

The Korean Neologism Investigation Project: Current Status and Key Issues

Kilim Nam, Soojin Lee, Hae-Yun Jung

Kyungpook National University

nki@knu.ac.kr, sjmano27@naver.com, haeyun.jung.22@gmail.com

Abstract

This paper reports on the Korean Neologism Investigation Project and discusses a number of unresolved issues related to neologism research. Since 1994, when the Korean government initiated the project, the use of the Internet and mobile phones has increased exponentially and the methods and scope of the investigation into Korean neologisms have been modified accordingly. This project consists of collecting all the neologisms that appear each year in news articles on the Naver portal using a Web-based neologism extractor (task 1), and examining the usage development of neologisms within the past decade using a Web crawler in order to determine whether those collected ten years ago are still in use (task 2). The extraction of new words is performed automatically but nonetheless requires manual identification. Since 2012, all the neologisms collected for task 1 have been added to the database of the partly crowdsourced online dictionary *Urimalsaem*, which is one of the main dictionaries of the Korean language, together with the more conservative *Standard Korean Language Dictionary* (SKLD). As a result of task 2, the neologisms that are still in continuous use after ten years can be considered as headword candidates for SKLD. At the outset in 1994, the methodology adopted for the project consisted of reading texts and searching for new words manually. Since then, the following methodological changes have been introduced: the construction of a large-scale corpus (2005), the use of the Web-based neologism extractor (2012) and Web crawler (2015), the use of a pattern-based extraction and the raising of the minimum threshold of frequency occurrence for neologism candidates to three (2016). This paper addresses issues that remain despite these adjustments, such as the

unsatisfactory precision and recall levels of automatic neologism detection, the difficulty of conducting a consistent frequency survey due to the dynamic nature of the Web as corpus, the identification of semantic neologisms that are not formal neologisms and the dependency on manual processes. Some of these can be approached in terms of Korean natural language processing or from a typological perspective of Korean as an agglutinative language. In their ten-year cycle investigation of neologism usage, Nam et al. (2016) have found that 75% of the neologisms survived after ten years. Whether this result constitutes a suitable criterion for lexicographic inclusion is also reexamined in the current study.

Keywords

Korean neologisms, neologism extraction, neologism usage investigation, headword candidates, *Urimalsem*, *Standard Korean Language Dictionary*

Introduction

This paper gives an overview of the history and current status of the Korean Neologism Investigation Project (KNIP) and addresses unresolved issues related to the extraction of Korean neologisms and their representation in dictionaries. The Korean government has been conducting the KNIP since 1994, the results of which have been published in annual reports titled *Korean Neologism Investigation of [year]*. Each report presents the neologisms of the relevant year according to frequency and in alphabetical order. All of the reports are accessible to researchers and the press; this does not mean, however, that the presented neologisms are to be included immediately in dictionaries. In fact, another task of this project is to investigate the usage trends of neologisms over a span of ten years, which allows us to distinguish the neologisms that survived from those that died out in order to extract a list of neologism headword candidates.

As yet, there has been little discussion amongst Korean academics working in lexicography of the criteria or policy for lexicographic treatment of neologisms. The *Naver Open Dictionary* service has

been available on Naver, the largest portal in Korea, since 2003 and the crowdsourced dictionary *Urimalsaem* (literally, ‘our language pond’) was put online in 2016; however, studies on users’ participation in the dictionary, including the writing of neologism entries, have remained rather limited. Moreover, most of the Korean online dictionaries, including the main Korean language dictionary *Standard Korean Language Dictionary*, are merely the online versions of print dictionaries, thereby adopting a rather conservative stance on neologisms.

The next section surveys the methodological changes in the investigation of neologisms as new resources such as Web and corpora were introduced. We will also discuss the significance as well as the limitations of the neologism usage trend investigation initiated in 2015.

Finally, the last section examines the status of high-frequency neologisms in the main Korean language dictionaries, that is, *Urimalsaem* and the *Standard Korean Language Dictionary*, but also in the dictionary services provided by Naver, namely *Naver Open Dictionary*, and the *Korea University Korean Dictionary*. This analysis will allow us to redefine the roles of linguists and lexicographers in the study of neologisms and their representation in dictionaries, and discuss the need for a general policy on neologisms in lexicography.

Korean Neologism Investigation and Neologism Usage Trends

Chronology and current status of the Korean Neologism Investigation Project. While the KNIP has been carried out for almost 25 years, the increasing use of the Internet and mobile devices has led to many changes in the methodology and scope of the project. Currently, there are two major tasks performed under this project. The first task is to collect annually all the neologisms of the year from online news articles using a Web-based neologism extractor. This task is performed semi-automatically as the automatic Web-based extractor is combined with manual identification. The other task consists in checking whether the neologisms collected ten years earlier are still in use and if so, examining their

usage trends in order to determine their status as headword candidates. This second task is performed using a Web crawler. The chronological development of the KNIP is shown in Table 1.

Table 1

Year	Project Development and Methodological Changes
1994	- Initiation of the Korean Neologism Investigation Project - No distinction between unrecorded words and neologisms ¹ - Manual extraction of neologisms
2002	Distinction between unrecorded words and neologisms
2005	Construction and utilization of a corpus to extract neologisms
2012	Development of a Web-based neologism extractor which identifies neologism candidates from 139 newspapers on Naver
2015	Beginning of the ten-year usage trend investigation for the neologisms extracted in 2005 and 2006
2016	Introduction of a pattern-based methodology of neologism extraction; raising of the minimum threshold of frequency occurrence for neologism candidates

METHODOLOGICAL CHANGES IN THE KOREAN NEOLOGISM INVESTIGATION PROJECT

As can be seen in Table 1, the neologisms were initially searched manually, with the researchers reading texts and looking for new words with the naked eye. Crucial methodological changes have been introduced since the 2000s, including the construction of a large-scale corpus (2005) and the use of the Web-based neologism extractor (2012) and Web-crawler (2015). Currently, the Web-based neologism extractor compares the Web-crawled corpus with existing dictionary entries to extract wordforms that

¹ As the project initially aimed to find new words to add to SKLD, the main criterion for extraction was simply to collect all the words that were not recorded in SKLD.

are not included in advance and create a list of neologism candidates. This list is manually checked in order to extract the neologisms of the relevant year. In order to minimize the manual work, we can search using specific syntactic patterns, such as patterns with quotation marks or strings including the word ‘*sin-e/sincwo-e*’ (neologism; coinage), as well as a list of ‘excluded words’, such as proper nouns, which is constantly updated. Nonetheless, the precision and recall scores for the automatic identification of neologisms remain somewhat unsatisfactory. At first, the accuracy of the extractor was approximately 0.5%. It then increased to 5% with the introduction of pattern analysis. Such a low accuracy level can be attributed not only to the inclusion of proper names, foreign words and free combination phrases to the candidate list, but also to the overanalysis of Korean morphemes carried out by the morphological analyzer of the neologism extractor.

Another issue related to the investigation of neologisms regards the frequency stability of the Web language. Since 2016, only neologisms that appear at least three times during the relevant period are considered to have been assimilated into the language and are listed in the annual report. However, it is not easy to provide stable frequencies, as the Web-crawled corpus used for the frequency analysis is a dynamic corpus, that is, subject to the changing nature of the Web. This issue can hardly be avoided since language increasingly depends on the Web and our basic data is in fact from the Web. Finally, the identification of semantic neologisms, as opposed to formal neologisms, is still heavily dependent on manual processes. Formal neologisms are new lexical items, encompassing both new wordforms and new multi-word expressions; they are, therefore, easily detected automatically. However, semantic neologisms refer to new meanings or new grammatical functions attributed to existing wordforms and cannot be detected without contextual analysis. This is a critical issue that the project has yet to resolve.

Some of the above issues can be approached in terms of Korean natural language processing, but some cannot be easily solved, including the issue of the automatic identification of semantic neologisms.

Despite such limitations, the investigation into Korean neologisms is a meaningful task in that the new wordforms are extracted from the most widely read online newspapers and the extraction is performed on a large scale. As for the issue of randomness pointed out by Rundell (2015), which arises when a dictionary user adds a new word to a crowdsourced dictionary², this will be illustrated with concrete examples later in the last section. Before that, the methodology of the neologism usage trend investigation is discussed in the next section.

Issues surrounding the neologism usage trend investigation. The objective of examining the usage trends of neologisms is to extract candidates for lexicographical inclusion by identifying the neologisms that have survived and those that have died out over a certain time span. Although Metcalf (2002, 152-164; 2004) and Barnhart (2007) have presented different criteria for the inclusion of neologisms in the dictionary, our investigation into Korean neologism usage trends focuses on online data only and thereby retains three criteria, which are detailed in (1) below.

(1) Criteria for identifying the surviving neologisms

- a. Total frequency over the ten-year cycle: 20 occurrences or more
- b. Number of articles in which it appears: ten articles or more
- c. Time span distribution: five years or more³

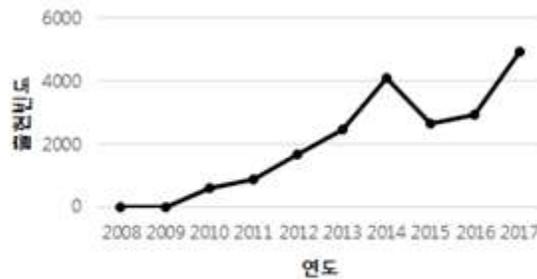
Nam et al. (2016) have analyzed the usage trends of the 2005-2006 neologisms and found that around 75% of the neologisms survived after ten years. The investigation of the 2007-2008 neologisms

² More specifically, crowdsourced dictionaries of the user-generated content type according to Rundell's typology of crowdsourced dictionaries.

³ This corresponds to the number of years a neologism should appear and need not be five consecutive years.

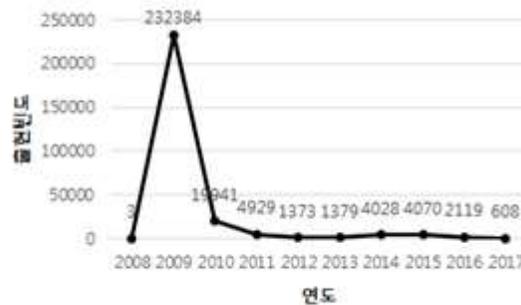
showed that only about 25% of them survived. The graphs in Figures 1 and 2 illustrate two cases of surviving neologisms which meet the above criteria regardless of whether the usage trend develops upwards or downwards.

Figure 1



USAGE TREND OF *YENKI-TOL* ‘ACTING-[I]DOL’⁴ FROM 2008 TO 2017

Figure 2



USAGE TREND OF *SINCONG-PHULLWU* ‘NEW TYPE-FLU’ FROM 2008 TO 2017

The primary issue regarding the methodology of the usage trend investigation is the relevance of the criteria mentioned above. First of all, our investigation only targets online mass media, thereby

⁴ A literal translation in English will be given next to each Korean neologism mentioned in this paper. Most Korean neologisms are blend words and some are formed with foreign words; when possible, omitted syllables are indicated in square brackets. For a full description of the neologisms’ meanings, see the Appendix.

excluding many other Web genres such as blogs, ‘café’⁵ posts and bulletin board posts. Therefore, the inclusion of personal text data from blogs or SNS (Social Networking Services) would imply the use of different quantitative criteria and yield different results⁶. While it seems rather appropriate to include and analyze blogs and cafés and not merely online mass media, it is also desirable to remain consistent and match the target data of this investigation with that of the initial project. Moreover, by limiting the analysis to mass media, we can avoid idiolects and calculate the frequencies of neologisms that are widely accessed.

Above all, our main concern is whether the results of our usage trend analysis mean that the surviving neologisms should be unconditionally included in the dictionary. In fact, there are other frequent external issues to consider. Metcalf (2002, 152-164; 2004) has suggested the FUDGE factors as a basis for the lexicographical inclusion of neologisms (Frequency, Unobtrusiveness, Diversity of users and situations, Generation of meanings and forms, and Endurance of concept). While our investigation

⁵ In Korea, a Web ‘café’ is a community Website on a particular topic, to which one subscribes to become a member and wherein members exchange information.

⁶ If we compare the occurrence frequencies of some wordforms in the mass media with those in blogs and cafés, there is hardly any direct proportionality. Nonetheless, we can find a constant correlation; that is, neologisms with high frequency in mass media also tend to be used at high frequency in blogs and cafés and the same applies (inversely) to neologisms with low frequency.

The following shows the frequency of three neologisms in order of ‘mass media//blog/café’.

a. *caynung-pongsa* ‘skill-volunteering’ 1,436//1,741/853

b. *kochwucang-nam* ‘chilli paste-man’ 159//436/159

c. *silpe-meni* ‘silver-money’ 20//25/44

aims to identify the surviving neologisms so as to include them in the dictionary, at present Korean dictionaries offer no clear policy regarding neologisms. In the next section, we examine the status of selected surviving neologisms in today's Korean lexicography and discuss issues related to the role of the lexicographer and lexicographical policy on neologisms.

The Inclusion of Neologisms in Korean Dictionaries

The 2005-2009 neologisms which are still in use. As mentioned above, the government-affiliated National Institute of the Korean Language has been conducting investigations of neologism usage trends/patterns on a 10-year basis since 2015. Thus, the 2015 investigation focuses on the neologisms collected in 2005 and 2006 and the 2016, 2017 and 2018 investigations on those collected in 2007, 2008 and 2009 respectively. In fact, each investigation also includes the neologisms collected beforehand and examines their usage patterns within the last decade. Thus, the 2018 investigation is concerned with the usage trends of neologisms that were collected in 2008 in particular but also those collected from 2005 to 2007, and which appeared more than 20 times, across ten articles or more, and over five years from 2008⁷ to 2017. In this paper, we checked whether the top ten neologisms of 2005, 2006, 2007, 2008 and 2009 (i.e. 50 neologisms in total), which are found to meet the aforementioned criteria, are recorded as headwords in *Urimal saem*, *Naver Open Dictionary* (NOD), *Standard Korean Language Dictionary* (SKLD), as well as *Korea University Korean Dictionary* (KUKD), the content of which is accessible online through the Naver dictionary platform. The following presents the top 50 neologisms from 2005 to 2009 which have been categorized as surviving neologisms as a result of the 2018 investigation. The neologisms are presented in (2) in order of high frequency, with absolute frequency shown in brackets.

⁷ The time frame for extracting neologisms of a particular year, say 2009, is not from January to December of that year but spans from July of the previous year to June of that year (e.g. July 2008 to June 2009).

(2) a. 2005 neologisms: *sopica-phihay-cwuuypo* ‘consumer’-‘detriment-notice’ (2815), *tatwungi-kacok* ‘multiple birth’-‘family’(2589), *cwumma-leylla* ‘middle-aged woman-[cinde]rella’ (2285), *sukhwul-phollisu* ‘school-police’ (2261), *kong-si-cok* ‘civil servant-exam-group’ (2242), *thelki-chwum* ‘shake-dance’ (1624), *hyem-hanlyu* ‘dislike-Korean wave’ (1615), *seykum-phokthan-lon* ‘tax-bomb-claim’ (801), *welkup-kokay* ‘salary-hill’ (616), *ansim-thayksi* ‘safe-taxi’ (589)

b. 2006 neologisms: *hwun-nam* ‘heartwarming-man’ (37426), *phuli-heku* ‘free-hug’ (31330), *koltu-misu* ‘gold-miss’(23519), *kiphu-thikhon* ‘gift-[emo]ticon’ (21983), *pepul-seypun* ‘bubble-seven’ (14168), *aitolpomi* ‘child-carer’ (9607), *hwun-nye* ‘heartwarming-woman’(8590), *saying-tong-seng* ‘bioequivalence’ (7907), *ssayng-el* ‘raw-face’ (7318), *toyncang-nye* ‘fermented soya bean paste-woman’ (7119)

c. 2007 neologisms: *cwuthayk-yenkum* ‘house-pension’ (30317), *wikhilikhusu* ‘wikileaks’ (29625), *em-chin-ttal* ‘mum-friend-daughter’ (20490), *kullaym-phing* ‘glam[orous]-[cam]ping’ (19164), *pokhap-hwansung-seynthe* ‘complex-transit-center’ (16182), *noin-cangki-yoyang-pohem* ‘elderly-long-term-convalescence-insurance’ (15407), *em-chin-a* ‘mum-friend-son’ (14056), *mopail-thwuphyo* ‘mobile-vote’ (13459), *pui-lain* ‘V-line’ (12619), *nolan-wusan-kongcey* ‘yellow-umbrella-exemption’ (11993)

d. 2008 neologisms: *thokhukhonsethu* ‘talk-concert’ (43260), *noksayk-sengcang* ‘green-development’ (30971), *noksayk-sanep* ‘green-industry’ (15439), *phuleyntulli* ‘friendly’ (14136), *saipe-moyok-coy* ‘cyber-insult-crime’ (6814), *pwulpich-chwukcey* ‘light-festival’ (6081), *cenca-congi* ‘electronic-paper’ (5453), *haksup-cito-yolyeng-hayselse* ‘teaching-guidance-tip-manual’ (4957), *kongceng-yehayng* ‘fair-travel’ (4778), *kunhwang-seylkha* ‘current situation-selfie’ (4773)

e. 2009 neologisms: *sincong-phullwu* ‘new type-flu’ (270834), *incung-syas* ‘evidence-shot’ (244510), *colipep* ‘recipe’ (32767), *sincong-inphullwueynca* ‘new type-influenza’ (32580), *cwuyo-cayngcem* ‘main-issue’ (25113), *yenki-tol* ‘acting-[i]dol’ (20281), *kkwul-pekci* ‘honey-thigh’(15850), *eysci-issta*

‘edgy-be’ (14499), *phwumcel-nam* ‘out of stock-man’ (13399), *twayci-inphullwueynca* ‘swine-influenza’(12679)

The neologisms in (2) were looked up in *Urimalsaem*, NOD, SKLD, and KUKD and the results are shown in Table 2. The number in brackets indicates the number of neologisms included in the dictionary for each year and each dictionary.

Table 2

	<i>Urimalsaem</i>	<i>Naver Open Dictionary</i>	<i>Standard Korean Language Dictionary</i>	<i>Korea University Korean Dictionary</i>
Total number of neologisms recorded as headword	37	21	1	6
2005 Neologisms	<i>tatwungikacok</i> , <i>kongsicok</i> , <i>hyemhanlyu</i> (3)	<i>cwummaleylla</i> , <i>kongsicok</i> (2)		
2006 Neologisms	<i>hwunnam</i> , <i>koltumisu</i> , <i>pepulseypun</i> ⁸ , <i>aitolpomi</i> ,	<i>hwunnam</i> , <i>phuliheku</i> ⁹ , <i>koltumisu</i> ,		<i>hwunnam</i> , <i>koltumisu</i> , <i>hwunnye</i> ,

⁸ Recorded as *pepul-seypun ciyek* ‘bubble-seven-area’.

⁹ Recorded as both *phuli-heku* ‘free-hug’ and *phuli-heking* ‘free-hugging’.

	<i>hwunnye, sayngtongseng, ssayngel, toyncangnye</i> (8)	<i>kiphuthikhon¹⁰, pepulseypun¹¹, hwunnye, ssayngel, toyncangnye</i> (8)		<i>ssayngel¹²</i> (4)
2007 Neologisms	<i>cwuthaykyenkum, wikhilikhusu, emchinttal, kullaymphing, pokhaphwansungseynthe, noincangkiyoyangpohem, emchina, mopailthwuphyo, puilain, nolanwusankongcey</i> (10)	<i>emchinttal, kullaymphing, emchina</i> (3)		<i>emchinttal, emchina</i> (2)
2008 Neologisms	<i>thokhukhonsethu, noksayksanep, saipemoyokcoy, pwulpichchwukcey, cencacongi, kongcengyehayng</i> (6)	<i>thokhukhonsethu, pwulpichchwukcey</i> (2)		

¹⁰ Recorded as *kiph-khon* ‘gif[t]-[emoti]con’.

¹¹ Recorded as *pepul-seypun ciyek* ‘bubble-seven-area’.

¹² Recorded as *sayng-el* ‘raw-face’.

2009	<i>sincongphullwu, incungsyas,</i>	<i>sincongphullwu,</i>	<i>colipep</i> ¹³	
Neologisms	<i>colipep,</i>	<i>incungsyas,</i>	(1)	
	<i>sinconginphullwueynca,</i>	<i>cwuyocayngcem,</i>		
	<i>cwuyocayngcem, yenkitol,</i>	<i>yenkitol, eysciissta,</i>		
	<i>kkwulpekci, phwumcelnam,</i>	<i>phwumcelnam</i> (6)		
	<i>twayciinphullwueynca</i> (10)			

STATUS OF SURVIVING NEOLOGISMS IN THE KOREAN LEXICOGRAPHY

As seen in Table 2, *Urimalsaem* included 37, NOD 21 and KUKD six out of the top 50 neologisms from 2005 to 2009. *Urimalsaem* and NOD share some similarity in that they provide user-generated content; yet, these two dictionaries are quite different in nature. *Urimalsaem* is a large-scale online dictionary based on SKLD, which has been compiled by the National Institute of Korean Language; therefore, it includes all the entries of SKLD. In addition, it provides dialectal forms, North Korean words, specialized terms, and archaic words which are not included in SKLD and were gathered from 2010 to 2016. Initially, *Urimalsaem* included a total of 1,109,722 headwords. Since its opening to the public, users have been able to record new headwords and write their definitions. Users with editor accounts review the proposed entries for potential inclusion. Once a new entry is reviewed and finalized in terms of lexicographical description, the new content is displayed on the general search page (Figure 3). Therefore, while many users can record the same word at the same time, only one entry is included since the whole process is supervised. Ordinary users can thus participate in the writing of the

¹³ *Colipep* ‘cooking method’, which was extracted as a neologism in 2009, was already included in *Standard Korean Language Dictionary*. It seems to have been considered as a neologism as a simplified substitute for the loanword *lesiphi* (recipe).

dictionary; nonetheless, it ultimately involves the expertise of lexicographers and linguists who have the authority of reviewing and editing newly recorded words. This makes a considerable difference from user-generated content dictionaries such as the *Urban Dictionary*¹⁴.

Figure 3



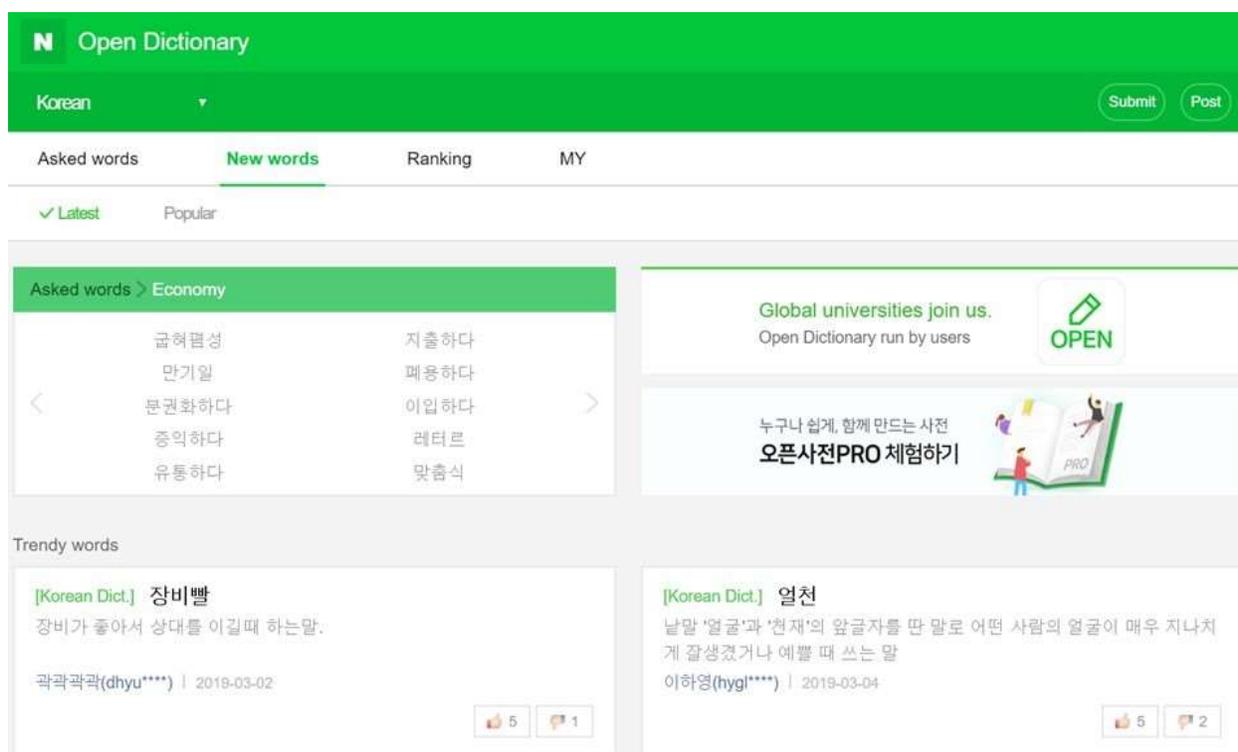
SEARCH SCREEN INTERFACE OF URIMALSAEM

NOD is a sub-service offered by Naver, which is one of the most popular portal sites used by Koreans. NOD started as a sub-category of the site’s forum on neologisms and trendy words. It is quite similar to the *Urban Dictionary* in that users can create many threaded discussions with the same word

¹⁴ The *Urban dictionary* is representative of user-generated content dictionaries, as it allows users to compile entire entries, from the definition to usage information and examples. They can also duplicate entries which have already been written by other users and entries are arranged in high ‘upvote’ order.

in the title, yet provide very different interpretations for it, and upvote discussion posts. Since the various threads on the same word are not posted all together in a single discussion and can be eventually pushed down by new discussions, information on that particular word is not easily retrieved. In that sense, NOD can hardly be called a dictionary. Now¹⁵, the home page of NOD only displays the top four or five items filtered either by ‘latest’ or ‘popular’ order and the remaining can be checked by clicking the ‘view more’ button. The search for words recorded in NOD is not performed through NOD itself, but through the regular Naver dictionary. To some extent, *Urimalsaem* and NOD have the advantage of including newly created words quite promptly.

Figure 4



INTERFACE OF NAVER OPEN DICTIONARY

Figure 5

¹⁵ As of March 2019.

NAVER Dictionary Translate User translation Encyclopedia

Dictionary Home English Korean Hanja Japanese Chinese French Spanish German More

Korean Dictionary

국어사전 *Beta* Advanced search

All | Word | Proverb/Idiom | Meaning | Examples | Spelling/Orthography

Word 3

엄친딸 -親-
 noun 능력이나 외모, 성격 등 거의 모든 조건이 완벽한 여자를 빗대어 이르는 말. '엄마 친구 딸'이 준 말로, 부모가 자식을 다른 아이들과 비교할 때 으레 친구의 자식을 언급하는 데에서 유래하였다.(→엄친아)
 Korea Univ. Korean Dict.

엄친딸 엄親딸 [엄친딸]
 noun '엄마 친구의 딸'이라는 뜻으로, 외모나 능력 등 모든 면에서 뛰어난 젊은 여자를 이르는 말.
 Urimalsae

엄친딸
 1. 엄마 친구 딸을 뜻하는 말. 특히 엄마 친구 딸 중 얼굴도 예쁘고 공부도 잘하는 등 거의 모든 것이 완벽한, 즉 ...
 → 그와양(5851****) | 2006-12-16

Sources

EXAMPLE OF THE SEARCH FOR THE NEOLOGISM *EM-CHIN-TTAL* 'MUM-FRIEND-DAUGHTER' IN NAVER KOREAN DICTIONARY

As for SKLD and KUKD, which are viewed by Koreans as linguistic authorities, they are almost completely closed to accepting neologisms. The online SKLD is basically the same as the paper dictionary which was compiled by the National Institute of Korean Language and published in 1999. It can be described as a prescriptive dictionary in the sense that Koreans usually consult it first when checking spelling and standard language. The online version was launched in 2002, and in 2008 the online dictionary was massively restructured and added a large number of words, reaching 511,160 headwords by 2014. Although entries are updated on a quarterly basis, only 30 to 40 updates at a time are performed on average, and correcting language information is the main type of update. Very few

new headwords are added. As a result, SKLD has kept a closed and conservative attitude towards the inclusion of neologisms as headwords. As for KUKD, it was compiled based on a Korean database of 100 million words and published in 2009 by the Research Institute of Korean Studies at Korea University. The dictionary includes 386,889 headwords and has no online version. Its content is nonetheless searchable online through search engines such as Naver and Daum, which provide dictionary services. As usage frequencies were taken into account during the compilation process, KUKD includes frequent neologisms of the time, such as *hun-nam* ‘heartwarming-man’ and *hun-nye* ‘heartwarming-woman’. However, since its publication it has not been updated and no more neologisms have been added. In that sense, it can be seen as a closed dictionary.

On the other hand, none of the following neologisms presented in (3) have been included in either of the four Korean dictionaries considered for this study.

(3) a. 2005 Neologisms: *sopica-phihay-cwuuyyo* ‘consumer’-‘detriment-notice’, *sukhwul-phollisu* ‘school-police’, *thelki-chwum* ‘shake-dance’, *seykum-phokthan-lon* ‘tax-bomb-claim’, *welkup-kokay* ‘salary-hill’, *ansim-thayksi* ‘safe-taxi’

b. 2008 Neologisms: *haksup-cito-yolyeng-hayselse* ‘teaching-guidance-tip-manual’, *kunhwang-seylkha* ‘current situation-selfie’

Although the surviving neologisms shown in (3) have been used in mass media data more than 500 times in ten years, they cannot be found even in *Urimalsaem* or NOD. One explanation could be given in terms of frequency. Except for *sopica-phihay-cwuuyyo* ‘consumer-detriment-notice’ and *sukhwul-phollisu* ‘school-police’, which were ranked 1st and 4th respectively in the high frequency ranking, the remaining six neologisms account for the lowest frequencies (6th or less in their respective year). Another – perhaps better – explanation is that these neologisms display high transparency as their meanings can be relatively easily inferred by looking at their components; therefore, they do not need to

be looked up in the dictionary to be understood. This could be the main reason why these words were not even recorded in the two user-generated content dictionaries surveyed for this study.

The 2018 neologisms. The latest survey carried out by the project corresponds to the collection of neologisms for the year 2018. We extracted a total of 434 words, which have been recorded in the annual report Korean Neologism Investigation of 2018. We checked whether the ten most frequent neologisms have been included as headwords in *Urimalsaem*, NOD, SKLD and KUKD.

Table 3

Freq. rank	Number of occurrences	Neologism	<i>Urimalsaem</i>	<i>Naver Open Dictionary</i>	<i>Standard Korean Language Dictionary</i>	<i>Korea University Korean Dictionary</i>
1	43500	<i>phanmwuncem-senen</i> ‘Panmunjeom-Proclamation’	Yes	No	No	No
2	34673	<i>mi-thwu-wuntong</i> ‘me-too-movement’	Yes	Yes – recorded as <i>mi-thwu</i>	No	No
3	3592	<i>ka-sim-pi</i> ‘price-psychological-ratio’	Yes	No	No	No

4	2053	<i>khophi-cenlyak</i> ‘nose bleed- strategy’	No	No	No	No
5	1515	<i>kimchi-phulimiem</i> ‘kimchi-premium’	Yes	No	No	No
6	1403	<i>peynsu-lwul</i> ‘fence-rule’	No	No	No	No
7	974	<i>thongcang-yoceng</i> ‘bank account- fairy’	No	No	No	No
8	845	<i>qheyleynsia</i> ‘querencia’	Yes	Yes	No	No
9	825	<i>sopi-yoceng</i> ‘consumer-fairy’	No	No	No	No
10	644	<i>pelcip-kyeycwa</i> ‘bee hive-bank account’	Yes	No	No	No

STATUS OF THE TOP TEN 2018 NEOLOGISMS IN THE KOREAN LEXICOGRAPHY

As shown in Table 3, six out of the top ten neologisms of 2018 have been included as headwords in *Urimalsaem*. Only the neologisms *mi-thwu* ‘Me Too’¹⁶ and *kheyleynsia* ‘querencia’ can be found in NOD and none is searchable in SKLD and KUKD. The four neologisms that have not been recorded in any of the four dictionaries examined are *khophi-cenlyak* ‘nose bleed-strategy’, *peynsu-lwul* ‘fence-rule’, *thongcang-yoceng* ‘bank account-fairy’, and *sopi-yoceng* ‘consumer-fairy’. Among those, *peynsu-lwul* ‘fence-rule’, which is a neologism used to indicate the distance men put between themselves and women in order to prevent accusations of molestation or sexual assault, has often been mentioned alongside the very frequent neologism *mi-thwu-wuntong* ‘me-too-movement’ but has not gained the status of headword. This example shows that although user-generated content dictionaries such as *Urimalsaem* tend to include high-frequency neologisms relatively quickly, this tendency lacks systematicity. The case of NOD, which has only included *mi-thwu* ‘me-too’ and *kheyleynsia* ‘querencia’, is a clear example of the randomness of headword inclusion in user-generated content. While even neologisms with high frequency are arbitrarily selected and recorded as headwords by users in crowdsourced dictionaries, the tendency becomes all the more serious for neologisms with low frequencies. The less frequent a neologism, the less likely it is to be recorded as a headword¹⁷. Rundell (2015: 4) has indeed pointed out that “a characteristic feature [of user-generated content] is its

¹⁶ The definition corresponding to that of the headword *mi-thwu-wuntong* ‘me-too-movement’ in *Urimalsaem* is provided as the second meaning of the headword *mi-thwu* ‘me-too’ in *Naver Open Dictionary*.

¹⁷ In order to further confirm this tendency, we also examined the 11th to 20th most frequent neologisms. Only three of them were included in at least one of the dictionaries and none of the ten least frequent neologisms were recorded in any of the dictionaries.

randomness” since the inclusion of new headwords “depends on the interests and preferences of users”. The results of our investigation show that a good number of neologisms live on in the dictionaries thanks to the participation of dictionary users; nonetheless, this participation has some limitations, including topic bias, lack of lexicographical expertise and, as mentioned above, randomness. Such limitations show that user-generated content might present some advantages for the lexicographical inclusion of neologisms but large-scale investigation into the Korean language, such as the one carried out by the National Institute of the Korean Language, is still needed to ensure the systematicity of the inclusion process and the representativeness of the neologisms included.

Conclusion

In this study, we presented a brief overview of the chronology and the current status of the Korean Neologism Investigation Project. We presented the major methodological changes that had been made since the project started about 25 years ago and discussed a number of issues to be solved as regards the extraction and lexicographical inclusion of neologisms. Although the development of methodology and tools contributed to massive improvements in the extraction of formal neologisms, the automatic identification of the neologisms is not yet wholly satisfactory. Moreover, great care is needed when investigating neologisms due to the changing nature of the Web which constitutes the source material for extracting neologism candidates. Presently, the identification of semantic neologisms is entirely reliant on manual performance and the development of our extraction tools is another issue to work on. In parallel with the annual extraction of neologisms, our project investigates the usage trends of these neologisms over a ten-year time span. Thus, we are able to determine which neologisms survived in language on the basis of quantitative criteria. Those that are found to be still in use are then considered as headword candidates. We examined the extent to which the main Korean language dictionaries accept neologisms. As could be predicted, prescriptive dictionaries such as SKLD and KUKD tend not to

include neologisms. Very few of the neologisms we investigated could be found in these dictionaries, regardless of the word frequencies. On the other hand, recently developed user-generated content dictionaries tend to include high-frequency neologisms. Nonetheless, as pointed out by Rundell (2015), this tendency is still dependent on users' preferences, thereby making the lexicographic inclusion of neologisms the product of arbitrary choices rather than a systematized selection. To overcome this issue of randomness, Korean linguists and lexicographers need to change their stance on neologisms by adopting a more systematic and active approach to the lexicographical inclusion of neological headword candidates.

References

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

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

Nam, Kilim, Soojin Lee and Hae-Yun Jung. 2016. The Life and Death of Neologisms: On What Basis Shall We Include Neologisms in the Dictionary? Paper presented at the 17th EURALEX International Congress, Tbilisi.

Rundell, Michael. 2016. Dictionaries and crowd-sourcing, wikis, and user-generated content. In *International Handbook of Modern Lexis and Lexicography*, edited by Patrick Hanks and Gilles-Maurice de Schryver. Berlin: Springer. Accessed online at http://michaelrundell.com/wp-content/uploads/2017/05/Crowdsourcing-and-lexicography_MR.pdf

Dictionaries

Korea University Korean Dictionary. <https://dic.daum.net/index.do?dic=kor>

Naver Open Dictionary. <https://ko.dict.naver.com/#/main>

Standard Korean Language Dictionary. <http://stdWeb2.korean.go.kr/main.jsp>

Urimalsaem. <https://opendict.korean.go.kr/main>

Appendix

Top ten neologisms per year from 2005 to 2009, in order of high frequency.

2005 neologisms

1. *sopica-phihay-cwuuyyo*: Notice issued by the Fair Trade Commission to prevent consumer detriment caused by unfair trade.
2. *tatwungi-kacok*: A family with many children.
3. *cwumma-leylla*: A married woman in her late 30s to late 40s who is economically independent, invests time and money for her personal growth, and actively participates in social activities.
4. *sukhwul-phollisu*: Specialist personnel, such as former police officers, who are assigned to schools and are in charge of protecting students and preventing violence in schools.
5. *kong-si-cok*: A noun designating people who are preparing for various civil service examinations.
6. *thelki-chwum*: A type of dance that consists of shaking the whole body.
7. *hyem-hanlyu*: A dislike of the Korean wave.
8. *seykum-phokthan-lon*: The claim against excessive taxes.
9. *welkup-kokay*: A situation wherein one has almost used up the salary of the previous month and has not yet received the next month's salary.
10. *ansim-thayksi*: A taxi operated by female drivers.

2006 neologisms

1. *hwun-nam*: A man who is not necessarily good-looking but radiates warmth and charisma. It can also refer to a handsome man.
2. *phuli-heku*: Free Hug (Campaign).
3. *koltu-misu*: An unmarried woman in her 30s who has a strong economic situation.
4. *kiphu-thikhon*: Emoticons that are gifted via SNS and can be used as a voucher.
5. *pepul-seypun*: A noun that designates seven areas in or near Seoul (Gangnam, Seocho, Songpa, Mokdong, Bundang, Pyeongchon and Yongin) where housing prices have dramatically soared.
6. *ai-tolpomi*: A babysitter.
7. *hwun-nye*: A woman who is not necessarily beautiful but radiates warmth and charisma. It can also refer to a pretty woman.
8. *sayng-tong-seng*: Short for *sayngmwulhakcek tongtungseng* ‘bioequivalence’.
9. *ssayng-el*: A bare skin face.
10. *toyncang-nye*: A pejorative way to describe a woman who is vain and only enjoys luxury and branded goods.

2007 neologisms

1. *cwuthayk-yenkum*: A type of monthly old age pension which the elderly can receive until death by putting their property up as collateral.
2. *wikhilikhusu*: Wikileaks.
3. *em-chin-ttal*: Short for *emma chinkwu ttal* ‘mum’s friend’s daughter’. A friend’s ‘ideal’ daughter that a mother always compares her own daughter to.

4. *kullaym-phing*: Fashionable and expensive camping sites, with tents including WI-FI, expensive furniture and electrical facilities.
5. *pokhap-hwansung-seynthe*:. A type of convenient transit station which connects trains, metros, coaches and buses, and also provides a massive parking lot.
6. *noin-cangki-yoyang-pohem*: Social insurance which provides long-term eldercare services, such as physical activity aid or housework help.
7. *em-chin-a*: Short for *emma chinkwu atul* 'mum's friend's son'. A friend's 'ideal' son that a mother always compares her own son to.
8. *mopail-thwuphyo*: Vote performed by voters registered onto an electoral roll on their mobile phones without going to the polling station.
9. *pui-lain*: The 'idealistic' shape of the face, that is, oval and narrower at the chin like the letter V.
10. *nolan-wusan-kongcey*: A system of tax income exemptions administered by the Korean Federation of Small and Medium Businesses in order to ensure life stability and help with business recovery when small businesses are facing hardship.

2008 neologisms

1. *thokhu-khonsethu*: A type of concert wherein the performers and the audience communicate with each other.
2. *noksayk-sengcang*: The development of eco-friendly industries and new growth technologies dealing with renewable energies.
3. *noksayk-sanep*: Eco-friendly industry that produces less carbon dioxide to avoid pollution and protect the environment.

4. *phuleyntulli*: English loanword ‘friendly’.
5. *saipe-moyok-coy*: An offence of insult committed in cyberspace.
6. *pwulpich-chwukcey*: An illumination festival where buildings are decorated with lanterns of various sizes and colors to create a fantastic and magical ambience. Simplified word for the Italian loanword *lwuminaliey* ‘luminarie’.
7. *cenca-congi*: Electronic paper.
8. *haksup-cito-yolyeng-hayselse*: A guideline for elementary, middle and high school textbooks drawn up by the Japanese Ministry of Education, Culture, Sports, Science and Technology.
9. *kongceng-yehayng*: A field trip in the countryside which aims to teach about and share with local people and to help locals make profits.
10. *kunhwang-seylkha*: A selfie or self-video that shows what you are doing at that moment.

2009 neologisms

1. *sincong-phullwu*: The 2009 flu pandemic.
2. *incung-syas*: A picture taken to prove one’s own statements.
3. *colipep*: Simplified word for the English loanword *leysiphi* ‘recipe’.
4. *sincong-inphullwueynca*: The 2009 influenza pandemic.
5. *cwuyo-cayngcem*: Simplified word for the English loanword *has isyu* ‘hot issue’.
6. *yenki-tol*: An idol acting in Korean dramas.
7. *kkwul-pekci*: A noun to describe attractive thighs. Short for *kkwulhepekci* ‘honey thighs’.
8. *eysci-issta*: A verb describing an original and unique item of clothing and/or makeup style.

9. *phwumcel-nam*: An eligible, competent man who is already married.

10. *twayci-inphullwueynca*: Swine influenza.

2018 neologisms

1. *phanmwuncem-senen*: The proclamation of hope for peace, prosperity and unification of the Korean peninsula made by the leaders of the two Koreas at the Panmunjeom Peace House after the Panmunjeom Inter-Korean Summit in April 27, 2018.

2. *mi-thwu-wuntong*: Me Too Movement.

3. *ka-sim-pi*: Short for *kakyektaypi simlicek mancokuy piyul* ‘ratio of psychological satisfaction to price’.

4. *khophi-cenlyak*: Military strategy which consists of attacking the opponent first.

5. *kimchi-phulimiem*: The phenomenon whereby Korean virtual currency prices are higher than global prices.

6. *pheynsu-lwul*: The principle whereby men put some distance between themselves and women so as to prevent accusations of molestation or sexual assault.

7. *thongcang-yoceng*: A person who manages personal finances by opening several bank accounts for each spending purpose in order to save more money.

8. *kheyleynsia*: From Spanish *querencia*, a place where one feels at peace and can relax without being disturbed by anyone.

9. *sopi-yoceng*: A person who spends a lot of money.

10. *pelcip-kyeycwa*: Business account under which several personal accounts are opened in order to deal with virtual currency transactions and which the banks seek to prevent.