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

Urban Knupleš[1] and **Diego Frassinelli**[2] and **Sabine Schulte im Walde**[1]
[1]Institute for Natural Language Processing, University of Stuttgart
[2]Department of Linguistics, University of Konstanz
{urban.knuples, schulte}@ims.uni-stuttgart.de
diego.frassinelli@uni-konstanz.de

Abstract

Humans tend to strongly agree on ratings on a scale for extreme cases (e.g., a CAT is judged as very concrete), but judgements on mid-scale words exhibit more disagreement. Yet, collected rating norms are heavily exploited across disciplines. Our study focuses on concreteness ratings and (i) implements correlations and supervised classification to identify salient multimodal characteristics of mid-scale words, and (ii) applies a hard clustering to identify patterns of systematic disagreement across raters. Our results suggest to either fine-tune or filter mid-scale target words before utilising them.

1 Motivation

Across disciplines, researchers have collected and exploited human judgements on semantic variables such as concreteness, compositionality, emotional valence, and plausibility. 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; WAR as extremely negative), we find considerable disagreement regarding human mid-range ratings, i.e., judging about semi-concreteness, semi-compositionality, semi-negativity. Presumably, conceptual *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, while focusing on the most prominent collection of concreteness ratings for English words (Brysbaert et al., 2014), henceforth *Brysbaert norms*. As basis for the Brysbaert norms, humans were asked to judge the concreteness (in contrast to abstractness) of English words on a 5-point rating scale from 1 (abstract) to 5 (concrete) regarding how strongly the participants thought the meanings of the targets can(not) be experienced directly through their five senses. Figure 1 illustrates the distribution of the mean concreteness ratings and standard deviations (SDs) across 25 raters and for the three word classes of nouns, verbs, and adjectives. These *croissant*[1] plots for ratings on a scale can exhibit "only a finite number of possible combinations of means and standard deviations" (Pollock, 2018): humans tend to agree on the extremes (→ low SD) and to disagree on intermediate *semi*-values (→ high SD).

In a first set of experiments, we analyse multimodal characteristics of the concreteness of target nouns in the Brysbaert norms (we provide additional materials for verbs and adjectives in the Appendix): 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/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 human judgements across the scale. We thus rely on the original per-participant ratings (i.e., 25 ratings per target) and apply exploratory cluster analyses to identify patterns of disagreement between the individual raters of targets with mid-scale ratings.

[1]We use this term due to the shape of the distribution plots.

Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL), pages 70–86
December 6–7, 2023.

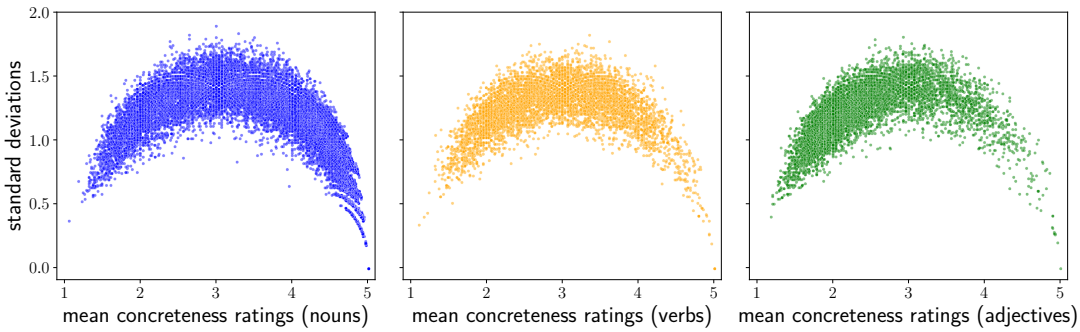

Figure 1: Croissant plots – Mean concreteness scores and standard deviations of ratings in Brysbaert et al. (2014).

Our contributions in this paper are two-fold. (i) We identify a range of target word characteristics that overall correlate with their degrees of concreteness ratings in different directions, and more specifically differ for mid-scale and extremely concrete or abstract target words. (ii) We identify a range of systematic disagreement patterns that clearly differ across target words with mid-scale mean ratings, thus pointing out fine-grained differences in judgements on semi-perception and suggesting to either filter or fine-tune mid-scale target words before utilising them in computational approaches.

In the remainder of this paper, we introduce previous related work (Section 2) and our concreteness targets (Section 3); we then report our analyses regarding general and mid-scale target characteristics (Section 4) and mid-scale disagreement patterns (Section 5).

2 Related Work

Collecting human judgements on a rating scale is a popular means of constructing concept-specific datasets across languages, research disciplines and (computational) linguistics tasks. Prominent example tasks and collections targeting semantic variables include compositionality ratings for compound–constituent relatedness (Reddy et al., 2011; Schulte im Walde et al., 2016; Cordeiro et al., 2019; Gagné et al., 2019; Günther et al., 2020, i.a.), affect ratings such as valence, arousal, dominance, emotion (Kanske and Kotz, 2010; Köper and Schulte im Walde, 2016a; Mohammad, 2018, i.a.), plausibility ratings (Wang et al., 2018; Eichel and Schulte Im Walde, 2023, i.a.), and concreteness ratings (Spreen and Schulz, 1966; Paivio et al., 1968; Algarabel et al., 1988; Della Rosa et al., 2010; Brysbaert et al., 2014; Köper and Schulte im Walde, 2016a; Bonin et al., 2018; Muraki et al., 2022, i.a.).

As a main motivation for collecting general conceptional ratings on a scale, Keuleers and Balota (2015) state that there is "no reason for words to be rated for every single experiment". Still, researchers across disciplines have pointed out problematic aspects of rating norms, because their reliability is unclear, especially when ratings have been collected via crowdsourcing or extrapolation (Keuleers and Balota, 2015; Mandera et al., 2015). Pollock (2018) describes the typical shape of ratings on a scale, pointing out that the mid-range concepts are the least agreed upon, and that the interpretation of the corresponding ratings conflates *semi*-properties and genuine disagreements. A mid-scale score in concreteness could thus refer to an average *semi*-perception (whatever this means), or to a specific *semi*-sense, such as vision, haptics, etc., as well as to disagreement about perceptual strength, or a combination of the above. Furthermore, many conceptual ratings have been collected by presenting the word in isolation without reference to the respective word class and out of context. For example, the Brysbaert norms rely on isolated target presentation, and part-of-speech information was added post-hoc from the SUBTLEX-US corpus (Brysbaert et al., 2012). Muraki et al. (2022) used the same setup as Brysbaert et al. (2014) but for multiword expressions, in which case part-of-speech ambiguity did not arise, but the targets were also presented out of context.

Despite these problems, ratings on a scale still remain the major strategy to collect human judgements on degrees of semantic variables, while alternatives such as best-worst scaling are available (Kiritchenko and Mohammad, 2017; Abdalla et al., 2023). The resulting norms are heavily exploited in state-of-the-art computational approaches; e.g., emotion and concreteness norms represent a crucial component in systems to detect figurative lan-

guage usage (Turney et al., 2011; Tsvetkov et al., 2014; Köper and Schulte im Walde, 2016b; Mohammad et al., 2016; Aedmaa et al., 2018; Köper and Schulte im Walde, 2018; Maudslay et al., 2020). The current study encourages researchers to distinguish between degrees of (dis)agreement of such norms, and to identify a meaningful way of exploitation, in particular for mid-scale ratings.

3 Concreteness Targets and Ratings

As materials for our experiments, we utilise the concreteness norms collected by Brysbaert et al. (2014), including approximately $40,000$ English target words.[2] The resource contains individual ratings by 25 participants on a 5-point scale ranging from 1 (abstract) to 5 (concrete), mean ratings and standard deviations. No context or part-of-speech (POS) were given; in a post-processing step, Brysbaert et al. (2012) added POS and frequency information from the SUBTLEX-US corpus.

We followed a further post-processing step suggested by Schulte im Walde and Frassinelli (2022), who assigned the most frequently occurring POS tag and frequency information to the target words using the ENCOW web corpus (Schäfer and Bildhauer, 2012; Schäfer, 2015), and then reduced the targets to a less ambiguous and less low-frequent subset by discarding words for which (i) the predominant POS did not represent at least 95% of all POS occurrences; (ii) the newly assigned ENCOW POS tag was not identical to the SUBTLEX-US POS tag, or (iii) for which the ENCOW target frequency was lower than $10,000$. Our subset includes $5,448$ nouns, $1,280$ verbs and $2,205$ adjectives, and is publicly available.[3]

4 Target Words: Characteristics

In our first set of experiments we analyse multimodal characteristics of our concreteness targets. After introducing these characteristics (Section 4.1), we start out with a holistic perspective by quantifying statistical relationships between degrees of concreteness and our selection of target characteristics (Section 4.2). We then zoom into differences in characteristics between mid-scale target words and extremely concrete/abstract target words, by applying a classifier that determines separability based on characteristics (Section 4.3).

4.1 Characteristics and Resources

Sense Perception Given that the original concreteness ratings in the Brysbaert norms rely on the raters' perceptions across senses, the most intimately connected set of characteristics explores the relationships between concreteness ratings and the five senses that were used in the task definition by Brysbaert et al. (2014) when collecting judgements for the concreteness norms. While Brysbaert et al. did not ask for a reference to specific senses rather than a general strength of sense perception, Lynott et al. (2020) collected judgements on specific senses (auditory, gustatory, haptic, olfactory, and visual) for the same targets as Brysbaert et al., using a scale from 0 to 5.

Emotion Dimensions Abstract words are considered to be more emotionally valenced than concrete words (Kousta et al., 2011; Vigliocco et al., 2014; Pollock, 2018). We thus explore emotion dimensions of our target words by using the NRC VAD Lexicon (Mohammad, 2018)[4] with ratings on valence, arousal, and dominance for over $20,000$ commonly used English words. The ratings were obtained by asking participants to judge the VAD strength of words using a best-worst scaling method. For each emotion dimension, the scores range from 0 (lowest VAD) to 1 (highest VAD).

Frequency and Ambiguity Frequency and ambiguity represent two standard dimensions influencing language processing and comprehension (Ellis, 2002; Baayen et al., 2016, i.a.). For frequency information, we rely on the target frequencies extracted from the ENCOW web corpus (see Section 3), containing $\approx$10 billion words. In order to distinguish between degrees of ambiguity of the targets, we rely on WordNet (Miller and Fellbaum, 1991; Fellbaum, 1998), a standard lexical semantic taxonomy for English word senses developed at Princeton University. WordNet organises words into classes of synonyms (*synsets*) connected by lexical and conceptual semantic relations. We looked up the number of noun and verb (but not adjective) target senses in WordNet version 3.0 and then used these WordNet ambiguity values if in the range $[1;6]$; targets with more than six senses in WordNet we assigned to a joint additional category.

[2] We disregard any two-word expressions.

[3] http://www.ims.uni-stuttgart.de/data/mid-scale

[4] https://saifmohammad.com/WebPages/nrc-vad.html

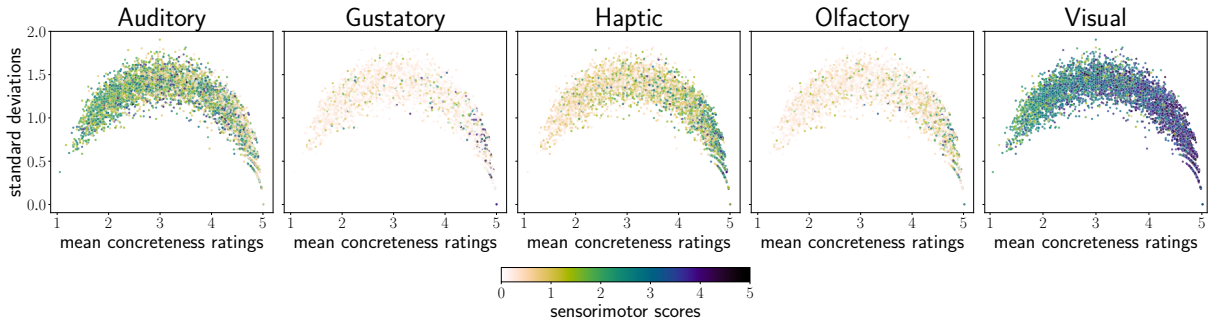

Figure 2: Mean noun ratings and standard deviations overlaid with the respective sense perception scores.

Free Word Associations Previous work suggested that free associations to abstract words differ from free associations to concrete words in terms of the number of types, thus pointing towards differences in conceptual semantic diversity. At the same time, associations to concrete words have been found weaker and more symmetric than for abstract words (Crutch and Warrington, 2010; Hill et al., 2014). The *Small World Of Words Project* SWOW (de Deyne et al., 2019)[5] provides large databases with free word associations across languages; for English, SWOW-EN includes more than 12,000 cue words with responses from over 90,000 participants. The associations were gathered from 2011–2018 by asking English speakers through crowd-sourcing to produce the first three response words that came to mind when presented with a cue word. We rely on SNOW-EN associations as indicators of diversity regarding our target words. Next to using only the first response R1, we aggregated the first two responses into a set R12, and all three responses into a set R123 to decrease sparsity, while accepting a minor association chain effect[6] (McEvoy and Nelson, 1982; Schulte im Walde and Melinger, 2008). We measured the diversity of responses by counting the number of types (i.e., the number of distinct associations that were produced across participants) in R1, R12, and R123, and normalised by the respective total numbers of response tokens.

Word Classes and Resource Coverage Table 1 provides an overview of how many of our targets are covered by the various resources across word classes. Note that from now on the main body of this paper will focus on nouns, and additionally we will refer to supporting evidence or differences regarding verb and adjective analyses in the text and in the Appendix.

	N	V	A
Targets in our subsets	5,448	1,280	2,205
Sense perception	5,440	1,280	2,202
Emotion	5,012	1,104	1,987
Frequency	5,448	1,280	2,205
Ambiguity[7]	5,400	1,277	–
Diversity: associations	3,501	780	1,255

Table 1: Coverage of target characteristics.

4.2 Holistic Perspective

Figure 2 visualises the relationships between mean noun concreteness ratings and standard deviations as introduced in Figure 1, in combination with heat maps indicating the rating strengths of auditory, gustatory, haptic, olfactory and visual perception (left to right).[8] Targets missing in a resource are plotted in grey. We can clearly observe an overall dominance of the visual perception (also see Table 5 in Appendix A for perception across senses), and that the strength of perception varies in different ways across the concreteness rating scale.

Table 2 informs us that visual, haptic, and olfactory sense perception (positively), as well as auditory (negatively), correlate with the noun concreteness scores. Regarding further target characteristics, the table reports a negative correlation with emotion regarding affect and dominance, as well as negative correlations with concept diversity regarding association types. The lexical characteristics do not show any correlations with concreteness.

[5] https://smallworldofwords.org/

[6] According to the association chain effect, the nth association response is supposedly associated to the (n-1)th association response rather than being associated to the target word; this effect might contaminate later association responses.

[8] Plots for further characteristics are in Appendix B.

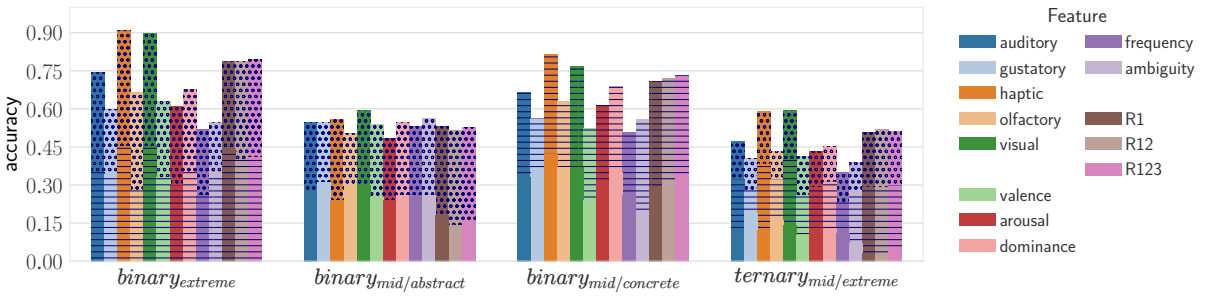

Figure 3: Results of classifications across characteristics and mid-scale/extreme experiments. The dotted and horizontal line patterns indicate the amount of abstract and concrete nouns correctly classified.

Target characteristics		ρ
Sense perception	Auditory	-0.28*
	Gustatory	0.01
	Haptic	0.58*
	Olfactory	0.29*
	Visual	0.61*
Emotion	Valence	-0.01
	Affect	-0.28*
	Dominance	-0.32*
Lexicon	Frequency	-0.00
	Ambiguity	-0.11*
Diversity: associations	R1	-0.33*
	R12	-0.41*
	R123	-0.43*

Table 2: Spearman's rank-order correlation coefficient ρ for the statistical relationships between degrees of concreteness and strengths of target noun characteristics; significance level is $p < 0.001$.

We thus conclude that overall the concreteness ratings of our target nouns[9] correlate to different degrees – and differing in negative vs. positive directions – with specific senses and also with further characteristics previously attributed to abstract vs. concrete concepts. This is our starting point for analysing whether any of these characteristics is particularly different for mid-scale target words and might have influenced their concreteness ratings.

4.3 Mid-Scale Peculiarities

We now investigate more specifically genuine characteristics of words that received mid-scale ratings, by zooming into differences in characteristics of mid-scale in contrast to extremely concrete/abstract target words, to maximise contrasts.

[9]See Tables 6–7 in Appendix C for verbs and adjectives.

Classification variants	Baseline	Accuracy
$binary_{extremes}$	0.50	0.98
$binary_{mid/abstract}$	0.50	0.75
$binary_{mid/concrete}$	0.50	0.93
$ternary_{mid/extremes}$	0.33	0.79

Table 3: Overall classification results (accuracy).

For this, we created three sets of 500 nouns each: the 500 most abstract nouns, the 500 most extreme nouns, and the 500 nouns with mean ratings closest to the rating-scale mean of 3 (with 250 nouns with mean $\leq$ 3 and 250 nouns with mean $>$ 3).[10] We then applied a Random Forest classifier and defined the following classification variants: a $ternary_{mid/extremes}$ condition where the classifier had to distinguish between the two extreme sets of 500 concrete and abstract targets from the mid-scale; $binary_{mid/abstract}$ and $binary_{mid/concrete}$ conditions to zoom into the individual mid-scale vs. extreme differences. As a control condition providing an upper bound for our classifiers, we included $binary_{extremes}$ where we classify only the extreme target sets with stronger differences between the two classes, while disregarding the mid-scale sets. The respective baselines are 50% for the binary classifications and 33% for the ternary classification.

The classifier used as features those target characteristics described and analysed in Section 4.2, separately and combined, in order to identify the characteristics that differ for mid-scale words in contrast to clearly abstract or concrete words. If a target word lacks a feature for a specific vari-

[10]We created several variants of mid-scale definitions, but given that neither modelling results nor insights differ strongly, we provide the variants in Appendix D.

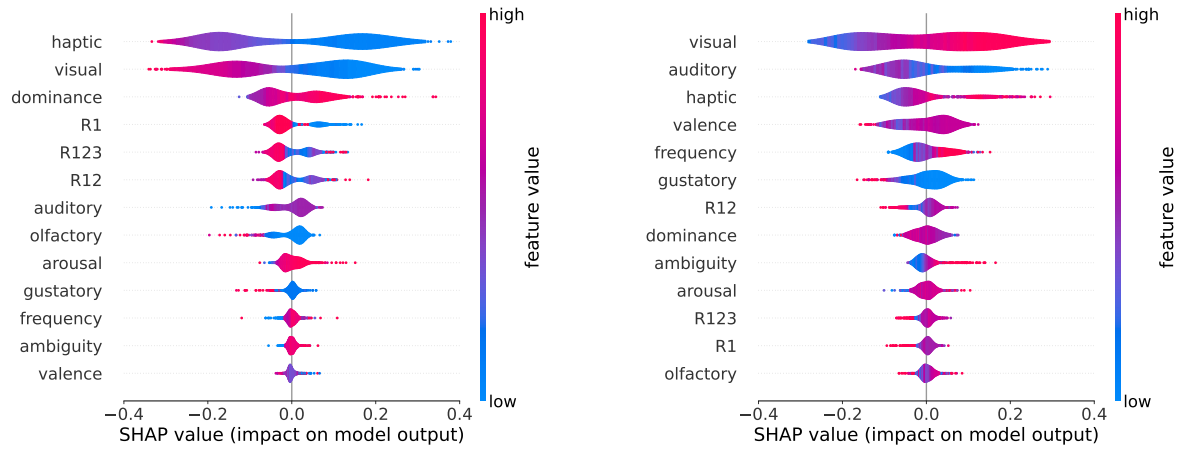

Figure 4: SHAP values – Importance of each feature for the output of the $binary_{mid/concrete}$ model (on the left) and the $binary_{mid/abstract}$ model (on the right). Extreme nouns are coded as negative, mid-scale nouns as positive.

able, we assigned 0 as the respective feature value. We applied 10-fold cross-validation and report the average accuracy score. The classification results using all the features at the same time are shown in Table 3. Figure 3 shows the results per feature type. As expected, the $binary_{extremes}$ classifications show the best results, with auditory, haptic, and visual sense perception as well as association diversity representing the strongest characteristics, in accordance with their overall correlation strengths in Section 4.2. The $ternary_{mid/extremes}$ results look like a miniature version of the $binary_{extremes}$ results with regard to accuracy across feature types, only on a lower scale (given the extra class). The results for the $binary_{mid/abstract}$ and $binary_{mid/concrete}$ conditions are lower than for $binary_{extremes}$, as predicted, because the contrasts on the concreteness scale are less strong. Also, we observe an interesting difference between the two conditions: targets with mid-scale ratings are distinguished better from targets with extremely concrete in comparison to extremely abstract ratings ($\rightarrow$ higher accuracy); at the same time, feature contributions in $binary_{mid/concrete}$ are similar to those in $binary_{extremes}$ and $ternary_{mid/extremes}$, while their contributions in $binary_{mid/abstract}$ are more uniform.

To further understand the differences between these two conditions, we inspected the contribution of each feature to the models' output using Shapley Additive Explanations (SHAP; Lundberg and Lee, 2017). Figure 4 shows the importance – as the magnitude of change – of each variable in predicting the concreteness scores of concrete (left plot) and abstract (right plot) nouns vs. mid-scale nouns. The colours of the violin plots indicate the values of the features. For the $binary_{mid/concrete}$ model, the three most important features for the classification are haptic, visual, and dominance, in that order. Conversely, for the $binary_{mid/abstract}$ model, the most important features are visual, auditory, and haptic. Notably, visual and haptic features emerge as the most informative in both cases. Associations, instead, show a relatively small contribution to the performance of the classifier when together with other feature types (as opposed to Figure 3).

An analysis of the colour-coded information (i.e., the value of each feature) supports our previous evidence. In the left plot in Figure 4, we can see a clear distinction between concrete nouns that are characterised by high (magenta) visual and haptic values, and mid-concreteness nouns characterised by low (blue) visual and haptic values. Conversely, in the right plot in Figure 4 the visual and haptic nature of abstract versus mid-scale nouns exhibits less pronounced differences with magenta colour associated both with mid-scale (positive) and abstract (negative) nouns.

We thus infer from our classification experiments that mid-scale target nouns are more easily distinguishable from extremely concrete in comparison to extremely abstract targets, with regard to our set of features. In the next section, we will investigate why this is the case.

5 Mid-Scale Disagreement Patterns

In our final analyses, we zoom into the numerical characteristics of mid-scale mean ratings. If there was substantial agreement behind the *semi*-perception of a mid-scale target (i.e., if all human

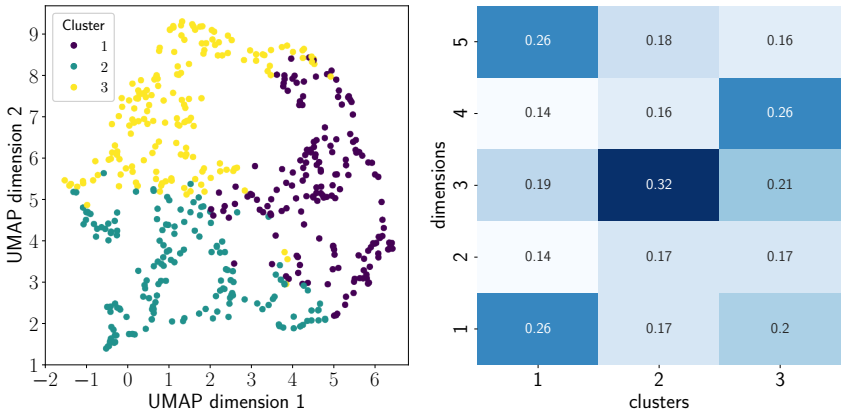

Figure 5: k-Means clustering ($k = 3$) of 500 mid-scale nouns based on original individual per-participant rating distributions. Cluster sizes are 170, 163, and 167. The heatmap shows the rating distributions of the centroid vectors.

raters had provided a rating of 3 or similar on the scale of 1 to 5), then we would see a standard deviation around 0 in the croissant plots in Figure 1. We however observe rather high standard deviations for targets with mean ratings of ≈3, thus indicating considerable disagreement across raters. The question we are asking is how these disagreements were triggered. We hypothesise that raters might have been influenced differently by their individual perceptions of target characteristics, and that we therefore find several patterns of disagreement across the mid-scale target words.

For this exploration of disagreement patterns, we make use of the original per-participant ratings in Brysbaert et al. (2014), and applied a standard k-means hard clustering approach to automatically assign the 500 mid-scale nouns to $k = 3$ clusters. As representations for the targets, we used 5-dimensional vectors with relative frequencies per rating categories $1, 2, 3, 4, 5$, based on the original individual ratings, e.g., the vector for the noun *discussion* is $\vec{v} = \langle 0.15, 0.07, 0.48, 0.15, 0.15 \rangle$, because 15% of the raters provided ratings of $1, 4$ or 5, while 7% judged it as 2, and 48% judged it as 3.

Figure 5 presents two perspectives on the resulting clusters with rather homogeneous cluster sizes $170, 163, 167$. On the left,[11] we can see that the three clusters are clearly separated, with relatively small overlapping areas, thus indicating that the underlying cluster features (i.e., the rating distributions) clearly differ. This is confirmed by the plot on the right, which shows the individual rating distributions (y-axis) of the three cluster centroids 1–3 (x-axis). The heatmap exhibits rather different patterns: in cluster 1, we find the strongest disagreements among raters, where each of the two extreme rating scores (1 and 5) were chosen by 26%, the mid-score by 19%, and the remaining scores are equally distributed over ratings 2 and 4 (14% each); in cluster 2, 32% of the raters judged the respective target nouns as 3 because they were completely undecided or they consciously chose a mid-scale *semi*-perception score, while the other raters decided for $1, 2, 4, 5$ with almost identical proportions (16–18%); finally, in cluster 3 we find a more uniform rating distribution, while a score of 4 was given by most of the raters (26%). Table 4 provides a few example targets for each of the three clusters, together with their rating distributions.

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	$\langle 0.23, 0.07, 0.23, 0.30, 0.17 \rangle$

Table 4: Examples of rating distributions for noun target words across clusters C.

Overall, Figure 5 thus provides very strong evidence in favour of our hypothesis that a mid-scale mean rating conflates rather different patterns of disagreements across human ratings. Figures 12

[11] We used UMAP (Uniform Manifold Approximation and Projection) for down-scaling our distributions to two dimensions (McInnes et al., 2018).

and 13 in Appendix E provide the respective plots for verbs and adjectives, where we find similar patterns of disagreement.

6 Discussion & Conclusion

We started out with the well-known observation that humans tend to strongly agree on ratings on a scale for extreme cases, but that judgements on mid-scale words exhibit more disagreement. This observation is well-described by the croissant-like shape of mean rating scores in relation to their standard deviations (cf. Figure 1). While individual studies have pointed out problems with such ratings on a scale (e.g., Kiritchenko and Mohammad (2017); Pollock (2018)) and also provided alternative settings (e.g., Kiritchenko and Mohammad (2017); Abdalla et al. (2023)), the scale-based norms are heavily exploited across disciplines, including state-of-the-art computational approaches.

In the current study, we first asked whether words with mid-scale concreteness ratings potentially exhibit specific characteristics that genuinely distinguish them from clearly concrete and clearly abstract words. The corresponding classification experiments and feature analyses demonstrated that mid-scale targets were indeed distinguishable from extreme targets with regard to a subset of the senses which were used as criteria for the concreteness–abstractness distinction (mainly visual and haptic), and also with regard to emotional dimensions and meaning diversity (implemented on the basis of association types). In this first set of experiments mid-scale targets therefore established themselves as genuine intermediate concepts. We also saw, however, that mid-scale nouns are more easily distinguishable from extremely concrete in comparison to extremely abstract nouns, and this asymmetry flips with regard to verbs and adjectives, presumably because their underlying rating distributions exhibit different skews (cf. the croissant plots in Figure 1 and the different mid-scale ranges in Figure 9 in Appendix D). So overall, words with mid-scale mean ratings represent rather genuine intermediate concepts regarding our implementations of features and analyses.

In the second part of our study, we investigated whether mid-scale ratings are generally agreed upon across raters, or whether raters disagreed regarding their *semi*-perception. Relying on explorative cluster analyses using the original per-participant rating distributions, we found clusters with obviously very different centroids. From this, we induce that a mid-scale rating mean of $\approx$3 conflates rather different yet systematic kinds of disagreements. This observation is in line with the mathematically-based observations by Pollock (2018) that "there is only a finite number of possible combinations of means and standard deviations", and at the same time it clearly demonstrated that mid-scale ratings indeed differ regarding their underlying rating combinations. So, on the one hand, our cluster analyses confirm a so-far rather theoretically-driven observation; on the other hand, we raise the question of whether and how this observation should influence the utilisation of ratings on a scale. We suggest two alternative routes: (i) either filter the norm targets and only keep those targets that are clearly attributable to one extreme, or (ii) fine-tune the mid-scale norm targets with regard to inherent disagreement patterns, because the set of mid-scale targets is itself rather inhomogeneous but nevertheless provides valuable information regarding specific differences in human perception.

Last but not least we would like to point out that inherent disagreements among human annotators are obviously not restricted to our particular focus on mid-scale ratings but represent a common issue under discussion across annotation tasks. In the past decade the field has moved from considering disagreements as pure noise towards zooming into disagreements in order to distinguish between noise and subjectivity, and to effectively exploit the value of disagreements in language modelling, see Alm (2011) and Uma et al. (2021) for a prominent opinion paper and a prominent survey, respectively. Our analyses and insights should be interpreted in the same vein: we attribute disagreements on concreteness mid-scale ratings to genuine intermediate concepts (see above) and suggest to take a fine-grained approach when utilising them in language modelling tasks and applications.

Limitations

Our study is targeting ratings on a scale but currently restricted to a selection of target properties and a specific case study on concreteness. Future work will explore additional target properties that might influence concreteness mid-scale ratings (such as the mass-count distinction and register) as well as characteristics of ratings on a scale in further collections and other languages than English.

Ethics Statement

For our study, we used and cited publicly available datasets and libraries. The resources do not contain any information that uniquely identifies individuals. Our research does not raise any immediate ethical concerns.

Acknowledgements

This research was supported by the Ad Futura Scholarship (305. JR) from the Public Scholarship, Development, Disability and Maintenance Fund of the Republic of Slovenia (Urban Knupleš), and by the DFG Research Grant SCHU 2580/4-1 (*MUDCAT – Multimodal Dimensions and Computational Applications of Abstractness*). We also thank the reviewers for suggesting additional perspectives regarding our analyses and interpretations.

References

Mohamed Abdalla, Krishnapriya Vishnubhotla, and Saif Mohammad. 2023. What Makes Sentences Semantically Related? A Textual Relatedness Dataset and Empirical Study. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 782–796, Dubrovnik, Croatia.

Eleri Aedmaa, Maximilian Köper, and Sabine Schulte im Walde. 2018. Combining Abstractness and Language-specific Theoretical Indicators for Detecting Non-Literal Usage of Estonian Particle Verbs. In *Proceedings of the NAACL 2018 Student Research Workshop*, pages 9–16, New Orleans, LA, USA.

Salvador Algarabel, Juan Carlos Ruiz, and Jaime Sanmartin. 1988. The University of Valencia's Computerized Word Pool. *Behavior Research Methods, Instruments, and Computers*, 20(4):398–403.

Cecilia Ovesdotter Alm. 2011. Subjective Natural Language Problems: Motivations, Applications, Characterizations, and Implications. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 107–112, Portland, Oregon, USA.

R. Harald Baayen, Petar Milin, and Michael Ramscar. 2016. Frequency in Lexical Processing. *Aphasiology*, 30(11):1174–1220.

Patrick Bonin, Alain Meot, and Aurelia Bugaiska. 2018. Concreteness Norms for 1,659 French Words: Relationships with other Psycholinguistic Variables and Word Recognition Times. *Behavior Research Methods*, 50:2366–2387.

Marc Brysbaert, Boris New, and Emmanuel Keuleers. 2012. Adding Part-of-Speech Information to the SUBTLEX-US Word Frequencies. *Behavior Research Methods*, 44:991–997.

Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014. Concreteness Ratings for 40 Thousand generally known English Word Lemmas. *Behavior Research Methods*, 64:904–911.

Silvio Cordeiro, Aline Villavicencio, Marco Idiart, and Carlos Ramisch. 2019. Unsupervised Compositionality Prediction of Nominal Compounds. *Computational Linguistics*, 45(1):1–57.

Sebastian J. Crutch and Elizabeth K. Warrington. 2010. The Differential Dependence of Abstract and Concrete Words upon Associative and Similarity-based Information: Complementary Semantic Interference and Facilitation Effects. *Cognitive Neuropsychology*, 27:46–71.

Simon de Deyne, Danielle J. Navarro, Amy Perfors, Marc Brysbaert, and Gert Storms. 2019. The "Small World of Words" English Word Association Norms for over 12,000 Cue Words. *Behavior Research Methods*, 51:987–1006.

Pasquale A. Della Rosa, Eleonora Catricala, Gabriella Vigliocco, and Stefano F. Cappa. 2010. Beyond the Abstract–Concrete Dichotomy: Mode of Acquisition, Concreteness, Imageability, Familiarity, Age of Acquisition, Context Availability, and Abstractness Norms for a Set of 417 Italian Words. *Behavior Research Methods*, 42(4):1042–1048.

Annerose Eichel and Sabine Schulte Im Walde. 2023. A Dataset for Physical and Abstract Plausibility and Sources of Human Disagreement. In *Proceedings of the 17th Linguistic Annotation Workshop*, pages 31–45, Toronto, Canada.

Nick C. Ellis. 2002. Frequency Effects in Language Processing: A Review with Implications for Theories of Implicit and Explicit Language Acquisition. *Studies in Second Language Acquisition*, 24(2):143–188.

Christiane Fellbaum, editor. 1998. *WordNet – An Electronic Lexical Database*. Language, Speech, and Communication. MIT Press, Cambridge, MA, USA.

Christina L. Gagné, Thomas L. Spalding, and Daniel Schmidtke. 2019. LADEC: The Large Database of English Compounds. *Behavior Research Methods*, 51:2152–2179.

Fritz Günther, Marco Marelli, and Jens Bölte. 2020. Semantic Transparency Effects in German Compounds: A Large Dataset and Multiple-Task Investigation. *Behavior Research Methods*, 52:1208–1224.

Felix Hill, Anna Korhonen, and Christian Bentz. 2014. A Quantitative Empirical Analysis of the Abstract/Concrete Distinction. *Cognitive Science*, 38:162–177.

Philipp Kanske and Sonja A. Kotz. 2010. Leipzig Affective Norms for German: A Reliability Study. *Behavior Research Methods*, 42(4):987–991.

Emmanuel Keuleers and David A. Balota. 2015. Megastudies, Crowdsourcing, and Large Datasets in Psycholinguistics: An Overview of Recent Developments. *The Quarterly Journal of Experimental Psychology*, 68(8):1457–1468.

Svetlana Kiritchenko and Saif M. Mohammad. 2017. Best–Worst Scaling More Reliable than Rating Scales: A Case Study on Sentiment Intensity Annotation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*, pages 465–470, Vancouver, Canada.

Maximilian Köper and Sabine Schulte im Walde. 2016a. Automatically Generated Affective Norms of Abstractness, Arousal, Imageability and Valence for 350 000 German Lemmas. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 2595–2598, Portorož, Slovenia.

Maximilian Köper and Sabine Schulte im Walde. 2016b. Distinguishing Literal and Non-Literal Usage of German Particle Verbs. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 353–362, San Diego, CA, USA.

Maximilian Köper and Sabine Schulte im Walde. 2018. Analogies in Complex Verb Meaning Shifts: The Effect of Affect in Semantic Similarity Models. In *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 150–156, New Orleans, LA, USA.

Stavroula-Thaleia Kousta, Gabriella Vigliocco, David P. Vinson, Mark Andrews, and Elena Del Campo. 2011. The Representation of Abstract Words: Why Emotion Matters. *Journal of Experimental Psychology: General*, 140(1):14–34.

Scott M. Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 4765–4774. Curran Associates, Inc.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2020. The Lancaster Sensorimotor Norms: Multidimensional Measures of Perceptual and Action Strength for 40,000 English Words. *Behavior Research Methods*, 52:1–21.

Pawel Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2015. How useful are Corpus-based Methods for Extrapolating Psycholinguistic Variables? *The Quarterly Journal of Experimental Psychology*, 68(8):1623–1642.

Rowan Hall Maudslay, Tiago Pimentel, Ryan Cotterell, and Simone Teufel. 2020. Metaphor Detection Using Context and Concreteness. In *Proceedings of the 2nd Workshop on Figurative Language Processing*, pages 221–226, Seattle, Washington (online).

Cathy L. McEvoy and Douglas L. Nelson. 1982. Category Name and Instance Norms for 106 Categories of Various Sizes. *American Journal of Psychology*, 95:581–634.

Leland McInnes, John Healy, Nathaniel Saul, and Lukas Grossberger. 2018. Umap: Uniform manifold approximation and projection. *The Journal of Open Source Software*, 3(29):861.

George A. Miller and Christiane Fellbaum. 1991. Semantic Networks of English. *Cognition*, 41:197–229.

Saif M. Mohammad. 2018. Obtaining Reliable Human Ratings of Valence, Arousal, and Dominance for 20,000 English Words. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*, Melbourne, Australia.

Saif M. Mohammad, Mohammad Salameh, and Svetlana Kiritchenko. 2016. How Translation Alters Sentiment. *Journal of Artificial Intelligence Research*, 55:95–130.

Emiko J. Muraki, Summer Abdalla, Marc Brysbaert, and Penny M. Pexman. 2022. Concreteness Ratings for 62 Thousand English Multiword Expressions. *PsyArXiv*.

Allan Paivio, John C. Yuille, and Stephen A. Madigan. 1968. Concreteness, Imagery, and Meaningfulness Values for 925 Nouns. *Journal of Experimental Psychology (Monograph Supplement)*, 76(1/2):1–25.

Lewis Pollock. 2018. Statistical and Methodological Problems with Concreteness and other Semantic Variables: A List Memory Experiment Case Study. *Behavior Research Methods*, 50:1198–1216.

Siva Reddy, Diana McCarthy, and Suresh Manandhar. 2011. An Empirical Study on Compositionality in Compound Nouns. In *Proceedings of the 5th International Joint Conference on Natural Language Processing*, pages 210–218, Chiang Mai, Thailand.

Roland Schäfer. 2015. Processing and Querying Large Web Corpora with the COW14 Architecture. In *Proceedings of the 3rd Workshop on Challenges in the Management of Large Corpora*, pages 28–34, Mannheim, Germany.

Roland Schäfer and Felix Bildhauer. 2012. Building Large Corpora from the Web Using a New Efficient Tool Chain. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 486–493, Istanbul, Turkey.

Sabine Schulte im Walde and Diego Frassinelli. 2022. Distributional Measures of Abstraction. *Frontiers in Artificial Intelligence: Language and Computation 4:796756. Alessandro Lenci and Sebastian Pado (topic editors): "Perspectives for Natural Language Processing between AI, Linguistics and Cognitive Science".*

Sabine Schulte im Walde, Anna Hätty, Stefan Bott, and Nana Khvtisavrishvili. 2016. G$_h$ost-NN: A Representative Gold Standard of German Noun-Noun Compounds. In *Proceedings of the 10th International Conference on Language Resources and Evaluation*, pages 2285–2292, Portorož, Slovenia.

Sabine Schulte im Walde and Alissa Melinger. 2008. An In-Depth Look into the Co-Occurrence Distribution of Semantic Associates. *Italian Journal of Linguistics. Alessandro Lenci (guest editor): From Context to Meaning: Distributional Models of the Lexicon in Linguistics and Cognitive Science*, 20(1):89–128.

Otfried Spreen and Rudolph W. Schulz. 1966. Parameters of Abstraction, Meaningfulness, and Pronunciability for 329 Nouns. *Journal of Verbal Learning Behavior*, 5:459–468.

Yulia Tsvetkov, Leonid Boytsov, Anatole Gershman, Eric Nyberg, and Chris Dyer. 2014. Metaphor Detection with Cross-Lingual Model Transfer. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 248–258, Baltimore, MD, USA.

Peter D. Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and Metaphorical Sense Identification through Concrete and Abstract Context. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, UK.

Alexandra N. Uma, Tommaso Fornaciari, Dirk Hovy, Silviu Paun, Barbara Plank, and Massimo Poesio. 2021. Learning from Disagreement: A Survey. *Journal of Artificial Intelligence Research*, 72:1385–1470.

Gabriella Vigliocco, Stavroula-Thaleia Kousta, Pasquale Anthony Della Rosa, David P. Vinson, Marco Tettamanti, Joseph T. Devlin, and Stefano F. Cappa. 2014. The Neural Representation of Abstract Words: The Role of Emotion. *Cerebral Cortex*, 24:1767–1777.

Su Wang, Greg Durrett, and Katrin Erk. 2018. Modeling Semantic Plausibility by Injecting World Knowledge. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 303–308, New Orleans, Louisiana.

A Dominance of Perception across Targets

Table 5 shows how many of our target words (nouns, verbs, adjectives, overall) were perceived predominantly by any of the human senses auditory, gustatory, haptic, olfactory, visual, according to the analyses by Lynott et al. (2020).

	Auditory	Gustatory	Haptic	Olfactory	Visual	Total
N	610	199	102	38	4,491	5,440
V	269	8	27	4	972	1,280
A	341	31	64	7	1,759	2,202
all	1,220	238	193	49	7,222	8,922

Table 5: Distribution of dominant perceptual modalities of our target words, based on Lynott et al. (2020).

B Visualisations of Rating Characteristics for Nouns[12]

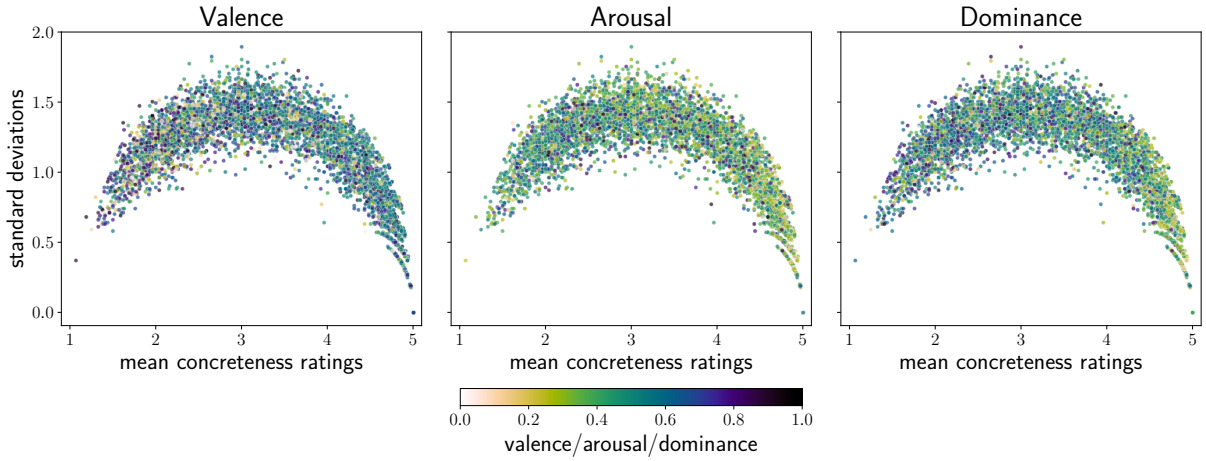

Figure 6: Mean noun ratings and standard deviations overlaid with the respective VAD scores.

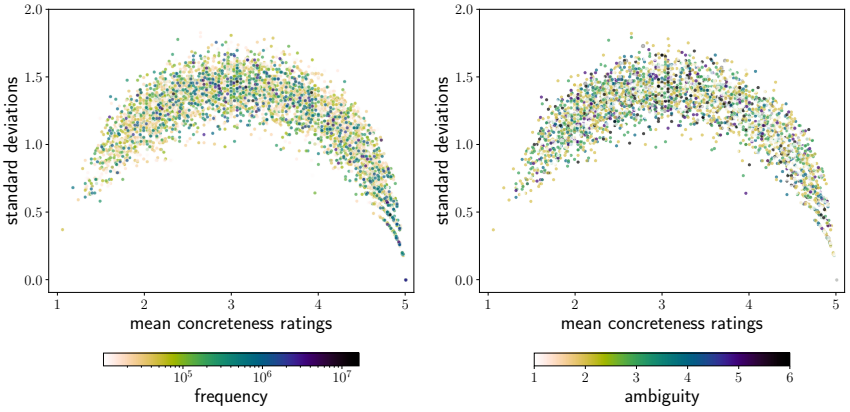

Figure 7: Mean noun ratings and standard deviations overlaid with heatmaps of the respective $\log_{10}$-scaled frequency and ambiguity values.

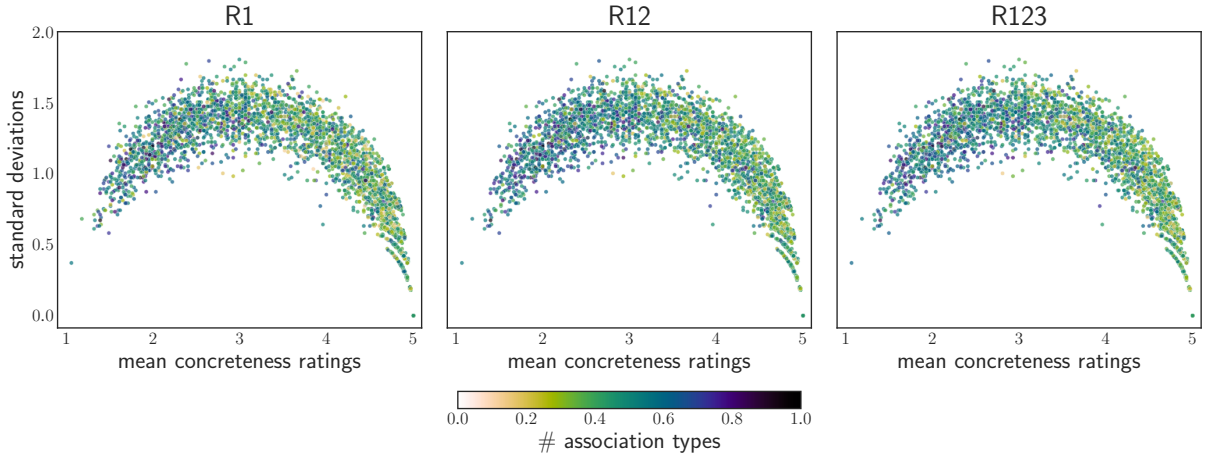

Figure 8: Mean noun ratings and standard deviations overlaid with a normalised number of the association types in the sets R1, R12, and R123.

[12]The corresponding visualisations of rating characteristics for verbs and adjectives are publicly available from http://www.ims.uni-stuttgart.de/data/mid-scale.

C Correlations between Target Characteristics and Concreteness: Verbs and Adjectives

Target characteristics		ρ
Sense perception	Auditory	-0.28*
	Gustatory	-0.09*
	Haptic	0.47*
	Olfactory	0.01
	Visual	0.47*
Emotion	Valence	-0.11*
	Affect	0.04
	Dominance	-0.15*
Lexicon	Frequency	-0.01
	Ambiguity	0.13*
Diversity: associations	R1	-0.30*
	R12	-0.31*
	R123	-0.31*

Table 6: Spearman's rank-order correlation coefficient ρ for the statistical relationships between degrees of concreteness and strengths of target *verb* characteristics; significance level is $p < 0.05$.

Target characteristics		ρ
Sense perception	Auditory	-0.37*
	Gustatory	-0.01
	Haptic	0.35*
	Olfactory	0.04
	Visual	0.39*
Emotion	Valence	-0.03
	Affect	-0.07*
	Dominance	-0.08*
Lexicon	Frequency	-0.04
Diversity: associations	R1	-0.28*
	R12	-0.32*
	R123	-0.31*

Table 7: Spearman's rank-order correlation coefficient ρ for the statistical relationships between degrees of concreteness and strengths of target *adjective* characteristics; significance level is $p < 0.05$.

D Mid-Scale Definitions, Ranges and Classifications across Word Classes

Intuitively, the interpretation of mid-scale targets refers to somewhere in the middle of the mean concreteness ratings plots that we have presented in Figure 1, in contrast to extremely abstract targets on the left and extremely concrete targets on the right. Accordingly, we suggest three ways of capturing this intuition, given that the number of targets per part-of-speech (POS) and also the ranges of ratings and their skewness differ across POS. We created three sets of 500 mid-scale noun targets accordingly, and also three sets of 200 mid-scale verb and 200 mid-scale adjective targets.

Mid-Scale-Mean The mid-scale score is defined as the mean value on the rating scale, which is 3 in our scale $[1; 5]$. Mid-scale targets are then defined as those words whose mean ratings are closest to 3.

Mid-Scale-Median Given that the rating distributions differ across POS and with regard to their left vs. right skews, the mid-scale score is defined as the median, in our case: 3.54 for the nouns, 2.47 for the verbs, and 2.19 for the adjectives. Mid-scale targets are then defined as those words whose mean ratings are closest to these medians.

Mid-Scale-Median-SD Incorporating disagreement between raters, we refine the mid-scale-median taking into account as mid-scale targets only those words whose mean ratings are closest to the median and whose standard deviations are > 1.4.

In all three cases, we selected an equal number of targets with mean ratings above and below the respective mid-scale score. Figure 9 provides the mean-rating ranges of our mid-scale targets across these three mid-scale definitions, based on the respective 500/200/200 mid-scale noun/verb/adjective targets. The same figure shows the mean-rating ranges of the extremely concrete and extremely abstract targets, relying again on sets of 500/200/200 targets. We can see that the mid-scale ranges clearly differ across definitions and POS. Table 8 shows the classification results (accuracy) across these mid-scale definitions, word classes and target set constellations. Figures 10 and 11 zoom into the classification results of verb/adjective targets per feature type and for the mid-scale mean definition, as done for nouns in Figure 3.

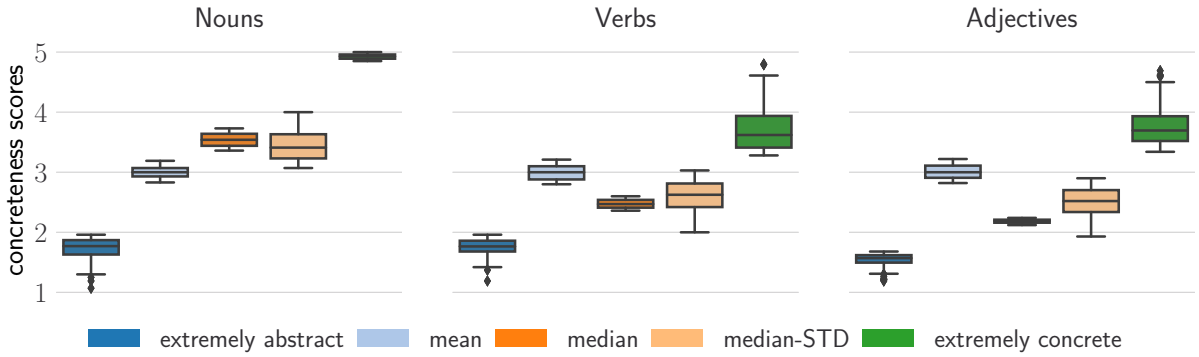

Figure 9: Distributions of concreteness scores across mid-scale definitions and POS.

		Mid-Scale Definition		
		Mean	**Median**	**Median-SD**
nouns	$binary_{extremes}$	0.98	0.98	0.98
	$ternary_{mid/extremes}$	0.79	0.82	0.82
	$binary_{mid/concrete}$	0.93	0.91	0.91
	$binary_{mid/abstract}$	0.75	0.83	0.82
verbs	$binary_{extremes}$	0.90	0.90	0.90
	$ternary_{mid/extremes}$	0.63	0.64	0.65
	$binary_{mid/concrete}$	0.64	0.78	0.78
	$binary_{mid/abstract}$	0.81	0.65	0.73
adjectives	$binary_{extremes}$	0.94	0.94	0.94
	$ternary_{mid/extremes}$	0.67	0.67	0.67
	$binary_{mid/concrete}$	0.68	0.86	0.81
	$binary_{mid/abstract}$	0.84	0.55	0.71

Table 8: Results of the classifications across mid-scale definitions and target set constellations.

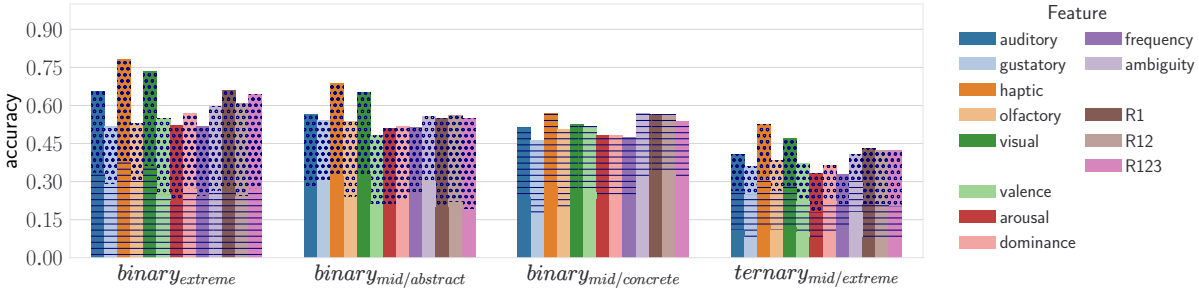

Figure 10: Results of classifications across characteristics and mid-scale/extreme experiments. The dotted and horizontal line patterns indicate the amount of abstract and concrete *verbs* correctly classified.

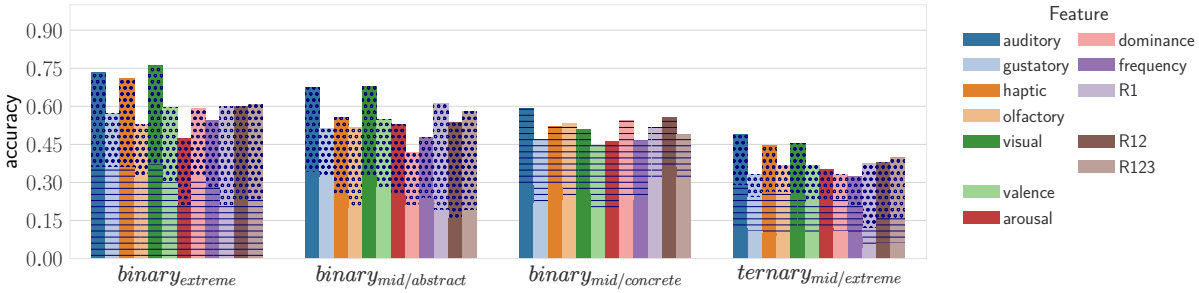

Figure 11: Results of classifications across characteristics and mid-scale/extreme experiments. The dotted and horizontal line patterns indicate the amount of abstract and concrete *adjectives* correctly classified.

E Mid-Scale Disagreement Patterns in Verb and Adjective Rating Distributions

Figures 12 and 13 present the clusters and the heat maps of rating distributions of the cluster centroids for verbs and adjectives. The clusters are based on the same k-Means clustering setup as those for nouns in Section 5.

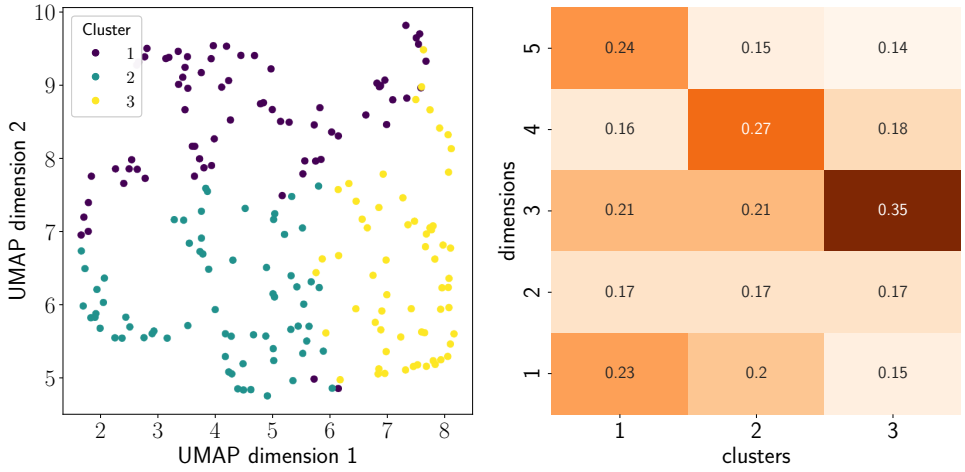

Figure 12: k-Means clustering ($k = 3$) of 200 mid-scale verbs based on original individual per-participant rating distributions. Cluster sizes are 71, 68, and 61. The heatmap shows the rating distributions of the centroid vectors.

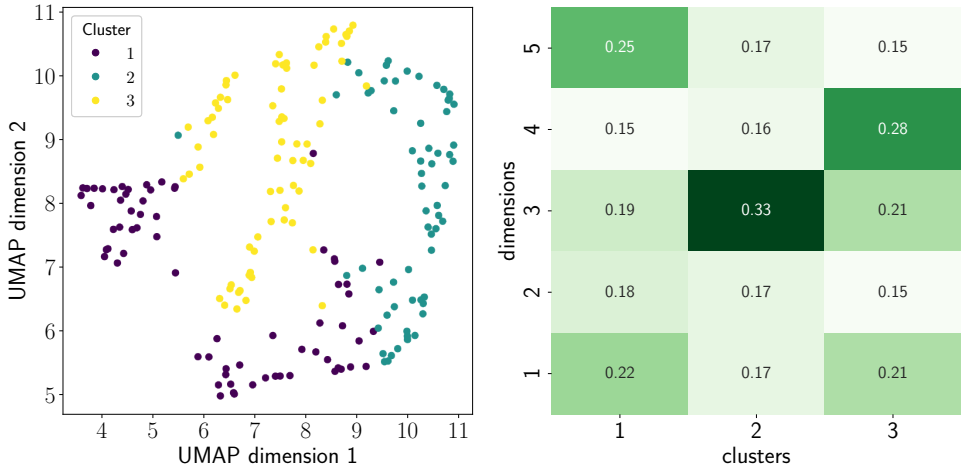

Figure 13: k-Means clustering ($k = 3$) of 200 mid-scale adjectives based on original individual per-participant rating distributions. Cluster sizes are 68, 62, and 70. The heatmap shows the rating distributions of the centroid vectors.