Context window and polysemy interpretation:
A case of Korean adverbial postposition *–(u)lo*

Seongmin Mun (Université Paris Ouest Nanterre La Défense; simon.seongmin.mun@gmail.com)
Gyu-Ho Shin (University of Hawaii at Manoa; gyuhoshin@gmail.com)

Construal of a polysemous word occurs in conjunction with a series of words, delivering various frame-semantic meanings (Goldberg,2006) and yet purporting similar interpretations (Harris,1954). In this regard, context window—a range of words surrounding a target word, affecting the determination of its characteristics—is drawing attention to the computational understanding of combinatorial properties of words.

We ask how context window addresses polysemy interpretation in Korean, a language typologically different from the major Indo-European languages investigated for this task. We report computational simulations regarding how various context window sizes address polysemy of *-(u)lo*, which manifests polysemy due to its multiple functions mapped onto one form. We used the Sejong corpus, with semantic annotations of this postposition cross-verified by three native speakers of Korean (κ=0.95). Employing a distributional semantic model (Harris,1954), we devised an unsupervised learning algorithm by combining Singular Value Decomposition with Positive Pointwise Mutual Information. We measured model performance through accuracy rates that the model classified test sentences by the functions of *-(u)lo*, with manipulation of context window from one to ten. For this purpose, we used the similarity-based estimate (Dagan.et.al.,1993) by calculating cosine similarity scores between *-(u)lo* and its co-occurring content words.

Our model achieved the highest classification accuracy rate in the window size of one, and the accuracy rates decreased as the window size increased. This trend aligns with advantages of small window sizes (Bullinaria&Levy,2007). Considering that a narrower range of context window relates more to syntactic than to sematic information (Patel.et.al.,1997), our model may have employed structural, more than semantic, characteristics of tri-grams (word-target-word) for the best classification performance. Given the networks of interlinked clusters of words and symbolic units in human cognition (*construct-i-con*; Goldberg,2006), our findings shed light on relations between a polysemous word and an abstract schema including the word, represented as context window, in addressing word-level polysemy.

References

Bullinaria, John A & Levy, Joseph P. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39(3), 510–526.

Goldberg, Adele. E. 2006. Constructions at work: The nature of generalization in language. *Oxford: Oxford University Press*.

Harris, Zellig S. 1954. Distributional Structure. *WORD*, 10(2-3), 146-162.

Ido Dagan, Shaul Marcus, & Shaul Markovitch. 1993. Contextual word similarity and estimation from sparse data. *In Proceedings of the 31st Annual Meeting of the Association for Computational Linguistics*, Columbus, 164-171.

Patel, Malti, Bullinaria, John A. & Levy, Joseph P. 1997. Extracting semantic representations from large text corpora. *Proceedings of the 4th Neural Computation and Psychology Workshop*, London, 199–212.