How does context window size address polysemy of adverbial postposition -(u)lo in Korean?
Mun, Seongmin1 & Shin, Gyu – ho2
Science of Language, UMR 7114 MoDyCo – CNRS, University Paris Nanterre1
Department of Linguistics, University of Hawaii at Manoa2

Keywords: context window, polysemy, adverbial postposition

Intended construal of a polysemous word occurs within a range of words, which deliver various frame-
semantic meanings (Goldberg, 2019) and yet purport similar meanings (Harris, 1954). In this regard,
context window—a range of words surrounding a target word, which affects the determination of
characteristics of the word—is drawing attention to the computational understanding of
combinatorial properties of words in human language (MacDonald & Ramscar, 2001).

We pose a question as to how context window size applies to polysemy of a function word such as a
postposition in Korean, a language typologically different from the major Indo-European languages
that have been investigated for this task. We report a computational simulation that explores how
various sizes of context window account for polysemy of -(u)lo, which manifests polysemy due to its
various functions mapped onto one single form (Choo & Kwak, 2008).

For this purpose, we used the Sejong corpus (Kim et al., 2007; 90% for training and 10% for testing),
with semantic annotation of this corpus cross-verified by three native speakers of Korean (κ = 0.95).
Employing a distributional semantic model (Baroni et al. 2014.), we devised an unsupervised learning
algorithm by combining Singular Value Decomposition with Positive Pointwise Mutual Information
(Turney & Pantel, 2010). Cosine similarity scores of -(u)lo and its co-occurring content words were re-
scaled through the min-max normalisation (Luai et al., 2006). Using these scores, model performance
was measured through the rate of accuracy that the model classified instances of the test set involving
the six functions of -(u)lo under manipulation of context window size from one to ten.

Overall, our model achieved the highest rate of classification accuracy in the window size of one, and
the rates of accuracy decreased as the window size increased. The global trend of accuracy that the
model demonstrated is consistent with previous research that shows advantages of small window
sizes (Bullinaria & Levy, 2007). A narrower range of context window relates more to syntactic
information than to semantic information (Patel et al., 1997). This invites an interpretation that our
model may have employed structural, rather than semantic, characteristics of tri-grams (i.e., a word-
target-word sequence) for the best performance in classification. 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) for addressing word-level polysemy.

References

Baroni, Marco, Georgiana Dinu & Germán Kruszewski. 2014. Don’t count, predict! a systematic
comparison of context-counting vs. context-predicting semantic vectors. In Proceedings of the
52nd Annual Meeting of the Association for Computational Linguistics. 238–247.
Bullinaria, John A & Joseph P. Levy. 2007. Extracting semantic representations from word co-
University Press, Cambridge, UK.
University Press.

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


