Contextualized Sensorimotor Norms: multi-dimensional measures of sensorimotor strength for ambiguous English words, in context

Anonymous ACL submission

Abstract

Most large language models are trained on linguistic input alone, yet humans appear to *ground* their understanding of words in sensorimotor experience. A natural solution is to augment LM representations with human judgments of a word's sensorimotor associations (e.g., the Lancaster Sensorimotor Norms), but this raises another challenge: most words are ambiguous, and judgments of words in isolation fail to account for this multiplicity of meaning (e.g., "wooden *table*" vs. "data *table*"). We attempted to address this problem by building a new lexical resource of contextualized sensorimotor judgments for 112 English words, each rated in four different contexts (448 sentences total). We show that these ratings encode overlapping but distinct information from the Lancaster Sensorimotor Norms, and that they also predict other measures of interest (e.g., relatedness), above and beyond measures derived from BERT.

1 Introduction

Most large language models (LMs) are trained on linguistic input alone. This approach may be fundamentally limited when it comes to language understanding (Bender and Koller, 2020; Bisk et al., 2020; Tamari et al., 2020), as the meaning of a word arguably depends on factors beyond which words it co-occurs with. In particular, humans appear to *ground* a word's meaning in a rich network of sensorimotor associations (Pulvermüller, 1999; Bergen, 2012; Bergen and Feldman, 2008; Barsalou, 1999; Winter and Bergen, 2012; Barsalou, 2008; Glenberg and Kaschak, 2002). For example, our understanding of the word "table" incorporates not just the words that frequently co-occur with "table", but also our embodied experience of tables: how they look, how they feel, which parts of our body we use to interact with them, and more. If human-like language understanding depends on grounding words in non-linguistic associations (Harnad, 1990), then LMs trained on text alone will never reach human levels of understanding (Bender and Koller, 2020).

One promising solution is to link an LM's representations—based on distributional statistics alone—to human judgments of a word's sensorimotor associations, such as the Lancaster Sensorimotor Norms (Lynott et al., 2019) (hereafter LS Norms). The LS Norms provide human ratings about the extent to which an isolated word (e.g., "table") is strongly associated with various *sensory modalities* (e.g., Vision vs. Touch) and *action effectors* (e.g., Hand/Arm vs. Foot/Leg). Recent work (Kennington, 2021; Wan et al., 2020b,a) has found that integrating these norms improves the performance of language models on several NLP tasks, such as GLUE (Wang et al., 2018) and metaphor detection (Wan et al., 2020a).

Despite the promise and early success of this approach, it faces a key limitation: resources like the LS Norms typically contain just a single set of judgments for each word. In practice, however, many words are *ambiguous* (Rodd et al., 2004; Haber and Poesio, 2021). In English, anywhere from 7% (Rodd et al., 2004) to 15% (Trott and Bergen, 2020) of words have multiple, unrelated meanings—and as many as 84% are polysemous, i.e., they have multiple, related meanings (Rodd et al., 2004). For example, the word "table" may refer to a piece of furniture or to a database organized into rows and columns. Further, even very similar uses of a word, like "lemon", in its fruit-denoting sense, evoke different sensorimotor associations in different contexts (e.g., "She peeled the lemon" vs. "She put the lemon in the bag") (Yee and Thompson-Schill, 2016; Elman, 2009; Trott et al., 2020). Accordingly, there is evidence that ratings of sensorimotor strength or concreteness can vary considerably depending on whether a word is presented alone or in context (Scott et al., 2019), or as a function of

which context a word is presented in (Reijnierse et al., 2019). This suggests that any effort to *ground* words should account for the fact that most words are ambiguous, with dynamic, context-sensitive meanings subject to construal.

In Section 2, we first describe related resources, as well as work on grounding large LMs using psycholinguistic resources and multimodal input. In Section 3, we introduce the Contextualized Sensorimotor Norms (CS Norms), a dataset of sensorimotor judgments about ambiguous words in context. In Section 4, we provide descriptive statistics about the CS Norms, as well as comparisons to other factors such as the *dominance* of a particular sense. In Section 5, we show that a metric derived from the CS Norms—the Sensorimotor Distance between two contexts of use—improves our ability to predict contextualized relatedness judgments, above and beyond a similar metric derived from BERT (Devlin et al., 2019). Finally, in Section 6, we discuss limitations of these norms, as well as avenues for future research.

2 Related Work

2.1 Related Resources

Norms presenting aggregated semantic judgments about words date back at least to the early 1980s; the MRC database contains judgments of both concreteness and imageability for just under 9000 English words, each presented in isolation (Coltheart, 1981). Later, the Brysbaert concreteness norms expanded this dataset to approximately 37,000 English words (Brysbaert et al., 2014b); concreteness ratings have also been collected for Dutch (Brysbaert et al., 2014a), Croatian (Ćoso et al., 2019), and more.

Judgments of concreteness or overall sensorimotor strength are limited in that they do not account for which sensorimotor features are particularly salient. More recently, researchers have collected ratings about multiple semantic features for each word, including its sensorimotor associations (Lynott et al., 2019), as well as even more fine-grained judgments within each modality (e.g., for Vision, whether the referent is Fast or Slow; for Touch, whether it is Hot or Cold) (Binder et al., 2016). Of these, the largest dataset is the Lancaster Sensorimotor Norms (Lynott et al., 2019), which includes 11-dimensional judgments for about 40,000 English words. This approach has been extended to other languages, such as French (Miceli et al., 2021) and Dutch (Speed and Brysbaert, 2021). Again, in each case, the words were presented without context.

Finally, several datasets have collected concreteness judgments about words in context (Scott et al., 2019; Reijnierse et al., 2019). However, to our knowledge, no dataset includes judgments about *which* sensorimotor features are particularly salient in different linguistic contexts.

2.2 Grounding LMs with Psycholinguistic Resources

Recent work in NLP has begun to incorporate these psycholinguistic resources. One approach attempts to predict these judgments about concreteness or salient sensorimotor features from LM representations, with varying degrees of success (Thompson and Lupyan, 2018; Turton et al., 2020; Chersoni et al., 2020; Utsumi, 2020). Another approach uses sensorimotor features to augment the ability of an LM on an applied task, such as the GLUE benchmark (Kennington, 2021) or metaphor detection (Wan et al., 2020b). These experiments suggest that sensorimotor features do improve performance on specific tasks, though as mentioned in Section 1, they are limited in that the sensorimotor features themselves were obtained for words in isolation.

2.3 Grounding LMs with Multimodal Input

An alternative approach is to ground LM representations more directly in multimodal input. Most of this work has emphasized the visual modality, linking words to static images (Kiros et al., 2018; Su et al., 2020) or video (Zellers et al., 2021). This paradigm shows considerable promise, though it is naturally limited by resource constraints; obtaining reliable multimodal data and aligning it to language can be both time-consuming and costly.

2.4 Summary

There is considerable interest in *grounding* among both psycholinguists and NLP practitioners. To that end, psycholinguists have developed large linguistic resources, which some NLP researchers have used to improve LMs.

Still, one limitation of the majority of existing resources is that they do not contain judmgents about different sensorimotor features for words in different contexts. Because most words are ambiguous, this makes it difficult to know which meaning the sensorimotor judgments reflect, which in turn reduces the precision and utility of these resources.

3 Contextualized Sensorimotor Norms

Our primary goal was to collect sensorimotor judgments about ambiguous words, appearing in controlled sentential contexts. We used sentences from the RAW-C (Relatedness of Ambiguous Words–in Context) dataset (Trott and Bergen, 2021). RAW-C contains relatedness judgments for 672 English sentence pairs, each containing the same target word (e.g., "bat") in either the same meaning (e.g., "furry bat" vs. "fruit bat") or different meaning (e.g., "furry bat" vs. "wooden bat"). There were 448 unique sentences in total (112 target words, with 4 sentences each).

Rather than collecting judgments about sentence pairs, we were interested in judgments about the sensorimotor associations evoked by a word in a particular sentential context . This also provided a more direct analogue to the Lancaster Sensorimotor Norms (Lynott et al., 2019), in which participants observed a particular lexical item (e.g., "bat") and provided ratings about its associated sensory modalities (e.g., Vision) or action effectors (e.g., Hand/Arm).

3.1 Participants

Our goal was to collect a minimum of 10 judgments per sentence. Thus, we recruited participants until each sentence had at least 10 observations, after applying the exclusion criteria.

A total of 377 participants were recruited through UC San Diego's undergraduate subject pool for Psychology, Cognitive Science, and Linguistics students. Participants received class credit for participation. After excluding non-native speakers of English, participants who failed to pass the bot checks, and participants whose inter-annotator agreement score was sufficiently low (see Section 3.3 below), we were left with 283 participants. Of these, 223 identified as female (47 male, 8 non-binary, and 5 preferred not to answer). The mean self-reported age was 20.4 (median = 20, SD = 2.98), and ranged from 18 to 43.

3.2 Procedure

We adapted the procedure directly from Lynott et al. (2019), with the main modification being that participants now saw words in sentential contexts. As in Lynott et al. (2019), participants were randomly assigned to one of two Judgment Types: 1) Perception, in which they provided ratings about a word's associated sensory modalities (Vision, Hearing, Touch, Interoception, Smell, and Taste); and 2) Action, in which they rated a word's assocated action effectors (Hand/Arm, Foot/Leg, Mouth/Throat, Head, and Torso). In total, 132 participants were assigned to the Perception Judgment Type, and 151 were assigned to the Action Judgment Type.

After giving consent, participants answered two bot check questions (e.g., "Which of the following is not a place to swim?"). They were then told that they would read a series of sentences, each containing a bolded word (e.g., "It was a wooden **table**"), and that their task was to rate the degree to which they experienced the concept denoted by that word with either six sensory modalities (in the Perception Judgment Type) or five action effectors (in the Action Judgment Type). Ratings ranged from 0 (not at all experienced with that sense/effector) to 5 (experienced greatly with that sense/effector).

Each participant rated approximately 60 sentences overall, randomly sampled from the set of 448 sentences. No participant saw the same target word twice. On each trial, the sentence was displayed at the top of the page, with the target word bolded. Underneath the sentence, the instructions read: "To what extent do you experience WORD:" (for Perception) or "To what extent do you experience WORD by performing an action with the:" (for Action), where "WORD" was replaced with the target word. Underneath the instructions were six (for Perception) or five (for Action) rating scales, corresponding to each possible sensory modality or action effector. For the Action Judgment Type, the scale was accompanied by a labeled diagram of the body, as in Lynott et al. (2019).

To reach the target of 10 respondents to each word in both Action and Perception tasks, we collected data in two stages. In the first stage (Group 1), participants were randomly assigned to either the Perception or Action Judgment Types, and the sentences they observed were randomly sampled from the set of possible sentences for each word. After we had collected responses from 264 participants in this way, there were still a number of sentences that had very few observations, simply by chance—as well as many with more than ten observations. Thus, in the second stage (Group 2), participants were assigned a mix of Low-N (sentences with fewer than 10 ratings) and High-N (sentences with 10 or more ratings) items. The goal was to speed data collection; to control for poten-

tial differences across groups, we compared their distributions of inter-annotator agreement scores, and found no evidence that the different data collection procedures induced different response behavior (see Section 3.3).

Finally, after providing ratings, participants reported their self-identified gender and age, as well as whether or not they were a native speaker of English.

The data collection was conducted online using JsPsych (De Leeuw, 2015).

3.3 Inter-Annotator Agreement

We sought to establish the degree to which different participants agreed about their ratings for each sentence, both to characterize the dataset and to exclude participants whose ratings diverged substantively from the rest of the sample. Following past work (Trott and Bergen, 2021), we used a leave-one-out scheme: for each participant, we computed the Spearman's rank correlation between that participant's responses and the mean ratings for those items from the rest of the sample (excluding the participant's ratings).

Importantly, we did this in two stages. First, we computed the distribution of agreement scores for the 264 participants in Group 1, i.e., the participants for whom each sentence was truly randomly sampled from the set of 448 sentences. Based on this distribution of inter-annotator agreement scores, we excluded a total of 18 participants, whose scores were more than two standard deviations below the mean for that Judgment Type. Among the final set of 246 participants in this group, the mean inter-annotator rank correlation was 0.47 for Action judgments (SD = 0.1) and 0.64 for Perception judgments (SD = 0.11).

Then we considered the 39 participants from Group 2, who provided ratings for a restricted set of sentences, i.e., sentences which either had below 10 judgments from Group 1 (low-N) or had more than 10 judgments from Group 1 (high-N). For each participant in Group 2, we compared the ratings for the high-N items to the mean response for those items among Group 1. After excluding participants with low inter-annotator agreement, we were left with a total of 37 participants in Group 2. The mean rank correlation was 0.5 for Action Judgments (SD = 0.11) and 0.64 for Perception judgments (SD = 0.1).

Finally, we combined the set of inter-annotator agreement scores from both groups, and constructed a linear regression with Rank Correlation as the dependent variable, and main effects of Judgment Type (Action vs. Perception) and Group (Group 1 vs. Group 2), as well as their interaction. There was no significant difference in agreement across groups ($p > .1$), but agreement was significantly higher for Perception ratings than Action ratings [$\beta = 0.17, SE = 0.01, p < .001$].

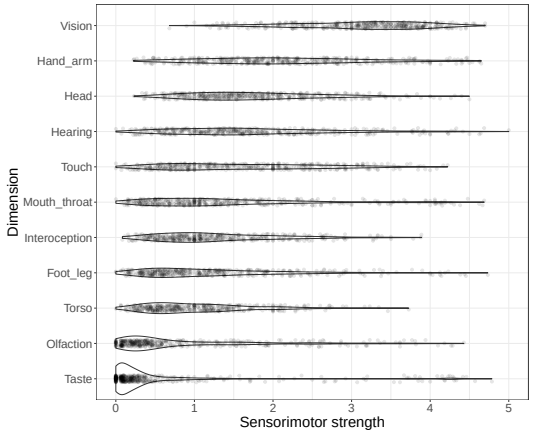

Figure 1: Distribution of mean sensorimotor strength judgments for each dimension. As in past work (Lynott et al., 2019), judgments are highest for the Vision dimension, and lowest for Olfaction and Taste.

3.4 Creating the Norms

Once we had obtained a minimum of ten ratings per sentence (per judgment type), we averaged across these ratings to produce a mean and standard deviation for each dimension. For example, the sentence "He saw the furry **bat**" would contain the mean (and standard deviation) of judgments about the salience of each sensorimotor feature.[1]

4 Characterizing the Contextualized Sensorimotor Norms

Our first goal was to characterize the Contextualized Sensorimotor Norms (CS Norms). The norms provide an 11-dimensional vector for each sentential context in which a word appears: the mean sensorimotor strength for 11 dimensions (6 sensory modalities, and 5 action effectors) for a target word in a given context.

[1]The norms (along with analysis code and a Data Sheet) are included in a .zip file as as Supplementary Data. A link to a public GitHub repository will be added once the anonymity period is over.

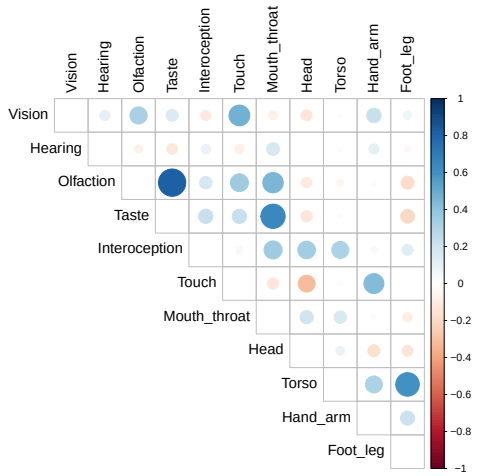

Figure 2: Pearson's correlation coefficients between the sensorimotor strength of each feature.

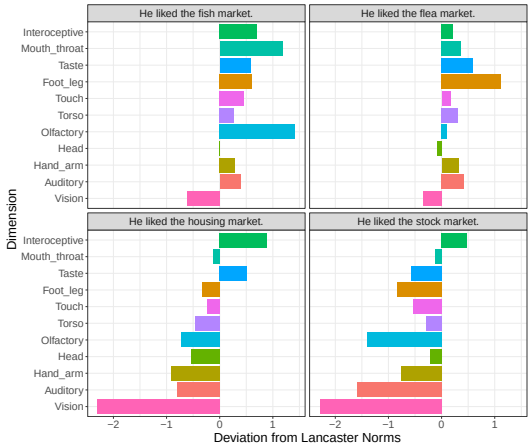

Figure 3: Deviation between the contextualized sensorimotor norms and the Lancaster Sensorimotor Norms for for the word "market", faceted by the distinct sentential context in which "market" appeared.

4.1 Comparing Sensorimotor Dimensions

As a first step, we visualized the distribution of sensorimotor judgments for each dimension (see Figure 1). Consistent with the original LS Norms (Lynott et al., 2019) and work on the English lexicon more generally (Majid, 2020), judgments tended to be highest for the Vision dimension, and lowest for Olfaction and Taste.

We then asked which dimensions were correlated with which other dimensions. Consistent with past work (Lynott et al., 2019), we found particularly strong positive correlations between Olfaction and Taste, as well as Foot/Leg and Torso; we also found a strong positive correlation between Taste and Mouth/Throat (see Figure 2).

4.2 Variance Across Contexts

A key motivation for the CS Norms was to account for potential variation within each word in which sensorimotor features were most salient across distinct sentential contexts.

We quantified this variation by normalizing the sensorimotor features for each context of use to the mean norms for that word from the LS Norms. For example, the LS norms have a single 11-dimensional vector for the word "market"; for each of the four sentential contexts in which "market" appeared, we calculated the difference in mean ratings across our norms and the LS Norms. This provides an estimate of the degree to which the human judgments were impacted by the sentential context, as opposed to a representation of the word's meaning in isolation (as in the LS Norms).

Figure 3 depicts these deviations from the LS Norms for a specific word, "market". This word was chosen because it displayed particularly high variation in its overall sensorimotor strength across contexts. Interestingly, the deviations from the LS Norms appear to track the two senses of the word being profiled. The two sentences corresponding to the *location* sense of "market" (i.e., "fish market" and "flea market") appeared to be closer to the LS Norms (i.e., the deviations were smaller on average); the notable exceptions were the *Olfactory* and *Mouth/Throat* dimensions for the "fish market" context, and the *Foot/Leg* dimension for the "flea market" context. Both deviations make sense: a salient property of fish markets is their smell and the fact that food is involved, thus evoking the use of the mouth/throat; in turn, walking might be a particularly salient feature of flea markets (relative to our conception of "market" in isolation), evoking the use of feet and legs.

In contrast, the sentences corresponding to the the *financial* sense of "market" (i.e., "housing market" and "stock market") were considerably lower in sensorimotor strength across almost all dimensions, especially *Vision*. Again, this makes sense, given that this meaning is more metaphorical or abstract than the *location* meaning of "market": apart from representations of their performance, neither housing markets nor stock markets can be visually perceived in the way that fish markets and flea markets can.

4.3 Sensorimotor Strength vs. Sense Dominance

One well-documented property of ambiguous words is that their multiple meanings are not always balanced: one sense is sometimes more cognitively salient than the other. This is called *sense dominance*. The degree of dominance is known to play an important role in the processing of ambiguous words, particularly for homonyms: empirical evidence suggests that comprehenders almost always activate the more dominant sense of a homonym, even when the linguistic context supports the subordinate meaning (Rayner et al., 1994; Binder and Rayner, 1998; Duffy et al., 1988). Most relevantly, there is some evidence that dominance is positively correlated with concreteness (Gilhooly and Logie, 1980).

We investigated whether this finding replicated in the CS Norms dataset. Following Lynott et al. (2019), we created a composite variable called Contextualized Sensorimotor Strength, which measured the maximum strength across the 11 sensorimotor features for each context of use. We then asked whether Sensorimotor Strength was significantly predictive of Sense Dominance, which we had measured for each sentence pair in the original RAW-C dataset (Trott and Bergen, 2021).

Using the *lme4* package (Bates et al., 2015) in R, we built a linear mixed effects model with Dominance as a dependent variable, fixed effects of Contextualized Sensorimotor Strength, a random intercept for each word, and two covariates reflecting the sensorimotor strength for each *word* (i.e., from the Lancaster Sensorimotor Norms dataset). This model explained significantly more variance than a model omitting only the Contextualized Sensorimotor Strength $[\chi^2(1) = 18.38, p < .001]$. Consistent with past work (Gilhooly and Logie, 1980), contexts of use with higher sensorimotor strength were also rated as more dominant $[\beta = 0.26, SE = 0.06, p < .001]$.

This finding does not explain *why* more concrete meanings are more dominant than meanings with less sensorimotor strength. It could be that people communicate about those meanings more frequently. Alternatively, their relative primacy in acquisition might impact psychological dominance; earlier learned meanings are also more dominant (Gilhooly and Logie, 1980).

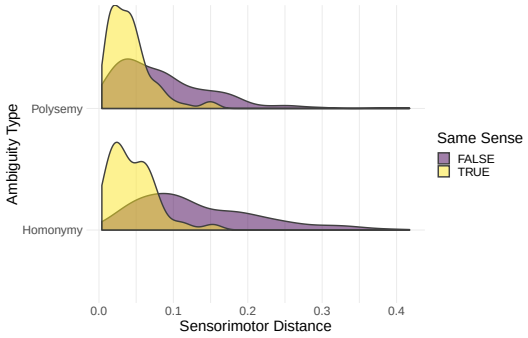

Figure 4: Distribution of sensorimotor distances as a function of same/different sense, as well as the type of ambiguity. Same sense uses have more similar sensorimotor associations than different sense uses.

4.4 Sensorimotor Distance

Another question that arises concerns the relationship *between* contexts of use. Because each context of use is associated with a vector, the similarity or dissimilarity between these contexts can be quantified by calculating the distance (e.g., the cosine distance) between these vectors (Wingfield and Connell, 2021). Thus, we calculated the cosine distance—referred to here as the Sensorimotor Distance—between the vectors corresponding to each sentence pair for each word (672 sentence pairs total). Larger distances reflect more dissimilar contexts of use, while smaller distances reflect more similar contexts.

We then asked whether Sensorimotor Distance was correlated with other psychologically relevant features, such as whether the two contexts of use corresponded to the same sense or different senses. Based on the preliminary findings in Section 4.2, we predicted that different sense uses would have less similar sensorimotor features.

Indeed, as depicted in Figure 4, Sensorimotor Distance was considerably larger for Different Sense than Same Sense contexts. The addition of Sense Boundary to a mixed effects model predicting Sensorimotor Distance improved model fit beyond a model with only Distributional Distance and Ambiguity Type (and random intercepts for words) $[\chi^2(1) = 34.86, p < .001]$. This is also consistent with Figure 3, in which the two *location* senses of "market" were more similar to each other than either was to the two *financial* senses.

5 Predictive Utility

We were also interested in the predictive utility of the information provided by the CS Norms, above

and beyond other commonly used factors. That is, to what extent do these ratings encode information that large language models, such as BERT, fail to capture?

As a first step, we sought to predict the *relatedness* of sentence pairs. RAW-C contains relatedness judgments for each unique sentence pair within each of the 112 words, with a total of 672 sentence pairs (Trott and Bergen, 2021). It is also annotated for whether the two contexts of use correspond to the same or different sense (Sense Boundary), and whether the relationship type is one of homonymy or polysemy (Ambiguity Type).

Past work (Trott and Bergen, 2021) has found that relatedness is negatively correlated with the cosine distance between BERT's contextualized embeddings for the target word in each sentence; here, we call this measure the *Distributional Distance*. Yet Distributional Distance falls short compared to human inter-annotator agreement: it underestimates the relatedness of same sense sentence pairs, and overestimates the relatedness of different sense homonyms (Trott and Bergen, 2021).

We asked whether a linear mixed effects model equipped with those previous factors (Distributional Distance[2], Sense Boundary, Ambiguity Type, and their interaction, as well as random intercepts for words) could be improved by the addition of Sensorimotor Distance (see Section 4.4). Indeed, Sensorimotor Distance significantly improved model fit $[\chi^2(1) = 36.74, p < .001]$. As expected, Sensorimotor Distance was negatively associated with Relatedness $[\beta = -1.81, SE = 0.22, p < .001]$: words with more dissimilar sensorimotor vectors were rated as less related, on average.

We also compared the Akaike Information Criterion, or AIC, of a number of different models predicting Relatedness. Three of the models corresponded to the two measures of distance, i.e., containing either Distributional Distance (derived from BERT's contextualized embeddings), Sensorimotor Distance (derived from the CS Norms) or both. One model contained only a fixed effect of Sense Boundary. The remaining two models contained either every factor listed above (along with an interaction between Sense Boundary and Ambiguity Type), or every factor but Sensorimotor Distance.

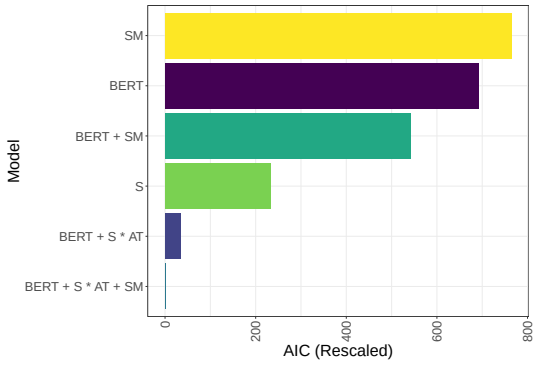

Figure 5: Rescaled AIC values for models predicting Relatedness using an assortment of factors: Sense Boundary (S), Ambiguity Type (AT), Distributional Distance (BERT), and Sensorimotor Distance (SM). A lower AIC score corresponds to better model fit.

Crucially, although Sense Boundary was by far the best isolated predictor of Relatedness, the inclusion of Sensorimotor Distance consistently improved model fit. This indicates that the CS Norms capture information that is at least partially independent from the information encoded by the other factors in the model.

6 Discussion

Embodied experience appears to be crucial for how humans learn and understand language (Bergen, 2012; Pulvermüller, 1999; Barsalou, 1999), yet most large language models (LMs) are exposed to linguistic input alone (Bender and Koller, 2020). One solution is to augment LM representations with psycholinguistic resources, such as human judgments of the sensorimotor features associated with a word (Lynott et al., 2019). However, this approach must also contend with the challenge of lexical ambiguity. Words mean different things in different contexts (Rodd et al., 2004; Trott et al., 2020), yet many lexical resources collect judgments about words in isolation.

We attempted to address this challenge by collecting judgments about the salience of various sensory modalities (e.g., *Vision*) and action effectors (e.g., *Torso*) for the same English word, in distinct sentential contexts (e.g., "flea *market*" vs. "housing *market*"). We called this dataset the Contextualized Sensorimotor Norms (CS Norms).

These contextualized norms capture variance in sensorimotor associations beyond the information already provided by the Lancaster Sensorimotor

[2]Distributional Distance was calculated by taking the cosine distance between the final layers of BERT's contextualized embeddings for the target word in each sentence, using the `bert-embedding` package (https://pypi.org/project/bert-embedding/).

Norms (Figure 3). We also replicated past work (Gilhooly and Logie, 1980) suggesting that the psychological dominance of a meaning is correlated with its sensorimotor strength. Third, we found that the sensorimotor distance between contexts of use was correlated with the existence of sense boundary (see Figure 4). Finally, in Section 5, we demonstrated the predictive utility of the CS Norms above and beyond large LMs such as BERT: the sensorimotor distance between two contexts of use predicted human judgments of relatedness, above and beyond a similar measure derived from BERT.

6.1 Limitations

This dataset is not without limitations.

First, it is restricted in size and breadth: 448 sentences (112 words, with 4 sentences each), in English only. In contrast, the Lancaster Sensorimotor Norms contain judgments of almost 40,000 English words (Lynott et al., 2019), and have now been extended to French (Miceli et al., 2021), Dutch (Speed and Brybaert, 2021), and more. Having demonstrated the utility of the CS Norms on a small subset of English words, one obvious direction for future research would be to expand this dataset—including more words, more sentences per word, a wider variety of sentences (i.e., both experimentally controlled and naturalistic sentences), and additional languages. Similarly, existing datasets on lexical ambiguity (Haber and Poesio, 2021; Karidi et al., 2021) could be augmented with sensorimotor judgments.

Second, as others have noted (Bender and Koller, 2020; Bisk et al., 2020; Tamari et al., 2020; Borghi et al., 2019), *grounding* goes beyond sensorimotor associations. Linguistic meaning is also grounded in social experience and interaction. Recent work has attempted to incorporate these social aspects of grounding, either by integrating social information into distributional models (Johns, 2021) or simply by including more dimensions in the grounded feature representations (Binder et al., 2016).

Finally, recent work has enjoyed some success in learning grounded feature vectors directly from LM representations, typically for words rated in isolation (Turton et al., 2020; Chersoni et al., 2020; Utsumi, 2020). One question is whether contextualized embeddings, derived from a large LM such as BERT, are sensitive enough to capture the fine-grained distinctions that the CS Norms encode across sentential contexts for the same word.

7 Conclusion

We have presented a novel resource: human judgments about the strength or salience of various sensorimotor features for 112 English words, each appearing in four distinct sentential contexts. This resource was extended from past work (Trott and Bergen, 2021), and thus also contains information about the relatedness *between* sentential contexts for the same word. We provided several demonstrations of the dataset's utility, above and beyond judgments of these words in isolation (Lynott et al., 2019), as well as large LMs such as BERT (see Section 5).

8 Ethical Considerations

All responses from human participants were anonymized before analyzing any data. Further, the final, publicly available dataset has collapsed across subject-level responses for each sentence.

All participants provided informed consent, and were compensated in the form of class credit. The project was carried out with IRB approval.

Finally, we have attempted to ensure dataset quality by: 1) removing responses from participants who failed bot checks; 2) removing participants whose inter-annotator agreement scores were more than two standard deviations below the average; and 3) collecting at least ten ratings per sentence, per judgment type, as in past work (Lynott et al., 2019).

Acknowledgements

References

Lawrence W. Barsalou. 1999. Perceptual symbol systems. *Behavioral and Brain Sciences*, 22(4):577–660. Publisher: Cambridge University Press.

Lawrence W. Barsalou. 2008. Grounded Cognition. *Annual Review of Psychology*, 59(1):617–645. _eprint: https://doi.org/10.1146/annurev.psych.59.103006.093639.

Douglas Bates, Martin Mächler, Ben Bolker, and Steve Walker. 2015. Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, 67(1):1–48.

Emily M. Bender and Alexander Koller. 2020. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, Online. Association for Computational Linguistics.

Benjamin Bergen and Jerome Feldman. 2008. 16 - Embodied Concept Learning. In Paco Calvo and Antoni Gomila, editors, *Handbook of Cognitive Science*, Perspectives on Cognitive Science, pages 313–331. Elsevier, San Diego.

Benjamin K. Bergen. 2012. *Louder Than Words: The New Science of How the Mind Makes Meaning*. Basic Books, New York, NY, USA.

Jeffrey R Binder, Lisa L Conant, Colin J Humphries, Leonardo Fernandino, Stephen B Simons, Mario Aguilar, and Rutvik H Desai. 2016. Toward a brain-based componential semantic representation. *Cognitive neuropsychology*, 33(3-4):130–174.

Katherine S Binder and Keith Rayner. 1998. Contextual strength does not modulate the subordinate bias effect: Evidence from eye fixations and self-paced reading. *Psychonomic Bulletin & Review*, 5(2):271–276.

Yonatan Bisk, Ari Holtzman, Jesse Thomason, Jacob Andreas, Yoshua Bengio, Joyce Chai, Mirella Lapata, Angeliki Lazaridou, Jonathan May, Aleksandr Nisnevich, Nicolas Pinto, and Joseph Turian. 2020. Experience grounds language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8718–8735, Online. Association for Computational Linguistics.

Anna M Borghi, Laura Barca, Ferdinand Binkofski, Cristiano Castelfranchi, Giovanni Pezzulo, and Luca Tummolini. 2019. Words as social tools: Language, sociality and inner grounding in abstract concepts. *Physics of life reviews*, 29:120–153.

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*, 46(3):904–911.

Emmanuele Chersoni, Rong Xiang, Qin Lu, and Chu-Ren Huang. 2020. Automatic learning of modality exclusivity norms with crosslingual word embeddings. In *Proceedings of the Ninth Joint Conference on Lexical and Computational Semantics*, pages 32–38, Barcelona, Spain (Online). Association for Computational Linguistics.

Max Coltheart. 1981. The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology Section A*, 33(4):497–505.

Bojana Ćoso, Marc Guasch, Pilar Ferré, and José Antonio Hinojosa. 2019. Affective and concreteness norms for 3,022 croatian words. *Quarterly Journal of Experimental Psychology*, 72(9):2302–2312.

Joshua R De Leeuw. 2015. jspsych: A javascript library for creating behavioral experiments in a web browser. *Behavior research methods*, 47(1):1–12.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Susan A Duffy, Robin K Morris, and Keith Rayner. 1988. Lexical ambiguity and fixation times in reading. *Journal of memory and language*, 27(4):429–446.

Jeffrey L Elman. 2009. On the meaning of words and dinosaur bones: Lexical knowledge without a lexicon. *Cognitive science*, 33(4):547–582.

Kenneth J Gilhooly and Robert H Logie. 1980. Meaning-dependent ratings of imagery, age of acquisition, familiarity, and concreteness for 387 ambiguous words. *Behavior Research Methods & Instrumentation*, 12(4):428–450.

Arthur M Glenberg and Michael P Kaschak. 2002. Grounding language in action. *Psychonomic bulletin & review*, 9(3):558–565.

Janosch Haber and Massimo Poesio. 2021. Patterns of polysemy and homonymy in contextualised language models. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2663–2676, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D: Nonlinear Phenomena*, 42(1):335 – 346.

Brendan T Johns. 2021. Distributional social semantics: Inferring word meanings from communication patterns. *Cognitive Psychology*, 131:101441.

Taelin Karidi, Yichu Zhou, Nathan Schneider, Omri Abend, and Vivek Srikumar. 2021. Putting words in BERT's mouth: Navigating contextualized vector spaces with pseudowords. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10300–10313, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Casey Kennington. 2021. Enriching language models with visually-grounded word vectors and the Lancaster sensorimotor norms. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 148–157, Online. Association for Computational Linguistics.

Jamie Kiros, William Chan, and Geoffrey Hinton. 2018. Illustrative language understanding: Large-scale visual grounding with image search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 922–933, Melbourne, Australia. Association for Computational Linguistics.

Dermot Lynott, Louise Connell, Marc Brysbaert, James Brand, and James Carney. 2019. The lancaster sensorimotor norms: multidimensional measures of perceptual and action strength for 40,000 english words. *Behavior Research Methods*, pages 1–21.

Asifa Majid. 2020. Human olfaction at the intersection of language, culture, and biology. *Trends in Cognitive Sciences*.

Aurélie Miceli, Erika Wauthia, Laurent Lefebvre, Laurence Ris, and Isabelle Simoes Loureiro. 2021. Perceptual and interoceptive strength norms for 270 french words. *Frontiers in Psychology*, 12:2018.

Friedemann Pulvermüller. 1999. Words in the brain's language. *Behavioral and brain sciences*, 22(2):253–279.

Keith Rayner, Jeremy M Pacht, and Susan A Duffy. 1994. Effects of prior encounter and global discourse bias on the processing of lexically ambiguous words: Evidence from eye fixations. *Journal of memory and language*, 33(4):527–544.

W Gudrun Reijnierse, Christian Burgers, Marianna Bolognesi, and Tina Krennmayr. 2019. How polysemy affects concreteness ratings: the case of metaphor. *Cognitive science*, 43(8):e12779.

Jennifer M Rodd, M Gareth Gaskell, and William D Marslen-Wilson. 2004. Modelling the effects of semantic ambiguity in word recognition. *Cognitive science*, 28(1):89–104.

Graham G Scott, Anne Keitel, Marc Becirspahic, Bo Yao, and Sara C Sereno. 2019. The glasgow norms: Ratings of 5,500 words on nine scales. *Behavior research methods*, 51(3):1258–1270.

Laura J Speed and Marc Brybaert. 2021. Dutch sensory modality norms. *Behavior Research Methods*, pages 1–13.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pre-training of generic visual-linguistic representations.

Ronen Tamari, Chen Shani, Tom Hope, Miriam R L Petruck, Omri Abend, and Dafna Shahaf. 2020. Language (re)modelling: Towards embodied language understanding. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6268–6281, Online. Association for Computational Linguistics.

Bill Thompson and Gary Lupyan. 2018. Automatic estimation of lexical concreteness in 77 languages. In *The 40th annual conference of the cognitive science society (cogsci 2018)*, pages 1122–1127. Cognitive Science Society.

Sean Trott and Benjamin Bergen. 2020. Why do human languages have homophones? *Cognition*, 205:104449.

Sean Trott and Benjamin Bergen. 2021. RAW-C: Relatedness of ambiguous words in context (a new lexical resource for English). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 7077–7087, Online. Association for Computational Linguistics.

Sean Trott, Tiago Timponi Torrent, Nancy Chang, and Nathan Schneider. 2020. (Re)construing meaning in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5170–5184, Online. Association for Computational Linguistics.

Jacob Turton, David Vinson, and Robert Smith. 2020. Extrapolating binder style word embeddings to new words. In *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, pages 1–8, Marseille, France. European Language Resources Association.

Akira Utsumi. 2020. Exploring what is encoded in distributional word vectors: A neurobiologically motivated analysis. *Cognitive Science*, 44(6):e12844.

Mingyu Wan, Kathleen Ahrens, Emmanuele Chersoni, Menghan Jiang, Qi Su, Rong Xiang, and Chu-Ren Huang. 2020a. Using conceptual norms for metaphor detection. In *Proceedings of the Second Workshop on Figurative Language Processing*, pages 104–109, Online. Association for Computational Linguistics.

Mingyu Wan, Baixi Xing, Qi Su, Pengyuan Liu, and Chu-Ren Huang. 2020b. Sensorimotor enhanced neural network for metaphor detection. In *Proceedings of the 34th Pacific Asia Conference on Language, Information and Computation*, pages 312–317, Hanoi, Vietnam. Association for Computational Linguistics.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. 2018. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium. Association for Computational Linguistics.

Cai Wingfield and Louise Connell. 2021. Sensorimotor distance: A fully grounded measure of semantic similarity for 800 million concept pairs.

Bodo Winter and Benjamin Bergen. 2012. Language comprehenders represent object distance both visually and auditorily. *Language and Cognition*, 4(1):1–16.

Eiling Yee and Sharon L Thompson-Schill. 2016. Putting concepts into context. *Psychonomic bulletin & review*, 23(4):1015–1027.

Rowan Zellers, Ximing Lu, Jack Hessel, Youngjae Yu, Jae Sung Park, Jize Cao, Ali Farhadi, and Yejin Choi. 2021. Merlot: Multimodal neural script knowledge models. *arXiv preprint arXiv:2106.02636*.