek). The study will enable the reader to gain the knowledge, skills, and tools to apply social network analysis in all social sciences, ranging from anthropology and sociology to business administration and history. Kang(2010)'s study investigated Constructing Networks of Related Concepts Based on Co-occurring nouns. A method of constructing a network of related words is proposed in this study and Network representation is provided by means of Pajek.
As a result of the previous studies, network studies are mostly that analysis is phenomenon through process of visualization.

2.2 Symbolic words

In this section we examine previous studies about symbolic words. Chae(2002) studied the meaning, usage and morphological properties of onomatopoeia used in novel, headlines, poems and on-line chatting language. According to Chae(2002), Onomatopoeia in headlines is used in the form of root. While it is not used in that form in other kinds of texts. This property is peculiar to morphologically Isolating Languages. Onomatopoeia in headlines has metaphorically extended meaning. It can be used without subjects or predicates. In that case, it has condensed meaning and implies meanings of subjects or predicates. Shin and Park(2005) investigated the properties of symbolic words in newspaper headline based on present-day scholarship. Most studies of the symbolic words have been handled the morphology.

3 Research Methods

It will advance the existing lexical network research, by visualizing and objectifying the semantic properties of symbolic words on the statistical basis.
First of all, let us think about the collocation. The collocation is the way that some words

occur regularly whenever another word is used. In linguistics, collocation is defined as a sequence of words or terms that co-occur more often than would be expected by chance. What has actually happened, and converts the difference between the two into a number which indicates the strength of the collocation. Therefore the higher the number, the stronger the collocation. From the point of frequency information by year and by topics, we get for each word its co-occurring words within a paragraph along with statistical significance measure (G2).

We have two different methods of identifying aspect of symbolic words. The first is network analysis, and the second is keyword analysis.

3.1 Target corpus

The corpus used is Trends21 corpus, a corpus of Korean newspaper texts, comprising around 400 million words.

·Newspapers: Chosun, Joongang, Hankyoreh, Donga

·Period: 2000 ~2009

Trends21 corpus was divided into two major groups of headlines and contents. It is interesting to compare the headline corpus and the content corpus. We carefully compared the headline corpus with the content corpus. Let's take a look at the table1.

Table 1: Top 20 list, Rank by frequency (descending order)

	Head line		*Content*	
rank	G2	Word	G2	Word
1	23913.83	안	11735.81	및
2	21867.63	못	9766.42	특히
3	14186.27	왜	8497.98	현재
4	4728.06	**쑥쑥**	4990.74	물론
5	4401.35	**후끈**	3821.92	각각
6	3949.62	**껑충**	3249.66	한편
7	3838.24	**들썩**	2811.59	더욱
8	3380.00	**북적**	2722.39	또는
9	3049.94	다시	2452.26	이미
10	2698.00	확	2201.57	불과
11	2420.32	**톡톡**	1998.89	또한
12	2419.80	**꿈틀**	1873.53	모두
13	2401.30	줄줄이	1656.57	달리
14	2376.85	**뚝**	1646.81	가장
15	2225.59	**활짝**	1618.49	매우
16	2163.85	**삐걱**	1607.90	다만
17	2149.36	**깜짝**	1466.70	거의
18	2078.09	**휘청**	1422.52	전혀
19	2017.84	**쏙쏙**	1393.30	훨씬
20	2005.81	곧	1378.14	주로

How do symbolic words of headline corpus compare with symbolic words of content corpus? The table1 can show us word frequency. According to the table1, we found more symbolic words in the headline corpus.

3.2 Studied words

The word list is derived from headlines.

Table 2: The list of symbolic words, Rank by frequency (descending order)

Rank	*G2*	*Word*
1	4728.06	쑥쑥
2	4401.35	후끈
3	3949.62	껑충
4	3838.24	들썩
5	3380.00	북적
6	2420.32	톡톡
7	2419.80	꿈틀
8	2376.85	뚝
9	2225.59	활짝
10	2163.85	삐걱
≈	≈	≈
20	1214.71	쾅

4 Aspects of symbolic words

4.1 Network analysis

For the application of methodology, we analyze "쑥쑥". "쑥쑥" is the most frequently used symbolic word.

Table 3: The list of co-occurred word of "쑥쑥", Rank by frequency (descending order)

Rank	*Freq*	*Word*
1	75	매출(sales)
2	43	성적(grade)
3	34	인기(popularity)
4	31	창의력(ingenuity)
5	27	교육(education)
6	25	기업(company)
7	24	시장(market)
8	24	주가(share price)
9	23	공부(study)
10	22	실력(ability)
≈	≈	≈

Based on nodes, links, and their connectivity indexes - density, degree, and centralization, we had been able to retrieve and cluster related words forming the network with co-occurred words of "쑥쑥".

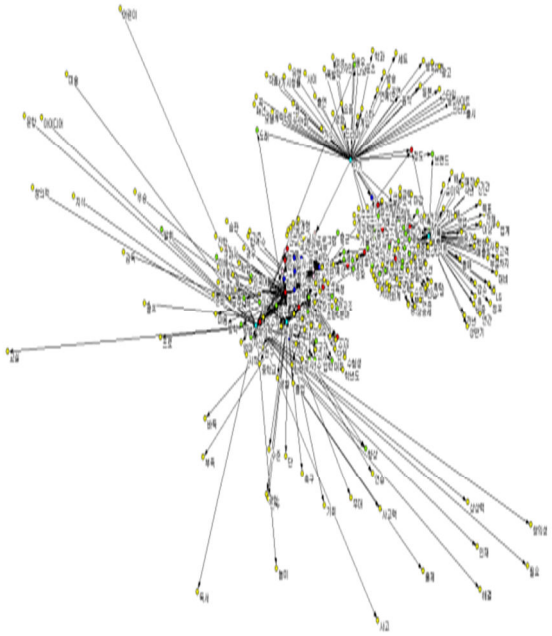

Table 3: the cluster list of "쑥쑥"

Rank	*freq*	*Word*
1	5	대학(university
2	5	교사(teacher
3	5	학교(school)
4	5	수학(math)
5	5	학생(student)
6	5	시험(exam)
7	5	능력(ability)
8	5	영어(English)
9	4	교육(education)
10	4	아이(kid)
11	4	학년(grade)
12	4	입시(entrance examination)
13	4	논술(essay test)
14	4	국내(domestic)
15	4	회사(company)
16	4	상품(product)
17	4	수업(class)
18	4	학습(study)
19	4	과목(course)
20	4	평가(evaluation)

Co-occurred words of "쑥쑥" were assorted into two groups. These are education and economy.

Education: 대학, 교사, 학교, 수학, 학생, 시험, 영어, 교육, 아이, 학년, 입시, 논술, 수업, 학습, 과목

Economy: 국내, 회사, 상품

4.2 keyword analysis

"후끈" and "화끈" are synonyms. Then how can the synonyms "후끈" and "화끈", as they were used, best be defined? Dictionaries are not much help on what are meant by the words.

Table 3: the keyword list of "후끈"

Rank	*freq*	*Word*
1	142	열기(heat)
2	102	경쟁(competition)
3	43	시장(market)
4	40	월드컵(World Cup)
5	39	마케팅(marketing)
6	27	인터넷(internet)
7	23	부동산(property)
8	22	논쟁(argument)
9	22	아파트(apartment)
10	18	수도권(capital area)

Table 4: the keyword list of "화끈"

Rank	*freq*	*Word*
1	18	프로야구 (professional baseball)
2	14	삼성(Samsung)
3	13	골(goal)
4	11	축구(soccer)
5	10	프로농구 (professional basketball)
6	9	공격(attack)
7	9	오늘(today)
8	9	프로축구 (professional football)
9	8	경기(game)
10	8	액션(action)
11	7	NBA

Results of the comparisons showed that there is a distinct difference between the two.

5 Conclusions

We examine various aspects in the use of symbolic words in newspapers.
As the result of analysis using network, The aspects of "쑥쑥" are come out related to education, popularity and economy. And the result of analysis using keyword, we found a large discrepancy between two symbolic words. From this result of analysis, It's very important to provide with possibility of extraction of new information based on Corpora.

References

Catford, J. C.(1988). A Practical Introduction to Phonetics. Oxford: Clarendon Press.

Kang, Beom-Mo. (2008). “Building Corpora and Making Use of Frequency (Statistics) for Linguistic Descriptions”, Korea University press.

Nooy, W., Mrvar, A. & B. Vladimir. (2005). Exploratory Social Network Analysis with Pajek, Cambridge University Press.

Stuart, K. and A. Botella (2009). Corpus Linguistics, Network Analysis and Co-occurrence Matrices, International journal of English studies.

Oh, Kwan-suk. (2009). A network Analysis on the Measurement of Community Elite's Influence : focus on the Comparative Analysis of UCINET and PAJEK., community research

Chae, Wan. (2002). A study of Onomatopoeia from a Text grammatical Point of View. The society of Korean Language and Literature 132. 5-384

Shin, Woo-Bong, Kim, Il-Hwan, Kim, Hung-Gyu. (2010). A study of Spatial Nouns and Network Analysis Based on corpus. Text linguistics 29. 221-250

Kang, Beom-mo. (2010). Constructing Networks of Related Concepts Based on Co-occurring Nouns . The society of Korean Semantics 32. 1-28.

Kim, Hung-Gyu et al. (2010). Trends21. Korea University: Research institute of Korean studies.