

# Literally Concrete or Figuratively Abstract? Multilingual Concreteness Norms for Verb-Object Expressions

Urban Knuples<sup>1</sup>

Diego Frassinelli<sup>2</sup>

Alexander Fraser<sup>3</sup>

Sabine Schulte im Walde<sup>1</sup>

<sup>1</sup>Institute for Natural Language Processing, University of Stuttgart

<sup>2</sup>Center for Information and Language Processing, LMU Munich

<sup>3</sup>School of Computation, Information and Technology, TU Munich

{urban.knuples, schulte}@ims.uni-stuttgart.de

frassinelli@cis.lmu.de

alexander.fraser@tum.de

## Abstract

While existing concreteness norms primarily target words in isolation, little attention has been paid to concreteness in context. To address this, we systematically collect multilingual concreteness ratings using Best-Worst Scaling (BWS) for 5,814 verb-direct object noun expressions in three languages with different degrees of resource availability: English, German, and Slovene. We identify consistent patterns where the concreteness of verb-noun combinations is more strongly influenced by the nominal object than the verb. Through comparative analyses on an English subset, we demonstrate that BWS guarantees more reliable concreteness judgments than traditional rating scales. Expanding beyond our human-generated data, we use traditional and LLM-based automatic extrapolation methods to generate a large-scale multilingual resource of over 430,000 expressions. Additionally, we conduct a study examining the interaction between concreteness and literal vs. figurative judgments for a subset of 1,800 expressions in all three languages, along with example usage sentences. Our findings show that lower concreteness ratings correlate with figurative language, thus reinforcing the link between abstractness and figurativeness. All resources are available from <https://github.com/urbikn/multilingual-concreteness-vo>.

## 1 Introduction

Concreteness refers to the degree to which a concept’s meaning can be directly experienced

through our five senses (e.g., a *cat* is more concrete than *wisdom*). Concreteness norms have been widely used across various NLP tasks and languages, thus representing a fundamental resource for many state-of-the-art computational methods, such as metaphor processing (Turney et al., 2011; Tsvetkov et al., 2014; Köper and Schulte im Walde, 2016; Alnafesah et al., 2020; Maudslay et al., 2020; Piccirilli and Schulte im Walde, 2022a,b; Hülsing and Schulte im Walde, 2024; Khaliq et al., 2024) and embodied agents and robots (Cangelosi and Stramandinoli, 2018; Rasheed et al., 2018; Ichter et al., 2023).

Traditionally, concreteness ratings are collected in isolation, i.e., out of context. For quantifying the degree of concreteness of more complex constructions (up to sentence level), these individual word-level ratings are typically averaged. This approach however completely ignores the interactions between word classes and lexical characteristics in context, where we argue that concreteness is not a simple additive property (see Figure 1). For example, *carry implication* represents a verb-object expression that is perceived as rather abstract, while most humans judge the verb *to carry* in isolation as rather concrete. This limitation is particularly evident in metaphor research, where the interplay between concrete and abstract concepts plays a fundamental role for understanding figurative language (Lakoff and Johnson, 1980). Nevertheless, up to date only few studies across languages have explored concreteness ratings in multiword settings or broader contexts (Frassinelli and Schulte im Walde, 2019; Gregori et al., 2020; Montefinese et al., 2023; Muraki et al., 2023).

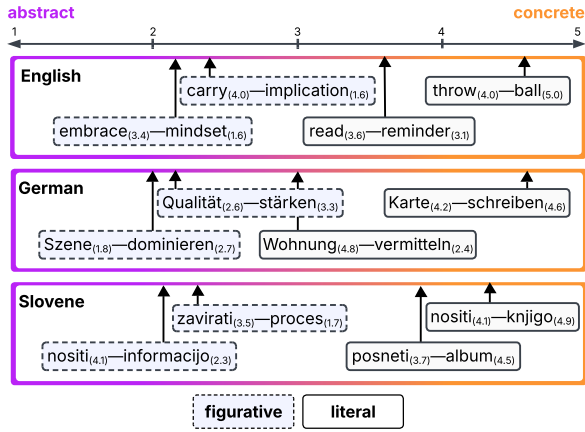


Figure 1: Examples of verb-object expressions for English, German, and Slovene with word-level (subscript) and expression-level (scale at top) concreteness ratings ranging from 1 (abstract) to 5 (concrete); dashed and solid box lines indicate a binary distinction between figurative and literal language (scale at bottom). An expression’s concreteness often diverges from the words’ average.<sup>1</sup>

As to our knowledge, no previous study has systematically worked bottom-up by paying attention to word functions and combinations of individual word-level concreteness.

Furthermore, conducting meaningful experiments using concreteness norms requires a very large vocabulary of normed words, but even extensive collections provide only partial coverage, thus limiting their applicability in large-scale NLP tasks. Given that a manual collection of human-annotated concreteness ratings is both time-consuming and expensive, various studies have developed methods to automatically generate concreteness scores on a large scale, aiming to expand the available resources while maintaining reliability and accuracy (Turney et al., 2011; Keuleers and Balota, 2015; Mander et al., 2015; Köper and Schulte im Walde, 2016; Conde et al., 2026), but as to our knowledge, no research has implemented extrapolation for multiword targets.

One final aspect we examine is *how* concreteness is quantified. Rating scales (RSs) are the predominant method for collecting concreteness norms, where annotators judge a word’s con-

<sup>1</sup>Translations for German and Slovene expressions: *Qualität stärken* (“strengthen quality”), *Karte schreiben* (“write card”), *Szene dominieren* (“dominate scene”), *Wohnung vermitteln* (“arrange apartment”), *zavirati proces* (“brake process”), *nositi knjigo* (“carry book”), *nositi informacija* (“carry information”), *posneti album* (“record album”).

creteness on an abstract-to-concrete scale. RSs however come with notable limitations in the form of higher annotator disagreement for mid-scale words (Pollock, 2018; Knupleš et al., 2023; Paisios et al., 2023). In contrast, we apply Best-Worst Scaling (BWS; Louviere and Woodworth (1991)), which offers a promising alternative where annotators select the most and least concrete items from a set.

Overall, this work addresses the lack of multilingual concreteness norms in broader linguistic contexts. We collect human judgments for verb-direct object expressions in English, German, and Slovene using BWS, thus ensuring a more reliable assessment of concreteness beyond single-word ratings. Given the interdisciplinary demand for large-scale concreteness norms and the high cost of human annotation, we generate a large-scale dataset of automatically extrapolated concreteness scores. We systematically compare a range of computational methods, from traditional machine learning methods to state-of-the-art large language models (LLMs), providing a robust solution for expanding concreteness norms across languages and linguistic contexts. Finally, to support further research on the connection between concreteness and figurative language, we annotate a subset of expressions as figurative or literal. To sum up, this work makes three key contributions:

1. We systematically investigate the role of concreteness in context by introducing novel concreteness norms for 5,814 verb-object noun expressions in English, German, Slovene.
2. We release a large-scale set of 431,262 silver-standard multilingual concreteness scores, automatically generated by traditional and LLM-based computational approaches.
3. We create a valuable resource for figurative language research by providing figurative judgments and example sentences for 1,800 verb-object expressions in English, German, and Slovene.

## 2 Related Work

**Concreteness Ratings** Over the past few decades, a substantial body of work has been dedicated to the collection of word-level concreteness norms across languages, e.g., German

(Lahl et al., 2009; Kanske and Kotz, 2010), Italian (Montefinese et al., 2014), Chinese (Yao et al., 2017), French (Bonin et al., 2018), and Croatian (Peti-Stantić et al., 2021), all containing between 1,000 and 6,000 words. Larger collections are available for Estonian (Proos and Aigro, 2023) and Dutch (Brysbaert et al., 2014a), including 36,000 and 30,000 words, respectively. The most famous and widely used collection of norms is by Brysbaert et al. (2014b), and contains concreteness ratings for approximately 40,000 English words.

The creation of these resources is both resource-intensive and time-consuming, resulting in many languages lacking even a basic collection of word-level concreteness norms. Researchers have thus employed extrapolation methods to predict concreteness ratings via monolingual machine-learning approaches (Mandera et al., 2015; Turney et al., 2011; Köper and Schulte im Walde, 2016) or cross-lingual transfer methods (Tsvetkov et al., 2013; Ljubešić et al., 2018).

Going beyond individual word ratings, researchers have placed more emphasis on the perception of concreteness in context. The CONCRETTEXT shared task (Gregori et al., 2020) introduced novel concreteness norms of words in context ratings for 550 Italian and 534 English sentences. More recently, Muraki et al. (2023) collected concreteness ratings for 62,000 English multiword expressions, ranging across different linguistic constructions (e.g., particle verbs, noun compounds, fixed expressions).

Nonetheless, there remains a significant gap in multilingual resources that explore concreteness in context in a more systematic and bottom-up way. The current study addresses this gap through the collection of concreteness ratings for the smallest meaningful units of context, specifically verb-direct object noun expressions.

**Rating Scales & Best-Worst Scaling** Concreteness norms have predominantly relied on rating scale approaches (i.e., Likert scale) for collecting human judgments, which however come with problematic aspects. While humans tend to agree on extremely concrete or abstract targets, the mid-scale range shows higher variance among annotators (Pollock, 2018; Knupleš et al., 2023; Paisios et al., 2023). These inconsistencies raise questions about the reliability of concreteness norms collected via rating scale approaches.

Best-Worst Scaling (BWS) (Louviere and Woodworth, 1991; Louviere et al., 2015) provides a viable alternative approach for collecting more reliable judgments across semantic variables. Unlike traditional rating scales, BWS annotators are presented with four items and asked to make comparative judgments in selecting the most and least representative item for the property of interest. In the context of collecting emotion intensity ratings for words and phrases, Kiritchenko and Mohammad (2017) reported significantly higher reliability values with BWS compared to rating scales, in particular for linguistically complex items, such as phrases. For these reasons, we consider BWS a promising method for collecting multilingual concreteness judgments for verb-object expressions.

### 3 Construction of Verb-Object Targets

We outline the steps to extract and select verb-object (v,o) target expressions for English, German, and Slovene. Objects are restricted to direct-object nouns; selection of verbs and nouns is balanced across frequency and concreteness ranges.

#### 3.1 Extracting Verb-Object Pairs

We begin by extracting sentences from three web corpora. For English, we use the ENCOW-16 corpus (Schäfer, 2015), containing  $\approx 10$  billion words; for German, we use the SdeWaC corpus (Faaß and Eckart, 2013), containing  $\approx 880$  million words; for Slovene we use the CLASSLA.sl corpus (Ljubešić and Kuzman, 2024), containing  $\approx 3$  billion words. From the latter, we discard low-quality text and incorporate missing syntactic dependency information using the CLASSLA-Stanza pipeline (Terčon and Ljubešić, 2023).<sup>2</sup>

We identify verb-direct object noun expressions using syntactic dependency criteria; for German, we rely on the SubCat-Extractor tool (Scheible et al., 2013) to obtain verb subcategorization. We disregard low-frequency and ambiguous combinations by only including words that (a) have one predominant POS ( $> 95\%$ ), (b) occur more than 10,000 times, and (c) are not proper names or pronouns. We then only retain (v,o) that (d) have both components included in the post-filtered words subset, and (e) occur more than 20 times for English and Slovene, and more than 5 for German.<sup>3</sup>

<sup>2</sup><https://pypi.org/project/classla/>

<sup>3</sup>The thresholds were chosen based on manual inspection.

Lang	a <sub>verb</sub>	c <sub>verb</sub>	a <sub>noun</sub>	m <sub>noun</sub>	c <sub>noun</sub>
EN	1.4–2.0	3.3–4.8	1.1–2.0	3.0–4.0	4.9–5.0
DE	1.8–3.1	3.1–4.9	1.3–2.8	2.8–3.7	4.1–5.0
SL	1.5–2.9	2.9–4.8	1.1–2.4	2.5–3.4	3.4–4.9

(a) Concreteness ranges for verb and noun categories.

Lang	Low	Mid	High
EN	20–33	34–71	72–20,406
DE	5–7	8–15	16–2,039
SL	20–34	35–81	82–3,414

(b) Frequency ranges across (v,o) categories.

Table 1: Concreteness and frequency ranges across individual word categories and (v,o) targets.

We obtain 290,514 English, 58,224 German, and 88,096 Slovene (v,o) expressions.

### 3.2 Selection of Target Pairs

To build a representative set of (v,o) targets, we balance the sampling of verbs and nouns across concreteness scores and frequency ranges. From our binned (v,o) candidates we randomly sample 111 pairs for English and Slovene, and 101 pairs for German, with the following characteristics.<sup>4</sup>

1. **Concreteness:** We categorize nouns as *highly abstract* (a), *mid-range* (m), or *highly concrete* (c) based on their concreteness ratings. Verbs we only categorize as *abstract* (a) or *concrete* (c) due to their low coverage. The individual ranges are shown in Table 1a. Note that the smaller availability of targets in the non-English norms in some cases leads to overlapping limits in concreteness ranges. The ratings we use in this study are extracted from various collections of norms. For English, we use the data from Knupleš et al.’s (2023), a balanced subset of 500 nouns and 200 verbs per bin sampled from Brysbaert et al. (2014b). For German, we use a balanced subset of 600 nouns and 140 verbs per bin sampled from Charbonnier and Wartena (2020). For Slovene, we collect new human judgments for 600 nouns and 198 verbs (see Section 4), due to the lack of existing word-level concreteness norms.

2. **Frequency:** We divide the (v,o) targets into three equally sized sets based on their joint

<sup>4</sup>The sample sizes represent the minimum number of available items extracted from the joint bins in each language.

frequencies:<sup>5</sup> *low frequency* (lf), *mid frequency* (mf), and *high frequency* (hf), using their normalized frequencies as measured on the original web corpora (see Section 3.1). The frequency ranges are shown in Table 1b.

To prevent possible over-representation of common high-frequency verbs (e.g., *want* appears in 4% of English (v,o) candidates), we apply weighted sampling using log-transformed inverse frequencies. In total, we obtain 1,998 English, 1,818 German and 1,998 Slovene (v,o) targets.

## 4 Collecting Concreteness Norms with Best-Worst Scaling

To collect concreteness ratings for our target expressions, we follow Kiritchenko and Mohammad (2017): we present participants with four expressions (a 4-tuple) at a time and ask them to select the most concrete and the most abstract targets. We randomly generate  $2N$  distinct 4-tuples (where  $N$  is the number of targets), such that no two targets within a tuple are identical, no two 4-tuples share more than two targets and each target appears in exactly eight different 4-tuples.<sup>6</sup> The final BWS responses are converted to real-valued concreteness scores using a simple *counting procedure* (Orme, 2009), for which the score  $s(i)$  of each item  $i$  is calculated as the normalized difference between the count of most concrete and most abstract selections, cf. Equation (1). We then linearly transform the scores from the original range  $[-1, 1]$  to a scale of 1 (abstract) to 5 (concrete).

$$s(i) = 2\left(\frac{\#best(i) - \#worst(i)}{\#overall(i)}\right) + 3 \quad (1)$$

**Crowd-sourcing Details** We recruit participants using the platform Prolific<sup>7</sup> and Google Forms<sup>8</sup> as the survey tool. Participants have to reside in the United States or United Kingdom for English, in Germany for German, and in Slovenia for Slovene, have matching citizenship, speak the respective language as their first and native language, and have an approval rating of 90–100%.

<sup>5</sup>In order to keep the number of parameters tractable, we do not additionally take the individual verb and object frequencies into account.

<sup>6</sup>We generate the  $2N$  4-tuples using scripts provided by Kiritchenko and Mohammad (2017).

<sup>7</sup><https://www.prolific.com>

<sup>8</sup><https://docs.google.com/forms/>

Lang	Type	# targets	# participants	# judgments	SHR ( $\rho$ )
EN	(v,o)	1,998	200	20,080	0.89
DE	(v,o)	1,818	114	18,180	0.91
SL	(v,o)	1,998	67	20,084	0.92
	n	600	51	5,300	0.94
	v	198	20	1,980	0.89

Table 2: BWS annotations: numbers of targets, participants, concreteness judgments and average split-half reliability (SHR) Spearman  $\rho$  correlations.

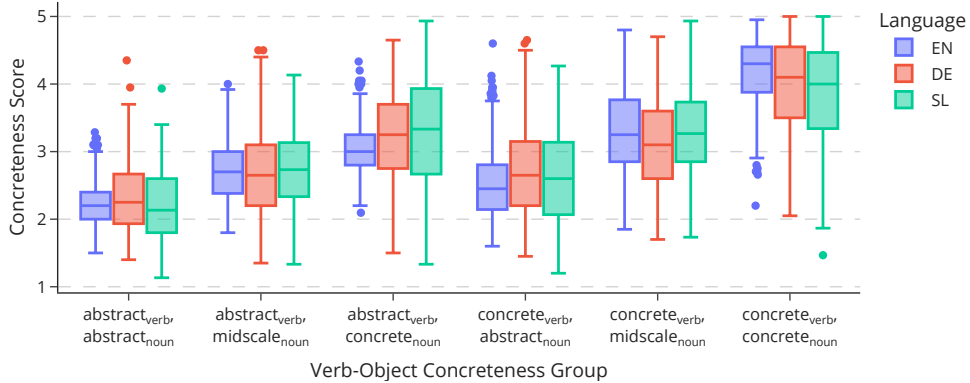


Figure 2: Verb-object BWS-based concreteness scores across concreteness groups and languages.

We create 117 surveys containing  $\approx 100$  of the 4-tuples drawn from the full set of  $2N$  4-tuples. Each survey is assigned to five annotators. With each target expression appearing in eight different 4-tuples across the survey, we collect 40 judgments per (v,o) expression. Following Brysbaert et al. (2014b), participants do not have any training to familiarize with the annotation task. For English and German, participants could only fill out one survey. For Slovene, participants were allowed to complete multiple surveys, due to the smaller participant pool for Slovene.

Each survey incorporates attention-check questions to ensure compliance and to measure the participants’ focus. We place eight manually selected checks evenly in each survey, with the first at the start. With three or more failed checks, we discard the survey responses of the corresponding participant. Participants are explicitly made aware of these attention checks and rejection criteria.

Participants were compensated 5.35€ for each completed survey. The median completion time was around 30 minutes. We recruited a total of 381 participants, who were on average 38.3 years old (18 minimum, 80 maximum). 199 participants identified as female, 181 as male, 1 participant preferred not to say.

**Results** We collect 58,344 judgments across English, German and Slovene for 5,814 (v,o) expressions, with  $\approx 40$  judgments per (v,o) expression (see Table 2 for an overview of our collection).

Figure 2 presents the BWS-based concreteness scores as score distributions across the six concreteness bins and our three target languages. When we inspect the median and inter-quartile ranges (IQR) in the distributions, we observe clear trends across languages. Zooming into the concreteness effects of nouns and verbs, we see that the concreteness of the noun strongly influences the (v,o) joint concreteness rating: irrespective of the concreteness of the verb, combinations with abstract nouns result in lower (v,o) concreteness ratings ( $\text{med}(a, a) = 2.2$ ;  $\text{med}(c, a) = 2.5$ ), while those with concrete nouns result in higher ratings ( $\text{med}(a, c) = 3.2$ ;  $\text{med}(c, c) = 4.2$ ). I.e., verbs play only a minor role, presumably because of their semantically more vague nature.

To assess the reproducibility of concreteness scores across multiple annotators, we compute split-half reliability (SHR) by randomly splitting annotations for each tuple into two groups, calculating separate scores for each group, and measuring Spearman’s rank-order correlation coefficient  $\rho$  (Siegel and Castellan, 1988). We compute SHR over 100 trials and report the average in Table 2.

We observe similarly high correlations across languages and target types, which indicates the reliability of our novel norms.

**Word-level Ratings in Slovene** Preceding the (v,o) collection, to even identify (v,o) expressions satisfying our sampling criteria (see Section 3.2), we collect word-level concreteness ratings for 600 nouns and 198 verbs following the same procedure described above,<sup>9</sup> with separate verb and noun surveys. In total we collect 7,280 word-level judgments produced by 61 participants (average age 29, range [21,63])<sup>10</sup> with 36 identifying as female and 25 as male, and a median completion time of 25 minutes. Once more, the SHR shows very high correlation scores both for nouns and verbs (see Table 2). We further compare our ratings against automatically generated scores by Ljubešić et al. (2018), obtaining a moderate Spearman correlation of  $\rho = 0.74$  ( $p < 0.001$ ) for 720 words. This aligns with prior work comparing human and automatic ratings for German ( $r = 0.83$ ) (Köper and Schulte im Walde, 2016) and Estonian ( $r = 0.71$ ) (Proos and Aigro, 2023).

## 5 English Norms: Comparing Best-Worst Scaling and Ratings on a Scale

To establish the reliability of our approach of using BWS for collecting concreteness ratings and gain additional insights into differences in collection approaches, we conduct a comparative analysis with rating scales (RS). We compare BWS with a 5-point Likert scale by collecting concreteness ratings for a subset of 1,823 English (v,o) targets.

**Crowd-sourcing Details** Participants are recruited through Amazon Mechanical Turk<sup>11</sup>. Each Human Intelligence Task (HIT), i.e., a single unit of work assigned to a worker on the crowdsourcing platform, includes 100 target items, plus 32 control items to ensure data reliability. To filter out low-quality work, we apply three criteria: (1) annotators have to rate at least 4 out of 8 duplicate instances consistently within a distance of 1, (2) a minimum standard deviation is required for control item ratings to avoid uniform responses, and (3) annotators are checked for providing expected ratings on control items.

<sup>9</sup>From our collected word targets, we randomly sample 66 verbs and 200 nouns per *lf*, *mf*, and *hf* frequency bins.

<sup>10</sup>In Table 2 we report number of participants separately.

<sup>11</sup><https://www.mturk.com/>

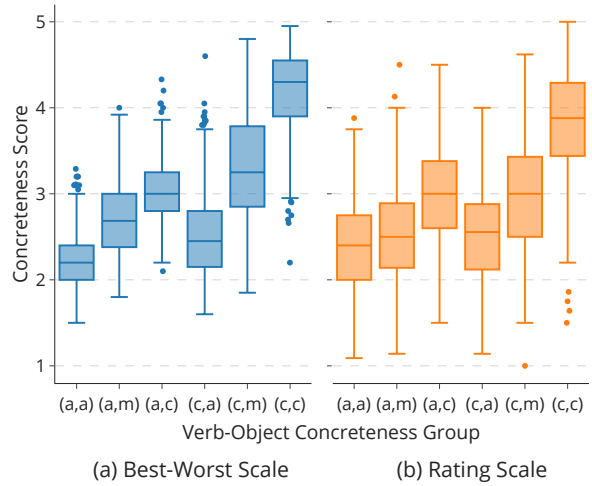


Figure 3: Comparison of concreteness scores for English (v,o) targets between BWS and RS collection approaches.

We redesign the previous BWS survey by instructing annotators to rate concreteness of individual (v,o) expressions on a scale from 1 (abstract) to 5 (concrete). We recruit 363 annotators residing in the United States or United Kingdom.

**Results** We collect a total of 16,817 ratings on a scale, obtaining  $\approx 9$  ratings for each (v,o) expression. We then computed an average rating for each expression. In our comparative analysis, we correlate the ratings derived from BWS and RS and obtain a moderate Spearman correlation of  $\rho = 0.64$  ( $p < 0.05$ ). In Figure 3 we compare the distributions of ratings from both approaches across each of the six concreteness bins. We can clearly see that RS ratings cluster closer to the mid-scale range than BWS-derived ratings. While both approaches show similar overall patterns (i.e., (a, a) (v,o) pairs were perceived rather abstract; (c, c) (v,o) pairs were perceived rather concrete; and all other categories show ratings in between), ratings derived from RS exhibit larger variability in responses across bins: the average standard deviation across concreteness bins is lower for BWS ratings ( $\bar{\sigma} = 0.48$ ) than for RS ratings ( $\bar{\sigma} = 0.58$ ). These observations complement prior work showing that BWS tends to minimize respondent variability (Louviere et al., 2015).

We further compare both approaches by measuring the extent to which the BWS and RS ratings agree or disagree on whether specific (v,o) targets are more or less concrete, by applying a simple pairwise measure: A pair  $(x_i, x_j)$  is agreed upon

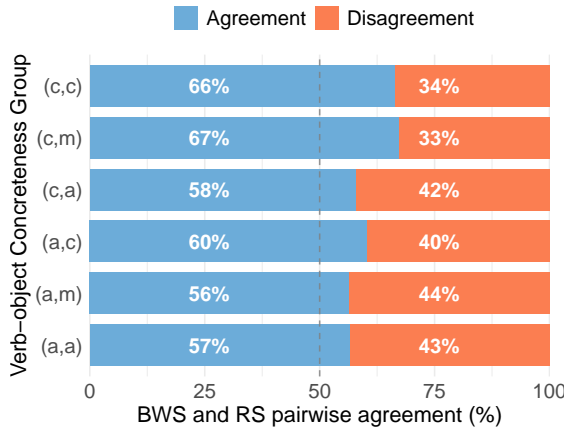


Figure 4: Agreement between BWS and RS.

if the concreteness ratings  $x_i >_{\text{BWS}} x_j \Leftrightarrow x_i >_{\text{RS}} x_j$  where  $i \neq j$ . We then calculate the proportion of agreeing pairs within concreteness bins as our measure of agreement. As shown in Figure 4, we see pairwise agreement in  $> 50\%$  cases across all bins. We also see that the choice between BWS and RS approaches leads to higher disagreement on judgments for abstract targets than for concrete targets, i.e., we find maximum disagreements in  $(a, m) = 44\%$  and  $(a, a) = 43\%$ , in contrast to the least disagreement in  $(c, c) = 34\%$ .

To evaluate the reliability of our RS ratings, we again calculate the average SHR over 100 trials. A Spearman correlation of  $\rho = 0.49$  shows lower reliability in comparison to judgments obtained with BWS ( $\rho = 0.89$ ). These results emphasize the complexity of our annotation task and its relationship to the specific annotation method. Similarly to insights from annotation work on sentiment intensity (Kiritchenko and Mohammad, 2017), the greater complexity and subjectivity in collecting concreteness ratings for (v,o) expressions results in an increased difference between RS and BWS annotation results.

## 6 Automatic Extrapolation of Multilingual Concreteness Norms

Given the complexity and cost of collecting large-scale concreteness norms, we explore prediction methods to expand our human-generated multilingual concreteness norms from 5,814 to over 430,000 (v,o) expressions. We then use the best-performing method to create an automatically generated multilingual corpus of concreteness ratings.

## 6.1 Experimental Setting

**Baseline** We implement the paradigm-based algorithm proposed by Turney and Littman (2003). The method compares a target expression to sets of concrete and abstract paradigm expressions, represented using embeddings. The concreteness score of a (v,o) combination is derived from the cosine similarity between a target expression and concrete paradigm items, minus the similarity with abstract paradigm items. Scores are then rescaled from  $[-1, 1]$  to  $[1, 5]$ .

While Turney and Littman (2003) experimented with Pointwise Mutual Information (PMI) and Latent Semantic Analysis (LSA) representations, i.e., static embeddings that do not capture context-specific representations, we employ contextualized embeddings to accurately capture the lexical-semantic effect the verb and noun exert on each other. Each (v,o) representation is the average of the two word embeddings (i.e., the verb and the noun) given an input (v,o) expression, excluding special tokens. The word embeddings themselves are extracted from the following monolingual encoder models: RoBERTa (Liu et al., 2019) for English, GBERT (Chan et al., 2020) for German, and SloBERTa (Ulčar and Robnik-Šikonja, 2021) for Slovene.

Following prior work (Turney et al., 2011; Köper and Schulte im Walde, 2016), we use a greedy supervised algorithm to select 20 concrete and 20 abstract paradigm expressions. The algorithm incrementally adds one expression at a time to each set, alternating between concrete and abstract, to maximize Spearman correlation with the ratings in the training data.

**Fine-tuning Models** As our second approach, we fine-tune the same monolingual models used in our baseline on a regression task with a mean squared error loss function, focusing on the CLS token. All models are fine-tuned for 10 epochs, on a batch size of 16, and a learning rate of  $3 \times 10^{-5}$  using the Adam optimizer.

**LLM Prompting** Due to recent interest in generating concreteness norms using Large Language Models (LLMs) (Martínez et al., 2025), we evaluate models on predicting concreteness ratings in both RS and BWS settings. Drawing from insights that LLMs performed better in BWS than RS settings for emotion intensity ratings (Bagdon et al., 2024), we assess this setup for concreteness

by presenting models with 4-tuples and prompting them to identify the most concrete and abstract expressions, converting these choices to ratings (as in Section 4). We assess model performance in zero-shot and few-shot scenarios: for RS, six expressions from distinct concreteness bins are provided; for BWS, we use three of our eight attention checks. We provide examples of our prompt formulations in Appendix B.2.

The selection of models is based on the following criteria: (1) being predominantly trained on the target language, i.e., employing monolingual models to limit exposure to cross-lingual influences,<sup>12</sup> and (2) having comparable model sizes within each approach. Accordingly, we use the following models: Falcon3-7B-Instruct (TII Team, 2024) for English, LeoLM-7B<sup>13</sup> for German, and GaMS-9B-Instruct (Vreš et al., 2024) for Slovene. Due to low performances of the monolingual models, we subsequently include the same evaluation of EuroLLM-9B-Instruct (Martins et al., 2025), a comparable multilingual LLM covering all three languages. For a full model overview, see Table 9 in Appendix B.1.

All methods are evaluated using Spearman correlation ( $\rho$ ) and Root Mean Square Error (RMSE) on a 10% held-out test set. For baseline and fine-tuning, we use 10-fold cross validation; for LLM prompting, we perform five runs with random seeds on the dataset. We report average RMSE and correlations using Fisher’s z-transformation.

## 6.2 Results and Extrapolation

The results for predicting (v,o) concreteness ratings are shown in Table 3. Fine-tuning monolingual models performed best across all languages, with the highest  $\rho$  (strong trend capture) and lowest RMSE (accurate score estimation). In contrast, LLMs reached lower RMSE but struggle to capture meaningful patterns (low  $\rho$ ). RS predictions yield higher  $\rho$  across languages compared to BWS, but still often below the baseline. Few-shot learning provided significant gains for RS, while not for BWS, thus highlighting the advantage of BWS in cases of no existing gold standard. Multilingual LLM exhibited lower  $\rho$  than monolingual models. On average, 4% of LLM responses failed to provide a valid output and were excluded.

<sup>12</sup>We acknowledge that most models, particularly LLMs developed for low-resource languages, include a smaller subset of English pre-training data.

<sup>13</sup><https://laion.ai/blog/leo-lm/>

		EN		DE		SL	
		$\rho$	RMSE	$\rho$	RMSE	$\rho$	RMSE
<b>Baseline</b>		0.75	3.05	0.77	3.04	0.49	3.07
<b>Fine-tuning</b>		<b>0.88</b>	<b>0.38</b>	<b>0.87</b>	<b>0.45</b>	<b>0.82</b>	<b>0.54</b>
<b>LLM Prompt</b>							
Mono.	→ RS; 0-shot	0.56	0.88	<i>-0.02</i>	1.56	0.30	1.00
	→ RS; F-shot	0.71	0.70	0.10	1.39	0.54	0.85
	→ BWS; 0-shot	0.75	0.57	<i>0.02</i>	0.96	0.51	0.76
	→ BWS; F-shot	0.65	0.62	<i>0.03</i>	0.96	0.52	0.75
Multi.	→ RS; 0-shot	0.08	1.44	0.05	1.43	0.13	1.25
	→ RS; F-shot	0.27	1.14	0.25	1.19	0.24	1.16
	→ BWS; 0-shot	0.17	0.86	<i>0.17</i>	0.89	0.09	0.95
	→ BWS; F-shot	<i>0.01</i>	0.93	<i>-0.01</i>	0.97	0.02	0.99

Table 3: Prediction results with RMSE and correlations ( $p < 0.05$ ), with  $\uparrow \rho$  and  $\downarrow$  RMSE implying better prediction results. LLM predictions are using a rating scale (RS) or Best-Worst Scaling (BWS) in zero-shot (0-shot) or few-shot (F-shot) example settings. **Bold** text designates best performance and *italic* statistical insignificance.

To investigate the underperforming LLM results, we analyze the mismatches between our concreteness norms and LLM-generated ratings. Starting with a quantitative view, we visualize rating distributions from both sources in Figure 5 across the three languages, showing ratings from the highest-performing models in each experimental setting. Comparing both prompting approaches in the LLM results section, we generally observe skewed distributions in rating values when using RS (top section) in contrast to distributions from BWS that resemble the shape of normal distributions (bottom section). When prompted with RS, the models indicate strong preferences to specific scores (e.g., the Slovene model with zero-shot in 90% cases using ratings 2, 3, or 4) or generate skewed distributions (e.g., the German model’s high occurrence of rating 1). These tendencies correspond to previously observed lower evaluation results in Table 3, i.e., leading to cases of negative or statistically insignificant correlations. Providing few-shot examples to the multilingual model produces a distribution more similar to the human norms.

In contrast, predicting with BWS shows closer alignment with our norms, especially when evaluating monolingual models. However, for cases of lower correlations in Table 3, the distributions show a lower spread and stronger concentration in the center, likely due to conflicting

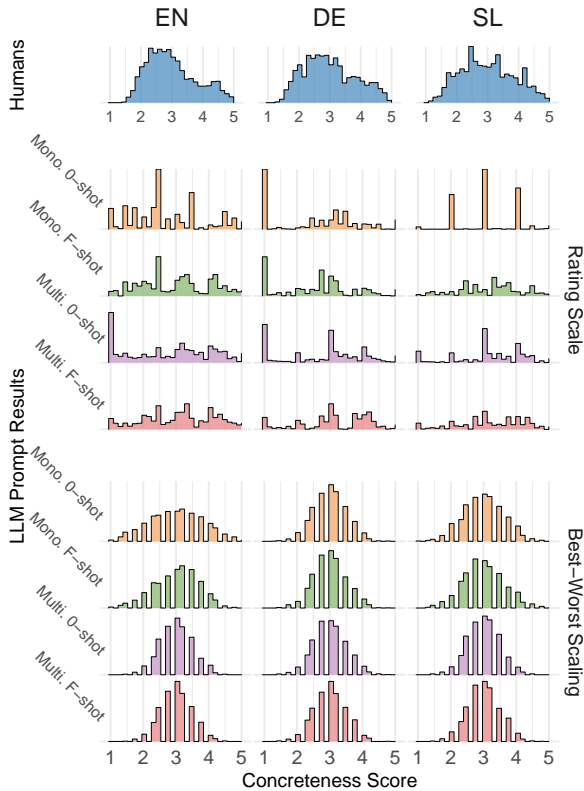


Figure 5: Distributional comparison of human and LLM-prompted concreteness ratings across experimental settings. The presented scores were produced in the highest performing model run.

BWS judgments that regress ratings towards middle values after applying the counting procedure, cf. Equation (1). Multilingual LLMs exhibit lower variance with sparse extremely abstract (1) or extremely concrete (5) ratings, thus explaining the monolingual LLM’s  $\uparrow \rho$  and  $\downarrow$  RMSE scores. Overall, these insights complement prior work on LLM-prompted BWS outperforming RS when evaluated with BWS-collected norms (Bagdon et al., 2024).

Continuing with a qualitative view in Table 4, we present the five strongest mismatches in terms of their  $\Delta$  between human (H) and LLM (L) ratings in BWS and RS settings, by selecting from the highest-performing monolingual 0-shot and monolingual F-shot models, respectively. Across languages, we observe that the  $\Delta$  values are consistently higher when using RS in comparison to predicting with BWS; some RS-based differences even reach  $\Delta > 3$ . Across approaches and languages, we find both abstract and concrete expressions among the strongest human–LLM mismatches. These disagreements reflect the LLMs’

misunderstandings of concepts across the concreteness spectrum that may be reflective of variations in connotative meaning.

Based on these results, we select the top-performing fine-tuned models to automatically generate concreteness ratings for all the 288,516 English, 56,406 German, and 86,340 Slovene (v,o) combinations<sup>14</sup> described in Section 3.1.

## 7 Figurative Meanings of Expressions

The availability of reliable, large-scale concreteness ratings across multiple languages and beyond words in isolation could be a game-changer for various NLP tasks involving figurative language detection, which has heavily made use of existing concreteness norms in the past (Turney et al., 2011; Tsvetkov et al., 2014; Köper and Schulte im Walde, 2016; Alnafesah et al., 2020; Maudslay et al., 2020; Piccirilli and Schulte im Walde, 2022a,b; Hülsing and Schulte im Walde, 2024, i.a.): according to Conceptual Metaphor Theory (CMT), metaphors involve a shift from concrete to abstract meanings (Lakoff and Johnson, 1980), so precise concreteness ratings support models to differentiate between literal and figurative language. Moreover, concreteness plays an important role in annotation frameworks, as a key criterion for distinguishing figurative in contrast to basic word meanings (Pragglejaz Group, 2007; Shutova and Teufel, 2010; Schulte im Walde et al., 2018).

In this last crowd-sourcing annotation study, we take a subset of the (v,o) targets and ask participants (1) to judge whether the (v,o) expression is literal or figurative (based on whether the expression meaning is clearly based on the meanings of its individual words) and (2) to provide an example sentence including the (v,o) expression. With this setup, we not only collect figurative judgments for our expressions but also collect use sentences containing them, thereby creating a valuable multilingual resource for the research community.

**Target Selection** For each language we randomly sample 100 (v,o) targets per concreteness bin from their respective interquartile range (see Figure 2). This results in 600 targets per language, or a total of 1,800 (v,o) expressions.

**Crowd-sourcing Details** Following the annotation setup and selection criteria described in Section 4, each participant is given a survey with 50

<sup>14</sup>Excluding combinations with human-generated ratings.

Language	(v,o) expression	H	L	$\Delta$
EN	hold liberty	2.1	4.3	2.1
	write ability	3.8	1.8	2.0
	qualify car	2.7	4.5	1.9
	preach faith	3.1	1.3	1.9
	write fantasy	3.8	2.0	1.8
DE	<i>Mund verschließen</i> 'close mouth'	4.8	1.8	3.0
	<i>Dokument schreiben</i> 'write document'	4.7	1.8	3.0
	<i>Widerstand unterschätzen</i> 'underestimate resistance'	1.8	4.5	2.8
	<i>Entwicklung verfolgen</i> 'track development'	2.0	4.8	2.8
	<i>Holz schneiden</i> 'cut wood'	5.0	2.3	2.8
SL	<i>pisati zgodbo</i> 'write story'	4.7	2.0	2.7
	<i>zaživeti življenje</i> 'live life'	1.7	4.3	2.6
	<i>imeti priključek</i> 'have connection'	4.0	1.5	2.5
	<i>pripraviti jajca</i> 'prepare eggs'	4.7	2.3	2.4
	<i>piti živce</i> 'drink nerves'	1.5	3.8	2.3

(a) Best-Worst Scaling

Language	(v,o) expression	H	L	$\Delta$
EN	develop acne	4.1	1.0	3.1
	prohibit student	3.4	1.0	2.4
	behold miracle	2.2	4.2	2.1
	eat word	3.0	5.0	2.1
	rebuild civilization	3.1	1.0	2.1
DE	<i>Schraube lösen</i> 'loosen screw'	4.9	1.0	3.9
	<i>Ring tragen</i> 'wear ring'	4.9	1.0	3.9
	<i>Mann schlagen</i> 'hit man'	4.9	1.0	3.9
	<i>Blut transportieren</i> 'transport blood'	4.8	1.0	3.8
	<i>Elefant töten</i> 'kill elephant'	4.8	1.0	3.8
SL	<i>piti živce</i> 'drink nerves'	1.5	4.5	3.0
	<i>odpreti dušo</i> 'open soul'	1.3	4.2	2.9
	<i>ponujati spekter</i> 'offer spectrum'	1.6	4.5	2.9
	<i>spremeniti predstavo</i> 'change performance'	1.9	4.8	2.9
	<i>spomniti izkušnja</i> 'remind experience'	1.7	4.5	2.8

(b) Rating Scale

Table 4: Top 5 expressions with the largest differences between human ratings (**H**) and LLM scores (**L**).

targets, plus an additional even distribution of five attention check questions. We present one target per page, including a task reminder, and assign five annotators per survey. See details of survey guidelines and examples in Appendix A.2.

**Results** Across our three languages we obtain a total of 9,000 judgments and 9,000 example sentences that include the 1,800 (v,o) expressions. The average length from the obtained example sentences is  $\approx 8$  words. To investigate the interplay between concreteness and figurative usage of our targets, Figure 6 visualizes the number of (v,o) expressions judged by the majority of the participants as “figurative” (in blue), “literal” (in red) or “unclear” (in yellow), where 3 vs. 2 we consider a not clear majority. The graphic highlights a clear relationship between the concreteness of the (v,o) pairs and their classification as figurative vs. literal language. Across our set of languages, more abstract expressions are judged more figuratively, while more concrete expressions are judged more literally, i.e., for concreteness we observe an increase for  $(a, a) < (a, m) < (a, c)$

and for  $(c, a) < (c, m) < (c, c)$ , which goes in parallel with an increase in figurativeness (degree of figurative language use). To provide an even clearer picture, Figure 7 shows the distributions of concreteness ratings and majority-assigned figurative labels across concreteness groups for English, German, and Slovene. Additionally in Table 5, we provide examples of (v,o) targets, their majority-assigned label, and one example sentence.

## 8 Conclusion

In this work, we investigated the perception of concreteness in context by zooming in on one smallest meaningful unit beyond word level, through a systematic collection of concreteness ratings for 5,814 verb-direct object noun expressions. To compare patterns across languages, we targeted the collection of ratings for three languages with varying degrees of resource availability, namely English, German, and Slovene. Due to significant costs associated with large-scale collections, we further automatically extended our multilingual concreteness ratings to over 430,000 expressions using advanced prediction methods.

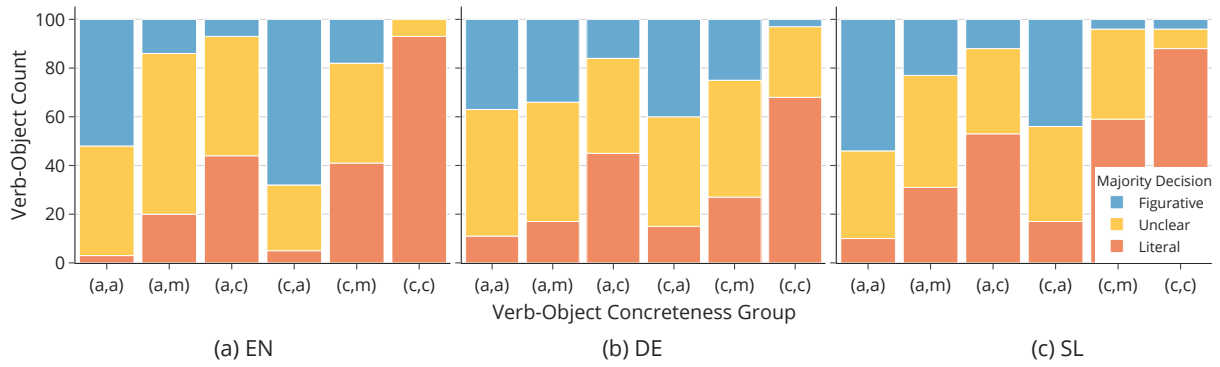


Figure 6: Figurative majority-assigned labels for (v,o) expressions across concreteness groups.

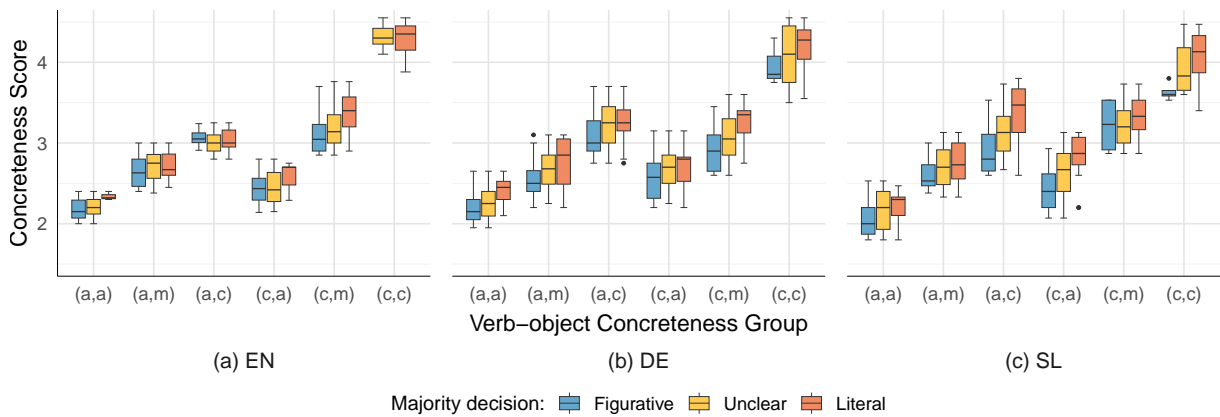


Figure 7: Distribution of concreteness ratings and majority-assigned figurative labels.

A noteworthy byproduct was the creation of concreteness norms for 798 Slovene words (verbs and nouns), addressing a resource gap in existing concreteness ratings.

Adopting a Best-Worst Scaling (BWS) approach, we reported similar trends in concreteness distributions across our three languages. Focusing on the functions of words in verb-object combinations, we found that the concreteness of the overall (v,o) expression was heavily affected by the concreteness of its nominal object. In a comparative analysis on a subset of English norms between BWS and ratings on a scale (RS), we found lower variability and higher reliability using BWS ratings, thus further establishing the reliability of our annotations.

Lastly, we examined the relationship between concreteness and figurative language by collecting literal vs. figurative judgments for 1,800 expressions across our three languages, along with 9,000 example sentences demonstrating their liter-

al/figurative usage. Our findings revealed that abstract verb-object combinations were significantly more likely to be perceived as figurative language, while combinations judged as concrete are predominantly interpreted as literal language. Our resources and insights support future research on concreteness and figurative language, with potential downstream applications in figurative language processing as well as further natural language understanding tasks.

Lang	(v,o) expression	Maj. decision	Example sentence
EN	<i>undermine religion</i>	(a, a) figurative	<i>Dogma only serves to undermine religion rather than uplift it.</i>
	<i>maximize energy</i>	(a, m) unclear	<i>With a fitness program he was able to maximize his energy.</i>
	<i>convince student</i>	(a, c) literal	<i>The teacher tried to convince the student to study harder.</i>
	<i>carry uncertainty</i>	(c, a) figurative	<i>She carried uncertainty when faced with a difficult problem.</i>
	<i>transmit pain</i>	(c, m) unclear	<i>The nerves in the body transmit pain.</i>
	<i>buy oil</i>	(c, c) literal	<i>I went to the supermarket to buy some oil.</i>
DE	<i>Wirklichkeit abbilden</i> "mirror reality"	(a, a) figurative	<i>Der Bericht konnte die Wirklichkeit nicht abbilden.</i> "The report could not mirror reality."
	<i>Bewegung bewirken</i> "cause movement"	(a, m) unclear	<i>Die Stabilisation half ihm nun größere Bewegungen im Arm zu bewirken.</i> "The stabilization now helped him to achieve larger movements in his arm."
	<i>Karte akzeptieren</i> "accept card"	(a, c) literal	<i>In diesem Geschäft wird jede Karte akzeptiert.</i> "Every card is accepted in this store."
	<i>Regelung tragen</i> "carry regulation"	(c, a) figurative	<i>Die Führungsebene sollte zuerst die Regelung tragen.</i> "The management level should first implement the regulation."
	<i>Kilometer schaffen</i> "accomplish kilometers"	(c, m) unclear	<i>Ich möchte heute mindestens 5 Kilometer schaffen.</i> "I want to manage at least 5 kilometers today."
	<i>Pflanzen sammeln</i> "collect plants"	(c, c) literal	<i>Wir sammeln Pflanzen für Tees und getrocknete Gewürze.</i> "We collect plants for teas and dried spices."
SL	<i>obogatiti dejavnost</i> "enrich activity"	(a, a) figurative	<i>Za privabljanje večjega števila turistov so se odločili obogatiti svojo dejavnost.</i> "To attract a greater number of tourists, they decided to enrich their activity."
	<i>odkriti primer</i> "discover case"	(a, m) unclear	<i>Policisti so odkrili primer kraje v bližnjem mestu in začeli preiskavo.</i> "Police officers discovered a case of theft in a nearby town and launched an investigation."
	<i>nasloviti pismo</i> "address letter"	(a, c) literal	<i>Pismo je naslovila na svojo prijateljico.</i> "She addressed the letter to her friend."
	<i>zavirati proces</i> "slow down (brake) process"	(c, a) figurative	<i>Kreme lahko zavirajo proces staranja, ne morejo ga pa ustaviti.</i> "Creams can slow down the aging process, but they cannot stop it."
	<i>odpreti polčas</i> "open half (time)"	(c, m) unclear	<i>Sodniki so odprli polčas za drugi del tekme.</i> "The referees opened the half for the second part of the match."
	<i>pisati skladbo</i> "write composition"	(c, c) literal	<i>Zaprł se je v sobo s klavirjem in papirjem in začel pisati novo skladbo.</i> "He locked himself in a room with a piano and paper and began to write a new composition."

Table 5: Examples of (v,o) expressions with majority-assigned labels and sampled example sentences.

## 9 Limitations

In this paper, we acknowledge several limitations that may impact the interpretation of our findings. In our collections of concreteness ratings, due to a relatively smaller participant pool for Slovenes, we allowed participants to participate in multiple studies. This participant overlap could potentially skew the data, as individual biases may be amplified when respondents evaluate multiple items.

Moreover, our sampling strategy involved utilizing three distinct large web corpora sourced from differing scopes. While all these corpora sample sentences from web sources, differences in the projects may introduce latent biases in our collection. Furthermore, web-sourced corpora may include unconventional or nonstandard verb-

object combinations that deviate from typical native usage, introducing variability into our collection. However across all languages, this study employs (among) the biggest general-language corpora, sampling pairs across a frequency spectrum representative of general language use.

During our evaluation of various prediction methods for automatically generating an extended multilingual corpus of concreteness ratings, we evaluate the performance of large language models of varying sizes. These disparities in model size influence the effectiveness and performance of concreteness prediction, potentially confounding our results. Nonetheless, we aimed to select the nearest appropriate models in terms of size and language coverage. Lastly, due to prompt brittleness in large language models, the variability in model responses to slight changes in input prompts may lead to inconsistencies, which could affect reliability and reproducibility of predictions.

## 10 Ethical Considerations

In the context of our annotation task, we collected concreteness and figurative language judgments using crowd-sourcing platforms. All crowd-workers were fairly compensated under the recommended guidelines of these platforms, and their participation was entirely voluntary. We did not collect any information that can link the data back to the participants. Before completing a survey, crowd-workers were informed that their responses would be used in a scientific publication. While we did reject annotations from workers who failed to meet the required attention checks, we did explicitly warn them about not getting paid if they fail the checks. We engaged in direct conversations with participants who reached out to clarify issues. We took the time to respond to all requests and made our decision-making transparent.

## 11 Acknowledgements

This research was supported by the DFG Research Grants SCHU 2580-4 *Multimodal Dimensions and Computational Applications of Abstractness* and FR 2829-8/SCHU 2580-7 *MeTRa-pher: Learning to Translate Metaphors*.

We also thank the reviewers and the action editor for their useful feedback on a previous version of this article, and our student researcher Sven Naber for supporting us in the collection of the English RS ratings.

## References

- Ghadi Alnafesah, Harish Tayyar Madabushi, and Mark Lee. 2020. [Augmenting neural metaphor detection with concreteness](#). In *Proceedings of the 2nd Workshop on Figurative Language Processing*, pages 204–210, Seattle, Washington (online).
- Christopher Bagdon, Prathamesh Karmalkar, Harsha Gurulingappa, and Roman Klinger. 2024. [“You are an expert annotator”](#): Automatic best-worst-scaling annotations for emotion intensity modeling. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 7924–7936, Mexico City, Mexico.
- 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, Michaël Stevens, Simon De Deyne, Wouter Voorspoels, and Gert Storms. 2014a. [Norms of age of acquisition and concreteness for 30,000 Dutch words](#). *Acta Psychologica*, 150:80–84.
- Marc Brysbaert, Amy Beth Warriner, and Victor Kuperman. 2014b. [Concreteness ratings for 40 thousand generally known English word lemmas](#). *Behavior Research Methods*, 64:904–911.
- Angelo Cangelosi and Francesca Stramandinoli. 2018. [A review of abstract concept learning in embodied agents and robots](#). *Philosophical Transactions of the Royal Society B: Biological Sciences*, 373(1752):20170131.
- Branden Chan, Stefan Schweter, and Timo Möller. 2020. [German’s next language model](#). In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 6788–6796, Barcelona, Spain (Online). International Committee on Computational Linguistics.
- Jean Charbonnier and Christian Wartena. 2020. [Predicting the concreteness of German words](#). In *Proceedings of the 5th Swiss Text Analytics Conference (SwissText) & 16th Conference on Natural Language Processing (KONVENS), CEUR Workshop Proceedings Vol. 2624*.
- Javier Conde, Gonzalo Martinez, Maria Grandury, Carlos Arriaga, Juan Haro, Sascha Schroeder, Florian Hintz, Pedro Reviriego, and Marc Brysbaert. 2026. [Updating the German psycholinguistic word toolbox with AI-generated estimates of concreteness, valence, arousal, age of acquisition, and familiarity](#). *Journal of Cognition*, 9(1):1–25.
- Gertrud Faaß and Kerstin Eckart. 2013. [SdeWaC – A corpus of parsable sentences from the web](#). In *Language Processing and Knowledge in the Web*, pages 61–68, Berlin, Heidelberg. Springer Berlin Heidelberg.
- Diego Frassinelli and Sabine Schulte im Walde. 2019. [Distributional interaction of concreteness and abstractness in verb–noun subcategorisation](#). In *Proceedings of the 13th International Conference on Computational Semantics*, pages 38–43, Gothenburg, Sweden.
- Lorenzo Gregori, Maria Montefinese, Daniele P Radicioni, Andrea Amelio Ravelli, and Rossella Varvara. 2020. [CONcreTEXT @ EVALITA2020: The concreteness in context task](#). In *Proceedings of the 7th evaluation campaign of Natural Language Processing and Speech tools for Italian (EVALITA 2020)*.
- Anna Hülsing and Sabine Schulte im Walde. 2024. [Cross-lingual metaphor detection for low- to high-resource languages](#). In *Proceedings of the 4th Workshop on Figurative Language Processing*, pages 22–34, Mexico City, Mexico.
- Brian Ichter, Anthony Brohan, Yevgen Chebotar, Chelsea Finn, Karol Hausman, Alexander Herzog, Daniel Ho, Julian Ibarz, Alex Irpan, Eric Jang, Ryan Julian, Dmitry Kalashnikov, Sergey Levine, Yao Lu, Carolina Parada, Kanishka Rao, Pierre Sermanet, Alexander T Toshev, Vincent Vanhoucke, Fei Xia, Ted Xiao, Peng Xu, Mengyuan Yan, Noah Brown, Michael Ahn, Omar Cortes, Nicolas Sievers, Clayton Tan, Sichun Xu, Diego Reyes, Jarek Rettinghouse, Jornell Quiambao, Peter Pastor, Linda Luu, Kuang-Huei Lee, Yuheng Kuang, Sally Jesmonth, Nikhil J. Joshi, Kyle Jeffrey, Rosario Jauregui Ruano, Jasmine Hsu, Keerthana Gopalakrishnan, Byron David, Andy

- Zeng, and Chuyuan Kelly Fu. 2023. Do as I can, not as I say: Grounding language in robotic affordances. In *Proceedings of The 6th Conference on Robot Learning*, volume 205 of *Proceedings of Machine Learning Research*, pages 287–318.
- 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. *Quarterly Journal of Experimental Psychology (2006)*, 68(8):1457–1468.
- Mohammed Abdul Khaliq, Diego Frassinelli, and Sabine Schulte im Walde. 2024. Comparison of image generation models for abstract and concrete event descriptions. In *Proceedings of the 4th Workshop on Figurative Language Processing*, pages 15–21, Mexico City, Mexico.
- Svetlana Kiritchenko and Saif 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 (Volume 2: Short Papers)*, pages 465–470, Vancouver, Canada. Association for Computational Linguistics.
- Urban Knupleš, Diego Frassinelli, and Sabine 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 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 70–86, Singapore. Association for Computational Linguistics.
- Maximilian Köper and Sabine Schulte im Walde. 2016. 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. 2016. 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.
- Olaf Lahl, Anja S. Göritz, Reinhard Pietrowsky, and Jessica Rosenberg. 2009. Using the World-Wide Web to obtain large-scale word norms: 190,212 ratings on a set of 2,654 German nouns. *Behavior Research Methods*, 41(1):13–19.
- George Lakoff and Mark Johnson. 1980. *Metaphors We Live By*. University of Chicago Press.
- Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.
- Nikola Ljubešić, Darja Fišer, and Anita Peti-Stantić. 2018. Predicting concreteness and imageability of words within and across languages via word embeddings. In *Proceedings of the Third Workshop on Representation Learning for NLP*, pages 217–222, Melbourne, Australia.
- Nikola Ljubešić and Taja Kuzman. 2024. CLASSLA-web: Comparable web corpora of South Slavic languages enriched with linguistic and genre annotation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation*, pages 3271–3282, Torino, Italia.
- Jordan J. Louviere, Terry N. Flynn, and A. A. J. Marley. 2015. *Best-Worst Scaling: Theory, Methods and Applications*. Cambridge University Press.
- Jordan J. Louviere and George G Woodworth. 1991. Best-worst scaling: A model for the largest difference judgments. Technical report.
- Paweł Mandera, Emmanuel Keuleers, and Marc Brysbaert. 2015. How useful are corpus-based methods for extrapolating psycholinguistic variables? *Quarterly Journal of Experimental Psychology*, 68(8):1623–1642.
- Gonzalo Martínez, Juan Diego Molero, Sandra González, Javier Conde, Marc Brysbaert, and Pedro Reviriego. 2025. Using large language

- models to estimate features of multi-word expressions: Concreteness, valence, arousal. *Behavior Research Methods*, 57(1):1–11.
- Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G.C. de Souza, Alexandra Birch, and André F.T. Martins. 2025. [EuroLLM: Multilingual language models for Europe](#). *Procedia Computer Science*, 255:53–62.
- 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).
- Maria Montefinese, Ettore Ambrosini, Beth Fairfield, and Nicola Mammarella. 2014. [The adaptation of the affective norms for English words \(ANEW\) for Italian](#). *Behavior Research Methods*, 46(3):887–903.
- Maria Montefinese, Lorenzo Gregori, Andrea Amelio Ravelli, Rossella Varvara, and Daniele Paolo Radicioni. 2023. [CONcreTEXT norms: Concreteness ratings for Italian and English words in context](#). *PLOS ONE*, 18(10):e0293031.
- Emiko J. Muraki, Summer Abdalla, Marc Brysbaert, and Penny M. Pexman. 2023. [Concreteness ratings for 62,000 English multi-word expressions](#). *Behavior Research Methods*, 55(5):2522–2531.
- Bryan Orme. 2009. Maxdiff analysis: Simple counting, individual-level logit, and HB. *Sawtooth Software, Inc.*
- Dmitri Paisios, Nathalie Huet, and Elodie Labeye. 2023. Addressing the elephant in the middle: Implications of the midscale disagreement problem through the lens of body-object interaction ratings. *Collabra: Psychology*, 9(1).
- Anita Peti-Stantić, Maja Anđel, Vedrana Gnjidić, Gordana Keresteš, Nikola Ljubešić, Irina Masnikosa, Mirjana Tonković, Jelena Tušek, Jana Willer-Gold, and Mateusz-Milan Stanojević. 2021. [The Croatian psycholinguistic database: Estimates for 6000 nouns, verbs, adjectives and adverbs](#). *Behavior Research Methods*, 53(4):1799–1816.
- Prisca Piccirilli and Sabine Schulte im Walde. 2022a. [Features of perceived metaphoricality on the discourse level: Abstractness and emotionality](#). In *Proceedings of the 13th International Conference on Language Resources and Evaluation*, Marseille, France.
- Prisca Piccirilli and Sabine Schulte im Walde. 2022b. [What Drives the use of metaphorical language? Negative insights from abstractness, affect, discourse coherence and contextualized word representations](#). In *Proceedings of the 11th Joint Conference on Lexical and Computational Semantics*, pages 299–310, Seattle, Washington, USA.
- Lewis Pollock. 2018. [Statistical and methodological problems with concreteness and other semantic variables: A list memory experiment case study](#). *Behavior Research Methods*, 50(3):1198–1216.
- Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1):1–39.
- Mariann Proos and Mari Aigro. 2023. [Concreteness ratings for 36,000 Estonian words](#). *Behavior Research Methods*.
- Nadia Rasheed, Shamsudin H.M. Amin, Umbrin Sultana, Abdul Rauf Bhatti, and Mamoona N. Asghar. 2018. [Extension of grounding mechanism for abstract words: Computational methods insights](#). *Artificial Intelligence Review*, 50(3):467–494.
- 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. Institut für Deutsche Sprache.
- Silke Scheible, Sabine Schulte im Walde, Marion Weller, and Max Kisselew. 2013. [A compact but linguistically detailed database for German verb subcategorisation relying on dependency parses from web corpora: Tool, guidelines and resource](#). In *Proceedings of the 8th Web as Corpus Workshop*, Lancaster, UK.

- Sabine Schulte im Walde, Maximilian Köper, and Sylvia Springorum. 2018. Assessing meaning components in German complex verbs: A collection of source–target domains and directionality. In *Proceedings of the 7th Joint Conference on Lexical and Computational Semantics*, pages 22–32, New Orleans, LA, USA.
- Ekaterina Shutova and Simone Teufel. 2010. Metaphor corpus annotated for source–target domain mappings. In *Proceedings of the 7th International Conference on Language Resources and Evaluation*, pages 3255–3261, Valletta, Malta.
- Sidney Siegel and N. John Castellan. 1988. *Non-parametric Statistics for the Behavioral Sciences*. McGraw-Hill, Boston, MA, USA.
- Luka Terčon and Nikola Ljubešić. 2023. CLASSLA-Stanza: The next step for linguistic processing of South Slavic languages. ArXiv preprint arXiv:2308.04255.
- TII Team. 2024. The falcon 3 family of open models.
- 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.
- Yulia Tsvetkov, Elena Mukomel, and Anatole Gershman. 2013. Cross-lingual metaphor detection using common semantic features. In *Proceedings of the First Workshop on Metaphor in NLP*, pages 45–51, Atlanta, Georgia.
- Peter Turney, Yair Neuman, Dan Assaf, and Yohai Cohen. 2011. Literal and metaphorical sense identification through concrete and abstract context. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 680–690, Edinburgh, Scotland, UK. Association for Computational Linguistics.
- Peter D. Turney and Michael L. Littman. 2003. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems*, 21(4):315–346.
- Matej Ulčar and Marko Robnik-Šikonja. 2021. SloBERTa: Slovene monolingual large pre-trained masked language model. *Proceedings of Data Mining and Data Warehousing, SiKDD*, pages 17–20.
- Domen Vreš, Martin Božič, Aljaž Potočnik, Tomaž Martinčič, and Marko Robnik Šikonja. 2024. Generative model for less-resourced language with 1 billion parameters. In *Language Technologies and Digital Humanities Conference*, page 485–511.
- Zhao Yao, Jia Wu, Yanyan Zhang, and Zhenhong Wang. 2017. Norms of valence, arousal, concreteness, familiarity, imageability, and context availability for 1,100 Chinese words. *Behavior Research Methods*, 49(4):1374–1385.

## A Details on Annotations

### A.1 Annotations using Best-Worst Scaling

**Survey** Figure 8 presents an example survey question. It asks participants to choose the *most concrete* and *most abstract* expression from sets of four verb-object noun expressions. We also provide a brief reminder of the distinction between abstract and concrete expressions above the question, and include an optional comment box below.

**Task reminder:**  
 A **concrete expression** refers to an event with which you can have an immediate experience through your senses (smelling, tasting, touching, hearing, seeing) and by performing or witnessing the event. An example of a concrete expression is *drive car*.

An **abstract expression** refers to an event or process you cannot experience directly through your senses or by performing the event. An example of an abstract expression is *deliver justice*.

Compare the four expressions below and select 1) the **most concrete** and then 2) \* the **most abstract** expression:

	sell hamburger	evaluate text	demolish notion	expose program
most concrete	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
most abstract	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Any comments?

Your answer \_\_\_\_\_

Figure 8: Example of English survey for collecting concreteness judgments using Best-Worst Scaling.

### Quality Checks for Verb-Object Expressions

Each survey incorporates attention-check questions to ensure compliance and measure participants’ focus. We place 8 attention checks evenly in each survey, with the first at the start after instructions. All attention check items listed in Table 6 were manually selected by the authors. With three or more failed checks<sup>15</sup>, we discard the survey responses of the corresponding participant. Participants are made aware of the attention checks and the rejection criteria.

### Quality Checks for Slovene Words

We apply the same approach described above for Slovene words. Attention check items are listed in Table 7. The same three-failure threshold applies.

<sup>15</sup>We changed our threshold from two to three after seeing that participants failing exactly two checks usually gave quality work, with failures mostly at the end.

Item 1	Item 2	Item 3	Item 4
throw ball	hold office	write thing	undermine principle
buy car	evaluate paper	assume office	justify existence
serve beer	want cookie	decide election	defy logic
read magazine	affect neck	develop telescope	unlock creativity
kill bird	evaluate paper	hold office	convey meaning
carry bag	tolerate salt	develop blood	ignite curiosity
mow lawn	vary text	know lake	generalize concept
sell hamburger	write thing	evaluate paper	develop strategy

(a) English

Item 1	Item 2	Item 3	Item 4
Ball werfen	Ansicht ablehnen	Karte akzeptieren	Grundsatz verstehen
Auto kaufen	Papier kritisieren	Amt einnehmen	Existenz nachweisen
Bier trinken	Kontakt herstellen	Wahl organisieren	Logik akzeptieren
Arzt rufen	Herz überprüfen	Zeit starten	Regelung zusammenfassen
Fisch töten	Papier studieren	Amt ablehnen	Bedeutung vermitteln
Bild tragen	Blut untersuchen	Trennung vorschlagen	Wirkung verursachen
Buch schreiben	Augenblick wählen	Spiel dominieren	Begriff definieren
Computer verkaufen	Wald schützen	Arme stärken	Strategie entwickeln

(b) German

Item 1	Item 2	Item 3	Item 4
metati žogo	opravljati funkcijo	napisati stvar	spremeniti idejo
kupiti avto	oceniti članek	prevzeti službo	ugotavljati obstoj
postreči pivo	izboljševati sistem	odločiti volitve	nasprotovati logiki
pogledati revijo	razvijati teleskop	objaviti prevzem	sprostiti ustvarjalnost
ubiti ptiča	imeti račun	pripraviti koncert	preučiti pomen
nositi torbo	izogibati soli	razviti kri	vzbuditi radovednost
kositi travnik	spreminjati besedilo	poznati jezero	sprejemati pravico
prodati hamburger	najti vrsto	zastopati šolo	razviti strategijo

(c) Slovene

Table 6: Attention checks for verb-object expression using Best-Worst Scaling, with highlighted **most concrete** and **most abstract** targets.

Item 1	Item 2	Item 3	Item 4
kobilica	človeštvo	kraja	hvaležnost
limonada	plača	jesen	neskončnost
srajca	obiskovalec	vljudnost	teorija
džip	gimnastika	oblast	ugled
mrož	predstavitev	bolezen	blaženost
oče	termin	ion	simbolika
sladkor	industrialka	rezervacija	dvoumnost
drevo	omega	ocenitev	slava

(a) Nouns

Item 1	Item 2	Item 3	Item 4
jesti	aktivirati	komentirati	razveljavljati
pisati	kisati	oznanjati	sanjati
peti	združevati	privlačiti	racionalizirati
ubijati	dekodirati	kopičiti	upravičevati
šivati	zagnati	izzarevati	zaupati
trgati	prižigati	premagati	razsvetljevati
kričati	tožiti	polarizirati	poosebljati
točiti	sprejemati	razjeziti	revolucionirati

(b) Verbs

Table 7: Attention checks for Slovene words using Best-Worst Scaling, with highlighted **most concrete** and **most abstract** targets.

## A.2 Details on Figurative Judgments

**Survey** The survey asks participants to rate whether verb-object noun expressions<sup>16</sup> are literal or figurative, and provide example sentences. As shown in Figure 9, task reminders and examples are shown above each question. We assign five annotators per survey and paid 5.70€ for a median time of  $\approx 32$  minutes.

**Tasks:**

1. Decide whether the event description is clearly based on the meanings of the two words. For example, the event "see bird" is **literally describing the event of seeing a bird**. In contrast, the event "grasp meaning" **does not describe someone literally grasping a meaning: the event meaning is figurative**, rather than literal.

Together with your decision regarding whether an event is literal or figurative, we also ask you:

2. Provide an example sentence including the two-word event, e.g. "see bird" -> *I saw a bird near the lake today.* or for "grasp meaning" -> *She quickly grasped the mathematical meaning.*

---

Decision for "illustrate complexity" \*

literal

figurative

---

Example sentence for "illustrate complexity" \*

The example sentence must include the two-word event

Your answer \_\_\_\_\_

Figure 9: Example of English survey question for collecting literal and figurative judgments.

**Quality Checks** For our figurative survey, we place five attention checks evenly distributed. Participants see the message “This is an attention check, please select **literal** for **paint wall**” in the respective language. Table 8 lists all attention check questions for each survey, with blue denoting figurative and orange denoting literal targets. Across languages, the literal targets remain aligned translations, while the figurative targets vary due to challenges in translations and preservation of equivalent syntactic constructions.

## B Details on Automatic Extrapolation of Concreteness Ratings

### B.1 Models

For all experiments in the paper, we use the models described in Table 9.

<sup>16</sup>To make the task more accessible for non-expert annotators, expressions are simplified into “two-word events.”

EN	DE	SL
steal show	Wirtschaft ankurbeln	nabirati izkušnje
paint wall	Wand bemalen	pobarvati steno
break heart	Leben einhauchen	prejeti milost
drive car	Auto fahren	voziti avto
catch fire	Frage anpacken	mešati zrak

Table 8: Attention checks regarding **figurative** and **literal** targets.

### B.2 Prompt Instructions

**Prompting with RS** As shown in Box B.2, the prompt provides instructions for rating the concreteness of expressions on a scale from 1.0 to 5.0, with examples of concrete and abstract expressions. The prompt translation is the same in German and Slovene.

**Prompting with BWS** As shown in Box B.2, the prompt provides instructions for selecting the most concrete and most abstract expression from a list of four expressions, with examples of concrete and abstract expressions. The prompt translation is the same in German and Slovene.

Model	Type	Number of Parameters	Language Support	URL
				<a href="https://huggingface.co">https://huggingface.co</a>
RoBERTa	Encoder	125M	English	<a href="#">/FacebookAI/roberta-base</a>
GBERT	Encoder	111M	German	<a href="#">/deepset/gbert-base</a>
SloBERTa	Encoder	110M	Slovene	<a href="#">/EMBEDDIA/sloberta</a>
Falcon3-7B-Instruct	Decoder	7B	English	<a href="#">/tiiuae/falcon3-7b-instruct</a>
LeoLM-7B	Decoder	7B	German	<a href="#">/LeoLM/leo-hessianai-7b</a>
GaMS-9B-Instruct	Decoder	9B	Slovene	<a href="#">/cjvt/GaMS-9B-Instruct</a>
			Multilingual	
EuroLLM-9B-Instruct	Decoder	9B	(incl. English, German and Slovene)	<a href="#">/utter-project/EuroLLM-9B-Instruct</a>

Table 9: List of models used in our concreteness score prediction and generation tasks.

### English Prompt Template (RS)

Background: A concrete expression refers to an event with which you can have an immediate experience through your senses (smelling, tasting, touching, hearing, seeing) and by performing or witnessing the event. An example of a concrete expression is “drive car”.

An abstract expression refers to an event or process you cannot experience directly through your senses or by performing the event. An example of an abstract expression is “deliver justice”.

Task: Rate the concreteness of the following expression on a scale from 1.0 to 5.0 (using one decimal place), where:

1.0 = very abstract  
5.0 = very concrete

Focus on only responding with floating-point number rating, and do not add any explanations. The values have to be within the 1.0 and 5.0 range.

Now for the following expression, provide a concreteness rating:

Expression: **“carry implication”**  
Rating: **[predicted score]**

### English Prompt Template (BWS)

Background: A concrete expression refers to an event with which you can have an immediate experience through your senses (smelling, tasting, touching, hearing, seeing) and by performing or witnessing the event. An example of a concrete expression is “drive car”.

An abstract expression refers to an event or process you cannot experience directly through your senses or by performing the event. An example of an abstract expression is “deliver justice”.

Task: Which of the four expressions is likely the MOST concrete and which of the four expressions is likely the MOST abstract.

Only provide the expression number. Do not repeat the text content.

Expression 1: **buy car**  
Expression 2: **evaluate paper**  
Expression 3: **assume office**  
Expression 4: **undermine principle**

Answer only with two numbers separated by a comma:

<most concrete expression number>,  
<most abstract expression number>  
**[number, number]**