**Motivation for Dataset Creation**:

*Why was the dataset created?*

There is increasing interest in *grounding* language models, e.g., with human judgments about the sensorimotor associations of particular words. To our knowledge, existing lexical resources do not account for lexical ambiguity––words might have very different sensorimotor associations in different contexts. Additionally, in the realm of psycholinguistics, there are still many open questions about the role of sensorimotor associations in the formation of sense categories, and how sensorimotor features affect the relative dominance of different senses. Thus, we created the dataset as a first step towards a resource that can be used to: 1) help ground and evaluate existing language models; and 2) help answer theoretical questions about the human mental lexicon.

*What other tasks could the dataset be used for?*

The primary NLP application is likely in evaluating how well large LMs encode information that reflects sensorimotor features. However, the dataset could theoretically be used to augment these models as well; it is relatively small (only 112 words, with 4 sentences each), so it’s unclear how successful this augmentation would be. But if an LM were augmented in this way, it could presumably be used for any other tasks for which LMs are currently used.

*Has the dataset been used for any tasks already?*

We have derived a measure of sensorimotor distance from the sensorimotor feature ratings, and used that measure to predict Relatedness. We have shown that sensorimotor distance predicts Relatedness above and beyond a measure derived from BERT. This work has not yet been published. Additionally, it is not so much a task involving LM training as one comparing *to* a LM’s predictive power.

*Who funded the creation of the dataset?*

The creators of the dataset are employees of a public university. It was not funded by any grant.

**Dataset Composition**:

*What are the instances?*

Each observation (or “row”) is a sentence containing a target word, along with ratings of sensorimotor strength for 6 sensory modalities (e.g., Vision or Touch) and 5 action effectors (e.g., Foot/Leg). Each row also contains the standard deviation around the mean human judgment for that sentence, as well as the number of people who rated that sentence in either the perception or action task.

*Are relationships between instances made explicit in the data?*

Sentences containing the same word can be identified using the “word” column.

*How many instances of each type are there?*

There are 112 words, with 4 sentences each (so 448 sentences total).

*What data does each instance consist of?*

Each instance is a row of a table, including: the sentence (a string), the word (a string), 11 columns reflecting the mean ratings for each sensorimotor dimension (all floats), 11 columns reflecting the standard deviation of those ratings, and two columns reflecting the count of people who rated that sentence.

*Does the data rely on external resources?*

No.

*Are there recommended data splits or evaluation measures*?

The dataset is quite small so leave-one-out cross-validation could be used. We have not yet used the dataset itself for any sort of evaluation. However, a logical point of evaluation would be predicting the 11 sensorimotor dimensions from some other linguistic representation (e.g., a contextualized embedding from BERT).

*What experiments were initially run on this dataset?*

We have used the sensorimotor features to predict the *dominance* of particular meanings, collected for a different dataset (Trott & Bergen, 2021: <https://arxiv.org/abs/2105.13266>). We have also used these features to predict *relatedness* of different meanings, also from that previous dataset.

**Data Collection Process:**

*How was the data collected?*

We collected the data online using JsPsych. Participants were recruited from the undergraduate Psychology, Linguistics, and Cognitive Science subject pool, and received class credit for participating.

*Who was involved in the collection process*?

The participants were undergraduate students, compensated in the form of class credit.

*Over what time-frame was the data collected*?

We collected the data between September of 2021 and the end of October 2021.

*How was the data associated with each instance acquired*?

Participants were presented with sentences and asked to provide ratings about the salience of various sensorimotor dimensions. The final dataset contains the mean responses across subjects for each unique sentences.

*Does the dataset contain all possible instances?*

Yes, every sentence is included.

*If the dataset is a sample, then what is the population?*

The dataset is a subset of the possible words in English and the sentences containing those words. We adapted the sentences from a previous dataset directly (Trott & Bergen, 2021: <https://arxiv.org/abs/2105.13266>). This makes the dataset limited in breadth: there are only 112 words, with only 4 sentences each, and all are in English.

*Is there information missing from the dataset and why?*

No.

*Are there any known errors in the data?*

Not to our knowledge.

**Data Preprocessing**

*What preprocessing was done?*

Responses were averaged across each individual subject for each sensorimotor dimension for each sentence.

*Was the “raw” data saved in addition to the cleaned data?*

Yes, and it is available upon request.

*Is the preprocessing software available?*

The Python and R code to clean the raw responses will be made available.

*Does this dataset procedure achieve the motivation for creating the dataset stated in the first section?*

We believe so.

**Data Distribution**

*How is the dataset distributed?*

We will host the dataset on GitHub.

*When will the dataset be released?*

We will publicly release the dataset when the anonymity period for the relevant reviews is over.

*What license is it distributed under*?

There is no license, but we request that you cite the paper in which the dataset will first be described/published.

*Any fees/restrictions?*

No.

**Dataset Maintenance**

*Who is supporting/maintaining the dataset?*

The first author of the paper (this will be de-anonymized when the review period is over).

*Will the dataset be updated*?

The dataset will be updated on GitHub if errors are found, along with a description on the GitHub repository of those changes.

*If the dataset becomes obsolete how will this be communicated*?

It is not clear how it would become obsolete, but if it did somehow, we would add an announcement on the GitHub and take the data itself down.

*Is there a repository to link to any papers that use this dataset?*

Not yet, as it is a new dataset.

*If others want to augment this dataset, is there a mechanism for them to do so?*

At the moment, there is no official mechanism or formal process put in place for this. Others are free to collect judgments for additional sentences and words, which could conceivably be merged with this dataset.

**Legal and ethical considerations**

*If the dataset relates to people or was generated by people, were they informed about the data collection?*

Responses were generated by people, who gave informed consent.

*If it relates to other ethically protected subjects, have obligations been met?*

N/A.

*If it relates to people, were there any ethical review applications/reviews/approvals?*

Yes, it was conducted with IRB approval.

*If it relates to people, were they told what the dataset would be used for and did they consent?*

Participants were informed that we were collecting information about the sensorimotor properties of various sentences/words. There is not an official mechanism in place for them to revoke their consent in the future.

*If it relates to people, could this dataset expose people to harm or legal action?*

No. All responses were anonymized.

*Does it unfairly advantage or disadvantage a particular social group?*

Not to our knowledge.

*Were people provided with privacy guarantees?*

Yes, all responses were anonymized and completely decoupled from any identifying information.

*Does the dataset comply with the GDPR and other relevant standards?*

Yes, we believe so.

*Does the dataset contain information that might be considered sensitive or confidential?*

No.

*Does the dataset contain information that might be considered inappropriate or offensive?*

No.