# Words that make SENSE:
# Sensorimotor Norms in Learned Lexical Token Representations

**Abhinav Gupta, Jesse Thomason, Toben H. Mintz**
University of Southern California
{abhinavg, jessetho, tmintz}@usc.edu

## Abstract

While word embeddings derive meaning from co-occurrence patterns, human language understanding is grounded in sensory and motor experience. We present SENSE (**S**ensorimotor **E**mbedding **N**orm **S**coring **E**ngine), a learned projection model that predicts Lancaster sensorimotor norms from word lexical embeddings. We also conducted a behavioral study where 281 participants selected which among candidate nonce words evoked specific sensorimotor associations, finding statistically significant correlations between human selection rates and SENSE ratings across 6 of the 11 modalities. Sublexical analysis of these nonce words selection rates revealed systematic phonosthemic patterns for the `interoceptive` norm, suggesting a path towards computationally proposing candidate phonosthemes from text data.

## 1 Introduction

Grounded cognition theory posits that humans rely on multimodal representations from perceptual, motor, and introspective experiences for cognitive activities (Barsalou, 2008) . Empirical studies suggest that comprehending action and perceptual concepts elicits rapid, automatic activity in modality-specific brain areas (Barsalou, 2008; Hauk et al., 2004). The Lancaster sensorimotor norms dataset quantifies these sensorimotor associations by averaging ratings from 3,500 participants for 39,707 English words across six perceptual modalities (`auditory`, `gustatory`, `visual`, `interoceptive`, `haptic`, and `olfactory`) and five action effectors (`hand/arm`, `foot/leg`, `head`, `torso`, and `mouth/throat`) (Lynott et al., 2019).

Systematic form-meaning associations called phonesthemes are sublexical units carrying consistent associations across words, for example, "gl-" in glitter, gleam, and glow (Bergen, 2004). Sublexical processing is crucial for unfamiliar words and language learning (Ubellacker and Hillis, 2022; Indefrey and Davidson, 2009), but whether neural language representations encode such correspondences is unclear (Abramova et al., 2013).

Current language models rely on word embeddings learned from co-occurrence patterns, and are not explicitly trained to encode sensorimotor information. Studies have found correlations between word embeddings and human embodied, perceptual and introspective experiences (Utsumi, 2020; Louwerse et al., 2008; Lenci et al., 2018). However, most prior work has focused primarily on emotional information (Lenci et al., 2018), evaluated small word sets (e.g., approximately 500 words (Utsumi, 2020)), and lacked human validation studies.

We address these limitations by developing SENSE, projecting from the learned lexical embeddings of over 34k Lancaster words to their 11 human-annotated sensorimotor norm dimensions, and explore these predictions through behavioral experiments nearly 300 participants.

## 2 Experiments

We consider three hypotheses:

**Hypothesis 1.** Lexical embeddings $\mathbf{e}_w$ of words $w$ implicitly encode sensorimotor norms $\mathbf{s}_w$, such that there exists a learned function $f : \mathbb{R}^d \rightarrow \mathbb{R}^{11}$ where $f(\mathbf{e}_w) \approx \mathbf{s}_w$ with lower error than a baseline predictor $f_{\text{baseline}}(\mathbf{e}_w) = \bar{\mathbf{s}}$, where $\bar{\mathbf{s}} = \frac{1}{N}\sum_{i=1}^{N} \mathbf{s}_i$.

**Hypothesis 2.** Projection model $f$ captures systematic form-meaning correspondences that generalize sensorimotor associations to nonce words $w^*$ such that $f(\mathbf{e}_{w^*})_m \approx \mathbf{s}^m_{w^*}$ for modality $m$, where $s^m_{w^*}$ is estimated with by human responses.

**Hypothesis 3.** Given $\mathbf{H_2}$ for modality $m$, character $n$-grams $c$ should exhibit systematic phonosthemic patterns where the human selection rate $P_H(w^* \rightarrow m \mid c \in w^*) \propto f(\mathbf{e}_c)_m$.

We present the **S**ensorimotor **E**mbedding **N**orm **S**coring **E**ngine (SENSE), which learns the projection function $f$ ($\mathbf{H_1}$). Then we conducted a human study to investigate whether $f$ can generalize its projected sensorimotor associations to nonce words ($\mathbf{H_2}$), and whether those projections maintain correlation with human ratings at the character $n$-gram level ($\mathbf{H_3}$). All associated data, code, and study protocols will be made publicly available.

### 2.1 SENSE Projections

The Sensorimotor Embedding Norm Scoring Engine (SENSE) projects lexical word embeddings onto the Lancaster Sensorimotor Norms. SENSE takes an embedding vector $\mathbf{e}_w \in \mathbb{R}^d$ as input and predicts sensorimotor norms $f(\mathbf{e}_w) \in \mathbb{R}^{11}$, where $0 \leq f(\mathbf{e}_w)_m \leq 1$ represents the predicted rating for word $w$ and modality $m$.[1]

We selected words and phrases present in each of the Word2Vec, GloVe, and Lancaster Norms vocabularies, averaging constituent word vectors for multi-word phrases. This selection yielded 34,110 aligned entries out of 39,707 Lancaster concepts, each with a sensorimotor vector and corresponding Word2Vec and GloVe lexical embeddings. For the BERT model, lexical embeddings were obtained as the CLS representation of each word or phrase passed through the model alone without sentential context. We randomly partitioned the selected words into training (70%), development (15%), and test (15%) sets.

We compared three architectures: a baseline predicting mean training set sensorimotor vector $\bar{\mathbf{s}}$; $k$-NN with $k$=5 using cosine similarity with weighted averaging to predict $\hat{s}_w$ from the neighborhood $s_{NN(e_w)}$; and a feed-forward neural network with one hidden layer, 64 or 128 neurons tuned on development set, ReLU activation, and 11-dimensional output, trained using Adam optimization with learning rate 0.001 for 10 epochs and batch size 128. We trained separate models for each embedding type with input dimensionality $d$ of 300 for Word2Vec, 100 for GloVe, and 768 for BERT CLS. All models were evaluated using MSE on the held-out test set.

### 2.2 Human Experiment

To conduct experiments related to $\mathbf{H_2}$ and $\mathbf{H_3}$, we selected the neural network architecture with BERT CLS embeddings as the final SENSE model due to consistent performance between learned embeddings and the flexibility to embed nonce words and character sequences.

| **Nonce word** $w^*$ | **Modality** $m$ | **SENSE** $\hat{s}^m_{w^*}$ |
|---|---|---|
| crilollering | Auditory | 1.00 |
| caeduseousness | Interoceptive | 0.79 |
| rehotes | Visual | 0.90 |
| sweadles | Hand/Arm | 0.75 |

Table 1: Sample pseudowords generated from Wuggy for the human study dataset with high corresponding modality norm predictions by SENSE.

We conducted an IRB-approved human study surveying 281 undergraduate students about which nonce words evoked specific sensorimotor associations, allowing us to compare these annotations to SENSE predictions. Nonce words were generated using the Wuggy Pseudoword Generator (Keuleers and Brysbaert, 2010), which creates pronounceable pseudowords by preserving sub-syllabic structure and transition frequencies of real English words. We used 63,975 seed words from Wuggy's lexicon, sourced from the 306,128-word Moby Word List (Ward, 2002), and generated 10 candidates per seed word using Wuggy's classic generator with a 2/3 sub-syllabic segment overlap ratio. We retrieved BERT CLS embeddings for all generated nonce words and used SENSE to compute $f(\mathbf{e}_{w^*})_m$ for each nonce word $w^*$ and modality $m$. To enforce lexical novelty, we excluded nonce words within Levenshtein distance 1 of, sharing stems with, or homophonous to Moby Dictionary entries. We then selected the 12 nonce words with highest $f(\mathbf{e}_{w^*})_m$ for each modality $m$, ensuring $f(\mathbf{e}_{w^*})_m > 0.5$ for all selections. Table 1 shows sample nonce words selected for each modality.

Using Qualtrics, we created a survey with 4 questions per modality asking: "Which 3 of the following nonsense words do you think most relate to the [modality]?" Each question presented 3 target words from the 12 selected for modality $m$ and 4 distractors words selected by $f(\mathbf{e}_{w*})_{m'} > 0.5, m' \neq m$ and with $f(\mathbf{e}_{w*})_m < 0.5$. For example, two of the options for Q1, "Which 3 of the following nonsense words do you think most relate to the sense of hearing?", were: "crilollering" ($w_1^*$) and "ancechuttos" ($w_2^*$), with $m = \texttt{auditory}$ and $m' = \texttt{gustatory}$, we have $f(\mathbf{e}_{w_1^*})_m = 1$ and $f(\mathbf{e}_{w_2^*})_m = 0$ while $f(\mathbf{e}_{w_2^*})_{m'} = 0.98$. The 3/7 ratio of nonce words with predicted modality $m$ in question means a random true positive label can be drawn with 42.9% probability, below chance. Each participant an-

[1] We normalize the Lancaster norms from their original $[0, 5]$ range per dimension to $[0, 1]$ per dimension; note that the norms are not unit vectors.

| Modality | Character Set ($c$) | $P_H$ |
|---|---|---|
| `Auditory` | in | 0.565 |
| `Auditory` | er | 0.547 |
| `Interoceptive` | ni | 0.579 |
| `Interoceptive` | tion | 0.538 |

Table 2: Example sub-lexical components $c$ along with the probability $P_H = P_H(w^* \to m_L \mid c \to w^*)$ of a nonce word containing $c$ the given sensory modality $m$.

swered two questions per modality, and questions were balanced across participants.

### 2.3 Sublexical Analysis

We investigated which sublexical components $c$, character $n$-grams, increased the probability of a word evoking a particular sensory modality by computing $P_H(w^* \to m_L \mid c \in w^*)$ as the mean participant selection rate across all words containing $c$. We aim to quantify the relationship between this selection rate and SENSE's ratings for the same components $f(\mathbf{e}_c)_m$.

The `interoceptive` and `auditory` modalities showed the strongest pseudoword correlations, so we focus the sublexical analysis on these two modalities. We extracted all character $n$-grams of length 2-4 from the 28 nonce words in each modality, retaining only character $n$-grams $c$ where $\mid (w^* | c \in w^*) \mid \geq 3$ to ensure sufficient recurrence for statistical reliability. To eliminate redundancy, we removed sub-string $n$-grams that appeared in the exact same word set as all their containing super-strings. For each retained character set $c$ within modality $m$, we computed $P_H(w^* \to m_L \mid c \in w^*)$, eliminating character sets $c$ where $P_H \leq P_H(w^* \to m_L)$. This yielded 13 character sets for the `interoceptive` modality and 19 for `auditory`. Table 2 presents some of the selected character sets $c$ for which $P_H - P(w^* \to m_L) > 0.10$.

We then tokenized each character sets, obtaining $e_c$ and predicted SENSE ratings for each isolated character set $f(\mathbf{e}_c)_m$. We then computed Pearson correlations between $P_H(w^* \to m_L \mid c \to w^*)$ and $f(\mathbf{e}_c)_m$.

## 3 Results

SENSE results demonstrate that word embeddings encode sensorimotor information ($\mathbf{H_1}$), the human study shows that SENSE predictions generalize to nonce words in alignment with human behavioral judgments ($\mathbf{H_2}$), and sublexical analysis reveals systematic form-meaning associations in the `interoceptive` modality ($\mathbf{H_3}$).

| Embedding | Baseline | KNN | Neural Net |
|---|---|---|---|
| **Word2Vec** | 0.028 | 0.015 | 0.016 |
| **GloVe** | 0.028 | 0.018 | 0.017 |
| **BERT CLS** | 0.028 | 0.020 | 0.016 |

Table 3: Overall performance of different model architectures (Baseline, KNN, Neural Network) $\text{MSE}_{\text{avg}}$, across 3 different embedding types (Word2Vec, GloVe, BERT CLS). For each sensorimotor modality $m_j$, we calculated $\text{MSE}_{m_j} = \frac{1}{N}\sum_{w_i=1}^{N}(f(\mathbf{e}_{w_i})_{m_j} - \mathbf{s}_{w_i}^{m_j})^2$ and evaluate $\text{MSE}_{\text{avg}}$ as the unweighted average MSE

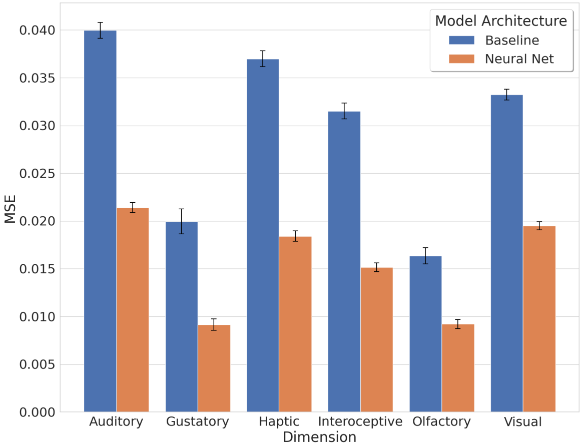

Figure 1: Mean squared error (MSE) for the six perceptual modalities comparing SENSE against the baseline model predicting mean training sensorimotor vector $\bar{\mathbf{s}}$ for all inputs. Error bars represent standard error. The neural network substantially outperforms the baseline across all modalities, with lowest errors for `gustatory` and `olfactory` and highest for `visual` and `auditory`.

### 3.1 SENSE

For each sensorimotor modality $m_j$, we calculated

$$\text{MSE}_{m_j} = \frac{1}{N}\sum_{w_i=1}^{N}(f(\mathbf{e}_{w_i})_{m_j} - \mathbf{s}_{w_i}^{m_j})^2,$$

and evaluated overall performance $\text{MSE}_{\text{avg}}$ as the unweighted average MSE across all 11 modalities.

Table 3 presents $\text{MSE}_{\text{avg}}$ across three embedding types $\mathbf{e}$ (Word2Vec, GloVe, BERT CLS) and three architectures $f$ (Baseline, KNN, Neural Network). Both KNN and neural architectures substantially outperformed the baseline, demonstrating learned projection functions can predict word-specific sensorimotor profiles rather than defaulting to average sensorimotor ratings based on their training data.

Figure 1 compares $\text{MSE}_{m_j}$ for the SENSE projections from BERT CLS versus the baseline av-

| **Modality** | **Correlation ($r$)** |
|---|---|
| `Interoceptive` | 0.73*** |
| `Auditory` | 0.69*** |
| `Torso` | 0.57** |
| `Visual` | 0.56** |
| `Gustatory` | 0.54** |
| `Hand/Arm` | 0.43* |

Table 4: Significant correlations between human selection rate and SENSE predictions, sorted by $r$ value. Five modalities (Foot/Leg, Olfactory, Haptic, Head, Mouth) showed non-significant correlations. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$.

erage for the 6 perceptual modalities. Notably, the lowest error was observed for the `gustatory` and `olfactory` modalities across all three embeddings, indicating that taste and smell related concepts are well-captured by word co-occurrence patterns. In contrast, `visual` and `auditory` dimensions showed the highest errors, but errors across all modalities were modest, suggesting that sensorimotor grounding is encoded in distributional semantics despite these models being trained solely on co-occurrence patterns.

Paired $t$-tests comparing per-word MSE between SENSE and the baseline model revealed significantly lower errors for the neural network across all 11 modalities, as well as overall at $p < .001$, providing strong support for $\mathbf{H_1}$.

### 3.2 Human Study

For each nonce word $w^*$ and modality $m$, we computed the human selection rate (proportion of participants selecting it as one of the seven given nonce words that evoked sensorimotor modality $m$) and SENSE's rating $f(\mathbf{e}_{w^*})_m$. Table 4 presents Pearson correlations between these measures for each modality that showed statistically significant correlations $r$ between human selection rate and SENSE predictions, sorted by $r$ value. Overall, SENSE predictions for 6 of the 11 modalities significantly correlated with human judgements, supporting $\mathbf{H_2}$.

Figure 2 illustrates the correlation for words in the `interoceptive` dimension between human selection rate and SENSE ratings, showing that nonce words SENSE rated highly were also frequently selected by human participants, while low-rated nonce words were rarely selected.

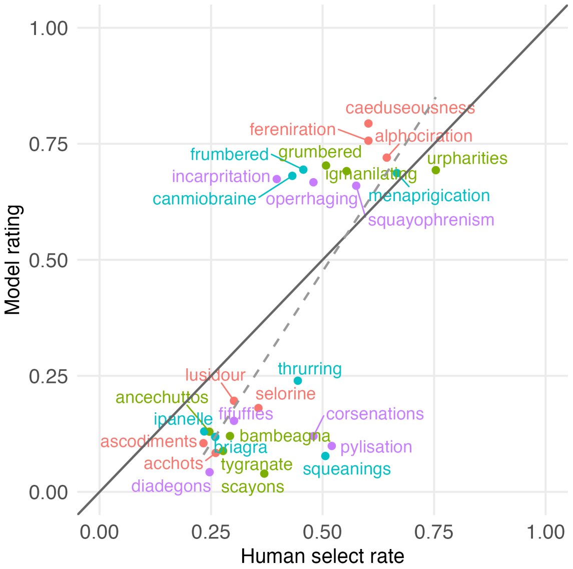

Figure 2: Correlation between the rate of human selection vs SENSE rating for nonce words shown to the humans under the `Interoceptive` category ($r = 0.73$).

### 3.3 Sublexical analysis

Since `interoceptive` and `auditory` modalities showed the strongest pseudoword correlations between human selection rates of nonce words and SENSE's ratings ($r > 0.65$), we focused the sublexical analysis on these two modalities and find partial support for $\mathbf{H_3}$.

`Interoception` showed a significant positive correlation between $P_H(w^* \to m_L \mid c \to w^*)$ and $f(\mathbf{e}_c)_m$ ($r = 0.630, p = .007$), indicating that character sets systematically associated with `interoceptive` experiences in human judgments are also captured in the model's embeddings. However, `auditory` showed no significant relationship ($r = 0.267, p = .562$), suggesting that strong whole-word sensorimotor associations do not always decompose into systematic character-level phonosthemic patterns.

## 4 Conclusion

We developed SENSE, a projection architecture that maps word embeddings onto the 11-dimensional Lancaster Sensorimotor Norms, demonstrating that distributional representations encode sensorimotor information with low prediction error across Word2Vec, GloVe, and BERT CLS embeddings. A behavioral study with 281 participants validated these predictions, showing strong correlations between human judgments and SENSE ratings for nonce words evoking specific sensorimotor experiences, with the strongest alignment in `interoception` ($r = 0.73, p < .001$). Sublexical analysis revealed systematic phonosthemic patterns for `interoceptive` experiences, establishing a computational framework for detecting form-meaning associations. This work bridges distributional semantic models and embodied cognition theory, providing evidence that language models trained solely on text capture the sensorimotor grounding essential to human language comprehension. However, the systematic form-meaning associations made by embedding models vary by modality, which provides oppurtinites to explore mimicing sublexical processing in language models.

## Limitations

This work has several limitations that suggest directions for future research. First, we evaluated only English words, limiting claims about cross-linguistic generalizability of sensorimotor encoding in distributional models. Second, our analysis examines orthographic forms rather than phonological representations. Phonesthemes are traditionally phonological phenomena, yet our analysis relies on written text due to the text-only training data underlying word embeddings and LLMs. The form-meaning associations we detect might be stronger with phonological representations that capture actual sound patterns. Third, while our sublexical analysis revealed systematic patterns for interoceptive phonesthemes, time constraints prevented conducting a targeted behavioral study to validate these specific form-meaning correspondences. Fourth, our behavioral validation used 281 undergraduate participants from a single university, which may not fully represent the broader population's sensorimotor associations. Finally, BERT's contextualized architecture was used to generate static word embeddings without sentence context, which may not fully leverage BERT's representational capabilities.

## References

Ekaterina Abramova, R. Fernández, and Federico Sangati. 2013. Automatic labeling of phonesthemic senses. In *Annual Meeting of the Cognitive Science Society*.

Lawrence W Barsalou. 2008. Grounded cognition. *Annu. Rev. Psychol.*, 59(1):617–645.

Benjamin K Bergen. 2004. The psychological reality of phonaesthemes. *Language*, 80(2):290–311.

Olaf Hauk, Ingrid Johnsrude, and Friedemann Pulvermüller. 2004. Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, 41(2):301–307.

P. Indefrey and D.J. Davidson. 2009. Second language acquisition. In Larry R. Squire, editor, *Encyclopedia of Neuroscience*, pages 517–523. Academic Press, Oxford.

Emmanuel Keuleers and Marc Brysbaert. 2010. Wuggy: A multilingual pseudoword generator. *Behavior Research Methods*, 42(3):627–633.

Alessandro Lenci, Gianluca E. Lebani, and Lucia C. Passaro. 2018. The emotions of abstract words: A distributional semantic analysis. *Topics in Cognitive Science*, 10(3):550–572.

Max M. Louwerse, Arthur C. Graesser, Danielle S. McNamara, and Shulan Lu. 2008. Embodied conversational agents as conversational partners. *Applied Cognitive Psychology*, 23(9):1244–1255.

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*, 52(3):1271–1291.

Delaney M. Ubellacker and Argye E. Hillis. 2022. Chapter 12 - the neural underpinnings of word comprehension and production: The critical roles of the temporal lobes. In Gabriele Miceli, Paolo Bartolomeo, and Vincent Navarro, editors, *The Temporal Lobe*, volume 187 of *Handbook of Clinical Neurology*, pages 211–220. Elsevier.

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

Grady Ward. 2002. Moby word lists.

## A Sample Lancaster Words and their sensorimotor ratings

Figure 3 contains examples from the Lancaster Sensorimotor Norms dataset, showing how words are rated across 11 sensorimotor dimensions: Auditory, Gustatory, Haptic, Interoceptive, Olfactory, Visual (perceptual), and Foot/leg, Hand/arm, Head, Mouth, Torso (action effectors).

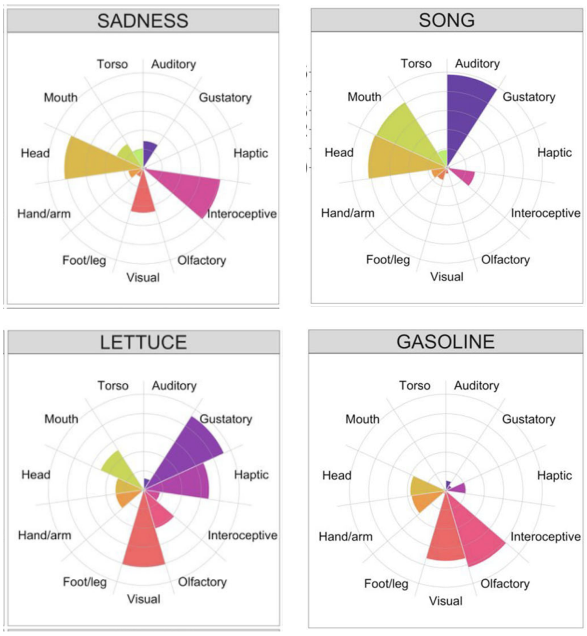

Figure 3: Sample words from the Lancaster Sensorimotor Norms dataset showing ratings across 11 dimensions (6 perceptual modalities and 5 action effectors).

## B Per-Modality Correlation Plots

Figures 4–13 show correlations between human selection rates and SENSE ratings for the 10 remaining modalities tested in the behavioral study.

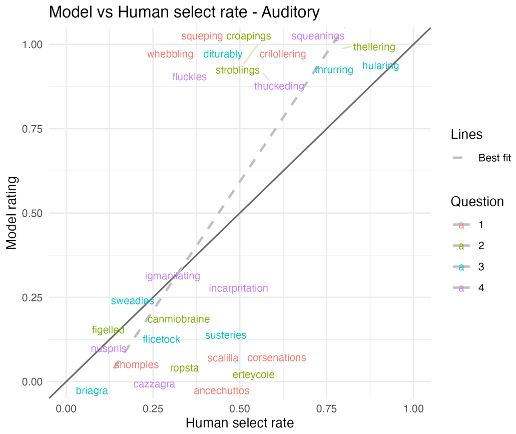

Figure 4: Correlation between human selection rate and SENSE ratings for nonce words in the `auditory` modality ($r = 0.69, p < .001$).

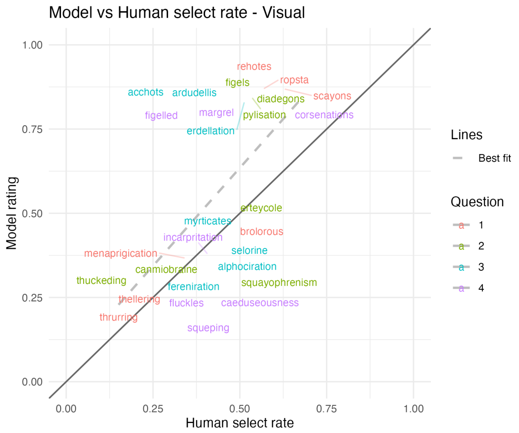

Figure 5: Correlation between human selection rate and SENSE ratings for nonce words in the `visual` modality ($r = 0.56, p = .002$).

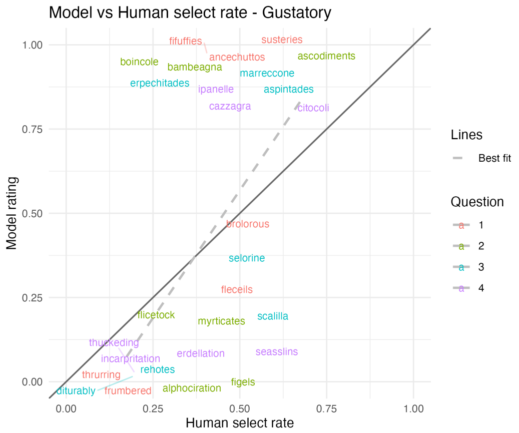

Figure 6: Correlation between human selection rate and SENSE ratings for nonce words in the `gustatory` modality ($r = 0.54, p = .003$).

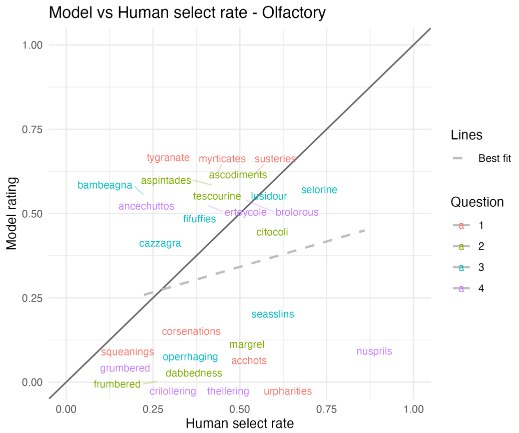

Figure 7: Correlation between human selection rate and SENSE ratings for nonce words in the `olfactory` modality ($r = 0.18, p = .349$).

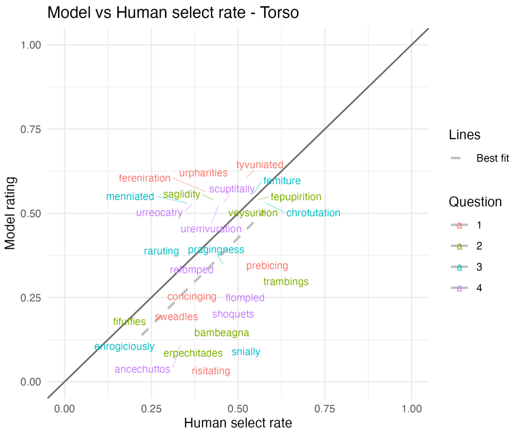

Figure 8: Correlation between human selection rate and SENSE ratings for nonce words in the torso modality ($r = 0.57, p = .002$).

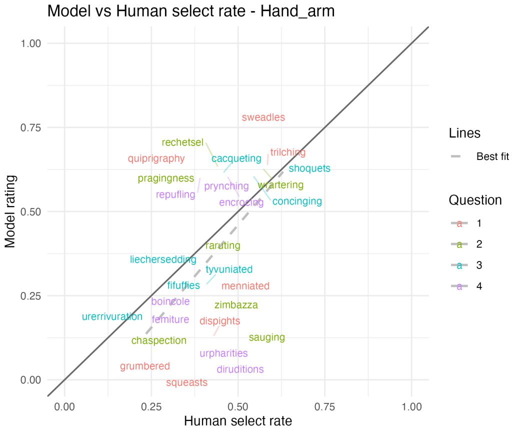

Figure 9: Correlation between human selection rate and SENSE ratings for nonce words in the hand/arm modality ($r = 0.43, p = .021$).

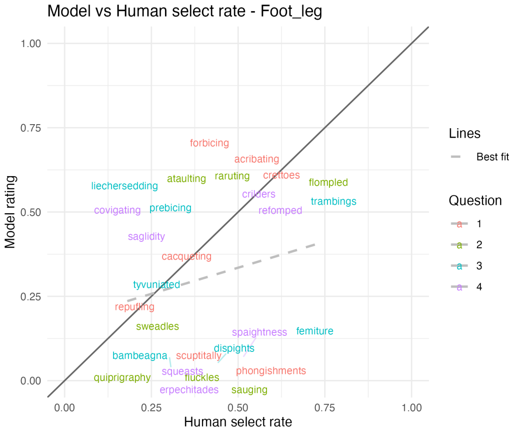

Figure 10: Correlation between human selection rate and SENSE ratings for nonce words in the foot/leg modality ($r = 0.25, p = .192$).

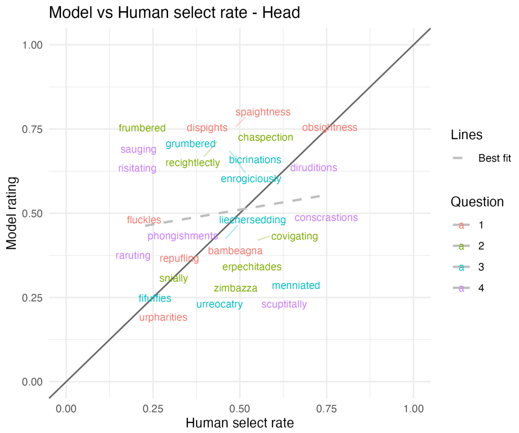

Figure 11: Correlation between human selection rate and SENSE ratings for nonce words in the head modality ($r = 0.10, p = .618$).

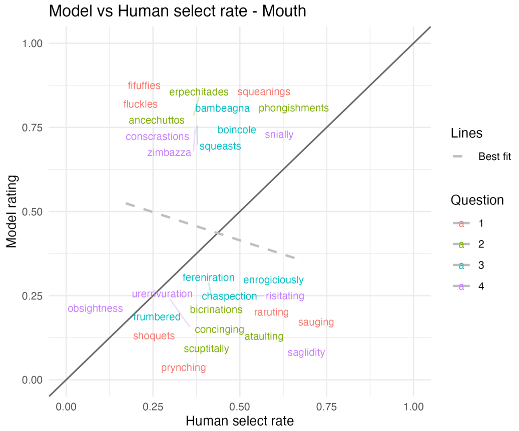

Figure 12: Correlation between human selection rate and SENSE ratings for nonce words in the mouth modality ($r = -0.14, p = .474$).

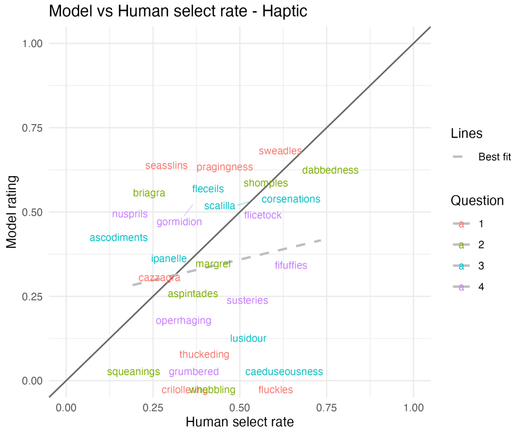

Figure 13: Correlation between human selection rate and SENSE ratings for nonce words in the haptic modality ($r = 0.16, p = .408$).