A computational approach to resolving the polysemy of postpositions in Korean

Seongmin Mun[1]
MoDyCo (UMR 7114), CNRS, Paris Nanterre University
seongmin.mun@parisnanterre.fr

The current on-going project is for the resolution of polysemy involving the Korean postpositions. Korean is a Subject-Object-Verb language, which marks case with dedicated postpositions [9]. In this research project, we investigate the polysemy of postpositions in Korean under the framework of Construction Grammar [5]. A postposition is defined as a function word indicating grammatical information to which it is associated [9]. As a form-function pairing, a postposition can be polysemous in that one form delivers multiple functions [4]. An adverbial postposition -(u)lo, for instance, is either directional or instrumental, the two major functions of this particle [2] (1, 2).

(1) -(u)lo as directional ('(I) went to the road.')

도로-(으)로 갔다.
tolo-(u)lo ka-ass-ta.
road-DIR go-PST-SE

(2) -(u)lo as instrumental ('(I) went by bicycle')

자전거-(으)로 갔다.
cacenke-(u)lo ka-ass-ta.
bicycle-INS use-PST-SE

We pose key questions as to what is polysemy of postpositions in Korean and how can computer identify the polysemy of the word? The meaning of a word in a sentence can be approximated by it's relation to the co-occurring words (dubbed the Distributional Hypothesis, [8]). It is thus assumed that we can identify the polysemy of a word based on information obtained from surrounding words and their network. This account has been implemented by way of NLP methods [6]. In this project, we use several NLP methods (such as SVD [3], PPMI SVD [12, 1], and SGNS [10, 11]) for the analysis of the Sejong corpus [7] which is made by large-scale corpus project in Korea in other to reveal the nature of polysemy involving postpositions in Korean. Currently, we are simultaneously progressing to develop the visualization and to make Gold standard from Sejong corpus. Gold standard includes target postpositions (-ey, -eyse, and -(u)lo) in the sentence and is designed to represent the functional semantic role of postposition (such as agent, experience, mental agent, companion, theme, location, direction, goal, final state, source, instrument, effector, criterion, purpose, content, etc.). With the visualization, we explore the information obtained from surrounding words and their network of a selected postposition. We will share the final system that can explore the distribution of polysemous postpositions and uses the distribution to automatically recognize the functional semantic role of postpositions in the upcoming conference.

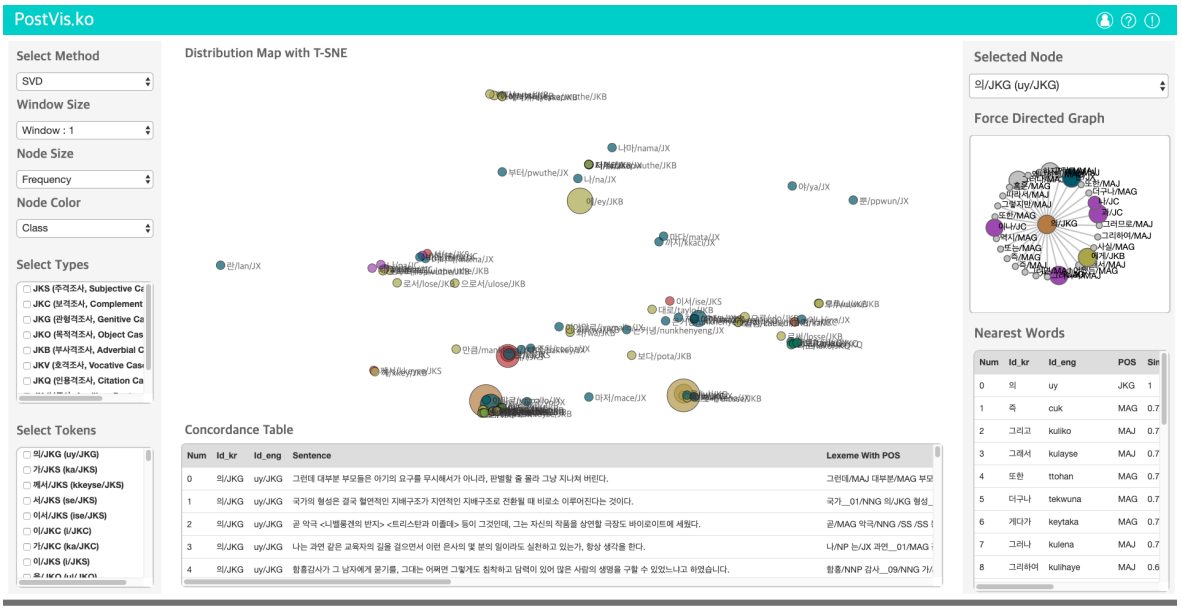

FIGURE 1 – PostVis.ko (URL : https ://seongmin-mun.github.io/VisualSystem/Major/PostVis/)

Références

[1] BARONI, M., AND LENCI, A. Distributional memory : A general framework for corpus-based semantics. *Computational Linguistics 36*, 4 (2010), 673–721.

[2] CHOO, M., AND KWAK, H. Y. Using korean. *Cambridge University Press, New York, NY.* (01 2008).

[3] ECKART, C., AND YOUNG, G. The approximation of one matrix by another of lower rank. *Psychometrika 1*, 3 (1936), 211–218.

[4] GLYNN, D., AND ROBINSON, J. A. Corpus methods for semantics. quantitative studies in polysemy and synonymy. *John Benjamins Publishing Company* (01 2014).

[5] GOLDBERG, A. Constructions : A construction grammar approach to argument structure. *Biblio-vault OAI Repository, the University of Chicago Press* (01 1995).

[6] HILPERT, M. Change in modal meanings. *Constructions and Frames 8*, 1 (2016), 66–85.

[7] HYO-PIL, S. The 21st sejong project : with a focus on selk(sejong electronic lexicon of korean) and the knc(korean national corpus). In *The 3rd International Joint Conference on Natural Langauge Processing* (2008).

[8] MCDONALD, S., AND RAMSCAR, M. Testing the distributional hypothesis : The influence of context on judgements of semantic similarity. In *In Proceedings of the 23rd Annual Conference of the Cognitive Science Society* (2001), pp. 611–6.

[9] SOHN, H.-M. The korean language. *Cambridge University Press, New York, NY.* (01 1999).

[10] TOMAS MIKOLOV, KAI CHEN, G. S. C., AND DEAN, J. Efficient estimation of word representations in vector space. In *In Proceedings of the International Conference on Learning Representations (ICLR)* (2013).

[11] TOMAS MIKOLOV, ILYA SUTSKEVER, K. C. G. S. C., AND DEAN, J. Distributed representations of words and phrases and their compositionality. *In Advances in Neural Information Processing Systems* (2013), 3111–3119.

[12] TURNEY, P. D., AND PANTEL, P. From frequency to meaning : Vector space models of semantics. *Journal of Artificial Intelligence Research 37*, 1 (2010), 141–188.