From Neologism Extraction to Dictionary Description: Methodological Issues in Corpus Balance, Word Unit Bias and LLM Assistance

Kilim NAM (nki@yonsei.ac.kr)
Affiliation: Korean Language and Literature, Yonsei University, Seoul, South Korea
Soojin LEE (sjmano27@naver.com)
Affiliation: Korean Language and Literature, Kyungpook National University, Daegu, South Korea
Hae-Yun JUNG (haeyun.jung.22@gmail.com)
Affiliation: Korean Language and Literature, Kyungpook National University, Daegu, South Korea

Corpus research, which was first pioneered in lexicography, has since developed into the methodology of corpus linguistics and has been expanded, refined, and eventually adopted by neological studies. 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 (Renouf 2013; Boussidan 2013; Nam et al. 2020; Freixa & Adelstein 2013; Klosa & Lüngen 2018). However, the spread of web languages and the emergence of large language models (LLMs) today 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 consisted in the semi-automated extraction from news corpora from 2005 to date, and explores ways to improve dictionary compilation as to reflect the dynamics of language from the cognitive perspective of discourse communities and individual speakers.

Chapter 2 examines the news media language bias in collecting neologisms[1] by using a Python programme to analyse 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, in order to compare the appearances, frequencies, domains and distributions of neologisms in different communication contexts. The analysis shows a significant bias of the news corpus toward public domains, such as politics, economics, and society, and scantily accounts for everyday language, neglecting expressions related to food or emotions for instance. A case in point is the particularly productive derivational suffix *-sulep-* 'be like', which is used to form adjectives in Korean. Adjectives constitute a part of speech that well expresses one's evaluations, attitudes, and emotions. The analysis of unregistered '*-sulep-*' derivational adjectives shows that there are 531 unregistered derivatives (35% of the total) across all web genres, 79% of which, however, are found only in online posts and/or IMs. While the '*-sulep-*' derivatives found in the news corpus are often related to politics or social issues (e.g. *yunsekyelsulepta* 'be very Yoon-Suk-Yeol-like'; *pheymisulepta* 'be feminist-like'), those from online posts and IM are often formed from bases denoting aspects of the daily life, such as food (e.g. *hansiksulepta* 'have a traditional Korean vibe'; *kokwumasulepta* 'be stifling [just as when eating sweet potato]'). In addition, the chapter discusses the value of such everyday language products as headword candidates.

Chapter 3 discusses the issues of the single-word and formal neologism bias. A comprehensive account of the lexicon of native speakers for a given era relies not merely on the identification of new forms but also on the analysis of their frequencies, distributions, and the discourse context in which they appear. This means that it has become crucial for the advancement of neologism research to develop a methodology for extracting semantic units such as phrases and collocations. Neological phrases are harder to identify than single-word neologisms and are often related to semantic neologisms, thereby being often dismissed from neologism extractions and headword selections. Instead, this study is to provide a closer look at phrase unit neologisms such as *kkwul ppalta* 'idle

time away [instead of working] (literally, 'suck honey')' that are deemed worth including in the dictionary.

Lastly, this study tests the recommendation of headwords and the compilation of dictionary microstructures for the neologisms presented here when prompting major Korean LLMs such as CLOVA by Naver. For now, it seems that CLOVA struggles not only to identify but also to define neologisms. For example, when prompted to recommend verbal neologisms or define a given neologism, it could only provide nominal forms and example sentences containing the lemma to define. In contrast, foreign LLMs such as Chat GPT could give an explanation of the lemmas it provided, along with pragmatic information such as 'slang' or 'informal' labels. Moreover, ChatGPT could provide verbal forms, although their quality and qualification as neologisms were somewhat questionable. This seems to point towards the inadequacy of the Korean data and existing dictionaries learning by LLMs for the identification and description of neologisms. Twenty years ago, Sinclair (2004:188-192) emphasized that the linguistics community needed to prepare for larger corpora in order to contribute to the future information society. In line with this, this study argues that Korean neologism research, to contribute to society in the modern age of LLMs, needs to think its methodology anew, turning to larger, balanced corpora and the contextualized extraction of semantic units for the dictionary to depart from the prescription of an idealised language and instead, be in tune with the dynamicity of actual language as spoken by native Koreans.

References

Boussidan, A. (2013), *Dynamics of semantic change: Detecting, analyzing and modeling semantic change in corpus in short diachrony*. Doctoral dissertation, Université de Lyon.

Freixa, J. & Adelstein, A. (2013). *Criterios para la actualización lexicográfica a partir de datos de observatorios de neología*. In *Comunicación presentada en Congreso Internacional El Diccionario: neología, lenguaje de especialidad, computación*, 28-30 October 2013, Mexico City, Mexico.

Klosa, A. & Lüngen, H. (2018). New German words Detection and description. In *Proceedings of the XVIII EURALEX International Congress Lexicography in Global Contexts* 17-21 July 2018, Ljubljana, Slovenia.

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

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

Sinclair, J. (2004). *Trust the text: Language, corpus and discourse*. Routledge.

[1] The list of unregistered words was monitored by using the list of headwords of the biggest Korean dictionary *Urimalsaem* as well as neologism lists collected from 1994 to present by the National Institute of Korean language and the Centre for Korean Language Information at Kyungpook National University.