new system, Oxford's New Words Prioritization Engine (NWPE), developed by Oxford Dictionaries to facilitate prioritization of large sets of candidate words by combining multiple sources of data in a single interface for analysis and by capturing human judgments about particular words so that they can be leveraged to improve future results.

Keywords: corpora, neologisms, prioritization

Using the Hypothes.is web annotation tool for neologism collection

Erin McKean

Erin McKean is the founder of Wordnik.com and works on open source strategy for Google. She was the editor-in-chief for American Dictionaries at Oxford University Press, and the editor of the *New Oxford American Dictionary, 2E.* She has written books on words and on dresses, was a regular columnist for *The Boston Globe* and the *Wall Street Journal*, and has served as an advisor to the American National Corpus, *American Speech,* and the Wikimedia Foundation.
erin@wordnik.com

Dictionary citation collection programs (sometimes called 'reading programs') involving both dedicated amateurs and paid professionals are not new, but have often required either cumbersome marking of print materials or creation of paper slips or access to private computer systems specific to individual projects. However, given the development and adoption of open standards for web annotation, citation collection by readers in and outside of dictionary programs can now be done easily without expensive proprietary tools or resorting to paper slips.

In this paper, we give an overview of Wordnik's reading program (currently in beta), which uses the free and open-source Hypothes.is web annotation tool to select, tag, and share citations from the open web directly for use on Wordnik.com. Using the Hypothes.is API, it is possible to import user-generated citations and their accompanying metadata directly into editorial workflows, including importing into KWIC corpora or other databases.

Since Wordnik is a radically inclusive dictionary (all words are eligible for inclusion), we discuss how this approach influences readers' marking of terms, and whether terms selected by readers are more likely to be typical neologisms (newly-coined words) or words overlooked by traditional dictionaries (e.g. jargon, slang, nonce, or other low-frequency words).

Keywords: dictionary users, web annotation, neologisms, hypothes.is, free-range definition

The Korean Neologism Investigation Project: Current status and key issues

Kilim Nam, Soojin Lee and Hae-Yun Jung

Kilim Nam has a PhD in Korean linguistics (on the copula *ida* structures in contemporary Korean, 2004) from Yonsei University (Seoul). She is a professor at the Department of Korean Language and Literature in Kyungpook National University (Daegu), has been the principal investigator of the Korean Neologisms Investigation Project since 2012, and is currently a board member of Korealex. Her research focuses on corpus linguistics and language performance.
nki@knu.ac.kr

Soojin Lee is a lecturer at the Department of Korean Language and Literature in Kyungpook National University (Daegu), where she obtained her MA (on academic keywords) and is doing her PhD. She has been a member of the research group for the Korean Neologisms Investigation Project since 2012. Her research interests include lexicography, lexicology as well as neology.
sjmano27@naver.com

Hae-Yun Jung received her MA in Korean Studies from SOAS (London) and is currently doing her PhD at Kyungpook National University (Daegu) under the supervision of Kilim Nam. Her PhD thesis is concerned with the treatment of pragmatic information in bilingual French-Korean lexicography, with particular attention to politeness. Her research interests include lexicography and cross-cultural pragmatics.
haeyun.jung.22@gmail.com

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. The two major tasks carried out within the scheme of the project consist of (1) collecting all the neologisms that appear each year in news articles on the Naver portal, and (2) investigating the usage development of neologisms within the past decade in order to determine whether those collected

ten years ago are still in use. These tasks are carried out using a web-based neologism extractor and a web crawler respectively. The extraction of new words is performed semi-automatically, since the automatic web-based neologism extractor is combined to manual identification. Since 2012, all the neologisms collected for task 1 have been added to the database of the online dictionary *Urimalsaem*, which became accessible to the public in 2016. *Urimalsaem* and the *Standard Korean Language Dictionary* (SKLD) are the main dictionaries of the Korean language. Both are state-run dictionaries, but have nonetheless distinct identities. *Urimalsaem* is a partly crowdsourced dictionary that enables contribution of dictionary users, while SKLD is a prescriptive dictionary for the use of standard language and grammar. 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 with the naked eye. Crucial methodological changes have been introduced since then, including the construction of a large-scale corpus (2005) and the use of the web crawler and web-based neologism extractor (2012). In 2015, a ten-year usage investigation for the neologisms extracted in 2005 and 2006 began. The following year, a pattern-based methodology of neologism extraction was introduced, and the minimum threshold of frequency occurrence for neologism candidates was increased to three. Despite these adjustments, the precision and recall levels of automatic neologism detection are still not satisfactory. Moreover, there are a number of other issues for improvement that are addressed in this paper, such as 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 morphological neologisms, and the dependency on manual processes. Some of these issues 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 only 75% of the neologisms survived after ten years. Whether this result constitutes a suitable criterion for lexicographic inclusion is also re-examined in the current study.

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

References

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

Nam, K., Lee, S., Jung, H.-Y. and Choi, J. 2016. The Life and Death of Neologisms: On What Basis Shall We Include Neologisms in the Dictionary? In *Proceedings of the XVIII EURALEX International Congress,* 389-393.

***Standard Korean Language Dictionary* [SKLD].** http://stdweb2.korean.go.kr/main.jsp.

***Urimalsaem*.** https://opendict.korean.go.kr/main.

New words in Japanese and the design of *UniDic* electronic dictionary

Teruaki Oka

Teruaki Oka graduated from Toyohashi University of Technology in 2010, and received his masters and PhD degrees in Engineering from Nara Institute of Science and Technology (NAIST), Ikoma, in 2012 and 2015, respectively. From 2015 to 2016 he was a Program-Specific Researcher at Kyoto University, and in 2016 he joined the National Institute for Japanese Language and Linguistics (NINJAL), where he currently serves as a Project Assistant Professor. His research interests are computational and corpus linguistics.
teruaki-oka@ninjal.ac.jp

The National Institute for Japanese Language and Linguistics (NINJAL) is involved in developing Japanese language corpora, including the Balanced Corpus of Contemporary Written Japanese, Corpus of Spontaneous Japanese, Corpus of Historical Japanese, and NINJAL Web Japanese Corpus. In the development processes we often encounter new words that are formed by using different character types (e.g., Hiragana, Katakana, Kanji) and their heterographs, with their combinations, even for writing a single word (e.g., *big*: おおきい, 大きい, オオキイ, ぉぉきぃ, 大キィ), which could be 'literal' (e.g., *as it was expected*: 矢張り), 'somewhat colloquial' (やっぱり), 'colloquial' (やっぱし), 'abbreviated' (やぱ), and so on. Thus, new words can appear as orthographic variants (おおきい vs. 大キィ), form variants (矢張り vs. やぱ) and new lemmas (such as エモい *emotional*), and be classified at these three levels (orthographic, form, lemma).

We apply a design policy called "hierarchical definition of word indexes" to register new words in *UniDic*, our electronic Japanese word dictionary, for annotating plain texts with morphological information. Using the hierarchical definition of word indexes, a single lemma (e.g., 矢張り) has its various word forms written in Katakana characters (e.g., 矢張り←ヤハリ, ヤッパリ, ヤッパシ, ヤパ) as its children, with each form having its orthographic variants as its children (e.g., ヤハリ←矢張り, やはり, ヤハリ). *UniDic* contains about 200 thousand lemmas and one million of their form and orthographic variants with rich morphological information (e.g., part of speech, lemmatized form, pronunciation, accent). To annotate morphological information in plain unsegmented texts, we select optimal records for character strings in the texts from UniDicDB, a word database system. The records and their morphological information are manually registered to UniDicDB when new words are detected during the annotation phase. We also employ UniDicExplorer, an annotator-friendly user interface capable of searching and registering words. Another feature is UniDicMA, a dictionary software for the morphological analyzer,