Encoding Semantic Scene Descriptions: Cross-Modal Representational Alignment from Language to Vision

Torrey Snyder

August 14, 2024

Abstract

Recent research in computational cognitive neuroscience has found that sentence embeddings of image captions can predict visual cortical responses to corresponding natural scenes. The study presented here expands upon this finding of cross-modal representational alignment by investigating how manipulating captions affects their neural predictivity of the brain's visual system. The primary manipulation of interest is semantic underspecification – in which disambiguating information is removed. For example, in comparison to the sentence "Alice threw a ball", the modified form "They threw a ball" is underspecified because the identity of the subject is unclear. In addition to this semantic intervention, this study also compared the effects of syntactic manipulations, such as scrambling word sequence. This study investigated how these manipulations affected sentence embeddings' alignment with fMRI brain data obtained from the Natural Scenes Dataset. Prior work has demonstrated that syntactic and semantic manipulations reduce neural alignment with brain regions involved in linguistic processing (Oota et al., 2024). This investigation examined how that alignment is affected cross-modally – to what extent can encoding models trained on manipulated caption embeddings predict the visual system's responses to the captions' corresponding images? Consistent with prior findings (Kauf et al., 2024), our findings revealed that syntactic manipulations, in the form of word scrambling, had a negligible effect on neural predictivity. On the other hand, semantic manipulations indeed resulted in a reduction in neural predictivity of the brain's vision network, suggesting that the representations learned by the sentence transformer model are more sensitive to semantic, rather than syntactic, information modifications.

1 Introduction

To construct rich semantic representations of natural scenes requires not only object recognition, but also a mechanism for encoding the relations between objects. Given that these interrelations can be explicitly articulated via language (in the form of image captions), this research aimed to investigate the representational alignment between sentence embeddings and neural activations of the brain's visual cortex in response

to the presentation of the captions' corresponding images. A recent study by Doerig et al. (2022) investigated the alignment between the representational spaces of language models and the visual cortex, building upon a breadth of research in cognitive computational neuroscience which implements representational similarity metrics to compare high-dimensional vectors derived from both brain recordings and ANNs (Kriegeskorte, 2015). In addition to these RSA-based approaches, encoding models to predict fMRI brain activity have become an increasingly frequent methodological technique in recent years, particularly as a tool for evaluating representational alignment.

The present study applied encoding models to investigate if caption manipulations affect their predictivity of neural activations in the brain's visual system in response to presentation of the captions' corresponding images. Enumerated below are the manipulations of interest:

1. scrambling
2. Subject Phrase (SP) underspecification
3. Verb Phrase (VP) underspecification
4. content word removal

By implementing manipulations (2) and (3), this research reveals the degree to which reducing image captions' referential specificity (how precisely informative a description is) perturbs LLM embeddings' alignment with visuo-semantic representations in the brain. Specifically, the subject/object caption manipulations reveal the role of referential specificity in semantic encodings. In comparison, the scrambling manipulations reveal the contribution of syntactic structure, consistently with previous findings from Kauf et al. (2024). Manipulation (4) involves the removal of all words belonging to a particular part-of-speech (POS) category (nouns/verbs). Recent work has explored how semantic underspecification affects the semantic embeddings generated by transformer models (Pezzelle, 2023). Investigating whether the cortical representation of a scene can be mapped from LLM embeddings (generated from underspecified image captions) provides an empirical assessment of the cross-modal representational alignment between language models and the brain. Regions of interest (ROIs) for the present study include the early visual cortex (V1-V4) and higher-level functional regions involved in scene and body recognition.

Prior work has demonstrated that syntactic manipulations of captions, such as scrambling, affect embeddings' representational alignment with neural activations in the brain's language network (Oota et al., 2024). Furthermore, recent research suggests that semantic content, more than syntactic structure, contributes to embeddings' representational alignment (Kauf et al., 2024). This evidence suggests that it is primarily object and action entities that encode the semantics of scenes. Meanwhile, function words, such as determiners, conjunctions, and prepositions, while necessary for grammaticality, do not convey semantic information. This work examines the effect of semantic manipulations on caption embeddings' predictivity of neural responses to corresponding visual scenes. Among these semantic manipulations are the comparison

between subject/object underspecification and noun ablation. We predicted that, consistently with the findings of (Kauf et al., 2024), semantic manipulations should have a more substantial effect on representational alignment than syntactic manipulations.

2 Background

Recently, neurolinguistics research has found that the human language network – a set of brain regions that are selectively and robustly activated during language processing (Fedorenko et al., 2011) – exhibits similar levels of activation in both word-order-manipulated and intact sentences (Kauf et al., 2024). It should be noted that the word-order manipulations in Kauf et al. (2024) preserve pointwise mutual information (an information theoretic metric that quantifies the degree to which words predict one another). In other words, information about local syntactic dependencies was conserved (Mollica et al., 2020). It has been argued that humans' insensitivity to these local manipulations of word order – as observed in psycholinguistic studies measuring subject's comprehension of scrambled texts – can be explained in terms of computational efficiency – given finite cognitive capacity, extracting the relevant meaning from a sentence should not be dependent upon the exact word sequence (Hahn et al., 2022).

Representations generated by language models, particularly transformer architectures (Devlin, 2018), can predict neural responses in the language network via regression-based encoding models (Caucheteux & King, 2022; Hosseini et al., 2022). It has been suggested that this correspondence stems from the convergence of the ANNs' linguistic representations with those in the brain, despite key differences in their architecture and learning mechanisms (Caucheteux & King, 2022). Prior research has shown that sentence embeddings generated from LLMs exhibit a significant degree of representational alignment with neural activations in the brain's temporal and inferior gyrus, regions implicated in language processing (Oota et al., 2024). Unlike one-hot encoding models, which merely capture whether each lexical item is present in a given sentence, sentence transformers generate contextualized sentence-level vector representations. These sentence transformer models are fine-tuned with the objective function of pairwise sentiment analysis, maximizing the cosine similarity for similar sentence pairs (labeled 0) and minimizing the cosine similarity for dissimilar sentence pairs (labeled 2) (Reimers & Gurevych, 2019). The contextualized embeddings encoded by transformers are an instantiation of the distributional hypothesis (Abrusán et al., 2018). Distributional semantics is based on the distributional hypothesis, which states that similarity in meaning results in similarity of linguistic distribution: words that are semantically related, are used in similar contexts (Harris, 1954). Distributional semantics approximates linguistic meaning with vectors summarizing the contexts where expressions occur (Baroni et al., 2012). Furthermore, research in distributional semantics suggests that distributional patterns can capture the meanings of content, but not function, words (Baroni et al., 2012; Abrusán et al., 2018; Boleda, 2020).

Recent work by Oota et al. (2024) provides evidence for the role of syntactic information in the alignment between the brain's language network and embeddings generated by LLMs. Oota et al. (2024) used a direct approach to eliminate information related to specific linguistic properties in BERT representations and observed how this

affected alignment with fMRI brain recordings. The study examined a range of linguistic properties including sentence length, syntactic properties (such as tree depth), and semantic features (such as number of subjects and objects). Oota et al. (2024) found that eliminating each linguistic property reduced brain alignment across all layers of BERT. This present work expands upon their findings of language models' alignment with the brain's language network by investigating the correspondence between language model embeddings and visual information processing in the brain.

Prior work by Pezzelle (2023) demonstrates that caption embeddings generated by the multimodal transformer model CLIP are highly sensitive to underspecification manipulations, finding that CLIP assigns much lower scores to underspecified descriptions compared to more detailed ones. Specifically, the study investigated the effects of quantity, gender, location, and object underspecification. Similarly, this present research takes inspiration from Pezzelle (2023) to implement an array of linguistic manipulations. Allen et al. (2022) collected a large-scale fMRI dataset for brain encoding. In that study, subjects viewed images of natural scenes from the MS COCO dataset - a dataset of richly annotated images (Lin et al., 2014). This dataset, called the Natural Scenes Dataset (NSD) was previously used by Doerig et al. (2022) in the training of brain encoding models. In the present study, we build upon the prior finding of Doerig et al. (2022) that caption embeddings are aligned with neural activations in the visual cortex, investigating the effects on encoding accuracies for ROIs ranging from the early visual cortex to high-level functional regions involved in place recognition. An additional key finding from the study by Doerig et al. (2022) was that representational alignment was not restricted to high-level areas, but to early visual areas as well, suggesting that sentence embeddings have a degree of representational similarity to purely visual encodings of scenes, in addition to multi-modal visuo-semantic encodings.

This current work is also motivated by a recent study by Kauf et al. (2024), which invested the contribution of lexical-semantic content vs. syntactic structure to the similarity between artificial neural network (ANN) language models and human brain responses in the language network. Kauf et al. (2024) used fMRI data from participants reading sentences, along with representations from GPT-2 language models. Kauf et al. (2024). also applied various manipulations to the original sentences, including word order changes, information loss, and semantic distance alterations. They found that lexical-semantic content, rather than syntactic structure, is the main driver of ANN-brain similarity in language processing. Word order manipulations had minimal impact on ANN-brain similarity, suggesting that syntax plays a secondary role compared to lexical semantics. Information loss manipulations showed that content words (nouns, verbs, adjectives) contribute more to ANN-brain similarity than function words. Semantic distance manipulations revealed that sentences with similar meanings to the originals maintained high ANN-brain similarity. Their findings challenge some traditional views in linguistics that emphasize the importance of syntax in language processing. Their results suggest that ANN language models like GPT-2 may capture aspects of human language processing primarily through lexical-semantic representations rather than syntactic ones.

Building upon the prior work of Doerig et al. (2022), the current study contributes to our understanding of how semantics is encoded in the brain by evaluating how representational alignment is affected by various linguistic manipulations, highlighting the

contribution of lexical semantic content. This work extends the prior work of Doerig et al. (2022) by assessing cross-modal representational alignment — caption embeddings were used to train encoding models to predict visual cortical responses to the captions' corresponding images. This project expands upon prior work in the Neuro-AI research domain that specifically investigates the alignment between representations in the brain and those learned by LLMs (Schrimpf et al., 2021; Kauf et al., 2024). Furthermore, this study builds on prior research using LLMs to provide insight into the computational mechanisms underlying semantic encoding (Toneva et al., 2022; Baroni, 2020).

3 Methods

3.1 Language Model

Unlike one-hot encoding models, which are limited by the curse of dimensionality and fail to capture similar sentiments between sentences, deep neural network-based language models exploit distributed encoding in order to construct language representations as sets of multiple semantic features. The transformer architecture (Vaswani, 2017) is the current state of the art for natural language processing. Sentence transformers process the whole input sequence (sentence) in parallel. Sentence-level embeddings are generated by calculatig an aggregate measure of all word embeddings through max-pooling (Reimers & Gurevych, 2019). These vector representations generated by the sentence transformer model SBERT used in the present study consist of 768 dimensions. The representational space learned by sentence transformers reflects the semantic relationship between embeddings – the cosine distance between embeddings of sentences that have a close semantic proximity is minimized.

MS-COCO captions were encoded using the sentence transformer model SBERT to generate sentence-level embeddings. To derive a fixed-sized sentence-level embedding SBERT adds a pooling operation to the output of BERT / RoBERTa. This pooling strategy computes the mean of all output vectors (MEANstrategy). In order to fine-tune BERT / RoBERTa, Reimers & Gurevych (2019) create siamese networks to update the weights such that the produced sentence embeddings are semantically meaningful and can be compared with cosine-similarity. This fine-tuning was performed using a mean squared-error (MSE) loss objective function, whereby the cosine similarity between the two sentence embeddings u and v is computed.

3.2 Data

To investigate the representational alignment of LLM embeddings of underspecified image captions, regression-based voxel-wise encoding models could be implemented to assess LLM embeddings' predictivity of cortical activation. The publicly available Natural Scenes Dataset provides fMRI data of 8 subjects viewing scenes from the MS COCO dataset (Allen et al., 2022). This methodology would enable evaluation of the cognitive plausibility of the representations learned by LLMs, specifically whether

abstract multi-modal semantic encodings in the brain's visuo-semantic network can be mapped from an underspecified form.

Caption embeddings will be mapped to voxelwise activations via a brain encoding model to evaluate representational alignment with neural activations in the visuo-semantic network. The mapping between the sentence embedding and each voxel will be learned via a ridge regression and used to predict brain responses to images in the test set (set of 1000 common images between all 8 subjects). The dimensions of this ridge regression correspond to the embedding space dimensions (768) plus one dimension corresponding to the brain response. The accuracy of the regression in one voxel implies that the neural activity in this voxel is correlated to the projection of the caption onto a vector in the embedding space (Arana et al., 2023). Fractional ridge regression (Rokem & Kay, 2020) was implemented. Regularisation methods such as this prevent overfitting by penalising the regression for large parameters (Tikhonov, 1963). This biases the regression towards sparse sets of parameters (sets with a low number of non-zero parameters), which are less prone to overfitting (Arana et al., 2023). In the present work, fMRI encoding models were trained using fractional ridge regression on stimuli representations from the Natural Scenes dataset.

Some brain regions specialize in the interpretation of visual inputs while others are more responsive to linguistic stimuli. We hypothesized that regions more proximal to the anterior temporal lobe, such as those involved in high-level visual processing tasks including object and scene recognition, would exhibit a higher degree of cross-modal representational alignment with sentence embeddings. The primary goal of each ridge regression-based encoding model is to predict fMRI voxel values, with expected levels of high correlation in late layers of the brain's vision network, in response to visual stimuli (images of natural scenes). A separate encoding model will be trained per subject (N=8). To train and test the performance of each encoding model, the fMRI data corresponding to the images unique to each subject will be used for training, and evaluation will be performed using the remaining fMRI data corresponding to the set of 1000 images observed by all participants. The semantic scene descriptions used to train the encoding models will be generated by SBERT.

3.3 Representational Alignment

The empirical evidence that alignment between artificial and biological neural networks improves generalization and transfer learning helps to justify representational alignment as a useful evaluation metric for LLMs (Sucholutsky et al., 2023). Representational alignment is the degree of correspondence between the representations learned by two information processing systems, which may be either biological or artificial (Sucholutsky et al., 2023). The degree of alignment between the brain and an ANN model is typically determined by comparing brain activity to the embeddings generated by the ANN model. This comparison is predicated upon the assumption that, if the brain is using the same semantic encoding mechanism as the model, then one should be able to map the embeddings to brain activity. While representational alignment is a widely used evaluation metric in cognitive science and neuro-AI (Sucholutsky et al., 2023), cross-modal representational alignment remains underexplored. Doerig et al. (2022) performed the first significant investigation of cross-modal align-

ment, using representational similarity analysis. This metric measures the similarity between two sets of numerical vectors, for example, language embeddings and voxel-wise activations in the brain. This similarity metric can be calculated by taking the pairwise (cosine) distances between vectors and calculating the Pearson's correlation between these distances. In this study, we assume that this mapping can instead be approximated via a ridge regression-based brain encoding model.

LLMs' "representational alignment" with the human brain can be defined by a brain encoding model's accuracy in predicting neural activation from LLM embeddings (Oota et al., 2024). Representational alignment is operationalised in this paper using brain encoding models. Given the SBERT representations r_l of the MS-COCO caption embeddings i and rv as the corresponding images represented in the brain's vision network, we calculate the Pearson's correlation between the actual and predicted voxel-wise activations. compute the pairwise cosine similarity between the representations. As such, this measures the degree of cross-modal alignment (RAlv) between image representations rv and SBERT embeddings rl. Ultimately, this approach facilitates a precise investigation into the representational code underlying the semantics of image processing.

3.4 Manipulations

3.4.1 Word Order Manipulations

Word order, as determined by a language's syntax rules, is an important cue used in language processing to understand the relations between words (Bever, 1970). However, psycholinguistics research has demonstrated has that language comprehension is highly robust to errors in linguistic input, such as word order errors, *provided that a plausible meaning can be recovered*. For example, in the case of the sentence *The woman gave the ball the girl*, people generally infer *The woman gave the girl the ball* to be the more plausible intended meaning. Therefore, syntactic information such as word order can be overriden in favor of a more plausible meaning (Levy et al., 2009). The syntactic manipulation investigated in the present study was implemented via a pairwise scrambling, whereby the positions of adjacent words were swapped. Additionally, a global scrambling manipulation – whereby all word positions are randomly shuffled – was also implemented.

3.4.2 Semantic Manipulations

The second class of manipulations targets the information conveyed in the subject and verb phrases of sentences. More specifically, we investigated the impact of underspecification, whereby the referential informativity of the caption is reduced by replacing either the subject or verb phrase with the underspecified forms "they" and "is doing something" respectively. To do this, we defined all words preceding the first verb occurrence in each caption as the Subject Phrase, and all succeeding words as the Verb Phrase. This was perfomed used the part-of-speech tagger (pos-tagger) in the NLTK library. The words in either the Subject or Verb Phrase were ablated from the caption and replaced with either of the underspecified forms "they" or "is doing something".

Note that these manipulations do not preserve Subject-Verb agreement. In cases of an original singular subject, the modified form "they" is inconsistent with a singular verb. Likewise, in cases of a plural subject, the modified form "is doing something" is inconsistent with a plural verb. However, given the high degree of similarity (as measured via cosine similarity) between sentence embeddings for the grammatical "They are doing something" and the ungrammatical form "They is doing something", it was observed that subject-verb agreement has a negligible impact on the vector representations generated by language models to encode sentence meanings. Therefore, the method for implementing these manipulations, as described above, set aside considerations of Subject-Verb agreement when modifying the input captions.

Table 1: Caption Manipulations

MANIPULATION	EXAMPLE
Original	a woman is throwing a ball
Pairwise Scrambled	woman a throwing is ball a
Global Scrambled	is a throwing woman ball a
Underspecified Subject Phrase	they is throwing a ball
Underspecified Verb Phrase	a woman is doing something

4 Results

In this section, we investigate what linguistic features (word order, referential specificity, semantic content, word count) contribute to representational alignment of regression-trained ANN-to-brain mapping models. More specifically, we trained brain encoding models on sentence embeddings of manipulated captions (with corresponding brain responses to matching images) and tested these models using held-out brain responses, corresponding to the shared images viewed by all 8 subjects. The brain predictivity scores are raw Pearson r values, rather than r values normalized by the noise ceiling.

We first investigated brain encoding performance on a control condition: mappings were between an unsorted (randomly permuted) list of captions and the ordered set of brain responses. As such, captions were not matched to the brain response to the corresponding image but were instead mapped to the response from another random image. As expected, the encoding models performed at near-chance level for this condition. Then we evaluated the effect of our two types of caption manipulations – manipulations of word order within the caption (word-order manipulations) and semantic underspecification (subject phrase & verb phrase) – on brain encoding models' capacity to predict neural responses, in comparison to performance with the original captions.

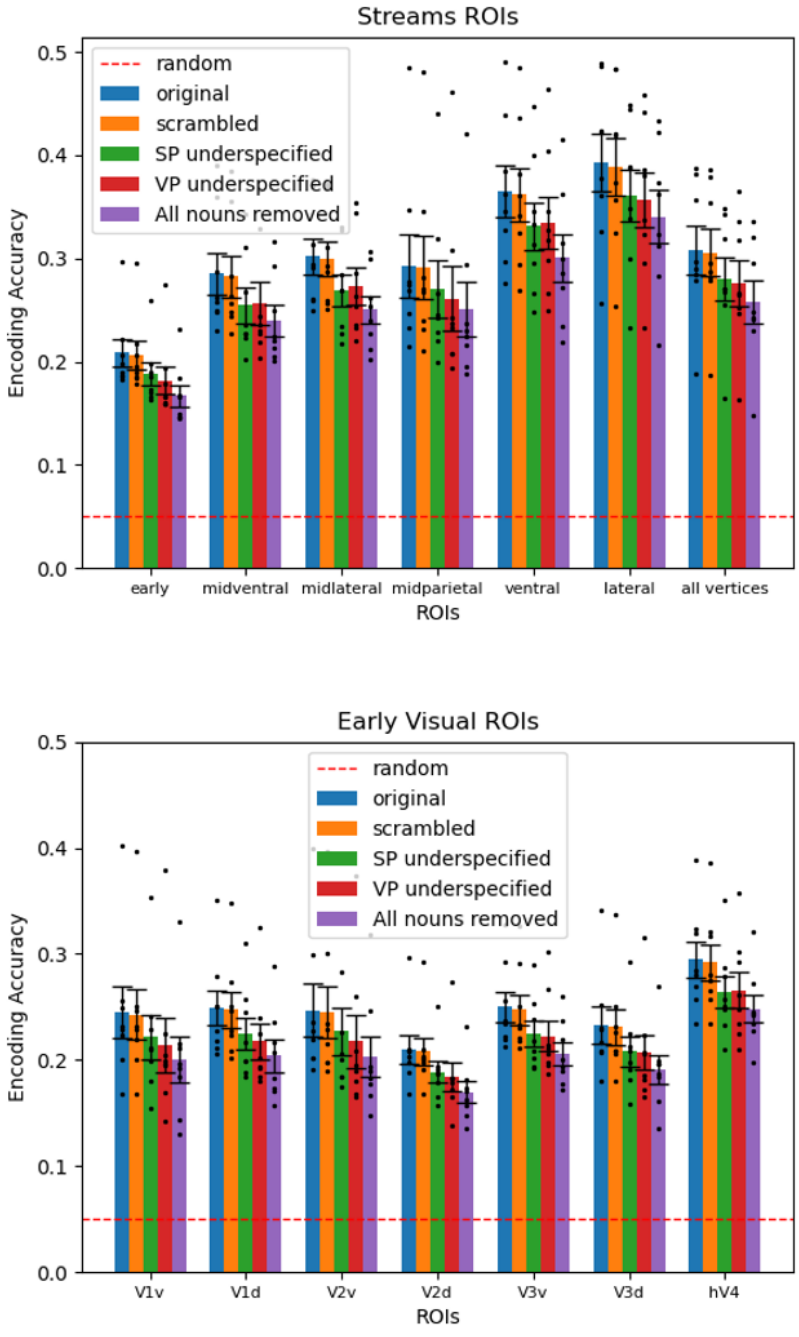

Figure 1: Comparison of neural predictivity across manipulation conditions. Included are individual data points denoting the encoding accuracy (Pearson's r) for each subject (n=8), with standard error bars. Note that the scrambled condition displayed here denotes the global scrambling condition. Consistent with the methodology implemented in Kauf et al. (2024), pairwise dependent-samples t-tests with Bonferroni correction procedure were performed to confirm statistical significance ($p < .05$) for SP, VP underspecified, and noun removal conditions. Dotted red line indicates average encoding accuracy across all ROIs for baseline (randomly matched NSD image-MS-COCO caption pairs) condition.

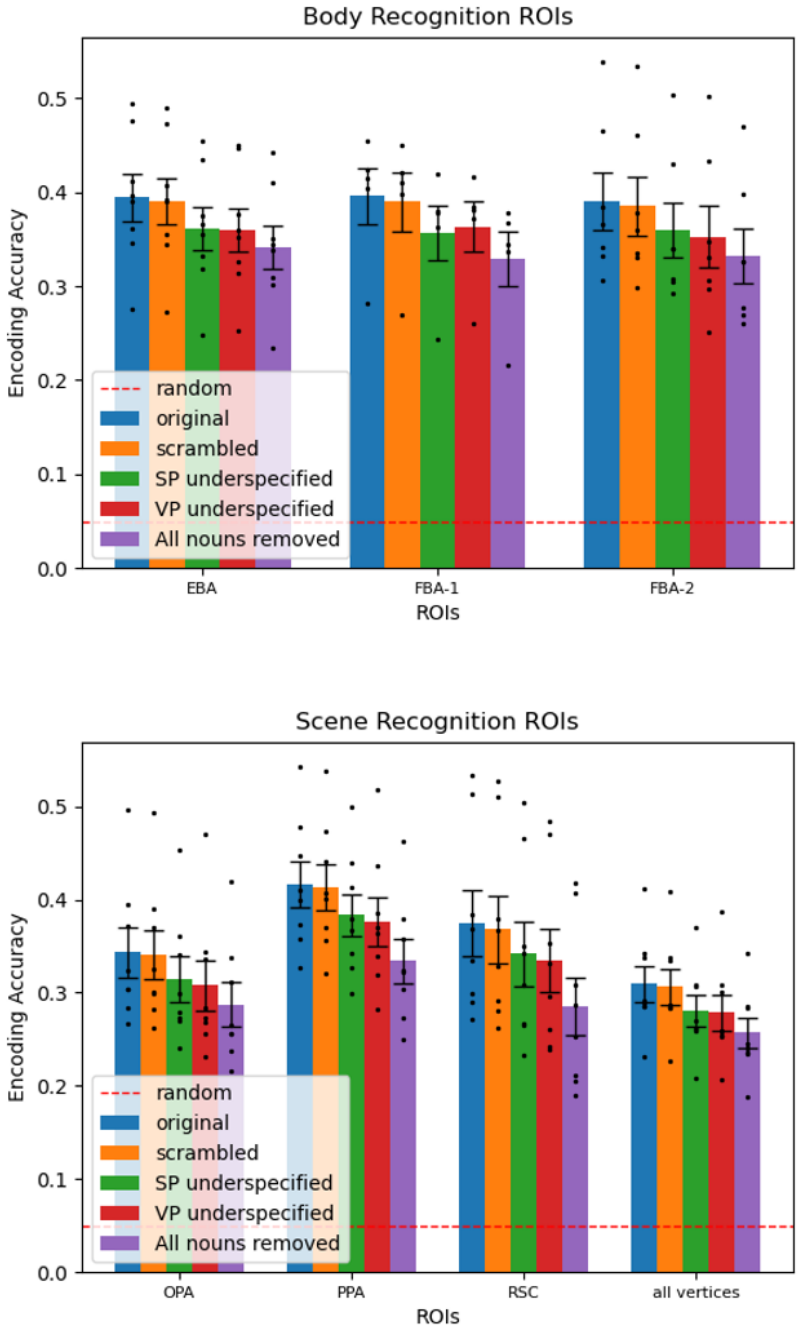

Figure 2: Barplots comparing neural predictivity across high-level visual ROIs for each manipulation condition.

4.1 Syntactic Manipulation

Word-order manipulations had negligible effect on brain predictivity in the visual cortex. This diverges from the observations of Kauf et al. (2024), which found that word-order manipulations had a small, yet statistically significant, impact on neural predictivity in the brain's language network. These varied findings could be explained by the different brain regions of interest in each study. While neural predictivity in the brain's language network may indeed be sensitive to word-order manipulations, based on the present findings, representational alignment with the brain's vision network may be unaffected by syntactic manipulations. These findings are therefore indicative of modality-specific distinctions in representational alignment, expanding upon the prior work of Doerig et al. (2022).

In addition to the pairwise scrambling manipulation described previously, the more severe global scrambling manipulation – whereby the positions of all words in the sentence are randomly shuffled – did not significantly affect neural predictivity. As observed in Fig.1, this manipulation still yielded an average Pearson's correlation of 0.30 across all voxels. It should be noted that this more extreme scrambling manipulation does not take into account point-wise mutual information (PMI) – an information-theoretic measure, defined by Mollica et al. (2020), quantifying local semantic dependency structure. The random shuffling manipulation implemented here is not designed to preserve this measure. Yet, even for this more severe disruption, the effect on brain predictivity was not significant. This finding is consistent with the prior observations of Kauf et al. (2024), which found that syntactic perturbations of language embeddings have a limited effect on neural predictivity.

4.2 Semantic Manipulation

Each of the semantic underspecification manipulations (Subject Phrase & Verb Phrase) led to a significant reduction in encoding accuracy relative to the Original condition. However, given the extent to which referential informativity is reduced in each of these manipulations, the finding that neural data can still be reliably mapped (with avg. Pearson's correlation of 0.28 for all voxels across both underspecification condiitons) from these underspecified captions is still somewhat surprising, especially since each manipulation results in information loss in the form of word removal from either the subject or verb phrase. One explanation for this limited reduction in encoding accuracy may be that these manipulations largely preserve content words.

Greater encoding accuracy was achieved in high-level functional regions involved in body and scene recognition. However, the overall pattern of results remained consistent with the results for early visual ROIS, whereby the semantic manipulations still yielded reductions in neural predictivity. As observed in the above Figures depicting encoding accuracy, semantic underspecification yields an average 9% reduction in neural predictivity across all voxels. To evaluate the validity of this speculative explanation, an additional manipulation was implemented, whereby all nouns were ablated from each caption (this condition is depicted in purple in above Figures 1 and 2). Indeed, this noun-ablation manipulation led to a more substantial decrease in predictivity values. Complete noun ablation resulted in an average 16% reduction in encoding

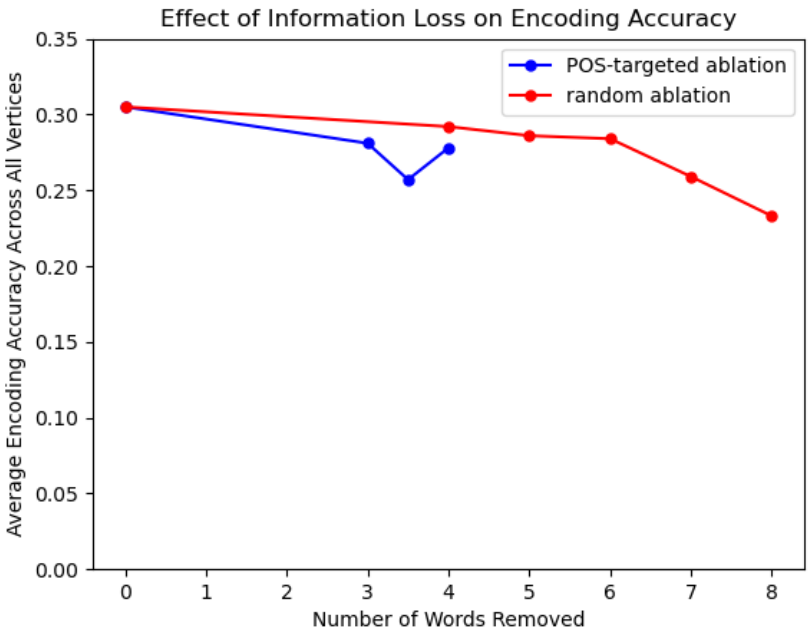

Figure 3: Effect of word removal on average neural predictivity across all voxels. As can be observed in the differences in line plots for the random and targeted information-loss manipulations (noun ablation, SP/VP underspecification), it is not the general removal of any words, but the specific ablation of content words that strongly reduce encoding accuracy. Individual data point at n=3 corresponds to SP underspecification condition, data point at n=3.5 corresponds to noun ablation condition, n=4 corresponds to VP underspecification condition. Note that, of these manipulations, noun ablation yields the strongest reduction in a similarly strong reduction in encoding accuracy is only observed in the random removal of 7 words.

accuracy across ROIs.

To ensure that the drop in predictivity for each of these semantic manipulations was not merely an artifact of the length of the caption, we also implemented additional information-loss manipulations whereby we ablated a progressively increasing number of words from the captions (n=4, 5, 6, 7, 8). As can be observed in Figure 3, the semantic underspecification and noun ablation manipulations had a more substantial effect on encoding accuracy than the random word-removal manipulations for n=4, 5, 6, even though these targeted information-loss manipulations on average only removed around 3 words from each caption. This comparison demonstrates that it is not merely the general loss of words, but rather the specific loss of content words (ex: nouns, verbs) that affect neural predictivity. The limited effect of word removal on brain predictivity suggesting an asymmetry in the distribution of content vs. function words present in the relatively short captions (average num. of words = 11) of the MS-COCO dataset.

5 Discussion

5.1 Contributions of Lexical-Semantic Content vs. Syntactic Structure to Representational Alignment

To evaluate the cognitive plausibility of LLMs, this research evaluated whether brain encoding models trained on underspecified embeddings can predict activity in semantic association networks in the brain. Investigating whether the intended meaning (operationalized here as neural activation of late-layer semantic association networks in the brain recorded as human subjects viewed corresponding images from the Natural Scenes Dataset) of an underspecified descriptive text can be mapped from the embeddings generated by LLMs has provided an empirical assessment of whether the representations learned by LLMs replicate neural semantic representations' invariance to underspecification. Human interpreters are capable of correctly inferring the intended referent from an underspecified utterance Frisson (2009). This suggests that activation of abstract multi-modal semantic encodings in the brain's high-level visuosemantic network is not affected by superficial manipulations of the specificity of a subject/object description, and may explain why underspecification has a smaller effect on encoding accuracy in comparison to stronger information-loss manipulations, such as noun ablation.

Recent studies in computational neuroscience have found that representations from transformer models align well with brain responses of humans processing linguistic input (Caucheteux & King, 2022; Schrimpf et al., 2021). However, precisely what features make language model embeddings align with the brain's semantic representations has been underexplored (Oota et al., 2024; Kauf et al., 2024), and research investigating cross-modal alignment with visual responses to images is even more limited (Doerig et al., 2022). Focusing on the SBERT model from the sentence transformers library, we investigated the effect of a diverse array of linguistic manipulations, including manipulations that affect sentence meaning (via underspecification or removal of content words) and those that primarily affect syntactic structure (carried by word order), on the neural predictivity of brain encoding models. Ultimately, we found that the removal of

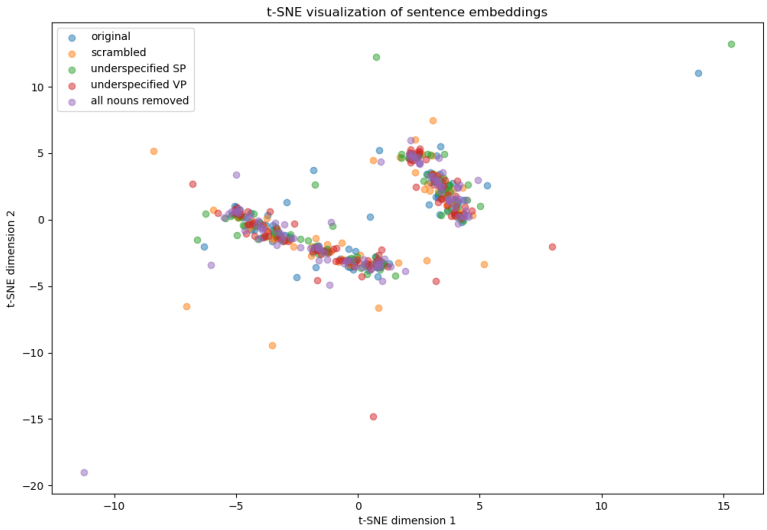

Figure 4: T-SNE Visualisation of SBERT Embedding Space. Plotted above are the embeddings for 100 randomly selected original MS-COCO captions, and their variants (scrambled, SP underspecified, VP underspecified, noun ablation). If the sets of embeddings were clustered into different manipulation groups, this would indicate that the SBERT embeddings were strongly affected by the manipulations. However, the close proximity of embeddings across manipulation groups observed here indicates that SBERT vector representations are largely unaffected by these manipulations. This would suggest that these vector representations are robust to perturbations of word order and referential specificity, and, as depicted in this visualisation, even the ablation of nouns. This observation is surprising – while the semantic content of captions is indeed preserved in the scrambled captions, it is conserved to progressively lesser extents in the SP/VP underspecification and noun ablation manipulations. Therefore, it was expected that the noun-ablated vector representations would be positioned at greater distances from the original (blue) vector representations. Indeed, preliminary analysis of the cosine similarity between the original and noun-ablated sets of MS-COCO caption embeddings validates this. Yet, this does not appear to be the case in the above visualisation. However, given that this dimensionality reduction technique flattens a 768-dimensional space into a 2-dimensional space, the distance between these clusters may be misleading (Wattenberg et al., 2016)

lexical-semantic content has a stronger reduction effect on representational alignment than perturbations of syntactic structure (implemented via word-order scrambling).

Comparatively subtle manipulations (ex: pairwise scrambling) had negligible impact on brain encoding accuracy. This finding has two possible explanations: either the embedding of the modified caption is very similar to that of the original versions, or neural activity in the visual cortex is not sensitive to these syntactic manipulations. Across semantic manipulations, encoding accuracy decreased the more information was removed, with embeddings of noun-ablated captions achieving an average reduction of neural predictity of 16% across ROIs.

Information-loss manipulations, rather than word-order manipulations elicit lower brain predictivity, suggesting that the ANN-to-brain encoding model performance is less sensitive to syntactic information, but rather relies on semantic information. These findings are consistent with evidence from neuroscience and computational linguistics. Past work in computational neuroscience points to the greater contribution of semantics, in comparison to syntax, to both the magnitude and distributed patterns of activation in the brain's language network as measured by fMRI (Fedorenko et al., 2016; Huth et al., 2016; Kauf et al., 2024). Meanwhile, research in NLP demonstrates that language models are not dependent upon syntactic information to achieve high performance on many current NLP benchmark tasks (Sinha et al., 2021).

In this study, we integrated neuroscientific and computational linguistics perspectives on the contribution of lexical-semantic content vs. syntactic information in representational alignment between sentence embeddings and neural responses in the brain. The word order manipulations implemented here were inspired by earlier work from Kauf et al. (2024) comparing the neural predictivity between naturalistic and scrambled inputs. Consistent with prior findings from Kauf et al. (2024) we observed that even completely randomized shuffling of word order, which disrupts local syntactic dependencies, leads only to a negligible decrease in brain encoding model performance. Further, we found that random word removal does little to decrease brain predictivity; instead, it is the targeted ablation of content words (such as verbs, nouns) that reduces representational alignment.

One possible explanation for these findings may be that syntactic information is not a strong contributor to neural representations in the visual cortex, though it may be a more influential component of semantic representations in the language network (Oota et al., 2024). This invariance of ANN-to-brain mapping models to these syntactic manipulations of the original captions might be explained by the vision network's comparative insensitivity to structure manipulations. While the language system is indeed sensitive to syntactic processing difficulty (Shain et al., 2024), the vision network may exhibit less sensitivity to sequential variations. The results presented here suggest that syntactic structure is not critical for cross-modal representational alignment between sentence embeddings and fMRI BOLD responses in the brain's visual cortex.

To further investigate whether the visual network indeed is sensitive to structural effects, alternative image-caption pairs could be used, such as the Winoground dataset (Thrush et al., 2022). In this dataset, caption structure is critical to interpretation, as in cases where the content (words) are held constant, but the word order is varied, yielding to distinct meanings, each corresponding to a different image (ex: a tree smashed into a car vs. a car smashed into a tree). However, this would also require additional data

collection, as there is currently no publicly available fMRI dataset for human subject responses to the Winoground image-caption pairs.

Since each manipulation affected representational alignment to varying degrees, we investigated possible explanations for these differences. Caption manipulations that led to lower brain encoding accuracy also led to more divergent representations in the SBERT embedding space (relative to the representations of intact sentences). As discussed above, we quantified the changes in the representational space across manipulation conditions relative to the intact (original) embeddings using the cosine similarity metric and observed that manipulations that resulted in a larger transformation of the embedding space (as determined by a lower cosine similarity), such as SP and VP underspecification, also yielded larger reductions in brain encoding accuracy. However, even the most severe manipulations (noun ablation) resulted in sentence embeddings that were still largely similar to the embeddings of the intact captions, and could still be mapped to human neural responses. One explanation for this finding may be that this information-loss manipulation is still too limited, and that the remaining verbs, adjectives and adverbs contained in the caption preserve substantial semantic information. A complete ablation of all content words should yield findings similar to those of Kauf et al. (2024), whereby representational alignment is substantially reduced.

To summarize, this study enabled us to identify features of linguistic stimuli (referential informativity, content words) that contribute to cross-modal representational alignment. These features affect the magnitude of brain encoding models' predictive accuracy. However, the strength of this effect was smaller than anticipated – when these features were manipulated such that we would not expect the resulting captions to carry much informative structure (as in the noun ablation manipulation), brain encoding models were still able to consistently predict neural activity in the visual cortex.

6 Conclusion

In this work, we investigated how manipulating sentence embeddings of image captions affects their representational alignment with neural responses (as measured with fMRI) during the viewing of natural scenes. To conduct this exploration, we evaluated which linguistic features (across two primary manipulation categories) reliably contribute to brain encoding model predictive accuracy. Consistent with prior work from Doerig et al. (2022), we found that the neural representation encoding the semantics of images aligns with the context-dependent word vectors generated by language models. We additionally found that the encoding accuracy of brain encoding models is significantly affected by manipulations of lexical-semantic content, rather than word order manipulations. Specifically, our underspecification manipulations reduced representational alignment to a greater extent than scrambling manipulations, though neural responses could still be reliably mapped (with a Pearson's correlation of 0.28 across all voxels). As expected, a more severe information-loss manipulation, involving the targeted removal of noun entities, resulted in a stronger reduction in representational alignment. These findings show that semantic, rather than syntactic, content contributes to representational alignment with the brain's visual cortex. This pattern of results suggests that the lexical-semantic content of an image caption is encoded in the brain's visual

cortex, particularly the high-level visual processing regions involved in scene recognition.

7 Limitations

As noted in previous sections, MS-COCO captions are rather short, averaging only 11 words. One implication of this brevity is that each caption contains a comparatively high proportion of function words, which merely signal grammatical relationships between words, and do not themselves convey semantic content. This may explain why the random word removal manipulation had a relatively small impact – it was likely that the removed words were function, rather than content, words. Therefore, little semantic information was lost in these manipulations. This explanation could also be applied to the semantic underspecification manipulations as well. While both underspecification manipulations each resulted in information loss via word removal (SP = avg. 3 words removed, VP = avg. 4 words removed), some of the words removed in the subject or verb phrases may have also been function words, such as determiners like 'a' or the', auxiliaries like 'do' or 'can', and demonstratives like 'this' or 'that'. In comparison, the more severe noun-removal manipulation (avg. 3.5 words removed), guaranteed that only content words (nouns) would be removed. This explains the greater reduction in representational alignment for the noun-removal manipulation, and indicates that it is specifically *content* words that contribute to representational alignment. These findings are consistent with the results of Kauf et al. (2024).

Perhaps the manipulations implemented here would have been more impactful on an alternative dataset, containing longer, complex sentences. Complex sentences, defined by the presence of dependent clauses, will contain a higher number of content words in the forms of additional nouns and verbs. While most existing crowd-sourced caption datasets, generated by human annotators, are characterized by this feature of low length, the introduction of multimodal models presents an opportunity for model-generated image captions. This would allow us to use the same set of images from the Natural Scenes Dataset, while generating lengthier captions.

An additional limitation with this study is one common to the cognitive computational neuroscience research programme, the issue of multiple realisability – similar representations do not imply similar mechanisms for realising those representations. In other words, representational alignment between language models and the brain does not imply that the mechanisms which generated these vector and neural representations are shared (Guest & Martin, 2023). Therefore, while the findings of this study enable us to learn what features (syntax, referential specificity, semantic content, word count, etc.) may contribute to ANN-to-brain mapping, further work is needed to elucidate the computational mechanisms implemented in the brain which construct these semantic representations.

References

Márta Abrusán, Nicholas Asher, and Tim Van de Cruys. Content vs. function words: The view from distributional semantics. In *Proceedings of Sinn und Bedeutung*, volume 22, pp. 1–21, 2018.

Emily J Allen, Ghislain St-Yves, Yihan Wu, Jesse L Breedlove, Jacob S Prince, Logan T Dowdle, Matthias Nau, Brad Caron, Franco Pestilli, Ian Charest, et al. A massive 7t fmri dataset to bridge cognitive neuroscience and artificial intelligence. *Nature neuroscience*, 25(1):116–126, 2022.

Sophie Arana, Jacques Pesnot Lerousseau, and Peter Hagoort. Deep learning models to study sentence comprehension in the human brain. *Language, Cognition and Neuroscience*, pp. 1–19, 2023.

Marco Baroni. Linguistic generalization and compositionality in modern artificial neural networks. *Philosophical Transactions of the Royal Society B*, 375(1791): 20190307, 2020.

Marco Baroni, Raffaella Bernardi, Ngoc-Quynh Do, and Chung-chieh Shan. Entailment above the word level in distributional semantics. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pp. 23–32, 2012.

Thomas G Bever. The cognitive basis for linguistic structures. *Cognition and the development of language*, 1970.

Gemma Boleda. Distributional semantics and linguistic theory. *Annual Review of Linguistics*, 6(1):213–234, 2020.

Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications biology*, 5(1):134, 2022.

Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

Adrien Doerig, Tim C Kietzmann, Emily Allen, Yihan Wu, Thomas Naselaris, Kendrick Kay, and Ian Charest. Semantic scene descriptions as an objective of human vision. *arXiv preprint arXiv:2209.11737*, 2022.

Evelina Fedorenko, Michael K Behr, and Nancy Kanwisher. Functional specificity for high-level linguistic processing in the human brain. *Proceedings of the National Academy of Sciences*, 108(39):16428–16433, 2011.

Evelina Fedorenko, Terri L Scott, Peter Brunner, William G Coon, Brianna Pritchett, Gerwin Schalk, and Nancy Kanwisher. Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences*, 113(41): E6256–E6262, 2016.

Steven Frisson. Semantic underspecification in language processing. *Language and linguistics compass*, 3(1):111–127, 2009.

Olivia Guest and Andrea E Martin. On logical inference over brains, behaviour, and artificial neural networks. *Computational Brain & Behavior*, 6(2):213–227, 2023.

Michael Hahn, Richard Futrell, Roger Levy, and Edward Gibson. A resource-rational model of human processing of recursive linguistic structure. *Proceedings of the National Academy of Sciences*, 119(43):e2122602119, 2022.

Zellig S Harris. Distributional structure. *Word*, 10(2-3):146–162, 1954.

Eghbal A Hosseini, Martin Schrimpf, Yian Zhang, Samuel Bowman, Noga Zaslavsky, and Evelina Fedorenko. Artificial neural network language models align neurally and behaviorally with humans even after a developmentally realistic amount of training. *BioRxiv*, pp. 2022–10, 2022.

Alexander G Huth, Wendy A De Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458, 2016.

Carina Kauf, Greta Tuckute, Roger Levy, Jacob Andreas, and Evelina Fedorenko. Lexical-semantic content, not syntactic structure, is the main contributor to ann-brain similarity of fmri responses in the language network. *Neurobiology of Language*, 5 (1):7–42, 2024.

Nikolaus Kriegeskorte. Deep neural networks: a new framework for modeling biological vision and brain information processing. *Annual review of vision science*, 1(1): 417–446, 2015.

Roger Levy, Klinton Bicknell, Tim Slattery, and Keith Rayner. Eye movement evidence that readers maintain and act on uncertainty about past linguistic input. *Proceedings of the national academy of sciences*, 106(50):21086–21090, 2009.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pp. 740–755. Springer, 2014.

Francis Mollica, Matthew Siegelman, Evgeniia Diachek, Steven T Piantadosi, Zachary Mineroff, Richard Futrell, Hope Kean, Peng Qian, and Evelina Fedorenko. Composition is the core driver of the language-selective network. *Neurobiology of Language*, 1(1):104–134, 2020.

SubbaReddy Oota, Manish Gupta, and Mariya Toneva. Joint processing of linguistic properties in brains and language models. *Advances in Neural Information Processing Systems*, 36, 2024.

Sandro Pezzelle. Dealing with semantic underspecification in multimodal nlp. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 12098–12112, 2023.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Ariel Rokem and Kendrick Kay. Fractional ridge regression: a fast, interpretable reparameterization of ridge regression. *GigaScience*, 9(12):giaa133, 2020.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118, 2021.

Cory Shain, Hope Kean, Colton Casto, Benjamin Lipkin, Josef Affourtit, Matthew Siegelman, Francis Mollica, and Evelina Fedorenko. Distributed sensitivity to syntax and semantics throughout the language network. *Journal of Cognitive Neuroscience*, 36(7):1427–1471, 2024.

Koustuv Sinha, Robin Jia, Dieuwke Hupkes, Joelle Pineau, Adina Williams, and Douwe Kiela. Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *arXiv preprint arXiv:2104.06644*, 2021.

Ilia Sucholutsky, Lukas Muttenthaler, Adrian Weller, Andi Peng, Andreea Bobu, Been Kim, Bradley C Love, Erin Grant, Jascha Achterberg, Joshua B Tenenbaum, et al. Getting aligned on representational alignment. *arXiv preprint arXiv:2310.13018*, 2023.

Tristan Thrush, Ryan Jiang, Max Bartolo, Amanpreet Singh, Adina Williams, Douwe Kiela, and Candace Ross. Winoground: Probing vision and language models for visuo-linguistic compositionality. *Proceedings of the IEEE/CVF COnference on Computer Vision and Pattern Recognition*, pp. 5238–5248, 2022.

Andrei Nikolaevich Tikhonov. On the solution of ill-posed problems and the method of regularization. In *Doklady akademii nauk*, volume 151, pp. 501–504. Russian Academy of Sciences, 1963.

Mariya Toneva, Tom M Mitchell, and Leila Wehbe. Combining computational controls with natural text reveals aspects of meaning composition. *Nature computational science*, 2(11):745–757, 2022.

Ashish Vaswani. Attention is all you need. *arXiv preprint arXiv:1706.03762*, 2017.

Martin Wattenberg, Fernanda Viégas, and Ian Johnson. How to use t-sne effectively. *Distill*, 1(10):e2, 2016.