https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2045 (Lexicofocus)

Detection and Description of Neologisms in Korean Lexicography: Methodological Issues in Corpus Balance, Word Unit Bias and LLM Assistance*

Kilim Nam, *Department of Korean Language and Literature, Yonsei University, Seoul, South Korea (nki@yonsei.ac.kr)* (*https://orcid.org/0009-0001-9358-2673*)

Soojin Lee, *International Exchange Department, Kyungpook National University, Daegu, South Korea (sjmano27@naver.com) (https://orcid.org/0009-0002-1720-1880)*
and
Hae-Yun Jung, *International Exchange Department, Kyungpook National University, Daegu, South Korea (haeyun.jung.22@gmail.com) (https://orcid.org/0009-0006-4837-2569)*

Abstract: This study explores the potential application of large language models (LLMs) in Korean neologism extraction and dictionary compilation while critically examining the limitations of existing methods, including the bias toward news-oriented data and morphological neologisms. By analysing data from news corpora alongside messenger and online post corpora, the study identifies significant limitations in current news-centred approaches, particularly in detecting the first occurrences and extracting neologisms related to everyday topics. Experimental results involving LLMs demonstrate their potential to address the limitations of news-biased neologism extraction by suggesting unregistered words from diverse web-based contexts. However, issues such as duplication and overgeneration persist. In tasks involving semantic neologism recommendation and dictionary microstructure creation, LLMs performed relatively well with high-frequency and news-biased topics when provided with additional contextual prompts, yet revealed limitations with low-frequency and non-news-biased neologisms. These findings suggest that the performance of current LLMs heavily relies on the diversity of training data and user-provided contextual information. The results of this study underscore the need for further investigation into

* A version of this paper was presented at the 6th Globalex Workshop on Lexicography and Neology (GWLN-6), held on 3 July 2024 at the University of Pretoria, Hatfield Campus, Pretoria, South Africa.

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2045 (Lexicofocus)

the critical challenges in neologism research, lexicography, and corpus linguistics, as well as the role lexicography might play in enhancing the performance of LLMs.

Keywords: LEXICOGRAPHY, NEOLOGISMS, UNREGISTERED WORDS, NEWS CORPUS, SEMANTIC NEOLOGISM, REPRESENTATIVENESS, BALANCE, LEXICOGRAPHIC DATA, MACRO-STRUCTURE, LARGE LANGUAGE MODELS

Opsomming: Die opsporing en beskrywing van neologismes in die Koreaanse leksikografie: Metodologiese kwessies rakende korpusbalans, vooroordeel teenoor woorddele en ondersteuning deur groottaalmodelle. In hierdie studie word die potensiële toepassing van groottaalmodelle (GTM'e) in Koreaanse neologisme-onttrekking en woordeboeksamestelling verken, terwyl die beperkings van bestaande metodes, insluitend die vooroordeel ten gunste van nuusgeoriënteerde data en morfologiese neologismes, krities ondersoek word. Deur nuuskorpusdata naas kitsboodskap- en aanlynposkorpora te ontleed, identifiseer hierdie studie beduidende beperkings in huidige nuusgesentreerde benaderings, veral in die opsporing van eerste voorkomste en die onttrekking van neologismes wat met alledaagse onderwerpe verband hou. Empiriese resultate in die gebruik van GTM'e demonstreer hul potensiaal om die beperkings van nuusbevooroordeelde neologisme-onttrekking aan te pak deur ongeregistreerde woorde uit diverse webgebaseerde kontekste voor te stel. Kwessies soos duplisering en oorgenerering bestaan egter steeds. In take wat aanbevelings rakende semantiese neologismes en woordeboekmikrostruktuurskepping behels het, het GTM'e relatief goed gevaar met hoëfrekwensie- en nuusbevooroordeelde onderwerpe wanneer hulle van addisionele kontekstuele prompe voorsien is, maar tekortkominge het ook duidelik by laefrekwensie- en nienuusbevooroordeelde neologismes geblyk. Hierdie bevindings dui daarop dat die prestasie van huidige GTM'e sterk op die diversiteit van opleidingsdata en gebruikersverskafde kontekstuele inligting staatmaak. Die resultate van hierdie studie beklemtoon die behoefte aan verdere ondersoeke na die kritieke uitdagings in neologisme-navorsing, die leksikografie en korpuslinguistiek, sowel as na die rol wat die leksikografie in die verbetering van die prestasie van GTM'e kan speel.

Sleutelwoorde: LEKSIKOGRAFIE, NEOLOGISMES, ONGEREGISTREERDE WOORDE, NUUSKORPUS, SEMANTIESE NEOLOGISME, VERTEENWOORDIGENDHEID, BALANS, LEKSIKOGRAFIESE DATA, MAKROSTRUKTUUR, GROOTTAALMODELLE

1. Introduction

Corpus research, which was pioneered in lexicography, has since developed into the methodology of corpus linguistics and has been expanded, refined, and eventually adopted by neological studies, among others. News corpora in particular have been used as authentic language data and a strong basis for the identification of institutionalised neologisms (as opposed to nonce words), the dating of their first appearances, and the investigation of their usage trends. Their status as the greatest resource for neologism collection and study has been abundantly discussed (cf. Renouf 2013; Boussidan 2013; Nam et al. 2020; Freixa and Adelstein 2013; Klosa and Lüngen 2018). However, the spread of web languages and

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2045 (Lexicofocus)

the emergence of large language models (LLMs) have a considerable impact on the creation and diffusion of neologisms, as well as on the application of language resources. In that sense, it has become crucial to re-examine not only the bias of [+formal] and [+written] news corpus, but also the neologism extraction methods centred on single-word units. This study critically reviews the methodology for Korean neologism research, which has relied on the semi-automated extraction from news corpora from 2005 to date, and explores ways to improve dictionary compilation to reflect the dynamics of language from the actual language use by different communities.

This paper is organised as follows. Section 2 reviews current methodologies for extracting neologism extraction and dictionary compilation and presents the research methodology used for this study. Three corpora have been compared and analysed, namely a news corpus of 500 million words, a 14-million-word corpus of online posts from forums and social media, and a 6-million-word instant messages (IM) corpus, which roughly span from 2020 to 2022.

Sections 3 and 4 test the potential of large language models (LLMs) for neologism detection and description as recent LLMs have begun to be discussed in language research and dictionary compilation. Their ability to identify and describe neologisms is reviewed to see if they can somewhat compensate for the limitations of existing neologism extraction methods through learning from large-scale data. Python programmes are used for corpus analysis and processing, and the detection of unregistered words draws on the largest Korean dictionary, *Urimalsaem*, as the accumulated list of neologisms to date. The study identifies three key issues related to the detection of neologisms and their lexicographic description, each of which is addressed in a section.

Firstly, Section 3.1 investigates issues related to the source genre and the news bias of neologisms. Based on domain, frequency, and distribution analyses, it is shown that neologisms mainly reflect news topics, such as politics, economy, and society, and comprise very few expressions pertaining to daily life, such as food-related terms or emotion terms. Despite their low frequencies in newspapers, everyday expressions that are frequently used by Korean speakers, especially those reflecting their attitudes or feelings on everyday topics, hold significant value as headword candidates. This section therefore argues for the need to diversify corpus resources beyond newspapers in order to collect neologisms that reflect native speakers' expressions and to build dictionaries that capture the dynamism of language.

Section 3.2 secondly examines how a comprehensive representation of the lexicon of a given era requires not only the identification of new forms, but also the analysis of frequencies, distribution, and discourse contexts. This entails the extraction and lexicographic description of semantic neologisms in addition to formal neologisms. Case studies are drawn upon to show that neologism research must be discussed in discourse context, which is closely related to phrase-level headwords, such as typical collocations and patterns.

Finally, Section 4 discusses the potential and limitations of LLMs in collecting and describing neologisms, both as an alternative to the issues raised in

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2045 (Lexicofocus)

the previous two sections, and as a new approach to lexicography. Although recent studies — including McKean and Fitzgerald (2023), Lew (2023, 2024), and De Schryver (2023) — have discussed the use of LLMs for lexicographic description, there has been little research discussing the current state of LLMs in the detection and description of neologisms or unregistered words. Two factors could explain this lack of research: the training data of LLMs is firstly not up-to-date, and secondly, neologisms only play a small part in lexicographic description. However, the detection of neologisms and unregistered words, which relies on corpora and lexicographers' intuition, still faces a number of challenges, such as the diversity in genres and topics, the exhaustiveness of detection (which tends to display a bias toward formal neologisms), and the subjectivity of neologism description. The section explores whether LLMs can contribute to the description of neologisms and unregistered words and enhance lexicographic productivity by means of concrete and extensive experiments. These experiments focus on the identification of neologisms, the recommendation of headword candidates, and the description of lexical microstructures. The experiments involve the main LLMs for Korean, namely the native Naver CLOVA[1] and the international model ChatGPT. The evaluation of their performances and the comparison between the two show how the volume of Korean data used in training, as well as the training based on existing dictionaries, can affect their ability to identify and describe neologisms.

More than twenty years ago, Sinclair (2004: 188-192) emphasised the need for the linguistics community to prepare for larger, unannotated raw corpora to contribute to the future information society. In the era of LLMs, it has become crucial to examine whether large corpora alone are sufficient for automatic identification of neologisms and unregistered words, and to balance corpora and LLMs as linguistic resources and methodologies in order to address written-language bias and complement human intuition. Thus, this paper argues that the meaningful contribution of neologism research to society in the era of LLMs entails building larger corpora, redefining corpus diversity and balance, developing new extraction methodologies, and compiling dictionaries that reflect the linguistic dynamics of native speakers.

2. Research background and methodology

2.1 Overview of the Korean Neologism Investigation Project

In terms of sources, the Korean Neologism Investigation Project (KNIP) is primarily focused on major daily newspapers, and in terms of methodology, it can be characterised as a semi-automatic approach to neologism extraction, that is, combining automatic extraction and manual collection of new form candidates. Table 1 summarises the types of sources and methodologies[2] used by the National Institute of Korean Language (NIKL) as well as by the Kyungpook National University (KNU) research team (highlighted in grey) in the neologism investigation project from its inception to date.[3]

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2045 (Lexicofocus)

Table 1: Sources and methods for the investigation of neologisms by the NIKL and KNU

	Time spans	Sources	Methodology
Phase 1	1994–2001	major national daily newspapers and magazines	manual extraction
Phase 2	2002–2004	major national daily newspapers and broadcasts	not mentioned
Phase 3	2005–2010	major national daily newspapers and broadcasts	corpus, semi-automatic extraction
Phase 4	2012–2019	around 130 online news websites	corpus (web-based extractor), semi-automatic extraction
Phase 5	2020 to date	5 online news websites	corpus, semi-automatic extraction

The issues raised here are twofold: (1) a bias toward the written language of the news genre as the target of the investigation, and (2) limitations in semi-automatic analysis based on 'forms', reflecting the constraints of the underlying corpus and extraction methodology, respectively. Despite the premise that in any corpus-based research, including corpus-based lexicography, corpus representativeness and corpus balance must be ensured, the collection of neologisms has relied on mass media. The main reason for this is that mass media provide evidence of 'institutionalised' neologisms, as opposed to nonce words. Major daily newspapers in particular provide information fulfilling crucial criteria, such as the date of first occurrence, frequency and distribution. Nonetheless, collecting news-based neologisms also presents limitations, including the restricted scope of neologisms, as these mostly revolve around public affairs, and the inability to identify the first occurrence and spread of some neologisms.

(1) Examples of news-biased neologisms
  a. Domain bias: *kinkupcaynansayngkyeyciwenkum*[4] ('emergency disaster relief support'), *cwungswucheng* ('Serious Crimes Investigation Agency'), and *panmwuncemsenen* ('Panmunjeom Declaration')
  b. First appearance and spread bias: *phathpwung* ('red bean pastry') (blogs: 2012; news: 2021), *mayneekkey* ('manners shoulder') (blogs: 20 occurrences in the past year; news: 0 occurrences)

Example (1a) provides typical examples of domain-biased neologisms when these are only extracted from news sources. Such neologisms focus on topics related to public affairs, such as politics, administration, law, economy, and society. In contrast, compared to neologisms concerned with public affairs, private communication texts tend to generate and spread more neologisms related to everyday

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2045 (Lexicofocus)

topics, such as food and hobbies, or expressions of personal feelings. As a result, the frequency of expressions such as those presented in (1a) is relatively low outside of news texts. *Phathpwung* ('red bean pastry') in (1b) first appeared in blogs in 2012 but was not mentioned in newspapers until 2021; furthermore, *mayneekkey* ('manners shoulder') was mentioned more than 20 times in blogs in the past year, but not once in newspapers. These examples show the limitations of news sources in fully capturing the first appearance and spread of neologisms while these rapidly spread in online media.

As for the limitation of form-centred neologism extraction, it correlates with the main characteristics of current neologism research methods. At present, the extraction of Korean neologisms consists of semi-automatic extraction of neologisms based on the comparative analysis of neologism candidates from a large-scale news corpus against existing lexical resources. The identification of neologisms based on such a form-centred comparative analysis usually means dismissing semantic neologisms or grammatical neologisms, wherein new meanings are created even though they do not involve new forms (Nam 2020; Nam 2021; Nam and An 2023). Therefore, phenomena such as semantic extension or acquisition of metaphorical meanings through new collocational combinations, as illustrated in Example (2), are more often than not overlooked in the extraction and lexicographic description of neologisms.

(2) Examples of semantic neologisms:
  a. *phokphwung* ('storm'): (1) storm warning, tropical storm; (2) storm (as in 'extreme') growth/tears/eating
  b. *konghwakwuk* ('republic'): (1) democratic republic, Hellenic Republic; (2) apartment/Seoul/real estate republic

The creativity of native speakers includes not only grammatical creativity as discussed since Chomsky (1957), but also morphological creativity and semantic creativity. Specifically, these refer the creation of new words or sentence structures, and the creation of new meanings through novel collocations or metaphors, respectively. In this context, fundamental methodological changes seem crucial to ensure that these creative meanings are no longer overlooked in neologism extraction and dictionary description. These two issues, namely the bias towards news language and the bias towards formal neologisms, are further discussed in Sections 3.1 and 3.2, respectively.

2.2 Methodology

This study focuses on large-scale neologism collection efforts combined with corpus use from 2005 to the present and surveys a total of 11 neologism datasets (2005, 2008–2010, and 2012–2022). In addition, the study examines two other corpora, which consist of distinct registers from the news and were predominantly used in previous neologism studies, to assess the adequacy of the linguistic

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2045 (Lexicofocus)

resources and methodology used for neologism collection. These corpora are the *Instant Messaging Corpus* (IM) and *Online Post Corpus* (OP) compiled by NIKL as part of the *Modu Corpus* (lit. 'All Corpora'). Additionally, the *Newspapers Corpus*, also from the *Modu Corpus*, was used for comparison. Table 2 shows the size of each corpus in *ecel* unit, which is the Korean word unit based on word spacing.[5]

Table 2: Size of the corpora under study in *ecel* (space unit word)

Source	Corpus	**Time Segment 1 (01-07-2019–30-06-2020)**	**Time Segment 2 (01-07-2020–30-06-2021)**	**Time Segment 3 (01-07-2021–30-06-2022)**	**Total**
NIKL	News	167 286 198	225 402 761	100 892 770	493 581 729
	IM	2 826 229	2 834 967	788 620	6 449 816
	OP	3 206 220	4 969 648	6 371 497	14 547 365

These corpora were analysed to identify issues related to the biases of linguistic resources predominantly used in previous neologism investigations. In other words, this study examines whether neologisms were omitted and whether certain neologisms first appeared in media or genres other than news outlets. The datasets shown in Table 2 actually have substantial differences in terms of time spans; as a result, the study focuses on a relatively short time frame of three years, during which all three datasets overlap. The selection of this time frame aligns with the characteristics of the neologisms focused on in this study. Neologisms constitute a highly dynamic lexical category, exhibiting emergence and disappearance within very short periods of time. As such, the analysis of the subtle dynamics of neologisms calls for a short-term diachronic approach,[6] as opposed to traditional diachronic studies looking into changes over centuries.

3. Korean neologism research: Current status and corpus linguistic issues

3.1 The bias of news language and challenges in neologism extraction

The advantage of using newspapers and news corpora for collecting neologisms lies not only in their widespread dissemination across the Korean-speaking community, but also in their capacity to reflect societal changes quickly. Nonetheless, it is evident that news data cannot fully represent the linguistic reality of Korean native speakers. Therefore, biases related to news data can also be found in the results of the neologism investigations that are based on such data. For instance, Lee (2022) analysed the semantic categories of neologisms from 2015 to 2021 and found a high proportion of neologisms in the categories of

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2045 (Lexicofocus)

Politics, Economy, and Society, correlating such results to the limitation of news text genre as the primary source for neologisms.

In the present study, the IM and OP corpora have been analysed for comparison with news sources, focusing on the date of first occurrence and topic diversity in order to identify biases related to written language. If a given neologism has been actively used in other domains long before it appeared in the news texts, it becomes evident that the timing of first occurrences needs to be revisited. In the same vein, if neologisms from news sources are biased toward topics pertaining to public affairs due to the [+public] nature of the news genre, the question of collecting neologisms that emerge and are used in [+private] domains also needs to be addressed.

As seen in Table 2, the *Newspaper Corpus*, *IM Corpus*, and *OP Corpus* were divided into three time segments of one year each. The second time segment (July 2020–June 2021) coincides with the time frame for the 2021 neologisms, and the third time segment (July 2021–June 2022) partially coincides with the time frame for the 2022 neologisms.[7] When analysing non-news genre data (i.e. instant messaging texts and online posts) separately, it appeared that a few neologisms collected in 2021 and 2022 occurred in the *IM Corpus* and the *OP Corpus* at earlier date than the extraction time frame, as shown in Table 3.[8]

Table 3: Neologisms of 2021 and 2022 that appear in IM and OP before News

	Neologism	First occurrence within Time Segment 1		First occurrence within Time Segment 2	
		IM	Online posts	IM	Online posts
2021 neologisms	*ttasangsang* ('double top stock')	✓			
	ssalmek ('items-for-rice')	✓			
	ccomccomttali ('teeny-weeny bits')	✓			✓
2022 neologisms	*phathpwung* ('red bean pastry')	✓			✓

Based on the News Corpus, the neologism *phathpwung* ('red bean pastry') was extracted within the time frame for the 2022 neologism collection. However, Table 3 shows it first appeared within the Time Segment 1, rather than Time Segment 2. Of these four neologisms, all but *ccomccomttali* ('teeny-weeny bits') share the common feature of being abbreviations, and only *ttasangsang* ('double top stock'), which falls under the domain of Economy, belongs to the semantic domains of Politics, Economy, and Society.

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2045 (Lexicofocus)

Table 4: First occurrence dates of *ttasangsang* ('double top stock'), *ssalmek* ('items-for-rice'), *ccomccomttali* ('teeny-weeny bits), *phathpwung* ('red bean pastry') in News and Blogs

	Neologism	Naver News in-links	Naver News out-links	Blogs
2021 neologisms	*ttasangsang* ('double top stock')	03-07-2020	06-07-2020	17-11-2019
	ssalmek ('items-for-rice')	10-07-2020	14-10-2021	07-08-2017
	ccomccomttali ('teeny-weeny bits')	31-01-2021	09-12-2020	27-02-2017
2022 neologisms	*phathpwung* ('red bean pastry')	09-12-2021	27-12-2020	13-02-2012

Table 4 shows the first occurrence dates of the four neologisms in Naver News in-link data as the timing criterion for neologism extraction and compares with their appearances in Naver News out-link data and blogs.[9] The 2022 neologism *phathpwung* ('red bean pastry') should be included within the time frame of the 2021 neologisms when expanding the scope to Naver News out-link data, even though the sources are still limited to the news genre. If the scope is further extended to blog data, the neologism should have been collected as a 2012 neologism as it appeared in blogs as early as February 2012. As the other three neologisms show the same tendencies, it could be assumed that the first occurrence dates of many other neologisms would be significantly earlier if the timing criterion was expanded to sources other than news.[10] Given the popularity and diversity of speakers in IM platforms (Huang and Nam 2023) and considering the multiplicity and receptivity of blogs as both interpersonal and mass communication (Smyk-Bhattacharjee 2009), web out-links need to be taken into consideration as a means to reflect the linguistic realities of native speakers that could potentially offer a more effective alternative to news media.

To explore the issue of written register bias further, derivatives based on the same suffix *-sulep-* are examined and compared in terms of occurrences and distribution across news, IM, and OP corpora. The suffix *-sulep-* is used to create adjectives meaning 'being like/having the quality of' the noun it is affixed to, and was chosen for its established high productivity, thereby providing a particularly suitable case for analysis. The suffix *-sulep-* ('-like') can be easily combined within rules and without significant restrictions with most types of nouns, including common nouns, proper nouns, native Korean words, and loanwords. As a result, many derivatives are nonce-words or short-lived expressions that do not undergo the process of institutionalisation. However, the regularity of derivation the suffix presents can differ depending on the genre or use domain.

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2045 (Lexicofocus)

Table 5 presents the number of occurrence types and the size of the corpus of unregistered -*sulep*- derivatives across news, IM, and OP corpora.

Table 5: Occurrence frequency for unregistered -*sulep*- derivatives by text genre

	News	IM	OP
Number of types	113	142	245
Corpus scale (in *ecel*)	220 million	10 million	10 million

The *News Corpus* is by far the largest corpus, followed by the *OP Corpus* and the *IM Corpus* in that order. However, the number of unregistered -*sulep*- derivatives is the lowest in the *News Corpus*, the largest number being found in the *OP Corpus*. In other words, many more -*sulep*- derivatives are created in a corpus that is much smaller than the *News Corpus*. This underlines the close correlation between the characteristics of the text genre and the creation of new words. Unlike newspapers, which tend to use a more formal and prescriptive language, online posts and instant messages allow users to deviate from norms freely, to play with language, and to create new expressions. Therefore, to capture a wide range of new words, it is necessary to look into diverse linguistic resources that take into account the characteristics of text genres and linguistic creativity, rather than solely relying on the size of the corpus.

(3) Examples of occurrences that are common to news and IM/OP genres (20 in total)
  a. *pwukhansulepta* ('North Korean-like') and *hankwuksulepta* ('South Korean-like')
  b. *pokswungasulepta* ('peach-like') and *ppangcipsulepta* 'bakery-like')
  c. *opesulepta* ('over-ish') and *cwacelsulepta* ('frustration-inducing')
  d. *salamsulepta* ('people-like') and *sencinkwuksulepta* ('developed country-like')

(4) Examples of occurrences exclusive to IM/OP genres (307 in total)[11]
  a. *tiolsulepta* ('Dior-like'), *meylukheylsulepta* ('Merkel-like'), and *sewulsulepta* ('Seoul-like')
  b. *chokhosulepta* ('chocolaty') and *haympekesulepta* ('hamburger-like')
  c. *kkamccaksulepta* ('surprise-inducing') and *ttiyongsulepta* ('boing-like')
  d. *kwiyepppoccaksulepta* ('cutie pie-like') and *kokwumasulepta* ('sweet potato-like')
  e. *cwummasulepta* ('auntie-like') and *cwung2sulepta* ('teenage-like'); *kemchalsulepta* ('prosecution-like') and *kwuninsulepta* ('military-like')
  f. *choposulepta* ('beginner-like') and *weneminsulepta* ('native-like')

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2045 (Lexicofocus)

g. *keyimsulepta* ('gaming-like'), *okhelthusulepta* ('occult-ish'), and *weypthwunsulepta* ('webtoon-like')
h. *kolmoksulepta* ('backstreet-like'), *khapheysulepta* ('café-like'), and *hyuyangcisulepta* ('vacation spot-like')

(5) Examples of exclusive 'Newspaper' types (74 in total)
a. *khakhaosulepta* ('Kakao-like') and *yunsekyelsulepta* ('Yoon Suk-Yeol-like')
b. *panillasulepta* ('vanilla-like') and *sukhonsulepta* ('scone-like')
c. *kammyengsulepta* ('impression-making') and *ongcolsulep* ('narrow-minded-ish')
d. *wisensulepta* ('hypocrisy-like') and *huymangsulepta* ('hope-like')
e. *haksayngsulepta* ('student-like') and *kongmwuwensulepta* ('civil servant-like')
f. *loksulepta* ('rock-like') and *khenchyulisulepta* ('country-like')

The *-sulep-* derivatives presented in Examples (3) to (5) have been grouped according to the base type *-sulep-* is suffixed to: (a) corresponds to proper nouns, (b) food-related terms, (c) emotion-related terms, and (d) object attributes or characteristics, all four categories being common to news, IM and OP genres. On the other hand, (4.e) refers to terms denoting [person] terms based on age or occupation, while (4.f) corresponds to [person] terms describing their personality or characteristics. Example (4.g) refers to contents consumed for hobbies or entertainment, and (4.h) groups terms whereby the suffix is combined with a place-related base. Just as (4.e) and (4.f), the bases in (5.e) represent [person] terms pertaining to profession, personality, or characteristics; the main difference is that (5.e) has a limited range of examples compared to the much more varied types observed in (4.e) and (4.f). Example (5.f) is somewhat akin to (4.g) in that the bases denote it subgenres of [music] as cultural contents to be enjoyed or consumed.

Among the types of *-sulep-* derivatives that appeared exclusively in the news genre, proper noun bases mainly pertain to politicians, celebrities, or companies' names, and emotion-related bases are terms frequently found in formal written language, such as *kammyeng* ('impression') and *ongcol* ('narrow-mindedness'). In contrast, the derivatives appearing only in non-news genres, often include combinations with food-related nouns — e.g., *chokho* ('choco') and *haympeke* ('hamburger') — and with terms related to leisure time — e.g. *keyiming* ('gaming'), *okhelthu* ('occult'), and *weypthwun* ('webtoon'). In addition, bases denoting emotions, such as the root *kkamccak* ('surprise') or the onomatopoeia *ttiyong* ('boing'), further highlight the difference from derivatives exclusive to the news genre.

These findings suggest that the thematic characteristics of the text genre can have a significant impact on the type of vocabulary collected. For instance, the news genre predominantly reflects topics such as politics, economics, and society, whereas IM and OP genres focus on everyday topics, such as food or hobbies. This confirms the need to diversify domains and genres beyond news sources to more fully represent the linguistic creativity of native speakers.

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2045 (Lexicofocus)

3.2 The bias of formal neologisms

From the start, KNIP has sought to detect semantic neologisms, defined as existing word forms used with a different meaning, along with formal neologisms. Accordingly, the early neologism reports marked headwords with the symbol ♣ or added phrases such as '♣ different meaning' or '※ different meaning' at the end of the entry to indicate that the headword was a semantic neologism, as shown in Example (6).

(6) *kancephwapep* 'indirect statement' (*1994 Neologism Report*, p. 3). 'In the early 1980s, Castro announced, by means of indirect statements made in state-run newspapers, the government's position that Cuban residents could emigrate overseas without interference from authorities.' (*Joseon Ilbo*, 94.08.21, p. 2) ♣ different meaning

In total, there are around 260 examples of semantic neologisms across the Neologism Reports, which represent a small number compared to formal neologisms. The yearly trends in the collection of semantic neologisms are presented in Table 6, with representative cases highlighted in grey.

Table 6: Number of semantic neologisms in the neologism reports

Year	1994	1995	2000	2001	2002	2003	2004	2005	2008	2009	2010	2012	total
Semantic neologisms	45	26	53	25	36	24	41	3	3	3	3	1	263

What can be immediately observed from Table 6, is the significant decrease in the number of semantic neologisms collected from Phase 3 onwards, that is, following the introduction of semi-automatic extraction in 2005, compared to the first and second phases, which heavily relied on manual collection. Furthermore, no more semantic neologisms have been collected from Phase 4 (2012–2019), when web-based automatic extraction was implemented, through to Phase 5 (2020–). While the development of Korean neologisms research can thus be divided into five phases, progressing from manual extraction of neologisms to extraction through static corpora and the use of dynamic time-series corpora, the interest in semantic neologisms has either stagnated or become even more passive.

In fact, the collection of semantic neologisms even seemed to retrogress, with more semantic neologisms included in the datasets during the manual collection phase than during the semi-automatic extraction phase. Moreover, the consideration of semantic neologisms has further weakened with the use of large-scale corpora, since semantic neologisms, having the same form as existing words, are ipso facto excluded from the list of neologism candidates, which is from the very first stage of the neologism investigation. Recent studies have nonetheless

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2045 (Lexicofocus)

looked into the automatic identification and detection of semantic changes in large corpora (cf. Boussidan 2013; Renouf 2013; Nam et al. 2018; Nam et al. 2019). While automatic and semi-automatic identification and detection of meanings have been proven feasible, it is still difficult to propose methods that can be directly applied to actual collection tasks. Nevertheless, the need remains to establish criteria for identifying the emergence and first occurrence date of new meanings, just as for formal neologisms. Continuous research and discussion are thus needed for determining what constitutes semantic novelty and for identifying the first occurrence of semantic neologisms.

4. Review of Korean LLMs for the lexicographic compilation of neologisms

4.1 Korean neologism research and experiments using LLMs

A number of experimental studies have recently been conducted to examine how LLMs can contribute to the automation of dictionary compilation and increase lexicographic productivity in line with lexicographic advancements, such as the use of corpora and computers, the development of editing tools, and the participation of dictionary users. McKean and Fitzgerald (2023) explore the potential and limitations of LLMs in lexicography by testing the performance of ChatGPT in headword recommendation and entry description. Other studies, including De Schryver (2023) and Lew (2023, 2024), discuss the impact and potential applications of LLMs in dictionary compilation extensively. A common conclusion is that although LLMs have shown considerable promise for automating lexicography, significant issues remain regarding accuracy, such as incorrect examples and hallucinations. However, most of these studies have not included discussions on the detection of neologisms and unregistered words, which play an important role in macrostructure expansion, nor on the description of neological headwords.

In light of this underexplored area, this section discusses the potential applicability of LLM-based dictionary compilation to the detection and description of Korean neologisms and unregistered words. As seen in Section 3, there are clear limitations in the collection and description of neologisms, having relied heavily on corpora and human intuition. Two scenarios could thus unfold from experimenting with LLMs in regard to the description of neologisms and unregistered words. Firstly, LLMs may be able to overcome or complement the limitations of existing neologism detection and description methodologies. In this case, LLMs could serve as a new methodology to address the shortcomings of traditional methodologies, namely the reliance on human intuition and corpora, thereby significantly improving lexicographic productivity. The second scenario is the opposite outcome: If LLMs are found unreliable in recommending and describing new and unregistered words, the role of lexicographers with 'intuition' and the construction of lexicographic databases will become even more critical. Moreover, the importance of lexicographic databases and descriptions

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2045 (Lexicofocus)

that can compensate for the weaknesses and limitations of LLMs will need to be discussed in greater depth.

For the experiments of detection and description of Korean neologisms or unregistered words,[12] this study uses OpenAI's ChatGPT (GPT-4 version), the most widely known generative AI and international model, and Naver CLOVA X (HyperCLOVA X version), a domestic model.[13] For both AI models, the experiments consisted of entering prompts on the user platforms.

In Section 4.2, LLMs are tested on the recommendation of neologism candidates spanning a broad range of genres to address the news bias discussed in Section 3.1. In Section 4.3, LLMs are prompted on the definition as well as the description of the semantic neologisms.

4.2 Experiments on news bias: Detection of neologisms and unregistered words

Section 3 examined the issue of genre bias in the semi-automatic extraction methods based on news-centred corpora traditionally used in neologism research, as demonstrated by the cases of unregistered *-sulep-* ('-like') derivatives. Here the question is whether LLMs trained on large-scale data encompassing various sources, such as encyclopaedias like Wikipedia, books, academic articles, news, blogs, forums, and other websites, can serve as an alternative.

To answer this, the two LLMs that are most utilised by Korean users (ChatGPT and CLOVA X) have first been tested to assess whether they can effectively perform the task of recommending neologisms and unregistered words. The query was for them to recommend a list of *-sulep-* derivatives not found in existing dictionaries, as shown in Example 7. In the case of ChatGPT, the LLM was trained on Korean data from the state-led *Standard Korean Language Dictionary* and the headword lists of the Korean dictionaries provided by Naver[14] and Daum[15] portals. Similarly, CLOVA X learned from the headword lists of the *Standard Korean Language Dictionary*, *Urimalsaem*, and the *Korea University Korean Dictionary*, all provided by Naver Dictionary Service.[16]

(7) Prompt:[17]
Make a list of 35 derivatives with the suffix *-sulep-* '-like'. Only include words that meet the following conditions:
- Not already included in an existing dictionary.
- Not AI-invented: they must be words that people have actually used.

(8) Generative AI answer: ChatGPT
a. {*chinkwu* 'friend', *kamseng* 'feeling', *hyensil* 'reality', *inkan* 'human', *kacok* 'family', *chengchwun* 'youth', *phulo* 'pro', *pinthici* 'vintage', *koyangi* 'cat', *hipsuthe* 'hipster', *wucwu* 'space', *semin* 'commoner', *tokthuk* 'unique', *ipeynthu* 'event', *centhong* 'tradition', *hyuka* 'vacation', *tosi* 'city', *yelceng* 'passion', *mohem* 'adventure', *chwimi* 'hobby', *yelceng* 'passion', *yehayng*

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2045 (Lexicofocus)

'travel', *milay* 'future', *cengcik* 'honesty', *yuhayng* 'fashion', *chopo* 'beginner', *tokse* 'reading', *cayen* 'nature'}*sulepta* '-like'

b. {*cayu* 'freedom', *sonye* 'girl', *phyenghwa* 'peace', *kamtong* 'emotion', *chincel* 'kindness', *sikol* 'countryside', *cayen* 'nature', *pimil* 'secret'}*sulepta* '-like'

Of the 35 derivatives suggested by ChatGPT, *yelceng* ('passion') is suggested twice, and 8 suggestions are already included in the Korean language dictionary *Urimalsaem* (8.b), reducing the list of suggestions to 26 unregistered derivatives. Additionally, 12 of them can be found in the *News*, *IM*, and *OP* corpora under study (Example 9).

(9) ChatGPT suggestions found as unregistered words in the corpus under study

a. OP: {*kamseng* 'feeling', *koyangi* 'cat', *milay* 'future', *pinthici* 'vintage', *wucwu* 'space', *centhong* 'tradition', *chengchwun* 'youth', *chopo* 'beginner', *phulo* 'pro'}*sulepta* '-like'

n. IM: *seminsulepta* 'commoner-like'

d. News: *milay* 'future', *yuhayng* 'fashion', *centhong* 'tradition', *chengchwun* 'youth', *hipsuthe* 'hipster'}*sulepta* '-like'

Out of these, {*milay* 'future', *centhong* 'tradition', *chengchwun* 'youth'}*sulepta* '-like' are common to both the *OP Corpus* and the *News Corpus*.

(10) ChatGPT suggested unregistered words with actual web usage[18]

{*chinkwu* 'friend', *hyensil* 'reality', *inkan* 'human', *kacok* 'family', *tokthuk* 'unique', *ipeynthu* 'event', *hyuka* 'vacation', *yelceng* 'passion', *mohem* 'adventure', *chwimi* 'hobby', *yehayng* 'travel', *cengcik* 'honesty', *tokse* 'reading'}*sulepta* '-like'

While {*tokthuk* 'unique', *ipeynthu* 'event', *cengcik* 'honest', *tokse* 'reading'}*sulepta* '-like' are likely nonce-words given their fairly low Google search frequencies, it is nonetheless worth noting that none of the 35 ChatGPT suggestions were "invented" by ChatGPT as actual forms used by Korean speakers. In other words, both Examples (9) and (10) can be considered as headword candidates, although further research on their actual frequencies and usage patterns may be needed.

(11) Generative AI Answers: CLOVA X

a. *kwiyemsulepta* 'lovable', *angcungsulepta* 'cute', *messulepta* 'stylish', *nekulewumsulepta* 'magnanimous', *alumtawumsulepta* 'beauty-like', *calangsulepta* 'proud', *mancoksulepta* 'satisfying', *tahayngsulepta* 'lucky', *kekcengsulepta* 'worrisome',

b. *cengtapta* 'affectionate'/*cengtapta* 'affectionate', *alumtawum* 'beauty'/ *alumtawum* 'beauty',

c. *sulkilopta* 'wise', *sulkilowum* 'wisdom'/*sulkilowum* 'wisdom',

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2045 (Lexicofocus)

d. *hayngpoksulewum* 'happiness'/*hayngpoksulewum* 'happiness'/*hayngpoksulewum* 'happiness', *salangsulewum* 'loveliness'/*salangsulewum* 'loveliness', *cayensulewum* 'naturalness'/*cayensulewum* 'naturalness', *poksulewum* 'fortune'/*poksulewum* 'fortune', *kwiyemsulewum* 'lovability'/*kwiyemsulewum* 'lovability', *angcungsulewum* 'cuteness'/*angcungsulewum* 'cuteness', *messulewum* 'stylishness'/*messulewum* 'stylishness'
e. *calangsulewum* 'proudness', *mancoksulewum* 'satisfaction', *tahayngsulewum* 'luckiness', *kekcengsulewum* 'worry'

The first issue with CLOVA X answer is that it included not only -*sulep*- derivatives but also -*tap*- (11.b) and -*lop*- (11.c) derivatives. While the latter two can be considered synonymous to -*sulep*- ('-like'), it was not the task requested, suggesting that CLOVA X is likely to include unwanted noise in the results when asked to recommend unregistered words or neologisms as headword candidates. The second issue is the high number of duplicates.[19] Although CLOVA X suggested a list of 35 items as requested by the prompt, 10 suggestions are actually repeated two or three times as can be seen in (11.b to 11.d). Another problematic aspect of CLOVA X results is that 6 items are -*sulep*-derivatives in the adjectival form -*sulepta* (11.a). Most of the other derivatives returned by CLOVA X are the nominalised form -*sulewum* (11.d and 11.e). This again leads to the issue of duplicates as shown in Example (12), where pairs such as *kekcengsulepta* 'worrisome'/*kekcengsulewum* 'worry' are simply two word forms of the same lexeme.

(12) Adjective/noun duplicates suggested by CLOVA X
kekcengsulepta 'worrisome'/*kekcengsulewum* 'worry', *kwiyemsulepta* 'lovable'/*kwiyemsulewum* 'lovability', *tahayngsulepta* 'lucky'/*tahayngsulewum* 'luckiness', *mancoksulepta* 'satisfying'/*mancoksulewum* 'satisfaction', *messulepta* 'stylish'/*messulewum* 'stylishness', *angcungsulepta* 'cute'/*angcungsulewum* 'cuteness', *calangsulepta* 'proud'/*calangsulewum* 'proudness'

Ultimately, after removing all the duplicates and noise, the -*sulep*- derivatives suggested by CLOVA X only amount to 13 items, listed in Example (13).

(13) CLOVA X suggestions of -*sulep*- derivatives
a. {*kekceng* 'worry', *kwiyem* 'love', *tahayng* 'luck', *mancok* 'satisfaction', *mes* 'style', *angcung* 'cuteness', *calang* 'pride'}*sulepta* '-like'
b. {*nekulewum* 'magnanimity', *alumtawum* 'beauty'}*sulepta* '-like'
c. {*pok* 'fortune', *salang* 'love', *cayen* 'nature', *hayngpok* 'happiness'}*sulewum* 'likeness'

Of the 13 suggestions in Example (13), only *kwiyemsulepta* ('lovable'), *nekulewumsulepta* ('magnanimous'), and *alumtawumsulepta* ('beauty-like') are unregistered words, the other 10 items being included in *Urimalsaem*. However, only *kwiyemsulepta*

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2045 (Lexicofocus)

('lovable') can be confirmed as being used by Korean speakers on the web. The other two derivatives are nominalised forms of adjectives and the combination of such forms with the suffix *-sulep-* is rather awkward in Korean. In fact, the search query of *nekulewumsulepta* ('magnanimous') in browsers returned no results and the few web occurrences of *alumtawumsulepta* ('beauty-like') seem to be artificially generated text rather than examples of natural Korean usage.

This experiment of querying recommendations of unregistered *-sulep-* derivatives showed that ChatGPT performed better than CLOVA X. Further investigation is yet needed to determine whether the results would differ with prompts that include additional information regarding the combinatory constraints of the suffix *-sulep-*.

4.3 Identification and description of semantic neologisms

To overcome the bias of formal neologisms, the adequacy of ChatGPT and CLOVA X was tested by prompting the tools to recommend 10 headword candidates that qualified as semantic neologisms (i.e., words that are already in the dictionary but are actually used with different meanings). Neither LLM could provide satisfactory suggestions, so they were tested again to examine their ability to identify different meanings of a word based on the given contextual information. To this end, the LLMs were provided with examples of different usages of the well-known semantic neologism *kokwuma* ('sweet potato'), collected from newspapers and blogs, as shown in Example (14).

(14) Prompt:
These days, the word *kokwuma* 'sweet potato' is used with two different meanings. Based on the following examples where *kokwuma* 'sweet potato' is used, write the definitions for the two different senses.
Examples:
- In the past, after harvesting barley, we planted *kokwuma*.
- Boil the *kokwuma* and peel them.
- It was a well-made and interesting drama, but it had a curious balance of '*kokwuma*' and 'cider'.
  (…)

(15) Generated AI answers
a. ChatGPT: Meaning of plant and food/figurative meaning (expressing frustration)
b. CLOVA X: Sweet potato as a crop produce/Sweet potato as a metaphor for a frustrating situation

As seen in Example (15), both models successfully distinguished the two meanings of *kokwuma* ('sweet potato'). However, as they also classified the examples according to the meaning of *kokwuma*, ChatGPT correctly classified the sentences under each meaning, whereas CLOVA X misclassified two sentences under the metaphorical sense instead of the literal sense and omitted one sentence. These results are based on the prompt provided in Example (16), which was given to both LLMs to generate definitions. Examples (17) to (19) compare the definitions of *kokwuma* drafted by the two LLMs and the dictionary entry written by a lexicographer in *Urimalsaem.*

(16) Prompt:
Write a dictionary entry for *kokwuma* 'sweet potato' as a headword for a Korean dictionary; make sure to include a definition and examples.

(17) Generated AI answer: ChatGPT
noun
1. [Plant] A perennial vine belonging to the morning glory family (…)
2. [Metaphorical expression] A term used metaphorically to describe a situation or person's behavior that feels frustrating and distressing. It is commonly used when a story develops slowly or a situation is not progressing.
   Examples:
   - [Plant] In the past, after harvesting barley, we planted *kokwuma*. | Boil the *kokwuma* and peel them.
     (…)
   - [Metaphorical expression] It was a well-made and interesting drama, but it had a curious balance of '*kokwuma*' and 'cider'.
     (…)

(18) Generated AI answer: CLOVA X
noun
1. [Plant] A perennial plant in the Convolvulaceae family (…).
   - In the past, after harvesting barley, we planted *kokwuma*.
   - This year, we planted three types of *kokwuma*.
2. [Crop produce] A root produce in the Convolvulaceae family (…). | Rich in carbohydrates and dietary fiber (…). | It is mainly eaten steamed or roasted (…)
   - Boil the *kokwuma* and peel them.
   - We still have leftover *kokwuma* from last year. I don't know why my mom gave us so many *kokwuma* ... I made *kokwuma* pizza to use them up.

(19) *Urimalsaem entry*
[001] [Plant] A perennial plant in the Convolvulaceae family (…)
   - The dewdrops on the *kokwuma* vines shone like beads. Yongnyeo pulled out a vine. (Han Seung-won, *Tidal Wave*)
     (…)

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2045 (Lexicofocus)

[002] The tuberous root of *kokwuma*. It is commonly elongated and oval-shaped, rich in starch, and used for food and industrial purposes.
- Boiling *kokwuma*.

  (…)

[003] A term used to describe a person who is inflexible and frustrating or a situation that is frustrating and not going as planned, likened to eating sweet potatoes and feeling choked.
- Recently, netizens have been using *kokwuma* to describe frustrating situations that feel stifling and *cider* to describe situations that provide a refreshing relief. (Sports Seoul, April 2016)

  (…)

As can be seen in Example (17), ChatGPT described separately the literal meaning and the metaphorical meaning of *kokwuma* ('sweet potato') in its dictionary entry proposal. Moreover, it reused the examples provided in the previous prompt. CLOVA X, on the other hand, omitted the dictionary description for the metaphorical sense of *kokwuma* and only wrote the definitions of *kokwuma* as a plant and crop produce, which correspond to the entries [001] and [002] in *Urimalsaem*. Just as ChatGPT, CLOVA X reused the examples provided in the previous prompt. Both models were successful in identifying the literal and metaphorical meanings of the well-established semantic neologism *kokwuma*. However, in terms of dictionary entry compilation, ChatGPT demonstrated better performance by accurately completing the task, although its sole reliance on the input examples could be seen as both a strength and a limitation. Conversely, CLOVA X failed to describe the metaphorical meaning of *kokwuma* in its dictionary entry proposal, and was thus unsuccessful in completing the task of describing the semantic neologism as prompted.

Finally, an additional experiment was conducted where no specific information was given to the LLMs besides prompting them to describe the lexicographic information for a neologism X and then reviewing the outputs. This was unlike the previous experiments where detailed constraints and context were provided in the prompts. The additional neologisms tested included the high-frequency semantic neologism *phokphwung* ('storm') and the low-frequency semantic neologism *yengcep* ('welcome'), the news-biased neologism *caychohwan* ('Reconstruction Excess Profit Refund System'), and the non-news-biased neologism *ccwulthayng* ('tense character', used in the gaming domain). As a result of these additional experiments, CLOVA X could identify and describe the meaning of the news-biased neologism *caychohwan*, but returned inadequate results for the other neologisms. ChatGPT generated relatively appropriate definitions for the high-frequency semantic neologism and the news-biased neologism, but yielded limited results, either lacking in information or containing errors, for the low-frequency semantic neologism and the non-news-biased neologism. In particular, the models failed to describe the figurative meaning of the low-frequency semantic neologism *yengcep*, and in the case of *ccwulthayng*, the initial

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2045 (Lexicofocus)

results did not align with the intent of the query. Only after providing additional contextual information did the models present appropriate results. This highlights that LLMs are highly context-dependent and heavily rely on the diversity and accuracy of their training data. Therefore, to increase the reliability of LLMs in detecting and describing neologisms, training must incorporate up-to-date data and reflect genre diversity. Additionally, improvements are needed to enable the models to infer meanings even in situations where contextual information is not provided.

Thus far, the potential for utilising LLMs in extracting and describing Korean neologisms have been examined. The experimental results provided by ChatGPT and CLOVA X indicate that LLMs show potential as useful tools for detecting unregistered words and neologisms that are often omitted from traditional news-centred corpora. However, issues such as noise in the list of recommended neologisms, limitations in describing semantic neologisms, and duplicate outputs remain challenges to be addressed. The performance of LLMs heavily depends on the recency and linguistic diversity of their training data, necessitating ongoing efforts to train on balanced datasets and improve their algorithms.

5. Conclusion

This study analysed the current state and issues in neologism research, which represents the lexical creativity of native speakers, and examined the potential introduction of LLMs as an alternative. Two main issues in current neological research and lexicographic description are discussed: firstly, the analysis centred on online news data, which is widely used in Korea and other linguistic regions, highlighting the issue of written language bias; and secondly, the challenge of fully automatic extraction and description of semantic neologisms, which remains impossible at present. This study discussed the limitations of traditional neologism analysis, by analysing a corpus composed of sub-corpora from news data, IM data, and OP. Additionally, the study aimed to assess the potential of two major LLMs commercially available in Korea as alternatives. The findings can be summarised as follows:

Firstly, the analysis of news, IM, and OP revealed that the current methodology for neologism extraction and lexicographic description, centred on online news sources, presents limitations in identifying the first occurrence date and detecting neologisms related to everyday discourses. In particular, an analysis of the patterns of -*sulep*- ('-like') derivatives demonstrated that the same suffix led to the formation of different derivative neologisms across the three corpora. This finding highlights the need to expand neologism extraction to a wider range of genres and use domains.

Secondly, it was confirmed that the process of neologism extraction based on large corpora did not contribute to the extraction and lexicographic description of semantic neologisms. When analysing the proportion of semantic neologisms

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2045 (Lexicofocus)

in Korean neologism extraction since 1994 and examining current dictionaries, it appeared that semantic neologisms remain a blind spot of neologism research.

Thirdly, this study evaluated the ability of LLMs to detect and describe neologisms from both macrostructural and microstructural perspectives. In terms of the macrostructure, ChatGPT demonstrated the ability to complement news bias in the headword recommendation process by providing various usage examples from spoken language, messengers, blogs, and so forth. However, both models exhibited duplication and overgeneration errors to varying degrees (approximately 26% for ChatGPT, and 63% for CLOVA X), confirming that the involvement of human lexicographers is still essential. Regarding the microstructure, the ability to identify new meanings and describe them was evaluated, though it should be noted that this evaluation was conducted on a very limited sample. For the task of recommending semantic neologisms, neither of the models produced satisfactory results for open-ended queries. In situations where example sentences were provided, ChatGPT performed relatively accurately, succeeding in identifying and describing the meanings of high-frequency neologisms and news-biased neologisms. However, the performance was much weaker for low-frequency neologisms and non-news-biased neologisms.

This study differs from other studies in its discussion of the lexicographic potential of LLMs by focusing on the limitations of current neologism research and targeting a broader scope of neologisms. Overall, LLMs have shown somewhat satisfactory results in extracting and describing high-frequency neologisms and semantic neologisms, which will contribute to improving the productivity of dictionary description of neologisms in the future. The handling of neologisms that appear in genres and usage domains beyond the scope of LLM training, as well as low-frequency neologisms, remains quite insufficient though, and the lexicographic description varies significantly depending on the model. This is related to the inherent issues in the learning mechanisms of LLMs. Additionally, this study aimed to examine the correlation between the volume of Korean data used for training and the performance of neologism description through a comparison of domestic and foreign models, with results showing that there was no significant difference between them. The results of this study suggest that future research in neology studies, as well as lexicography and corpus linguistics, is needed. Discussions on how lexicography can contribute to the improvement of LLM performance are also necessary.

Endnotes

1. Naver CLOVA is reported to have trained on 6 500 times more Korean data compared to ChatGPT (https://www.donga.com/news/article/all/20230227/118100451/1).
2. Regarding the Corpus Query System, the research team in Phase 3 used the 'Neologism Investigation Program'. In Phase 4, the 'Neologism Extractor', a tool of Urimalsaem available only to expert account holders, was used. In Phase 5, the research team has been using a Python programme to extract neologisms.

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2045 (Lexicofocus)

3. This table is a revised and supplemented version of Table 1 in Nam et al. (2022: 84), which adds the methodology of the neologism investigation conducted by the Language Information Research Center at KNU. The methodology follows a similar approach to the NIKL neologism research project after its termination.
4. Transcriptions of Korean follow the Yale Romanization System.
5. For the *IM Corpus*, only past conversations provided by the participants themselves were included in the corpus, and data collected by 'collection bots' or 'chatbots' were excluded to observe natural language use better. For the *Newspaper Corpus*, only a portion of the materials were included for comparison with other sources. However, if there is a need to examine the usage trends of a word over a longer time span, the entire corpus may be utilised.
6. The short-term diachronic research methodology has been described as a 'paradigm shift' in neologism research (Boussidan 2013) because previous studies on language change mainly focused on long-term trends over centuries. Although studies of language change have been conducted in traditional diachronic research, the development of corpora and the greater availability of language resources have brought more attention to short-term diachronic research. This method enables the study of the dynamics of a word or words, from the time the word or words first emerge, to the period of rapid frequency increase, the process of becoming systematised in the language, and the time when usage declines as the term fades from people's awareness. For a detailed discussion, see Boussidan (2013), Nam et al. (2019), and Lee (2024).
7. The time frame for the 2021 neologisms is 1 July 1 2020 to 30 June 2021, and for the 2022 neologisms 1 July 2021 to 30 June 2022. While Time Segment 1 partially overlaps with the investigation time frame for the 2020 neologisms, the corpus used in this study did not reveal any instances of 2020 neologisms appearing earlier than the investigation period. Therefore, only the 2021 and 2022 neologisms are considered here.
8. Words that also appeared in the *Newspaper Corpus* during the same time period are excluded.
9. Out-links redirect users to the homepages of news outlets and have different metadata structures depending on the content producer. Conversely, in-links connect to pages within Naver New, so their format, converted to Naver News' metadata structure, offers the advantage of being much simpler and more stable for data processing.
10. According to Nam et al. (2024), who examined the appearance trends of the 2023 neologisms collected from news sources in news and blogs, more than three quarters of the neologisms in the study appeared first in blogs and then in news sources.
11. These are the types that did not appear in newspapers but appeared in IM or OP. A total of 21 types that did not appear in newspapers were found in both IM and OP.
12. All experiments in this study were conducted in June 2024 and also involved Google's Gemini, in parallel to ChatGPT and CLOVA X. However, despite using the same prompts, Gemini frequently failed to properly address the requests or suggested unnatural vocabulary items that are unlikely to be created by native Korean speakers. Given the limitations of space and the need for concise discussion, the results from Gemini experiments are omitted from this paper.
13. Naver CLOVA X is a large-scale language model (LLM) AI released by Naver on 24 August 2023, which claims to have trained on more than 6 500 times more Korean data than ChatGPT at the time of development (February 2023). The comparison seems to have been made with GPT-3.5, the version preceding GPT-4, which is used in this study. Specific data on how much Korean data GPT-4 was trained on, compared to GPT-3.5, has not been disclosed.

https://lexikos.journals.ac.za; https://doi.org/10.5788/35-1-2045 (Lexicofocus)

14. https://www.naver.com/
15. https://www.daum.net/
16. https://ko.dict.naver.com/#/main
17. As a reminder, the answers generated by AI are variable depending on the prompt. The prompts used in this study were revised and refined several times to obtain answers suitable for the research objectives.
18. Only the forms -*sulepta* (basic form) and -*sulewun* (conjugated form) were searched.
19. Duplicates were also found in ChatGPT results, although less frequently; this suggests that duplicates are a common problem in AI-generated results and not specific to CLOVA X. However, in the experiments for this study, CLOVA X produced significantly more duplicate words than ChatGPT.

References

Dictionaries

Korea University Korean Dictionary. https://ko.dict.naver.com/#/main
Standard Korean Language Dictionary. https://stdict.korean.go.kr/main/main.do
Urimalsaem. https://opendict.korean.go.kr/main

Neologism reports

Kim, H. et al. 2005. *Neologisms of 2005*. Seoul: National Institute of Korean Language.
Lee, S. et al. 2010. *Neologisms of 2010.* Seoul: National Institute of Korean Language.
Nam, K. et al. 2012. *Neologisms of 2012*. Seoul: National Institute of Korean Language.
Nam, K. et al. 2013. *Neologisms of 2013*. Seoul: National Institute of Korean Language.
Nam, K. et al. 2014. *Neologisms of 2014*. Seoul: National Institute of Korean Language.
Nam, K. et al. 2015. *Neologisms of 2015*. Seoul: National Institute of Korean Language.
Nam, K. et al. 2016. *Neologisms of 2016*. Seoul: National Institute of Korean Language.
Nam, K. et al. 2017. *Neologisms of 2017*. Seoul: National Institute of Korean Language.
Nam, K. et al. 2018. *Neologisms of 2018*. Seoul: National Institute of Korean Language.
Nam, K. et al. 2019. *Neologisms of 2019*. Seoul: National Institute of Korean Language.
Nam, K. et al. 2021. *Neologisms of 2020. A New Language of the COVID-19 Pandemic*. Seoul: Hankwukmwunhwasa.
Nam, K. et al. 2022. *Neologisms of 2021. COVID-19: Are We Going Back to Normal?* Seoul: Hankwukmwunhwasa.
Nam, K. et al. 2023. *Neologisms of 2022. What New Words Have Koreans Created and Used in this Era of Generative AI?* Seoul: Hankwukmwunhwasa.
National Institute of Korean Language. 1994. *A Study of Neologisms*. Seoul: National Institute of Korean Language.
National Institute of Korean Language. 1995. *A Study of Neologisms*. Seoul: National Institute of Korean Language.
National Institute of Korean Language. 1996. *A Study of Neologisms*. Seoul: National Institute of Korean Language.
Park, Y. et al. 2000. *Neologisms of 2000*. Seoul: National Institute of Korean Language.

Park, Y. et al. 2001. *Neologisms of 2001*. Seoul: National Institute of Korean Language.

Park, Y. et al. 2002. *Neologisms of 2002*. Seoul: National Institute of Korean Language.

Park, Y. et al. 2003. *Neologisms of 2003*. Seoul: National Institute of Korean Language.

Park, Y. et al. 2004. *Neologisms of 2004*. Seoul: National Institute of Korean Language.

Soh, K. et al. 2009. *Neologisms of 2008*. Seoul: National Institute of Korean Language.

Soh, K. et al. 2010. *Neologisms of 2009*. Seoul: National Institute of Korean Language.

Other literature

Boussidan, A. 2013. *Dynamics of Semantic Change: Detecting, Analyzing and Modeling Semantic Change in a Corpus of Short Diachrony.* Unpublished PhD Dissertation. Lyon: Université de Lyon.

Chomsky, N. 1957. *Syntactic Structures*. The Hague: Mouton.

De Schryver, G.-M. 2023. Generative AI and Lexicography: The Current State of the Art Using ChatGPT. *International Journal of Lexicography* 36(4): 355-387.

Freixa, J. and A. Adelstein. 2013. Criterios para la actualización lexicográfica a partir de datos de observatorios de neología. Martínez, M.J., J.A. Hinojosa and F.J. Ruiz. (Eds.). 2013. *Proceedings of the Congreso Internacional El Diccionario: Neología, lenguaje de especialidad, computación, Mexico City, Mexico, 28–30 October 2013*: 1-12. Mexico: El Colegio de México.

Huang, Y. and K. Nam. 2023. Unregistered Words in Everyday Language and a Proposal for Their Optimal Lexicographic Microstructure. *Lexicography* 10(2): 94-116.

Klosa, A. and H. Lüngen. 2018. New German Words: Detection and Description. Čibej, J., V. Gorjanc, I. Kosem and S. Krek (Eds.). 2018. *Proceedings of the XVIII EURALEX International Congress: Lexicography in Global Contexts, EURALEX 2018, Ljubljana, Slovenia, 17–21 July 2018*: 559-569. Ljubljana: Ljubljana University Press.

Lee, S. 2022. A Study of 2015–2020 Neologisms in Korean Language — Focusing on Neologism of 'Society', 'Economy', and 'Politics/Policies'. *Eo Mun Lon chong (Korean Language and Literature)* 94: 99-131.

Lee, S. 2024. *A Corpus Linguistic Study of Korean Neologisms from a Short-term Diachronic Perspective.* Unpublished PhD Dissertation. Daegu: Kyungpook National University.

Lew, R. 2023. ChatGPT as a COBUILD Lexicographer. *Humanities and Social Sciences Communications* 10(1): 1-10.

Lew, R. 2024. Dictionaries and Lexicography in the AI Era. *Humanities and Social Sciences Communications* 11(1): 1-8.

McKean, E. and W. Fitzgerald. 2023. The ROI of AI in Lexicography. *Proceedings of the 16th International Conference of the Asian Association for Lexicography: Lexicography, Artificial Intelligence, and Dictionary Users, ASIALEX 2023, Seoul, South Korea, 22–24 June 2023:* 18-27. Seoul: Yonsei University.

Nam, K. 2020. Issues of Frequency in Korean Neologisms. *Korean Semantics* 68: 213-239.

Nam, K. 2021. Overcoming Issues in Frequency-based Extraction and Lexicographic Inclusion of Korean Neologisms: A Triangulation Approach. *Lexicography* 8(1): 3-31.

Nam, K. and J. An. 2023. A Lexicographical Study on Phrase-unit Serialised Neologisms [N + Seday 'Generation'] — From 'Computer Generation' to 'Corona Generation' to 'MZ Generation'. *Journal of Korealex* 42: 33-66.

Nam, K., S. Lee and J. Choi. 2018. Research Trends and Issues on Semantic Neology Using Web Corpus. *Journal of Korealex* 31: 55-84.

Nam, K., S. Lee and H.Y. Jung. 2020. The Korean Neologism Investigation Project: Current Status and Key Issues. *Dictionaries. Journal of the Dictionary Society of North America* 41(1): 105-129.

Nam, K., S. Lee and B. Kang. 2019. The Linguistics of 'Frame' in Korean. *Korean Semantics* 65: 135-163.

Nam, K., H. Song, J. An and B. Kang. 2024. Where Do Neologisms Originate and Spread? — A Critical Review of Neologism Extraction Focused on News Media. *Language Facts and Perspectives* 63: 5-35.

Nam, K., H. Song, J. Choi and S. Lee. 2022. *A Study on Modern Neologism.* Seoul: Hankukmunhwasa.

Renouf, A. 2013. A Finer Definition of Neology in English: The Life-cycle of a Word. Hasselgård, H., J. Ebeling and S. Oksefjell Ebeling (Eds.). 2013. *Corpus Perspectives on Patterns of Lexis*: 177-208. Amsterdam: John Benjamins.

Sinclair, J.M. 2004. *Trust the Text: Language, Corpus and Discourse.* London: Routledge.

Smyk-Bhattacharjee, D. 2009. *Lexical Innovation on the Internet — Neologisms in Blogs*. Unpublished PhD Dissertation. Zurich: University of Zurich.