Contextual Characteristics of Concrete and Abstract Words

Diego Frassinelli, Daniela Naumann & Sabine Schulte im Walde (University of Stuttgart)
frassinelli@ims.uni-stuttgart.de

The literature about conceptual representation has extensively debated about the nature of concrete concepts, but much less has been said about abstract concepts and the similarities and differences between these two classes (Murphy, 2002). According to Barsalou and Wiemer-Hastings (2005), concrete concepts are directly grounded in the sensory-motor system, while abstract concepts are mapped to concrete concepts in order to be processed. Distributional semantics represents a very powerful approach to investigate word meaning in a data-driven fashion: the distributional hypothesis states that we can infer the meaning of a word by looking at the linguistic context it co-occurs with (Turney & Pantel, 2010).

Hypothesis. The aim of this work is to quantitatively investigate similarities and differences between concrete and abstract words by looking at the concreteness ratings (CRs) of their respective linguistic contexts (co-occurrences). Based on the literature, both concrete and abstract words should primarily co-occur with other concrete words.

Materials. Noun-noun co-occurrences have been computed from the 16 billion tokens ENCOW14 English web corpus (Schäfer, 2015). For each noun, we extracted the CR (1=abstract to 5=concrete) from the Brysbaert et al. (2014) collection.

Study 1. After computing the cosine similarity between pairs of target words, we selected the 16 top nearest neighbors (NNs) of each target. We analyzed the relation between the CR of each target and the average CR of its NNs. A linear mixed effect (LME) analysis indicates that an increase in CR for the target corresponds to a significant increase in the average CR of its NNs ($\beta_{\text{concTarget}}=0.22$, $p<.001$).

Study 2. After grouping the targets in five sets according to their CRs, we averaged the CRs of their first 2 to 256 most frequent context words. A LME analysis shows a significant difference between the data in the five sets (e.g., increasing the concreteness of the target corresponds to a significant increase in the average concreteness of its context words, $\beta_{\text{concTarget}}=.189$, $p<.001$); a significant negative effect of frequency ($\beta_{\text{freqContext}}=-0.03$, $p<.001$) and a significant negative interaction between the two has been found ($\beta_{\text{concTarget:freqContext}}=-0.005$, $p<.001$).

Study 3. Each target is represented as a 9-dimensional concreteness vector (one dimension per CR): each dimension represents the sum of frequencies of each context word having a specific CR. 60% of the contexts of abstract words are abstract (CR<2), and 45% of the contexts of concrete words are concrete (CR>4).

Discussion. Overall, the three studies show consistent results. Concrete words predominantly co-occur with other concrete words, while abstract words co-occur with abstract words. The results for the concrete words are fully aligned with multiple studies in the literature. On the other hand, our results seem to disagree with the grounding hypothesis for abstract words: abstract words do not share the same context as concrete words. Further investigation is required to support our preliminary findings.

Barsalou L. W., & K. Wiemer-Hastings (2005). Situating abstract concepts. In D. Pecher & R. A. Zwaan (Eds.), *Grounding cognition: The role of perception and action in memory, language, and thought*, 129–163.

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods*, 46(3), 904–11.

Murphy, G. L. (2002). *The big book of concepts*. London: MIT.

Schäfer R., (2015). Processing and querying large web corpora with the COW14 architecture. CMLC-3.

Turney, P. D., & Pantel, P. (2010). From frequency to meaning: Vector space models of semantics. *Journal of Artificial Intelligence Research*, 37, 141–188.