# Bright Pink Flamingos and Bright Pink Secretaries: Meaning from Unexpected Input

**Harshada Vinaya**[1], **Ben D. Amsel**[1], **Marta Kutas**[1] **& Seana Coulson**[1]
[1]Department of Cognitive Science, University of California, San Diego

## Abstract

Understanding how different sources of our semantic knowledge contribute to language comprehension, especially under unpredictable conditions, remains a challenge. Using vision-language CLIP, and language-only GPT2 representations, we examined whether these information sources differentially predict single-trial EEG responses to expected (i.e., high cloze probability) and unexpected (i.e., near-zero cloze probability) words in constraining sentence contexts. We also assess whether these effects vary across semantic processing by analyzing 100-ms intervals from 200-700 ms post word onset. GPT2 provided stronger fits for both expected and unexpected words from 300 to 500 ms. CLIP explained variance beyond GPT2 from 300 ms for expected, and from 400 ms for unexpected words. For the unexpected words, CLIP performed better than GPT2 in the 500-600 ms interval. Results show that both pure distributional and vision-informed language semantic information explain unique variance in EEG responses, with systematic differences due to word predictability as well as processing time.

**Keywords: unexpected word processing, distributional semantics, embodied semantics**

## Introduction

Humans comprehend an extraordinary range of language with remarkable ease - be it literal or figurative, familiar or novel, real or fictional, expected or surprising, and a key goal of cognitive science of language is characterizing the representational structures that make such all-inclusive comprehension possible. Across a range of theories, these structures are argued to arise from a combination of our perceptual, motor, and affective experience with the world, as well as from the statistics of language, although accounts differ regarding the relative importance of these sources (Kemmerer, 2019; Winkielman et al., 2023).

Comprehension events when humans cope with expectancy violations can provide a powerful window into our representational structures, because they stress the system and reveal how meaning is constructed when conditions are suboptimal. Event-related potential (ERP) language research has documented neural differences in the processing of expected versus unexpected words, namely in the amplitude of the N400 ERP component (DeLong et al., 2011; Kutas and Federmeier, 2011). This research establishes that contextual expectations shape online comprehension, but leaves open intriguing questions about what information the brain draws on in the contextual integration of an unexpected word, and whether and how this information differs from that used when integrating a more expected word.

Some studies show that our embodied knowledge may be needed to process conceptual combinations that are unsupported by the statistical properties of language, such as novel compounds (Connell and Lynott, 2011), less accessible features (Hoenig et al., 2008) or false attributes of concepts (Vinaya et al., 2025). Vinaya et al., (2025) contrasted two *kinds* of information - distributional and sensorimotor - using representations derived from the GloVe model (Pennington et al., 2014) and the Lancaster Sensorimotor Norms (Lynott et al., 2020), respectively, and modeled electroencephalography (EEG) responses on a property verification task from the onset of the property term to 800 ms after. They showed that although both distributional and sensorimotor information measures made independent contributions, after 200 ms distributional provided a better fit than sensorimotor for the true trials (e.g., apple–red), but sensorimotor matched or outperformed distributional across a majority of time windows for the false trials (e.g., apple–black).

In this paper, we extend this investigation from isolated word-pairs to more naturalistic sentence processing. We employ contextual embeddings from a vision language model (VLM, CLIP here) to operationalize vision and language-informed semantic information, along with an architecturally matched text-only model (GPT2) to operationalize language-only distributional information. We then analyze 100-ms averaged event-related potentials (ERPs) elicited by expected and unexpected words in relatively high constraining sentence contexts from Amsel et al., (2015) to test whether and how the recruitment of each *kind* of information differs between the expected words and their unexpected counterparts.

### Vector-based Representational Structures

Representing meanings as embeddings learned from large-scale language models has become a valuable way of studying human semantic representations. Crucially, these models produce context-sensitive embeddings that allow word meanings to be characterized as functions of their linguistic context rather than as static entries. Such vector embeddings also make it possible to operationalize contributions of multimodal semantic knowledge by combining different sources of information (such as vision and language in CLIP) in a unified representational space where semantic relationships can be quantified and compared (Piantadosi et al., 2023).



In G. Boleda, I. Dautriche, M. Maldonado, C. Saldana and T. Torrent (Eds.), *Proceedings of the 48th Annual Conference of the Cognitive Science Society* 

In this study, we derive contextualized representations from the text encoder of CLIP-Large and from GPT2-Small to examine how vision-informed and purely distributional language information are differentially recruited for expected and unexpected word continuations in sentential contexts.

**CLIP-Large** CLIP stands for Contrastive Language–Image Pre-Training and is a model trained on 400 million image-caption pairs with the objective of maximizing the similarity of matching pairs and minimizing the similarity of mismatched ones (Radford et al., 2021). This training makes CLIP learn conceptual representations that are informed by both language statistics and visual knowledge. The CLIP architecture has separate text and image encoders and maintains distinct representations of the modalities until they are projected into a final shared embedding space. Previous work evaluating models using behavioral tasks designed to study embodied simulation in humans shows that CLIP's textual representations are sensitive to shape and color information, suggesting its image training may induce sensitivity to some visual knowledge (Jones et al., 2024).

**GPT2-Small** Measurements from the GPT2-small model were included as a purely distributional information counterpart. The choice of GPT2-Small is based on the observation that it is architecturally similar to CLIP's text encoder, with a comparable transformer depth, hidden dimensionality, and number of attention layers, but is trained only on text (Radford et al., 2021). The inclusion of embeddings from GPT2 may thus allow us to dissociate effects driven by distributional language statistics from those attributable to CLIP's multimodal (vision–language) training.

## The Present Study

We use CLIP and GPT2 models to capture the representational distance between the context and the **expected** and **unexpected** word continuations, and construct linear mixed effects regression (LMER) models to predict ERP responses for these words in sentences. To assess whether CLIP and GPT2 measures are differentially relevant for expected versus unexpected word processing over time, we analyze ERPs averaged for 100-ms windows starting 200-300 ms to 600-700 ms windows from the word onset. Model fits are compared using Akaike Information Criterion (AIC) scores within each time window and condition (Burnham and Anderson, 2004). We predict that while both GPT2 and CLIP measures may explain unique variance in EEG, they may recruit these sources differently for the expected and unexpected word conditions.

(a) For the expected words, GPT2-based information measure will provide a better model fit than CLIP-based measure.

(b) If unexpected words rely more on sensorimotor information as proposed, CLIP-based measures will improve model fit beyond GPT2-based measures for the unexpected words.

# Materials and Methods

**Materials** This study uses a subset of data from the work reported by Amsel et al., (2015); the specifics of the methodology can be found in that publication. Briefly, each participant read a total of 136 sentence pairs. The first sentence provided the context of the situation (e.g., "Amber was excited to go to the zoo with her family.") and the second continued to set the expectation for a noun (e.g., "She couldn't wait to see the bright pink"). The second sentence was completed with four different nouns, marking the four conditions, though individual participants only viewed a single version of each stimulus. The present study utilizes data collected in the **expected** condition, viz. the most likely response on a separately administered sentence completion task ("flamingos"), and from the **unrelated** condition, a noun that was not produced on the norming task and that was incongruent with the prior context ("secretary"). Each participant read 34 sentence pairs in each condition. As our aim is to contrast the processing of highly expected and unexpected words, the **unrelated** trials from Amsel et al., (2015) will be referred to below as the **unexpected** condition.

**Participants** Data were collected from thirty two undergraduates from a large public university in the United States. All participants were right-handed, with normal or corrected vision and between 18-30 years of age. Data from three participants were excluded from the analysis due to a large number of movement artifacts.

**Procedure** The participants were seated in a dimly lit, sound attenuated chamber 112 cm from the computer screen. They were instructed to 'read each passage as you normally would read any text' (Amsel et al., 2015). The first context sentence was presented in its entirety, while the second sentence was presented one word at a time, both in the center of the screen. Each word in the second sentence was presented for 200 ms with a 300 ms inter-stimulus interval.

**EEG Recording** EEG was recorded using 26 tin electrodes and referenced to the left-mastoid. Eye movements were monitored using electrodes placed on the outer canthus and the infraorbital ridge of each eye. The signal was digitized online at 250 Hz, with a bandpass of 0.01 to 100 Hz.

**EEG Preprocessing** EEG was re-referenced offline to the average of two mastoids and bandpass filtered from 0.1 to 32 Hz. Eye movement artifacts were corrected using maximum-likelihood ICA and no more than two components removed per participant to avoid overcorrection. Target-word epochs (e.g., "flamingos", "secretary") were extracted from 100 ms before the onset of the word to 700 ms after. All epochs were baseline corrected using the 100 ms interval before the word onset. All EEG pre-processing was performed using the MNE package in Python (Gramfort et al., 2013).

## Analysis

We modeled single-trial EEG across five successive 100-ms interval averages, beginning 200 ms after the onset of the target word to 700 ms later. The **expected** and **unexpected** trials were analyzed separately *a propos* the research questions related to differences in comprehension processes with or without strong expectation violations, and also to avoid an overly complex three-way interaction in the models. In light of the separate modeling, we report the lexical properties of items. The completions across **expected** and **unexpected** conditions were matched during study design length, word frequency, and orthographic neighborhood (Table 1, Amsel et al., 2015). We further examined concreteness and imageability ratings. For the unexpected words, concreteness had a mean of 4.70 (SD = 0.38), compared to 4.78 (SD = 0.26) for the expected condition (Brysbaert et al., 2014). Of the 272 unique words in both conditions, imageability ratings were available only for 230 words (Coltheart, 1981). From the available ratings, the imageability of the unexpected words had a mean rating of 577 (SD = 49.10), while the mean for the expected words was 591 (SD = 36.08). Overall, this suggested comparable distributions of these predictors across conditions. Because imageability ratings were unavailable for all of our materials, and because they were correlated with concreteness ratings (expected: $r = 0.52$, $p < .001$; unexpected: $r = 0.73$, $p < .001$), we chose not to include them in our statistical models of the EEG data.

Together, from 29 participants, we collected 1972 total trials: 986 from each condition. 24 trials were rejected based on an EOG (±250 μV) and EEG (±150 μV) artifact criterion, and 22 were further excluded following visual inspection for residual artifacts (Luck, 2014). Overall, this resulted in a final dataset of 1926 epochs, 963 in **expected** and **unexpected** conditions each. We then fit five linear mixed-effects regression (LMER) models for each of our five 100-ms intervals and each condition, resulting in 25 models per condition. These models predicted averaged single trial EEG amplitude in the respective 100-ms intervals using a combination of logarithmic word frequency of the target word, and the CLIP and GPT2-based cosine distance measurements.

**EEG Measurements** We measured the mean EEG voltage in single-trial EEG epochs at each electrode in five successive 100-ms windows, starting 200-300 ms and ending 600-700 ms post word-onset. In our LMER models, we allowed our lexical-semantic predictors to interact with the X, Y, Z coordinates for each electrode (Winsler et al., 2018). Accordingly, we represented spatial information using three continuous variables: the horizontal (left/right) X axis, vertical (anterior/posterior) Y axis, and the Z axis (indexing depth, from top of the head to periphery).

**Predictors** The five LMER models we constructed, per condition and time window, employed the following three lexical-semantic predictors:

1. Log-transformed word frequency of the target word (e.g., "flamingos") was obtained from the SUBTLEX-US database (Brysbaert et al., 2012). For the **expected** condition, the mean was 2.76 (SD = 0.64), and for the **unexpected** condition, the mean was 2.88 (SD = 0.52).
2. CLIP-based cosine distance: Cosine distance is measured as $1 - \frac{\mathbf{A}\cdot\mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$ where A is the vector representation of the context (i.e., first and second sentence) and B is the vector representation of the context plus the target word (i.e., first and second sentence + target word). The distance between the context embedding and context-plus-target-word embedding extracted from CLIP model's text encoder was used as a measure of "representational shift" or change in embedding space caused by adding the target word, reflecting knowledge from visual information in conjunction with distributional language information. Vectors were obtained from the text encoder before projection and were mean-pooled across tokens in the input sequence.

   The pre-trained CLIP text encoder was accessed via the HuggingFace library, and embeddings were extracted from the final (12th) layer of the model. Raw cosine distances were small in magnitude, with a mean of 0.00020 (SD = 0.00008) for the **expected** condition, and a mean of 0.00020 (SD = 0.00009) for the **unexpected** condition.
3. GPT2-based cosine distance: We measured cosine distances using embeddings acquired from the last (12th) layer of the GPT2 small model (Radford et al., 2019), accessed via the HuggingFace library. Given GPT2's autoregressive architecture, the final token embeddings for the context and for the context-plus-target-word were used to compute cosine distances. The raw estimates of the GPT2-based cosine distances had a mean of 0.017 (SD = 0.008) for the **expected** condition and a mean of 0.019 (SD = 0.010) for the **unexpected** condition.

All predictors were z-scored for interpretability. We also examined the pairwise Pearson correlation within each condition data subset. Pearson correlations indicated a moderate negative association between CLIP-based cosine distance and logarithmic word frequency in the **expected** condition ($r = -0.33$, $p < 0.001$). All other pairwise correlations among GPT2-based cosine distance, CLIP-based cosine distance, and word frequency were small and non significant in both the **expected** and **unexpected** conditions.

**Competing Models** In our five LMER models predicting EEG voltages in each 100-ms window and each condition, the lexical-semantic predictors interacted with three continuous scalp-location dimensions: X, Y and Z, that enabled regressing the spatial aspects of the EEG data (Winsler et al., 2018). Moreover, an interaction of a predictor with each dimension informs about the nature of the topographic association the predictor has. For example, a predictor's positive estimate for interaction with the Y axis suggests more positive ERPs on the anterior sites compared to the posterior sites due to that specific predictor. All models also included random effects

for subjects.

1. The first is the **NULL** model that includes no lexical or semantic predictor, but only the scalp dimensions, and a subject-level random effect. This model serves as a reference for AIC visualization. The equation for the **NULL** model in LMER form is:

$$\text{voltage} \sim (X + Y + Z) + (1 \mid \text{subject})$$

2. The second model is the **BASE** or the baseline model that adds z-scored logarithmic word frequency to NULL.

$$\text{voltage} \sim (X + Y + Z) * \text{word\_frequency} + (1 \mid \text{subject})$$

3. The next is the **GPT2** model, which adds the z-scored GPT2-based cosine distance predictor to BASE.

$$\text{voltage} \sim (X + Y + Z) * (\text{word\_frequency} + \text{GPT2\_distance}) + (1 \mid \text{subject})$$

4. Then is the **CLIP** model, adding the z-scored CLIP-based cosine distance in the BASE model.

$$\text{voltage} \sim (X + Y + Z) * (\text{word\_frequency} + \text{CLIP\_distance}) + (1 \mid \text{subject})$$

5. The final is **GPT2 + CLIP** that add both GPT2 and CLIP-based scaled measures to the BASE model.

$$\text{voltage} \sim (X + Y + Z) * (\text{word\_frequency} + \text{GPT2\_distance} + \text{CLIP\_distance}) + (1 \mid \text{subject})$$

We use Akaike Information Criterion (AIC) scores for model comparison, a measure that penalizes model's likelihood according to the number of parameters in the model and discourages unnecessary model complexity. We take a reduction in AIC greater than 4 as evidence of improved model fit (Burnham and Anderson, 2004).

# Results

## Model Comparison

Figure 1 plots the model fit comparison using AIC values for both **expected** and **unexpected** conditions. AICs from all four models of interest, **BASE**, **GPT2**, **CLIP**, and **GPT2 + CLIP** are scaled relative to the **NULL** model's AIC by subtraction (i.e., $AIC_{\text{model}} - AIC_{\text{NULL}}$) for visualization. For each 100-ms window, we report which of the GPT2 and CLIP models provided a robustly better fit to the data, and whether the GPT2 + CLIP model improved over each of them.

1. 200-300 ms: In the **expected** condition, GPT2 provided the best fit to the data (ΔAIC to BASE = -5), but CLIP did not improve over BASE (ΔAIC to BASE = 2). In the **unexpected** condition, no model improved fit over BASE.

2. 300-400 ms: GPT2 provided a better fit than CLIP (ΔAIC to CLIP = −47) in the **expected** condition, and GPT2 + CLIP showed an improvement over GPT2 (ΔAIC = -7). In the **unexpected** condition, GPT2 improved the fit over CLIP (ΔAIC to CLIP = −13). However, GPT2 + CLIP did not improve over GPT2 (ΔAIC to GPT2 = 1)

3. 400-500 ms: In the **expected** condition, GPT2 again provided a better fit than CLIP (ΔAIC to CLIP = −37), and GPT2 + CLIP offered an improvement over GPT2 (ΔAIC to GPT2 = −16). Similarly, in the **unexpected** condition, GPT2 outperformed CLIP (ΔAIC to CLIP = −69), and GPT2 + CLIP over GPT2 (ΔAIC to GPT2 = −18).

4. 500-600 ms: GPT2 again outperformed CLIP (ΔAIC to CLIP = −12) and GPT2 + CLIP improved over GPT2 (ΔAIC to CLIP = −19) for the **expected** condition. For the **unexpected** condition, CLIP outperformed GPT2 (ΔAIC to GPT2 = −23) and GPT2 + CLIP improved over CLIP (ΔAIC to CLIP = −22) as well.

5. 600-700 ms: For the **expected** condition, GPT2 improved over CLIP (ΔAIC to CLIP = −35) and GPT2 + CLIP improved over GPT2 (ΔAIC to GPT2 = −22). For the **unexpected** condition, performances of GPT2 and CLIP were indistinguishable (ΔAIC = −1), but GPT2 + CLIP showed improvement over GPT2 (ΔAIC to GPT2 = −7).

## Predictors' Scalp Estimates

Interactions of lexical-semantic predictors and scalp- dimensions ["X", "Y", and "Z"] indicated how effects of these predictors varied across channels. We used the beta coefficients from these predictor-scalp dimension interactions in our GPT2 + CLIP model to estimate predictors' topographic effects. Figure 2 shows topographic estimates for each predictor at 100-ms time intervals where a model including that predictor showed a robust improvement over the NULL model; in cases where the BASE (word frequency) model showed no improvement over NULL, word frequency scalp distributions are not shown even though it is included as a control predictor in other models.

**Word Frequency** The estimated scalp distributions in Figure 2 reveal an anterior negativity as word frequency increases in the **expected** condition from 200-300 ms, and in the **unexpected** condition from 200-600 ms, that shift more frontocentral in 600-700 ms.

**GPT2** In the **expected** condition, Figure 2 shows increasing anterior positivity beginning at 200-300 ms that strengthens over time as GPT2-based cosine distance increases until the end of the epoch. For the **unexpected** condition, GPT2 distance elicited larger positivities on the centroparietal channels between 300 and 700 ms, with the strongest effect in 400-500 ms.

**CLIP** Figure 2 shows that in the **expected** condition, increasing CLIP-based distance is associated anterior positivities in 300-500 ms and more broadly distributed positivities in 500-700 ms. In the **unexpected** condition, CLIP-based cosine distance elicits more positive ERPs concentrated in the centroparietal channels from 400-700 ms, with the strongest effect in 500-600 ms.

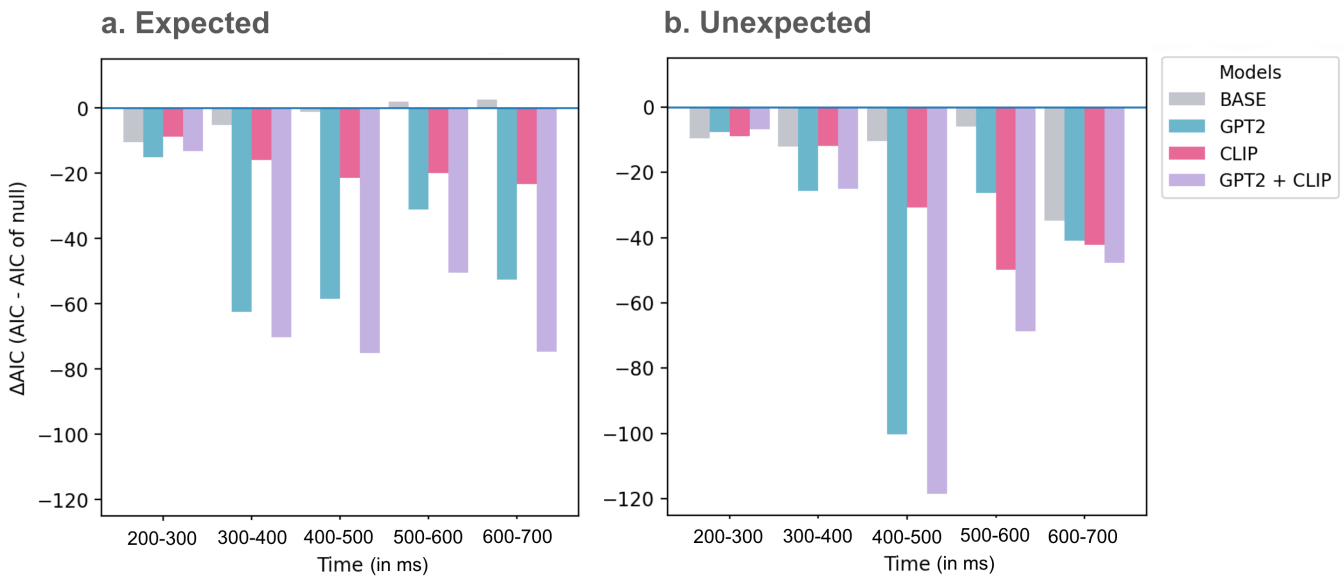


Figure 1: **EEG Model Comparison:** (a) shows the model comparison for the **expected** condition and (b) for the **unexpected** condition. The X-axis presents time, marking the five 100-ms windows, and the Y-axis presents scaled ΔAIC, i.e., AICs scaled to the NULL model AIC for each window and condition.

## Discussion

Using vision-informed CLIP model representations and language-distributional GPT2 representations, this study tested the proposal that distributional and sensorimotor information may differently contribute to processing of unexpected versus expected words, with sensorimotor information supporting unexpected word processing more than expected words. We fit LMER models for 100-ms averaged ERPs from 200 to 700 ms post word onset to also investigate whether CLIP and GPT2's relative contribution changes across semantic processing. We find that both GPT2 and CLIP-based measures make unique and context-sensitive contributions to expected and unexpected word processing. For both kinds of words, GPT2-based measure predicted ERPs better than CLIP-based measure in 300-500 ms; this effect continued for the expected words until the end of the epoch. However, after 500 ms, CLIP-based representations performed better or comparably to GPT2, only for the unexpected words.

### Multiple Representational Substrates

From 300 to 700 ms, overlapping the core semantic processing interval (viz. the N400 and P600 components), the best fit to the neural data in the **expected** condition is provided by the CLIP + GPT2 model, with GPT2-based measure fitting the data better than CLIP-based measure and CLIP explaining additional variance beyond GPT2. This pattern replicates findings from Vinaya et al., (2025), who showed that both GloVe and Sensorimotor norms explained unique variance for the true trials (e.g., apple-red), a parallel to the expected words condition here, with GloVe providing a better fit than Sensorimotor norms.

First, the prominence of purely distributional measure in predicting EEG responses to the **expected** condition is consistent with the idea that language-distributional information has a strong influence on semantic processing when contextual support is high (Connell, 2018; Ettinger and Linzen, 2016). Notably, however, CLIP's additional contribution shows that human semantic representations are informed both by a word's statistical regularities and by its associated visual properties. Even for highly predicted words, language-distributional information alone is therefore not sufficient for semantic processing, despite capturing substantial regularities about our world knowledge from linguistic input (Anderson et al., 2021; Fernandino et al., 2022; Jones et al., 2022). The topographic estimates of regression coefficients for GPT2-based measure differ from those of CLIP-based measure in the expected condition (Figure 2), further supporting the idea that these predictors may be associated partially non-overlapping neural resources.

In the **unexpected** condition, a similar pattern was observed from 300-500 ms, with GPT2 model providing a better fit than the CLIP model, suggesting distributional information makes dominant contributions to EEG responses in the earlier parts of semantic processing irrespective of contextual predictability. CLIP-based cosine distance explained additional variance beyond GPT2 from 400-500 ms, suggesting sensory information also contributes to meaning processing for unexpected words during the peak semantic processing, viz. the N400 component.

Word frequency is a particularly strong predictor for the **unexpected** words throughout our analyses windows, eliciting an

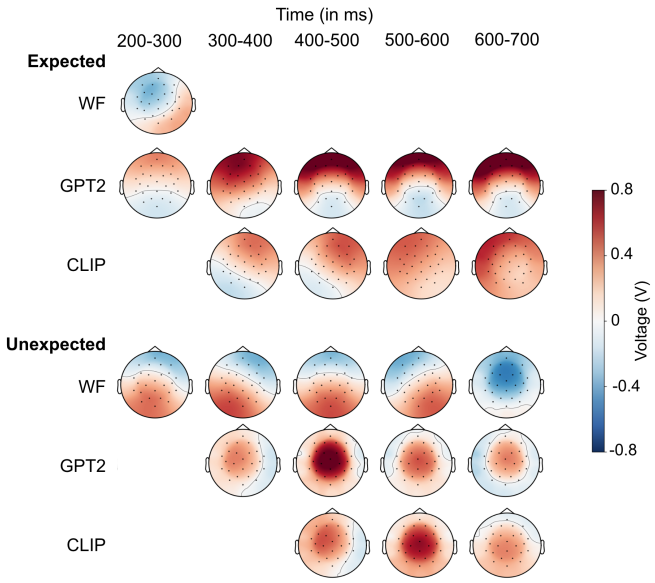


Figure 2: **Scalp Distributions of Predictor Estimates:** Estimate for each channel (i) were measured following the approach by Vinaya et al., (2025). For channel i, the estimate is measured as $\beta + \beta_x \cdot \text{Xi} + \beta_y \cdot \text{Yi} + \beta_z \cdot \text{Zi}$, where $\beta$ is the main effect coefficient for a predictor, $\beta_x$ the coefficient of interaction with scalp dimension X (and similarly, $\beta_y$ for Y, and $\beta_z$ for Z) and Xi, Yi, Zi are the coordinates of channel (i).

N400-like effect from 200-600 ms. Note that the effects look inverted (larger positive estimates in centroparietal channels during N400) due to the low-to-high coding of our continuous frequency predictor. The word frequency effect for the **unexpected** words is consistent with the reduced N400 amplitude as word frequency increases, especially in low-constraining contexts, such as in earlier compared to later word positions in sentences found in van Patten & Kutas et al., (1990).

## GPT2 Drives Early Lexical Access During N400

GPT2 model provides a better fit than CLIP model in 200-500 ms for the **expected**, and from 300-500 ms for the **unexpected** conditions. Theoretical accounts of the N400 component diverge on whether the component indexes difficulty in access or subsequent integration of word meaning with context, with evidence leaning towards the component reflecting access to stored information in long-term memory than post-access integration of the information (Lau et al., 2009).

GPT2 is trained to predict upcoming tokens, and thus the deepest (12th) layer used here, which is optimized for contextual prediction, may capture contextual dependencies that closely mirror the information available for lexical access (Radford et al., 2019). The presence of stronger GPT2 prediction in 300-500 ms, for both **expected** and **unexpected** words, suggest that these early lexical access mechanisms are sensitive to distributional information irrespective of the contextual predictability (Barsalou et al., 2008; Simmons et al., 2008).

## Do Visual Resources Help Unexpected Words?

One of the predictions we set out to test was whether sensory information confers an additional support for processing of unexpected words. In the **unexpected** condition, the GPT2 + CLIP model improved over the GPT2 model in 400-700 ms, suggesting CLIP explained unique variance even for words that could not be meaningfully integrated with the current context. CLIP provided better or comparable model fit performance than GPT2 after 500 ms for the **unexpected** words, such performance was not seen for the **expected** words. Moreover, CLIP's superior fit to data in 500-600 ms overlaps with the timing of enhanced sensorimotor contributions reported for false trials (e.g., apple-black) in Vinaya et al., (2025). This suggests that vision-informed semantic information contributed selectively during later stages of semantic processing when contextual predictions are violated. One possible interpretation is that unexpected words stress semantic retrieval processes, and more perceptually grounded semantic structure may be useful for accessing, evaluating, and integrating the unexpected word in the ongoing context.

## Conclusion

In this study, we tested whether vision-informed CLIP representations and language-distributional GPT2 representations differentially predict ERPs for expected and unexpected words in sentence contexts, and whether CLIP representations contribute more strongly to the processing of unexpected words. We operationalized sensorimotor semantics using CLIP and a distributional counterpart using GPT2, and modeled 100-ms averaged single-trial EEG responses separately for expected and unexpected words. GPT2-based representations provided stronger predictions for both expected and unexpected words from 300-500 ms. CLIP-based representations explained additional variance beyond GPT2-based representations for the expected words from 300 ms until the end of the epoch, and likewise for unexpected words from 400 ms until the end of the epoch. Critically, CLIP-based representations performed *better* than GPT2 500-600 ms after the onset of unexpected words. These findings suggest that both types of information contribute to semantic processing in a context-sensitive manner, and reveal a limited, temporally specific sensorimotor advantage for the processing of unexpected words.

## References

Amsel, B. D., DeLong, K. A., & Kutas, M. (2015). Close, but no garlic: Perceptuomotor and event knowledge activation during language comprehension. *Journal of Memory and Language*, *82*, 118–132. https://doi.org/10.1016/j.jml.2015.03.009

Anderson, A. J., Kiela, D., Binder, J. R., Fernandino, L., Humphries, C. J., Conant, L. L., Raizada, R. D. S., Grimm, S., & Lalor, E. C. (2021). Deep artificial neural networks reveal a distributed cortical network encoding propositional sentence-level meaning [Epub 2021 Mar 22]. *Journal of Neuroscience*, *41*(18),

4100–4119. https://doi.org/10.1523/JNEUROSCI.1152-20.2021

Barsalou, L. W., et al. (2008). Language and simulation in conceptual processing [Online edition, Oxford Academic, 22 Mar. 2012]. In M. de Vega, A. Glenberg, & A. Graesser (Eds.), *Symbols and embodiment: Debates on meaning and cognition*. Oxford.

Brysbaert, M., New, B., & Keuleers, E. (2012). Adding part-of-speech information to the SUBTLEX-US word frequencies. *Behavior Research Methods*, *44*(4), 991–997. https://doi.org/10.3758/s13428-012-0190-4

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known english word lemmas. *Behavior Research Methods*, *46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Burnham, K. P., & Anderson, D. R. (2004). Multimodel inference: Understanding aic and bic in model selection. *Sociological Methods & Research*, *33*(2), 261–304. https://doi.org/10.1177/0049124104268644

Coltheart, M. (1981). The mrc psycholinguistic database. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, *33A*(4), 497–505. https://doi.org/10.1080/14640748108400805

Connell, L. (2018). What have labels ever done for us? the linguistic shortcut in conceptual processing [Received 24 Aug 2016, Accepted 20 Apr 2018, Published online: 11 May 2018]. *Journal of Cognitive Psychology*, *30*(8), 1308–1318. https://doi.org/10.1080/23273798.2018.1471512

Connell, L., & Lynott, D. (2011). Modality switching costs emerge in concept creation as well as retrieval. *Cognitive Science*, *35*(4), 763–778. https://doi.org/10.1111/j.1551-6709.2010.01168.x

DeLong, K. A., Urbach, T. P., Groppe, D. M., & Kutas, M. (2011). Overlapping dual erp responses to low cloze probability sentence continuations. *Psychophysiology*, *48*(9), 1203–1207. https://doi.org/10.1111/j.1469-8986.2011.01199.x

Ettinger, A., & Linzen, T. (2016). Evaluating vector space models using human semantic priming results. *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, 72–77. https://doi.org/10.18653/v1/W16-2513

Fernandino, L., Tong, J., Conant, L. L., Humphries, C. J., & Binder, J. R. (2022). Decoding the information structure underlying the neural representation of concepts. *Proceedings of the National Academy of Sciences*, *119*(6), e2108091119. https://doi.org/10.1073/pnas.2108091119

Gramfort, A., Luessi, M., Larson, E., Engemann, D., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*.

Hoenig, K., Sim, E.-J., Bochev, V., Herrnberger, B., & Kiefer, M. (2008). Conceptual flexibility in the human brain: Dynamic recruitment of semantic maps from visual, motor, and motion-related areas. *Journal of Cognitive Neuroscience*, *20*(10), 1799–1814. https://doi.org/10.1162/jocn.2008.20123

Jones, C. R., Bergen, B., & Trott, S. (2024). Do multimodal large language models and humans ground language similarly? *Computational Linguistics*, *50*(4), 1415–1440. https://doi.org/10.1162/coli_a_00531

Jones, C. R., Chang, T. A., Coulson, S., Michaelov, J. A., Trott, S., & Bergen, B. (2022). Distributional semantics still can't account for affordances. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *44*.

Kemmerer, D. (2019). *Concepts in the brain: The view from cross-linguistic diversity*. Oxford University Press.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). *Annual review of psychology*, *62*, 621–647. https://doi.org/10.1146/annurev.psych.093008.131123

Lau, E., Almeida, D., Hines, P. C., & Poeppel, D. (2009). A lexical basis for n400 context effects: Evidence from meg. *Brain and Language*, *111*(3), 161–172. https://doi.org/10.1016/j.bandl.2009.08.007

Luck, S. J. (2014). *An introduction to the event-related potential technique* (2nd ed.). MIT Press.

Lynott, D., Connell, L., Brysbaert, M., Brand, J., & Carney, J. (2020). The Lancaster Sensorimotor Norms: Multidimensional measures of perceptual and action strength for 40,000 English words. *Behavior Research Methods*, *52*(3), 1271–1291. https://doi.org/10.3758/s13428-019-01316-z

Pennington, J., Socher, R., & Manning, C. (2014, October). GloVe: Global vectors for word representation. In A. Moschitti, B. Pang, & W. Daelemans (Eds.), *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532–1543). Association for Computational Linguistics. https://doi.org/10.3115/v1/D14-1162

Piantadosi, S. T., Muller, D. C. Y., Rule, J. S., Kaushik, K., Gorenstein, M., Leib, E. R., & Sanford, E. (2023). Why concepts are (probably) vectors. *Trends in Cognitive Sciences*, *27*(12), 1018–1032. https://doi.org/10.1016/j.tics.2023.08.011

Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. (arXiv:2103.00020). https://doi.org/10.48550/arXiv.2103.00020

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., & Sutskever, I. (2019). Language models are unsupervised multitask learners.

Simmons, W. K., Hamann, S. B., Harenski, C. L., Hu, X. P., & Barsalou, L. W. (2008). Fmri evidence for word association and situated simulation in conceptual processing. *Journal of Physiology-Paris*, *102*(1–3), 106–119.

Van Petten, C., & Kutas, M. (1990). Interactions between sentence context and word frequency in event-related brain potentials. *Memory & Cognition*, *18*(4), 380–393. https://doi.org/10.3758/BF03197127

Vinaya, H., Trott, S., Pecher, D., Zeelenberg, R., & Coulson, S. (2025). Words and worlds both: Dynamic effects of distributional and sensorimotor information in semantic processing. *Open Mind*, *9*, 2149–2174. https://doi.org/10.1162/OPMI.a.316

Winkielman, P., Davis, J. D., & Coulson, S. (2023). Moving thoughts: Emotion concepts from the perspective of context dependent embodied simulation. *Language, Cognition and Neuroscience*, *38*(10), 1531–1553. https://doi.org/10.1080/23273798.2023.2236731

Winsler, K., Midgley, K. J., Grainger, J., & Holcomb, P. J. (2018). An electrophysiological megastudy of spoken word recognition. *Language, Cognition and Neuroscience*, *33*(8), 1063–1082. https://doi.org/10.1080/23273798.2018.1455985