The Nature of Disagreements on Mid-Scale Ratings: A Case Study on the Abstractness-Concreteness Continuum

Urban Knupleš[1], Diego Frassinelli[2] & Sabine Schulte im Walde[1]
[1]Universität Stuttgart, [2]Universität Konstanz
urban.knuples@ims.uni-stuttgart.de, diego.frassinelli@uni-konstsanz.de,
schulte@ims.uni-stuttgart.de

Across disciplines, researchers have collected and exploited human judgements on semantic variables such as concreteness, compositionality, emotional valence, and plausibility (inter alia). Traditionally, those judgements are collected as a degree on a continuum between extremes. While humans tend to strongly agree on their ratings for extremes (e.g., a CAT is typically judged as extremely concrete; GLORY as extremely abstract; the compound CROCODILE TEARS as extremely non-compositional; a WAR as extremely negative), we find considerable disagreement regarding human mid-range ratings, i.e., judging about semi-concreteness, semi-compositionality, semi-negativity, etc. Presumably, semi-properties are not easily graspable, thus generating stronger disagreement among raters. Nevertheless, the collected norms are heavily exploited in state-of-the-art computational approaches, where the respective knowledge represents a crucial task-related component (such as concreteness information for figurative language detection, and emotional valence for sentiment analysis).

The current study provides a series of analyses on human mid-scale ratings (Knupleš et al., 2023), while focusing on the most prominent collection of concreteness ratings for English concepts (Brysbaert et al., 2014). In a first set of experiments, we analyse multi-modal characteristics of the concreteness of target nouns in the Brysbaert norms: perception strength for specific senses (auditory, gustatory, haptic, olfactory, visual), emotional dimensions (valence, affect, dominance), lexical properties (frequency, ambiguity) and association types as indicators of meaning diversity. We start with a holistic perspective via correlations between targets' concreteness and their characteristics, and then zoom into differences for words with mid-scale vs. extremely concrete or abstract mean concreteness ratings, by applying supervised classification and feature analyses. In a second set of experiments, we hypothesise that mid-scale ratings are due to different combinations of individual ratings across the scale. We rely on the original 25 participant ratings per target word and apply exploratory cluster analyses to identify patterns of disagreement between the individual raters. Our results suggest to either filter or fine-tune mid-scale targets before utilising them.

References: • Brysbaert, M., A.B. Warriner & V. Kuperman (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods* 64, 904–911. • Knupleš, U., D. Frassinelli & S. Schulte im Walde (2023). Investigating the nature of disagreements on mid-scale ratings: A case study on the abstractness-concreteness continuum. In *Proceedings of the Conference on Computational Natural Language Learning.*