Exploring idiomaticity with variant-based distributional measures and Shannon's entropy

Marco S. G. Senaldi
SNS, Pisa
marco.senaldi@sns.it

Gianluca E. Lebani
University of Pisa
gianluca.lebani@for.unipi.it

Alessandro Lenci
University of Pisa
alessandro.lenci@unipi.it

The goal of this research is to investigate whether we can take advantage of the syntactic and lexical fixedness of idiomatic expressions to devise corpus-based indices of idiomaticity and compositionality and whether these measures can actually predict human ratings of idiom syntactic flexibility.

First of all we describe a method for automatically distinguishing potential idioms from only literal combinations via compositionality indices that leverage the greater lexical rigidity of idioms. Starting from two sets of idiomatic and literal Italian verbal constructions and adjective-noun pairs, we generated a series of lexical variants out of them, replacing their constituents with semantically related words. We then represented both the original targets and their variants as vectors in a distributional space and calculated cosine similarity between a given target and its variants, expecting idiomatic vectors to result less similar to the vectors of their variants with respect to the literal expression vectors. All in all, this proved to be the case, showing that focusing on the limited exchangeability of the constituents is an effective way to compute the idiomaticity degree of a given word combination.

In the second part of our study, participants to a CrowdFlower questionnaire gave 1-7 acceptability scores to sentences containing Italian verbal idiomatic and literal combinations in different syntactic variants. We then modeled the human ratings with a hierarchical regression analysis via corpus-based measures computed for the same idioms. These included all the aforementioned compositionality indices and other formal flexibility measures which used Shannon's Entropy to calculate the idiom variability with regard to various parameters, such as the constituents morphology, the presence and type of determiners, etc. Promising results in this regression analysis support the cognitive plausibility of our computational indices to explain the way speakers process idioms.