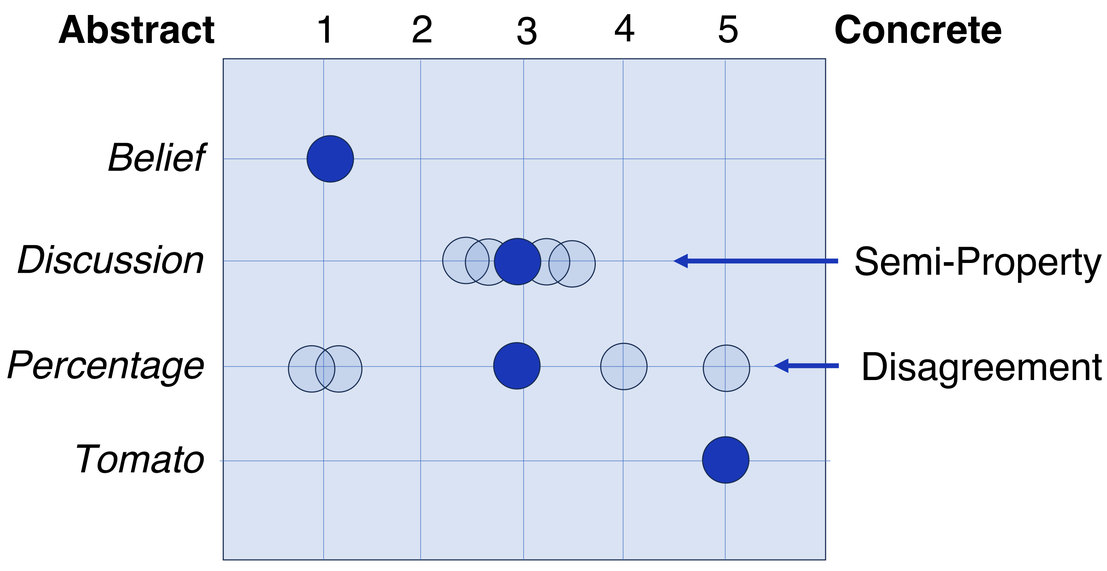
Investigating the Nature of Disagreements on Mid-Scale Ratings: A Case Study on the Abstractness–Concreteness Continuum

Urban Knupleš, Diego Frassinelli, and Sabine Schulte im Walde
University of Stuttgart & University of Konstanz

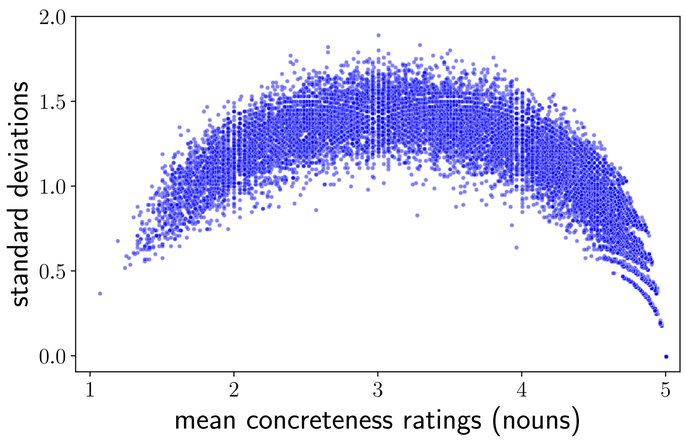
What are the properties of concepts with mid-scale ratings?

- High disagreement among raters → High standard deviation
- Distinct multimodal characteristics (**Study 1**)
- Specific patterns of disagreement (**Study 2**)

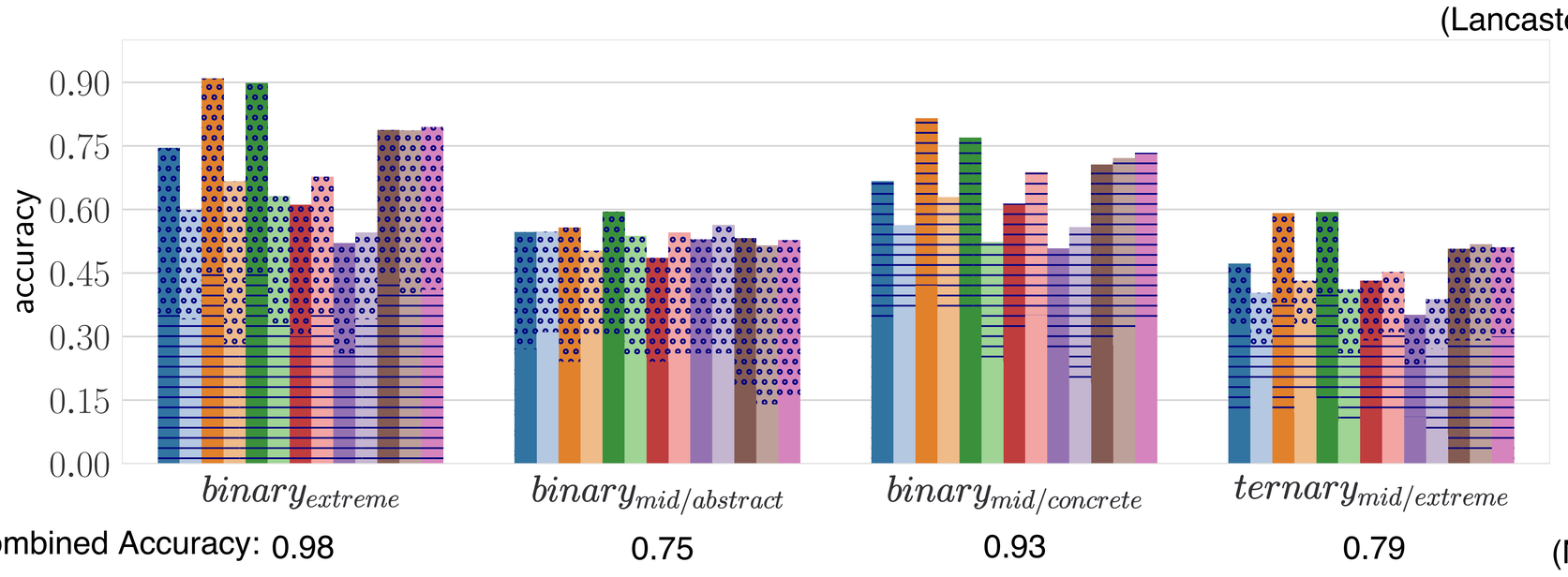


Materials

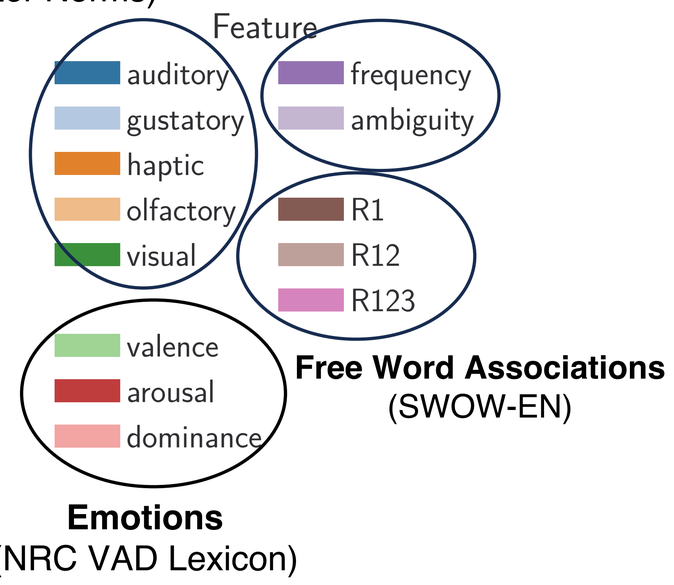
- Concreteness ratings for **1500 English Nouns** (*Brysbaert Norms*)
  - 500 extreme abstract (*1.07 - 1.71*)
  - 500 mid-scale (*2.90 - 3.31*)
  - 500 extreme concrete (*4.85 - 5.00*)
- Average scores (**Study 1**) and single ratings (**Study 2**)



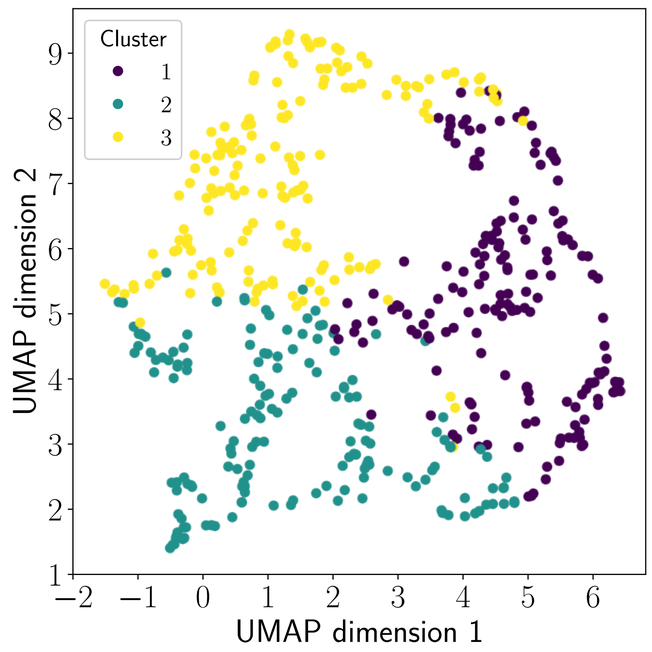
Study 1: Mid-Scale Peculiarities

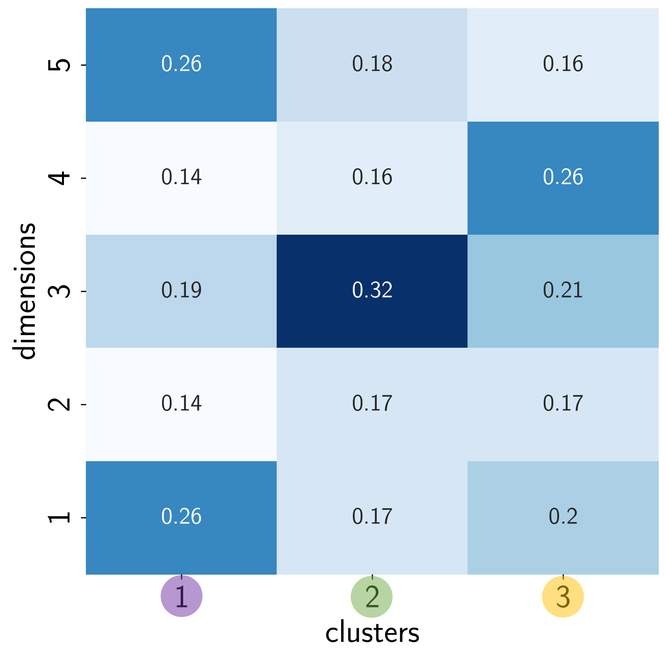


Combined Accuracy: 0.98 | 0.75 | 0.93 | 0.79



Study 2: Mid-Scale Disagreement Patterns





C	Target	Distribution
1	*definition*	$\langle 0.32, 0.11, 0.14, 0.11, 0.32 \rangle$
	hero	$\langle 0.22, 0.11, 0.26, 0.19, 0.22 \rangle$
	percentage	$\langle 0.40, 0.03, 0.10, 0.20, 0.27 \rangle$
2	*coward*	$\langle 0.17, 0.20, 0.30, 0.20, 0.13 \rangle$
	discussion	$\langle 0.15, 0.07, 0.48, 0.15, 0.15 \rangle$
	labor	$\langle 0.16, 0.12, 0.40, 0.12, 0.20 \rangle$
3	*booster*	$\langle 0.32, 0.07, 0.14, 0.29, 0.18 \rangle$
	election	$\langle 0.20, 0.10, 0.23, 0.27, 0.20 \rangle$
	hour	$\langle 0.23, 0.07, 0.23, 0.30, 0.17 \rangle$

How can we use them in computational modeling?

- **Exclude** them from the study → focus on extremes
- **Fine-tune** them according to disagreement patterns

References:
Brysbaert Concreteness Norms (*Brysbaert et al., 2014*) ✦ Lancaster Norms (*Lynott et al., 2020*) ✦ Encow Corpus (*Schäfer and Bildhauer, 2012*) ✦ WordNet 3.0 (*Miller and Fellbaum, 1991*) ✦ NRC VAD Lexicon (*Mohammad, 2018*) ✦ The Small World Of Words Project SWOW (*de Deyne et al., 2019*)