The Life and Death of Neologisms: On What Basis Shall We Include Neologisms in the Dictionary?

Kilim Nam, Soojin Lee, Hae-Yun Jung, Jun Choi
Kyungpook National University
e-mail: nki@knu.ac.kr, sjmano27@naver.com, haeyun.jung.22@gmail.com, c-juni@hanmail.net

Abstract
This research analyzes web crawling corpora and examines how many of the neologisms that are coined every year are dying out and how many endure. It seeks to grasp what implications the results of the analysis have for the inclusion of these neologisms in the dictionary. The Korean government initiated the investigation into neologisms in 1992 and has been supervising this research project ever since. Some 400 to 500 coinages that meet definite criteria are being extracted every year, compiled and printed out in the form of a glossary. This paper focuses on the years 2005 and 2006, for which 408 and 530 respectively, that is, 938 new words in total, were recorded. The study turns then to the analysis of the usage changes in the Korean mass media which these neologisms have been undergoing for the past decade. On a quantitative level, the investigation shows that 27% of those neologisms have been in consistent usage for the last ten years.
Keywords: neologisms; usage changes; web crawling corpus; frequency; news articles

1 Introduction
The *Korean New Words Investigation Project* was implemented to collect and record data on the contemporary Korean Language. This project has been carried out and surveys conducted since 1992. Our research consists in studying the new coinages that appear in the mass media within a year. We collect every year about 400 to 500 neologisms and we gather them into a glossary printed under the title *New Words of [year]*. In this study, we present how our investigation into neologisms is being conducted and discuss methodological and procedural issues. Finally, we propose how to use the results of such an investigation for supplementing dictionary entries.
A number of questions have been raised, which form the basis for our study. First of all, how many of the neologisms collected each year die out and how many endure? Second, as we examine the changes in neologism usage, what are the criteria for their extinction and survival? Third, what are the significance and limitations of frequency and statistical distribution when investigating the fluctuations of neologism usage, and how to overcome these limitations? Finally, how can the results of such investigations be utilized when including neologisms in the dictionary? In order to address these questions, we focus on the neologisms extracted in the years 2005 and 2006 and follow their evolution within a time frame of about ten years.

- *Object of study*: neologisms of year 2005 (408 words) and year 2006 (530 words), i.e., 938 words in total
- *Time frame*: from 2005 to date (for a period of 10 years or so)

2 Object and Methodology
The neologisms we investigate in this study are restricted to 'lexical neologisms' (i.e., new word forms). The *New Words Investigation System* allows us to extract automatically the new word forms that appear on the Web, but poses practical issues as it cannot automatically distinguish 'semantic neologisms' (i.e., existing word forms that assume a new meaning) and 'formal neologisms' (i.e., existing word forms that assume a new grammatical function) (Renouf 2013). There are several points to consider in order to investigate the changes in usage of neologisms over the past decade.

- *The extent of the materials*. Sufficient quantity and wide diversity of the mass media must be ensured in order to reflect extensively the changing usages of the neologisms. For the purpose of this study, we have built a corpus consisting of 136 Korean online news reports through Web crawling. By making use of the mass media, we could exclude one-time, personal coinages.
- *The reliability of the frequency analysis results*. While, needless to say, the frequency analysis must be performed accurately, one must keep in mind the gap between the number of online news articles in which neologisms appear and the actual usage frequency. In other words, as we build and analyze our Web crawling corpus, we need to take into consideration both the frequency of news articles and the usage frequency.

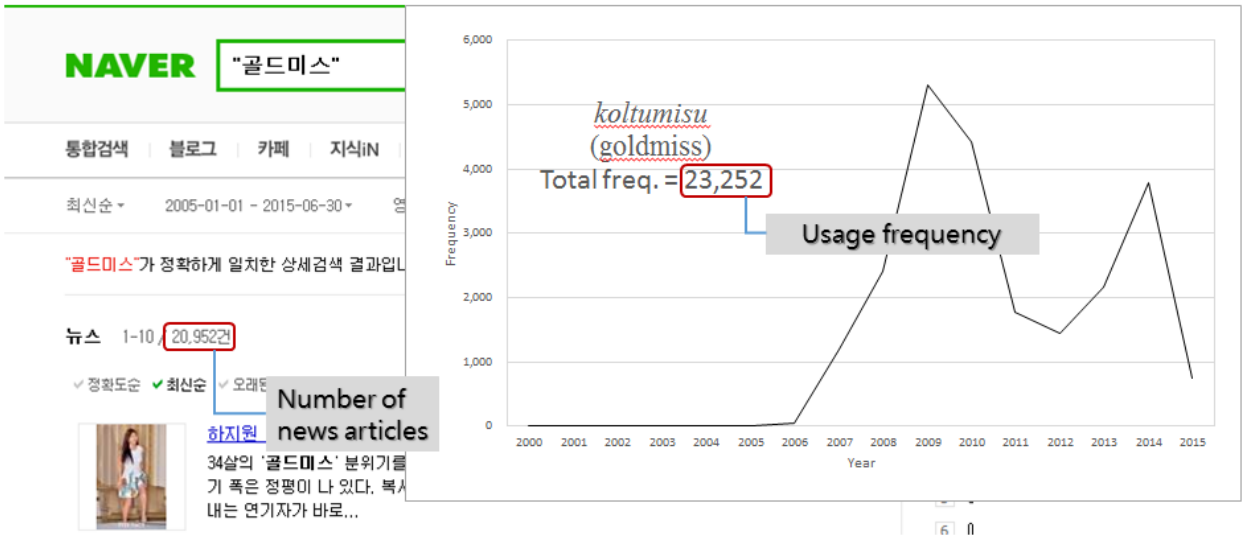

Figure 1: Frequency difference between news articles and actual usage.

- *The minimum time frame for investigating changes in usage*. In order for a neologism to be included in the dictionary, it has to be in usage continuously for a fixed minimum amount of time. In addition to usage frequency, it has also been necessary to take into account the usage distribution for the past ten years so as to exclude 'fashionable coinages' that were only in usage at the time they were created.

Several suggestions have been advanced for determining the relation of lexicography to the life cycle of neologisms. Metcalf suggested five criteria for the inclusion of new words in the dictionary, namely Frequency, Unobtrusiveness, Diversity of users and situations, Generation of meanings and forms, Endurance of concept, otherwise known as the "FUDGE rule" (2002: 152-164). Barnhart (2007) put forward the VFRGT criteria, where V is the number of forms of W[words]; F, the frequency of W; R, the number of sources in which W occurs; G, the number of genres in which W occurs; T, the time span over which W has been observed. Synthetizing the criteria defined in these earlier studies under the concept of "frequency diversity", Hsieh (2015) has argued that the frequency analysis based on diversity factors (diversity of users, genres, subjects, and media) was a crucial criterion in estimating the longevity of neologisms.

In this study, we have chosen to focus on the neologism frequency and yearly distribution factors to analyze the changes in their usage; that is to say, the neologisms that appear in the Web newspapers in 2005 and 2006 forming the starting point for our analysis, we calculated their overall usage frequency over the past ten years, the number of news articles in which they occur, and finally their distribution per year. As a result, a first list of candidate neologisms could be compiled. This list of candidates was then submitted to a team of lexicographers for review and potential inclusion in the dictionary as headwords.

3 Usage Change Analysis of the 2005-2006 Neologisms

The method and process of the usage change analysis are described in (1). 938 neologisms were collected from crawling online mass media articles. Homographs and partially identical forms were eliminated so as to extract new words only. Their frequency was then computed for each consecutive year. Table 1 shows the final results of four examples.

(1) *Investigation methodology and process:*
a. 938 neologisms searched through Web crawling, duplicate articles being excluded as only one address was retained in case of linked articles
b. Analysis of completely or partially identical forms
c. Frequency calculations

neologisms/year	2005	2006	2007	2008	2009	2010	2011	2012	2013	2014	2015	total
tangkeynangin	114	0	0	0	0	0	0	0	0	0	0	114
eylphalachi	1	0	0	0	0	0	1	0	0	0	1	3
kumsappa	0	2	4	2	2	7	30	85	658	167	143	1100
koltumisu	0	45	1188	2412	5303	4416	1770	1443	2158	3781	736	23252

Table 1: Four examples of neologisms showing their overall frequency and their frequency per year.
tangkeynangin: people who lead the public opinion by uploading a lot of posts on parties' homepages
eylphalachi (LPG + paparazzi): people who take pictures of petrol station selling faulty LPG and report them in order to receive compensation
kumsappa: people who quickly fall in love
koltumisu (goldmiss): single female in their thirties, who missed the age of marriage but is financially comfortable.

As seen in Table 1, the overall frequency and the annual frequency show no proportional relationship. The neologism *tangkeynangin* may have appeared more than a hundred times in the mass media for a year but fell into complete disuse from the following year. In the case of *eylphalachi*, the occurrences of the neologism after its first (unique) appearance are barely significant. There is only little chance, if any, that such neologisms will be included in the dictionary. On the other hand, neologisms such as *kumsappa* and *koltumisu* have been continuously in use since 2006 and therefore seem to be suitable candidate neologisms for lexicographical inclusion.

4 Results and Discussion: Frequency Diversity and Criteria for Lexicographical Inclusion

As a neologism enters the phase of common usage, the criteria for its inclusion in the dictionary must be determined. We excluded disyllabic duplicate forms and examined the remaining 915 neologisms featuring the 2005 and 2006 glossaries so as to compile a final list of candidate headwords. (2a), (2b), and (2c) below show the criteria for the inclusion of neologisms in the dictionary, and (a'), (b'), and (c') indicate the number of suitable candidates for each respective criterion.

(2) a. *Frequency of occurrences*: the neologism must appear 20 times or more.
a'. 342 neologisms meet this criterion.
b. *Number of news articles*: the neologism must appear in 10 articles or more.
b'. 374 neologisms meet this criterion.
c. *Annual distribution*: the neologism must appear at least in 5 years out of 10.
c'. 280 neologisms meet this criterion.

Suitable candidates should satisfy all three criteria. 107 and 143 neologisms in 2005 and 2006 respectively did so, constituting about 27% of the total 937 neologisms collected these two years. Usage changes of these 250 neologisms over the past decade were then thoroughly examined in

order to list them as candidate headwords for lexicographical inclusion. Table 2 compiles the top ten neologisms in descending order of frequency of occurrences, number of articles and total number of years in which they occurred.

Ranking	Neologism	Definition	Frequency of occurrences	Number of articles	Total number of years
1	*phuliheku* (freehug)	from the Free Hugs Campaign, in which strangers give hugs to people to make them feel good	23,249	13,673	10
2	*koltumisu* (goldmiss)	a single female in their thirties, who missed the age of marriage but is financially comfortable	22,810	9,446	10
3	*pepulseypun* (bubbleseven)	the seven districts of Seoul where property prices rose dramatically	20,572	9,357	10
4	*pankapaphat*	apartment which is much cheaper thanks to governmental aids	12,969	5,697	10
5	*aitolpomi*	a system that looks after children	8,940	5,596	10
6	*toyncangnye*	a woman who is vain and enjoys luxury and designer labels	8,452	4,805	10
7	*ssayngel*	a face without makeup	8,135	4,748	10
8	*hwunnye*	a female who is not particularly beautiful but is amiable	7,264	3,996	10
9	*sayngtongseng*	(mainly used in pharmaceutics) bioequivalence	6,948	3,768	10
10	*ssangchwunnyen*	a year when there seem to be two springs	4,793	2,960	10

Table 2: List of neologisms in descending order of frequency/number of articles/number of year.

The examples displayed in the above table show the significance of our investigation which follows up neologisms year after year. Most of these neologisms are related to the issues of time and reveal the introduction of new systems or concepts. If, later on, we were to analyze those neologisms out of above-cited 250 ones, which would survive after ten years, the first task would be to measure what intra- and extralinguistic factors would influence the neologisms' longevity.

In this study, we defined a set of three criteria for the inclusion of neologisms in the dictionary and established a list of 250 neologisms that met all of these criteria. However, from a broader perspective, another methodology could be applied just as our research results could be interpreted more flexibly. Indeed, the validity of classifying a wide range of neologisms according to frequency has to be investigated.

For how many years more, after the ten year investigation, should the neologisms we collect every year be consistently used? According to which quantitative criteria should we measure this consistency? What should the criteria be for assessing the usage frequency of the neologisms that are finally included in the dictionary? These questions remain to be examined from various angles in studies to come.

5 References

Barnhart, D.K. (2007). A Calculus for New Words. In *Dictionaries: Journal of the Dictionary Society of North America*, 28, pp.132-138.

Hsieh S. (2015). The Secret of Long-Living Words: Predicting the Lexical Age of Neologism with Big Data. In *Proceedings of the 9th Asialex International Congress, 25-27 June 2015*. Hong Kong.

Metcalf, A. (2002). *Predicting New Words: The Secrets of Their Success*. Boston: Houghton Mifflin.

Nam, K. (2015). *New Words of 2015*. Seoul: National Institute of Korean Language.

Renouf, A. (2013). A Finer definition of neology in English. In Hasselgård, H., Ebeling, J., Ebeling. S. O. (eds.) *Corpus perspectives on patterns of lexis,* pp.177-207.