When word order matters: human brains represent sentence meaning differently from large language models

James Fodor*, Carsten Murawski*, Shinsuke Suzuki†

July 2, 2025

Abstract

Large language models based on the transformer architecture are now capable of producing human-like language. But do they encode and process linguistic meaning in a human-like way? Here, we address this question by analysing 7T fMRI data from 30 participants reading 108 sentences each. These sentences are carefully designed to disentangle sentence structure from word meaning, thereby testing whether transformers are able to represent aspects of sentence meaning above the word level. We found that while transformer models match brain representations better than models that completely ignore word order, all transformer models performed poorly overall. Further, transformers were significantly inferior to models explicitly designed to encode the structural relations between words. Our results provide insight into the nature of sentence representation in the brain, highlighting the critical role of sentence structure. They also cast doubt on the claim that transformers represent sentence meaning similarly to the human brain.

Understanding how human language is processed and represented in the brain is a major scientific challenge. The past decade has seen a proliferation of work integrating theoretical approaches from linguistics and computer science with empirical data from neuroimaging studies in an effort to better understand how meaning is represented in the brain[1–7]. Most research has focused on evaluating vector-based models, in which the meaning of a word or phrase is represented as a vector of numbers. This approach forms the basis for large language models, which are neural networks based on the transformer architecture and trained to predict hidden tokens on very large corpora of natural text. Leading models such as GPT-4, Gemini, Llama, and Claude are highly versatile, capable of performing of generating grammatical and relevant responses to a wide range of queries and instructions[8–10]. The extensive linguistic capabilities of these models, along with their ability to acquire language competence from naturalistic data, has generated significant interest in their potential value as cognitive models of language processing in humans[11–13]. Studies have consistently found statistically significant correlations between brain activity and various semantic models, with several finding that transformers better explain brain activity compared to static word embedding models[14–17].

Most research comparing language models to brain activity has used stimuli that have not been selected to evaluate any specific linguistic hypothesis. While there are many benefits to utilising naturalistic stimuli in the study of language[18–21], such stimuli have the disadvantage that they may not adequately sample the linguistic phenomena of most interest[19], and do not control for variables crucial for contrasting the representations of different models[22]. A particular challenge is distinguishing whether language models are predictive of brain activity solely due to word-level (lexical) semantic information, or whether they also incorporate representations of sentence structure in a manner comparable to the brain. Direct comparison of static word embeddings with contextualised transformer embeddings is insufficient to resolve this issue, because contextualised embeddings also capture polysemy and other semantic phenomena not directly related to sentence structure. Another limitation of existing studies is that establishing that features extracted from large language models are predictive of brain activity does not necessarily provide much insight about what information these features encode or how such information is utilised by the brain[23–25]. A final limitation of existing studies is that encoding techniques are best suited to use with vector representations of language, making it difficult to conduct comparisons with graph-based or other approaches specialised for explicitly representing sentence structure.

Here, we present results from an fMRI study in which 30 participants read isolated sentences and answered simple questions about their meaning. We also collected a separate dataset of behavioural ratings of all pairwise comparisons of the same set of sentences. First, we developed a hand-crafted set of sentences designed specifically to control for the confound of lexical similarity, allowing for clearer inferences about how sentence-level information is represented by the brain. Second, we conduct model comparison using representational similarity analysis (RSA), which involves comparing pairwise similarity scores for voxel activations and semantic models.

*The Centre for Brain, Mind and Markets, Faculty of Business and Economics, The University of Melbourne, Australia

†Faculty of Social Data Science, Hitotsubashi University, Japan

This technique extracts information about the patterns of similarity of model representations, thereby providing additional insight into the nature of brain semantic representations beyond voxelwise predictability. Furthermore, RSA facilitates comparison between dissimilar types of representations, thereby allowing us to compare a wider range of computational models, including both vector-based and graph-based models. While this technique has been used extensively at the level of individual words, we are not aware of any previous research using RSA for model comparison at the sentence level[26–30].

1 Results

1.1 Stimuli and models

Our hand-crafted sentences were carefully designed to reveal the role of sentence structure in semantic representation. Illustrative example sentences are shown in Figure 1a, along with the design matrix indicating the different types of sentence comparisons we considered. This matrix exhibits a block diagonal structure, which is a result of including six subsets of sentences, each sharing a core vocabulary of words which are then rearranged to create systematic variations of sentence structure while preserving high lexical similarity. Within these six diagonal blocks, we distinguish between 'on-diagonal' and 'off-diagonal' sentence pairs. On-diagonal sentence pairs (depicted in shades of blue) have sentence elements simply added or removed. By contrast, the off-diagonal sentence pairs (depicted in light green) have sentence elements interchanged to vary sentence meaning while keeping most of the constituent words the same. This approach builds on our previous work using behavioural data[31], where we showed that such methods allow for effective dissociation of lexical similarity from overall similarity in sentence meaning. The primary objective of the present study is to analyse the brain representations of the block diagonal sentences extracted during an fMRI reading task, and compare these to representations derived from a variety of computational models of sentence meaning to determine which models best match brain representations.

We next computed the representations for each sentence using a range of computational models. We analysed four distinct approaches to semantic representation. The first was a simple 'Mean' model, consisting of the element-wise averages of static word embeddings of each word in the sentence. Since this model ignores the position of words within a sentence as well as their grammatical role, it serves as a baseline incorporating only lexical information. The second class consists of embeddings extracted from various transformer neural networks. Results for the 'Transformers' model are computed by averaging results over six different transformer models, with the details given in Methods. Both Mean and Transformer models are vector-based approaches, as they represent the meaning of a sentence with a vector of numbers[32]. By contrast, 'Graph' models are based on a nested graph formalism constructed in accordance with a semantic parsing paradigm. Here we selected Abstract Meaning Representation (AMR) as a widely-used exemplar of this approach to semantic representation[33]. Finally, we analysed a 'Hybrid' model, which includes components from both vector-based and graph-based formalisms. Building on our previous work[31], our Hybrid model uses a semantic parser to tag each word based on its semantic role, and then constructs a separate vector embedding for each semantic role. All four models are summarised in Figure 1b.

Having constructed the model representations for our sentences, we next computed the similarities between all sentence pairs, using these data to construct RSA matrices for all four computational models. As shown in Figure 1c, the block diagonal structure corresponding to the six sentence subsets is clearly visible. Sentence pairs within these blocks have higher similarity owing to sharing many words in common, as per our design. More importantly, the RSA matrices also illustrate clear differences between how the four models represent sentences. In particular, the 'swapped' off-diagonal sentence pairs are accorded high similarities by the Mean model, much lower similarities by the Graph and Hybrid models, and intermediate similarities by the Transformer models (OpenAI embeddings shown for illustration). These differences are consistent with our previous findings that transformers are less sensitive to changes in sentence structure than hybrid or graph models. Here we aim to test which pattern of representational similarities best matches data collected using neuroimaging during a sentence reading task. The full set of RSA matrices for all models is shown in Supplementary Information Figure S1.

1.2 fMRI results

To evaluate how well each model describes sentence processing in the brain, we collected fMRI data from 30 participants while reading each of the 108 sentences. Our experimental pipeline is depicted in Figure 1d, with additional details given in section 3. We presented each sentence four times, with randomly interposed questions incorporated as an attention check. Voxel data were analysed using GLMSingle, an algorithm which fits a hemodyamic response function to each voxel and then estimates the response of that voxel to each stimulus. We selected a subset of voxels for further analysis based on their stability score, which is computed as the average correlation of voxel activity across repetitions of the same stimulus[2,34,35]. We analysed stable voxels within two regions of interest: the language network[36], and the entire cortex less the primary visual cortex. Model fit was assessed using representational similarity analysis, with higher correlations indicating that the corresponding model represents the set of stimuli more similarity to the brain.

We performed representational similarity analysis in

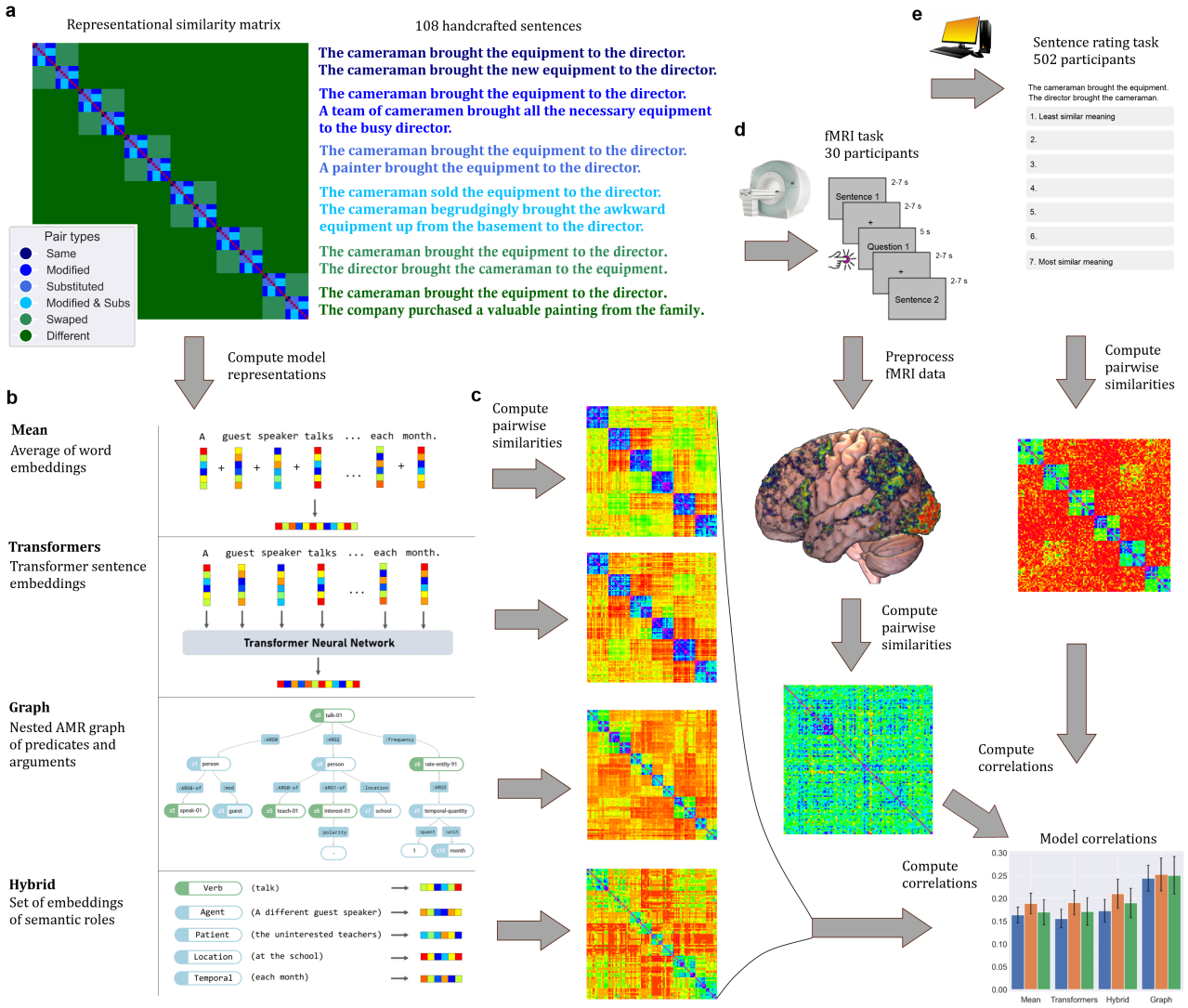

Fig. 1: Summary of study methods for constructing stimuli, computing model representations, and collecting fMRI and behavioural data. a) We construct 108 handcrafted sentences, designed to enable systematic variation in sentence meaning while controlling for lexical similarity. Here we show the corresponding 108×108 design matrix colour-coded with the type of each sentence pair. Sentence pairs in the six blocks along the diagonal are the primary pairs of interest in this study. **b)** All sentences were encoded using each of the four computational models of sentence meaning which we examine in this study. **c)** We then computed representational similarity matrices of the 108 stimuli for each of the four models. More similar sentence pairs are shown in blue, and less similar in red. **d)** Study pipeline for the fMRI experiment, in which participants were presented one sentence at a time for 2-7 seconds depending on sentence length. Multiple choice comprehension questions were interspersed randomly to assess attention. After scanning, data was processed and brain activity patterns were used to compute a neural representational similarity matrix for each participant. Correlations were then computed between the model and brain RSA matrices. **e)** Study pipeline for behavioural experiment, in which online participants were each shown 112 sentence pairs and asked to rate their semantic similarity. Ratings were averaged over participants to compute a similarity matrix. The correlation was then computed between the model and behavioural RSA matrices.

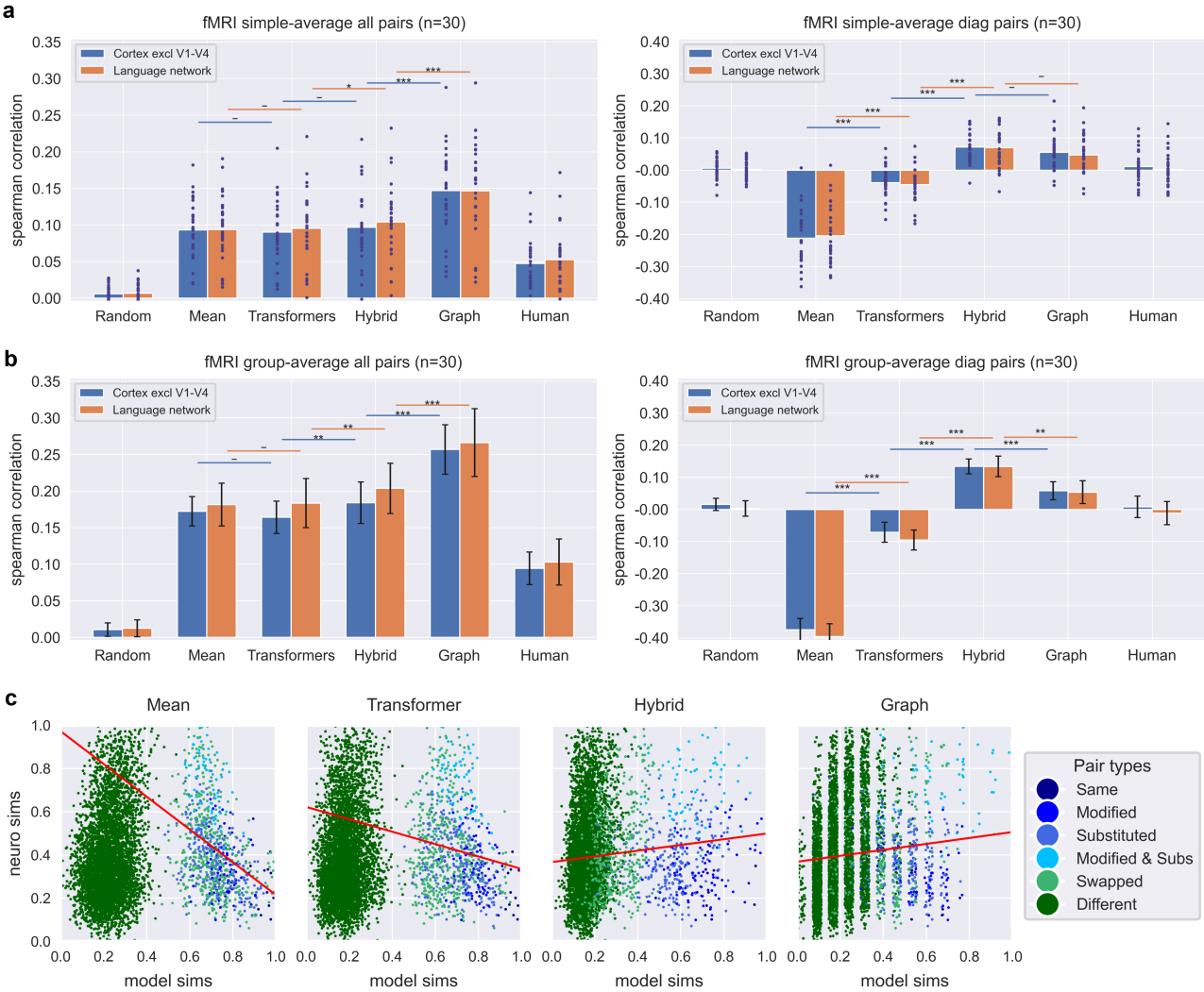

Fig. 2: Model correlations with brain activity for all sentence pairs and the block-diagonal subset of sentence pairs. Partial correlations between RSA matrices of five computational models (Random, Mean, Transformers, Hybrid, and Graph) and the brain RSA matrix, controlling for differences in sentence length. 'Human' refers to behavioural ratings. Blue bars indicate inclusion of all stable cortical voxels (excluding visual regions V1-V4), and green bars indicate inclusion of only stable voxels in the language network. Notation for statistical significance: * for p<0.05, ** for p<0.01, and *** for p<0.001, with Bonferroni correction for three independent comparisons. **a)** Partial correlations for each individual participant shown as blue dots, with the simple average over individual correlations shown as a bar. **b)** Partial correlations computed using the group-averaged RSA matrix. Error bars show 95% confidence intervals calculated by bootstrap resampling over participants. **c)** Scatterplots showing the relationship between model similarities (horizontal axis) and group-average neural similarities (vertical axis) for all four computational models. Each dot corresponds to a single pairwise similarity, scaled to between 0 and 1 for visualisation. While all sentence pairs are shown for comparison, regression lines (red) are computed over the block diagonal pairs only.

two different ways. In the simple-average approach, we computed the RSA correlation for each participant separately and then took the average. In contrast, the group-average approach involves first averaging the RSA matrix across participants, and then computing the RSA correlation for this group-averaged matrix[26,27,37]. In each case, we computed the Spearman partial correlation across all 5,778 sentence pairs and also across the 918 block diagonal sentence pairs, controlling for differences in sentence length. The full set of results for all 17 models tested are shown in Supplementary Information Figure S2. Here we discuss results of the four models of main interest.

We first consider correlations computed using all sentence pairs, as shown in Figure 2a. In language network voxels, all models show positive correlations, with relatively small differences between models. For the simple-average method, the differences in correlation were not significant when comparing the Mean and Transformers models ($\Delta\rho = 0.001, t = 0.686, p = 0.4981$), or the Hybrid and Transformer models ($\Delta\rho = 0.009, t = 2.720, p = 0.0109$). However, the Graph model had a significantly higher correlation compared to the Hybrid model ($\Delta\rho = 0.043, t = 7.393, p < 0.0001$). Similar results were found using the group-average method (shown in Figure 2b), but with higher absolute values. The fact that all models show positive correlations when evaluating all sentence pairs is unsurprising, since most sentences can be differentiated from one another using purely lexical differences, which all models are sensitive to.

We now consider correlations computed using only the block diagonal sentence pairs, which are designed to be more difficult for computational models to distinguish owing to high lexical similarity. Here our results are noticeably different. For the simple-average of voxels within the language network, we found a correlation of -0.204 for the Mean model. This comparatively large negative correlation indicates that brain representations of sentences differ significantly from representations constructed considering only lexical similarity, providing evidence that brain representations of sentences are highly sensitive to sentence structure. The Transformers model achieves a correlation of -0.045, which is significantly higher than the Mean model ($\Delta\rho = 0.159, t = 14.287, p < 0.0001$), though the negative sign indicates that transformers still poorly match brain similarities. The Hybrid model achieves the highest correlation of 0.070, much larger than the Transformers model ($\Delta\rho = 0.115, t = 8.150, p < 0.0001$). The Graph model shows similar results to the Hybrid model, with a correlation of 0.047 ($\Delta\rho = -0.023, t = -1.783, p = 0.0851$). Results were very similar using the group-average method, though generally correlations had higher absolute values. The results for Mean, Transformer, and Hybrid models were all consistent with our preregistered predictions based on previous work with a separate behavioural dataset[31], though we did not make a prediction for the Graph model. In all cases, results are very similar whether computed over the entire cortex (excluding V1–V4) or focusing just on the language network.

To better understand the origin of such large differences in correlations, we plotted neural similarities against the similarities derived from all four computational models (see Figure 2c). For both the Mean and Transformer models, the blue 'modified' and 'substituted' sentence pairs are accorded comparable similarities to the light green 'swapped' sentence pairs. By contrast, the Hybrid and Graph models generally accord 'swapped' sentence pairs as having distinctly lower similarity than 'substituted' and 'modified' sentence pairs. This is easiest to see on the Hybrid subplot of Figure 2c, where the 'swapped' sentence pairs are noticeably to the left of the 'modified' and 'substituted' sentence pairs. Such a difference indicates that the Hybrid and Graph models have a greater ability to discriminate sentence pairs that are lexically similar but structurally different (due to interchanged semantic roles). This leads to sentence similarities which are in better accord with brain similarity data, and thereby drives the positive RSA correlations. These results indicate that when keeping lexical similarity roughly constant, as is the case for the block diagonal sentence pairs, brain similarity patterns are best explained by models that explicitly represent sentence structural elements, namely the Hybrid and Graph models. The Mean model, which completely ignores such structure, explains brain representations the worst, with Transformer models doing better than the Mean but still poorly overall.

We also conducted an analysis of RSA correlations for each layer of the Llama 3 transformer model. We chose this for analysis as a larger, more recent architecture with a large number of layers. As shown in Figure 5,layers 0 and 1 had large negative correlations more similar to the Mean-CN model, while layers 2 and 3 had slightly positive correlations closer to that of the Hybrid model. Layers 4 and on had more moderate negative correlations, with a slight downward trend over later layers. This pattern was largely similar for both the set all all pairwise comparisons and the set of block diagonal comparison pairs, though in the latter case correlations remained essentially constant from around layer 4 onwards. The corresponding RSA matrices (see Figure 5c) show clear differences in representation across layers, though the significance of these patterns is difficult to interpret. We found only modest differences across layers of the AMRbart and ERNIE transformers (see Supplementary Information Figures S6 and S7).

We next compared representations across different brain regions. In addition to the language network and visual cortex (V1–V4), we also considered several regions previously demonstrated to show activity in response to language stimuli, namely the dorsomedial prefrontal cortex, the dorsolateral prefrontal cortex, the posterior cingulate cortex, and the precuneus. The primary somatosensory cortex (S1) is also included as a comparison of a brain region expected to show little response to linguistic stimuli. As shown in Figure 3a, the RSA matri-

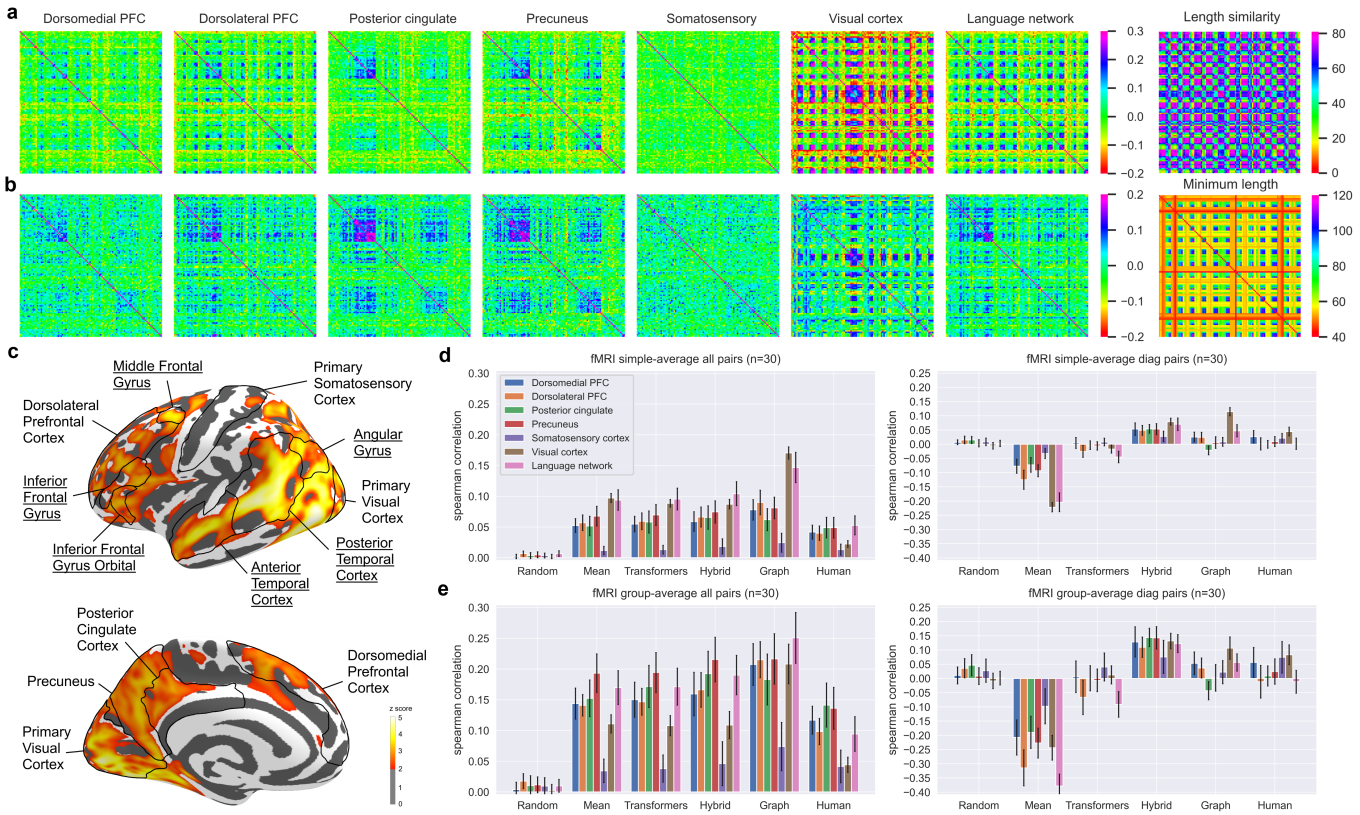

Fig. 3: Comparison of sentence representations and model correlations across brain regions. **a)** RSA matrices for various cortical regions, computed controlling for differences in sentence length. **b)** RSA matrices for various cortical regions, computed controlling for differences in sentence length and minimum sentence length. **c)** Searchlight RSA for the Hybrid model using 8mm radius showing cortical regions of interest, with those part of the language network underlined. RSA correlations are thresholded at z=2. **d)** Partial correlations controlling for differences in sentence length by cortical region, with each individual participant shown as blue dots, and the simple average over individual correlations shown as a bar. **e)** Partial correlations controlling for differences in sentence length computed using the group-averaged RSA matrix, shown by cortical region. Error bars show 95% confidence intervals calculated by bootstrap resampling over participants.

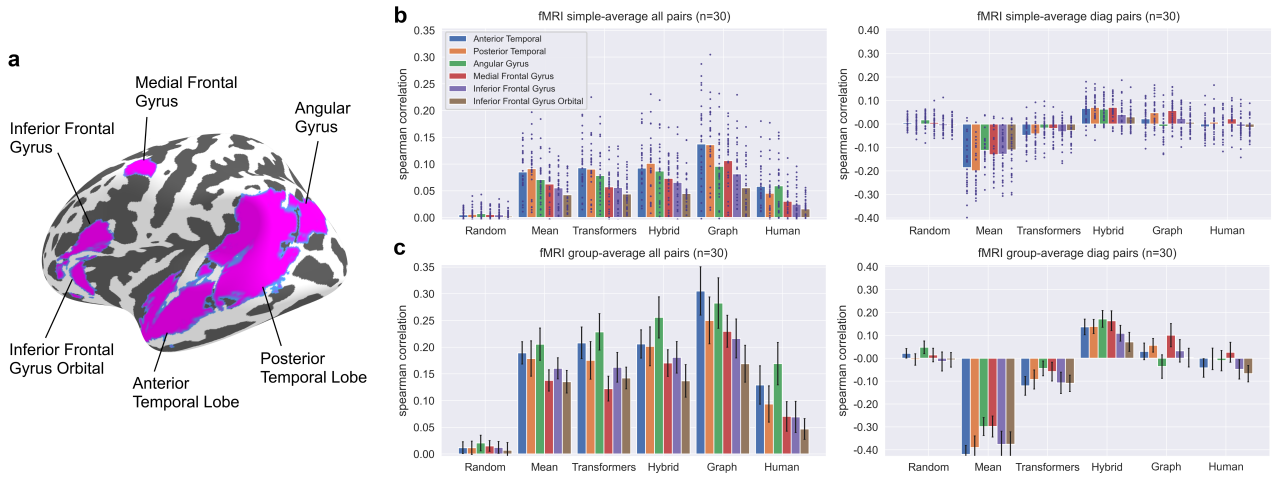

Fig. 4: Comparison of model correlations across subregions of the language network. a) Regions within the language network. **b)** Partial correlations controlling for differences in sentence shown by language network region, with each individual participant shown as blue dots, and the simple average over individual correlations shown as a bar. **c)** Partial correlations controlling for differences in sentence length computed using the group-averaged RSA matrix, shown by language network region. Error bars show 95% confidence intervals calculated by bootstrap resampling over participants.

ces for most of these regions show a very robust grid-like pattern not explained by the type of sentence pair in the design matrix. This effect is not explained by differences in sentence length, as the RSA matrices already control for this variable (shown on the right of Figure 3a). Upon further investigation, we identified the grid-like pattern as resulting from consistently high brain similarity of sentence pairs in which both sentences are relatively long, as measured by the number of characters. This is evident by visual comparison with the 'minimum length' RSA matrix on the right of Figure 3b, which shows the shortest length of the two sentences in each pair. In Supplementary Information Figure S4, we show that our main results are qualitatively similar when additionally controlling for the 'long sentences effect'. After regressing out this effect using the minimum sentence length for each sentence pair (Figure 3b), we recovered a block diagonal structure comparable to the original design matrix shown in Figure 1a, most clearly visible in the language network.

To more clearly visualise the location of the brain regions responsible for encoding sentence information in common with the computational models, we conducted an RSA-searchlight analysis. This involves computing the RSA correlation between each model and the voxel activations within an 8mm sphere surrounding each voxel within a cortical mask. The results (see Figure 2c) show significant correlations throughout the language network, including regions of the temporal lobe, the angular gyrus, and the frontal. Significant correlations are also evident in the posterior cingulate cortex, precuneus, and the visual cortex, with sporadic pockets throughout the dorsolateral and dorsomedial frontal cortical regions. In Figure 3c-d we show the correlations for each model in each region. We observe low correlations for the somatosensory cortex, generally high correlations for the language network, and intermediate correlations for all other regions. For block diagonal sentence pairs, the Hybrid model has similar correlations across all regions, while the Graph model has the highest correlation in the visual cortex, but still positive correlations in the language network. We find similar results when additionally controlling for minimum sentence length, as shown in Supplementary Information Figure S5.

We also performed an analysis comparing the representation of each subregion of the language network, the locations of which are depicted in Figure 4a. We found a similar overall pattern of results within all subregions, with consistently positive correlations for the entire set of pairwise comparisons. The magnitude of the correlations varied across subregions, with the highest values observed for the anterior and posterior temporal lobe, and lower values for all frontal regions (see Figure 4b). For the set of block diagonal sentence pairs, all subregions showed the same pattern as our main results, with a negative correlation for the mean model, modest negative correlations for transformer models, and positive correlations for the hybrid model. These findings support previous results indicating that all subregions of the language network are sensitive to lexical, syntactic, and compositional aspects of language, without any obvious specialisation across subregions[38,39]. We find little difference when additionally controlling for minimum sentence length, as shown in Supplementary Information Figure S8.

1.3 Behavioural results

To supplement our neuromaging data, we also collected a set of behavioural data consisting of semantic similarity judgements. As illustrated in Figure 1e, we recruited 502 participants using an online platform, each of whom was presented with a set of 102 sentence pairs selected randomly from all 5,770 unique sentence pairs. Participants were asked to rate each sentence pair for semantic similarity on a scale of 1-7. Ratings were averaged over participants and scaled to between 0 and 1 for comparison with model similarities. The normalised human sentence similarity ratings ranged from 0 to 0.962, with mean=0.484 and SD=0.171 for block diagonal sentence pairs, and mean=0.072 and SD=0.071 for all other sentence pairs. The average standard deviation of similarity scores for each sentence pair computed across participants was equal to 0.244 for block diagonal sentence pairs and 0.106 for all other pairs. This is comparable to the 0.19 adjusted average standard deviation of the SICK sentence similarity dataset[40], and 0.216 for the STS3k dataset[31]. The split-half reliability with the Spearman-Brown correction was 0.938 for the entire dataset, 0.954 for the block diagonal sentence pairs, and 0.715 for all other pairs, indicating high levels of agreement between participants.

We evaluated the fit between behavioural data and each computational model in the same manner as for the fMRI data. For the full set of sentence pairs (Figure 6a left), the Mean and Transformer models performed best with correlations of 0.510 and 0.568 respectively ($\Delta\rho = 0.049, t = 11.327, p < 0.0001$). The Hybrid model had a lower correlation relative to the Transformers ($\Delta\rho = -0.093, t = -16.432, p < 0.0001$)., and the Graph model the lowest of all ($\Delta\rho = -0.044t = -6.306, p < 0.0001$). This pattern was reversed in the case of the block diagonal sentence pairs (Figure 6a right), with the Mean model having by far the lowest correlation of 0.437. Transformers had a much higher correlation of 0.639 ($\Delta\rho = 0.188, t = 22.449, p < 0.0001$), as did the Hybrid model with a correlation of 0.698 ($\Delta\rho = 0.045, t = 3.765, p = 0.0001$). The Syntax model had an intermediate correlation of 0.533, lower by than the Hybrid model ($\Delta\rho = -0.145, t = -12.371, p < 0.0001$). This pattern of results is comparable to that we observed for our fMRI data, though with much higher correlations across all models owing to the much reduced noise in behavioural ratings compared to fMRI voxel data. Correlations for all computational models are shown in Supplementary Information Figure S3.

As before, we show scatterplots of the human ratings

plotted against model similarities (Figure 6b). While all four models broadly follow the ordering of human ratings along the 45-degree line, both the Mean and OpenAI transformer models place the 'swapped' sentence pairs below the line, meaning that these sentence pairs are accorded higher similarity ratings by the models than by humans. By contrast, the Hybrid and Syntax models place 'swapped' sentence pairs above the 45-degree line, meaning that they accord these sentence pairs lower similarities than humans do. These results indicate that for this set of stimuli, the Mean and OpenAI transformer models are less sensitive to variations in sentence structure than human raters, while the Hybrid and Syntax models are slightly more sensitive to such structure than the human raters.

2 Discussion

In this paper we present, to our knowledge, the first fMRI evaluation of models of sentence representation that utilises stimuli specifically designed to distinguish the effects of lexical semantics from sentence structure. We also present the first quantitative comparison of static word embeddings, transformer neural networks, semantic parsing graphs, and hybrid models of sentence representation using a unified framework. In our neuroimaging experiment we found that over the block diagonal sentence pairs (the subset of sentence pairs designed to test for sensitivity to sentence structure), considering voxels in the language network, the Mean model had a strong negative correlation, the Transformers model a smaller negative correlation, and the Hybrid model a modest positive correlation. We found similar (though less pronounced) differences in our behavioural experiment. These findings provide two major contributions to our knowledge of sentence representation in the brain. First, we show that controlling for lexical similarity illuminates the brain's sensitivity to sentence structure in a way that is not evident when the lexical confound is present. Second, the success of our Hybrid model provides novel insight into how sentence structure is represented in the brain, indicating the importance of semantic role and highlighting limitations of representations derived from transformer models.

Previous studies analysing sentence processing in the brain have used a variety of controlled stimuli to isolate the mechanisms of semantic composition. One method involves randomly shuffle the order of words within a sentence, thereby preserving lexical semantics while varying overall sentence meaning[41,42]. A second method involves constructing 'jabberwocky' sentences, which involve nonsensical words placed in grammatically well-formed sentences[38,43,44]. These stimuli are designed to control for syntactic structure or sentence form while manipulating sentence meaning. In both cases, the objective is typically to use jabberwocky or shuffled sentences as a control condition in which composition is prevented, thereby providing a baseline for sentences in which composition occurs[45]. Our study differs from these approaches in that we aim to preserve, rather than prevent composition. Instead, we control for lexical similarity while constructing semantically meaningful sentences with differing meanings.

Our results indicate that transformer representations do not adequately incorporate sentence structure in a brain-like manner. In particular, we show that when evaluated against both the brain and behavioural data, transformers are insufficiently sensitive to 'swapping' of semantic roles (see Figure 6 and Figure 2), ranking such sentence pairs as more similar than do human participants (in the behavioural data) or brain representations (in the fMRI data). This effect was very robust, with negative correlations observed for block diagonal sentence pairs for all eight transformers we studied (see Supplementary Information Figure S2). By contrast, the Hybrid model is by design highly sensitive to such alterations of semantic roles, which better matches the pattern of brain similarities than other models. Indeed, for the behavioural data we find that the hybrid model is actually more sensitive to these 'swapped' sentences compared to human participants, who rate their similarity in between that of the Hybrid and Transformer models.

Several previous studies have found that voxelwise encoding models trained using features extracted from transformers are able to better predict brain activity than static word embedding models which ignore sentence structure[15–17,46]. However, interpreting these findings is difficult because there is no established method for determining which model features drive these correlations[47]. Indeed, some studies have found that even features from untrained transformers can achieve high voxelwise correlations[15,48], casting doubt on whether the transformer features which drive brain correlations are linguistically relevant. Similarly, other studies using shuffled sentences to remove information about sentence structure have found this results in only modest reductions in voxelwise correlations[49,50]. An analyses which better controlled for various confounds found that most variance explainable by transformers was accounted for by static word embeddings and word rate[51]. Our results complement these findings, showing that in cases where sentence structure is critical, transformers are insufficiently sensitive to structural aspects of sentence meaning. In cases where transformers have been found to have an advantage, this may be due to their greater ability to contextualise polysemous word meanings based on the presence of other words, rather than their ability to represent sentence structure.

We analysed the fMRI data in two different ways: computing correlations for each individual participant and taking a simple average, and also computing a group-averaged RSA matrix and then computing the correlation. We found that both methods yielded a very similar pattern of results, but with the group-averaged correlations having about twice the magnitude of the simple-average correlations (see Figure 2). For instance, the

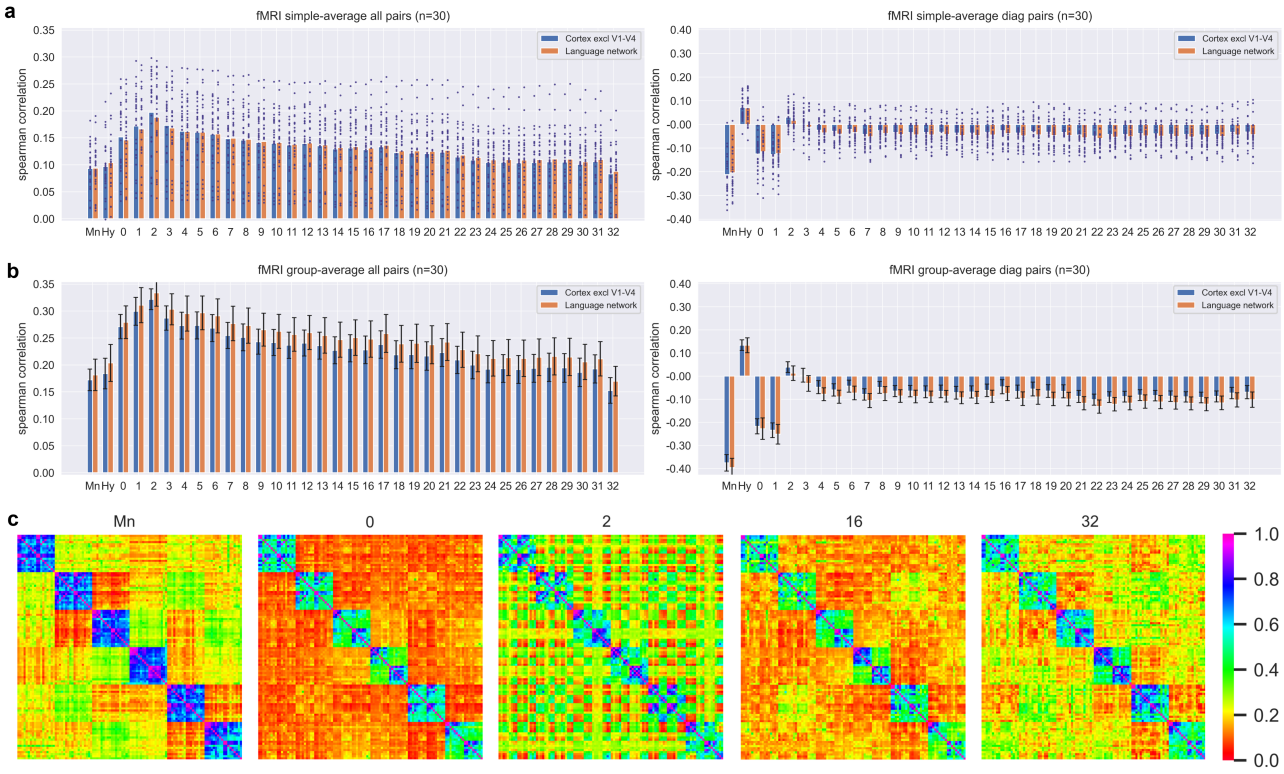

Fig. 5: Average correlations between RSA matrices of each layer of Llama 3 and brain RSA matrix of each participant. Mean (Mn) and Hybrid (Hy) models are also shown for comparison. **a)** Partial correlations for each individual participant shown as blue dots, with the simple average over individual correlations shown as a bar. **b)** Partial correlations computed using the group-averaged RSA matrix. Error bars show 95% confidence intervals calculated by bootstrap resampling over participants. **c)** RSA matrices for the Mean model (Mn) along with selected layers of the Llama 3 model, computed controlling for differences in sentence length.

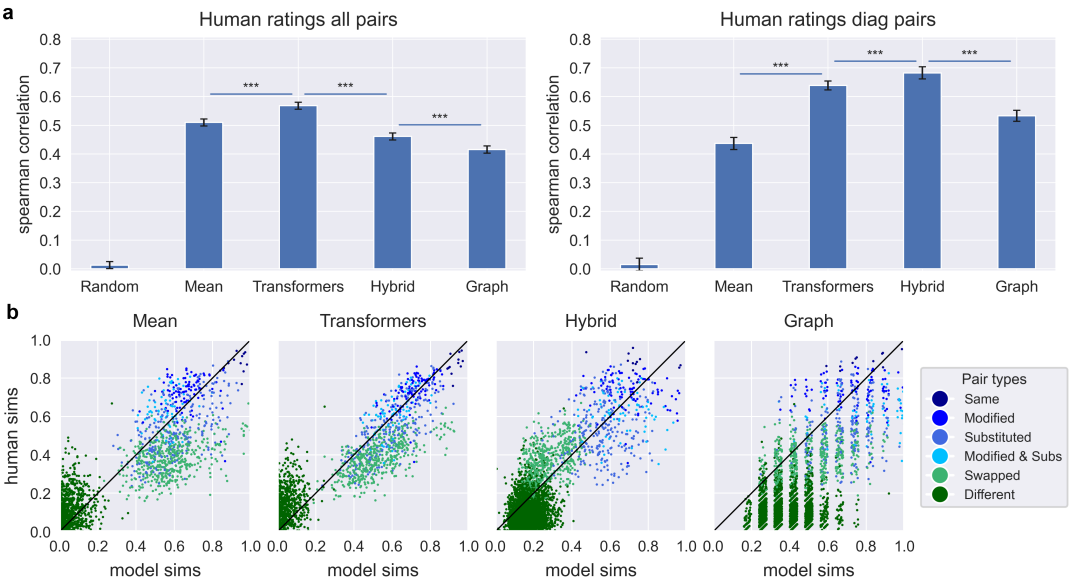

Fig. 6: Behavioural ratings of sentence similarity show similar results to fMRI results, but with higher absolute correlations. **a)** (Left) Average correlations between RSA matrices of four computational models and human-rated similarities using all sentence pairs. (Right) Average correlations between RSA matrices of four computational models and human-rated similarities using only diagonal sentence pairs. **b)** Scatterplots showing the relationship between model similarities (horizontal axis) and human rated similarities (vertical axis) for all four computational models. Each dot corresponds to a single pairwise similarity, scaled to between 0 and 1 for visualisation. The 45-degree line (black) shows a hypothetical line of perfect fit between model and human similarities.

correlation over block diagonal sentence pairs for the Mean model is -0.204 when averaged over individual subjects, and -0.378 when computed using the group-averaged RSA. Likewise, the hybrid model correlation is 0.070 when averaged over individual subjects, and 0.122 when computed using the group-averaged RSA. These results are likely explained by the highly noisy results at the individual level, which is partly averaged-out when computing the group-averaged RSA matrix.

We also found a robust 'minimum sentence length' effect, in which sentence pairs consisting of two long sentences resulted in highly similar brain activity (see Figure 3a-b). This is not explained by similar length sentences eliciting similar brain activity, as the effect does not arise for pairs consisting of two short or two medium-length sentences. Furthermore, all RSA partial correlations already control for the similarity of sentence length. Though we are not aware of this result having been reported using RSA, previous studies using other methodologies have found that activation of the language network increases with sentence length[38,52–54]. The cause of this effect is unclear. It may partly be explained by visual similarity of longer sentences, however we observe no similar effect that might be expected for the visual similarity of short or medium sentences. Furthermore, the minimum length effect is also evident in many brain regions outside the visual cortex, including the language network and frontal regions Figure 3. We speculate that the effect may be driven by multiple causes, including increased cognitive processing or memory load for processing longer sentences, greater depth of processing elicited by semantically richer stimuli[55], or additional processing required for compositional combination of a larger number of sentence components. It is also possible that the structural similarity of longer sentences in our study, which all contain a similar set of semantic roles, results in similar brain representations even when the sentences do not have similar overall meanings. If so, this would indicate that extracting semantic features is important for brain processing of sentences even aside from lexical similarity. Further research will be required to disentangle the relative impacts of these distinct processes.

Our study has several limitations. First, our stimulus set consists of a relatively small selection of sentences, which follow broadly similar structure. Our aim in this study was to disentangle the effects of lexical similarity from structural similarity in realistic sentences, and as such we did not attempt to compile a representative sample of sentences from natural dialogue. In future work we hope to investigate the extent to which our results generalise to more complex and varied types of sentences. Second, the comparison between behavioural and fMRI data is somewhat difficult owing to the difference in task structure. In the behavioural experiment, participants viewed many pairs of related sentences, and explicitly asked to pay attention to differences in the words of each sentence. In contrast, in the fMRI task participants read one sentence at a time without an explicit comparison. Third, we analyse brain representations of sentence meaning over a single contiguous 3s interval. This is a substantial simplification of sentence processing, which takes place dynamically over time as words are successively integrated to form progressively more complex and structured representations[22,38,56–58]. While our approach is an important contribution, and builds upon previous studies comparing syntactic parse trees with brain data[52,59,60], additional work is needed to link model representations with the dynamic cascade of brain activity during sentence processing.

While our results show that transformers do not represent sentences in a manner comparable to the brain, it is likely that individual features within transformer embeddings do represent aspects of sentence structure. Indeed, large language models show clear capabilities of correctly interpreting sentence structure[61], and probing studies have found that transformers represent information about syntax and word order[62,63]. Nonetheless, the fact that transformers can encode and utilise structural information to perform linguistic tasks does not mean that they effectively utilise this information to construct a brain-like representation of sentence meaning. Our results indicate that despite the linguistic competencies of transformers, they do not combine syntactic and semantic information into an integrated sentence representation in a manner analogous to the human brain. Further research is needed to investigate exactly which features of sentence meaning are represented by large large models, and how they differ from those encoded in the human brain.

Our results provide important new insights about how sentence structure is represented in the brain. The simple Mean model, which ignores sentence structure, was a very poor match to brain activity when evaluated against the block diagonal subset of sentences (the sentence pairs designed to be difficult for models which do not represent sentence structure). While transformers were a much better match to brain activity than the Mean model, correlations were still negative, indicating that transformer representations were still a poor match to brain representation. In line with our preregistered prediction, we found that the Hybrid model best matched brain representations, thereby providing evidence that the brain incorporates structured information from semantic roles when representing sentence meaning. Evidently, such structure is not always adequately represented even in state-of-the-art transformer models. Our results highlight the importance investigating which semantic features are most important for representation of sentence meaning in the brain.

Acknowledgements

We would like to thank the staff at the Melbourne Brain Centre Imaging Unit for their assistance with collecting the fMRI scans. **Funding**: This research was supported by a University of Melbourne Graduate Research

Scholarship from the Faculty of Business and Economics (Fodor). **Author contributions**: Conceptualisation: J.F.; Methodology: J.F., C.M., S.S.; Investigation: S.S.; Formal analysis: J.F.; Visualisation: J.F.; Writing – original draft: J.F.; Writing – review editing: J.F., C.M., S.S. **Competing interests**: The authors declare that they have no competing interests. **Data and materials availability**: Data will be uploaded to OpenNeuro upon publication, with code available from github.

References

1. Mitchell, T. *et al.* Predicting human brain activity associated with the meanings of nouns. *Science* **320,** 1191–1195. ISSN: 0036-8075 (2008).
2. Just, M., Cherkassky, V. L., Aryal, S. & Mitchell, T. M. A neurosemantic theory of concrete noun representation based on the underlying brain codes. *PloS one* **5,** e8622. ISSN: 1932-6203 (2010).
3. Wehbe, L. *et al.* Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PloS one* **9,** e112575. ISSN: 1932-6203 (2014).
4. Huth, A. G., De Heer, W. A., Griffiths, T. L., Theunissen, F. E. & Gallant, J. L. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature* **532,** 453–458. ISSN: 1476-4687 (2016).
5. Pereira, F. *et al.* Toward a universal decoder of linguistic meaning from brain activation. *Nature communications* **9,** 1–13. ISSN: 2041-1723 (2018).
6. Günther, F., Rinaldi, L. & Marelli, M. Vector-space models of semantic representation from a cognitive perspective: A discussion of common misconceptions. *Perspectives on Psychological Science* **14,** 1006–1033. ISSN: 1745-6916 (2019).
7. Karamolegkou, A., Abdou, M. & Søgaard, A. Mapping Brains with Language Models: A Survey. *arXiv preprint arXiv:2306.05126* (2023).
8. Ouyang, L. *et al.* Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* **35,** 27730–27744 (2022).
9. Touvron, H. *et al.* Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
10. Team, G. *et al.* Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).
11. Bhatia, S., Richie, R. & Zou, W. Distributed semantic representations for modeling human judgment. *Current Opinion in Behavioral Sciences* **29,** 31–36 (2019).
12. Erk, K. The probabilistic turn in semantics and pragmatics. *Annual Review of Linguistics* **8,** 101–121 (2022).
13. Tuckute, G., Kanwisher, N. & Fedorenko, E. Language in brains, minds, and machines. *Annual Review of Neuroscience* **47,** 277–301 (2024).
14. Anderson, A. *et al.* Deep artificial neural networks reveal a distributed cortical network encoding propositional sentence-level meaning. *Journal of Neuroscience* **41,** 4100–4119. ISSN: 0270-6474 (2021).
15. Schrimpf, M. *et al.* The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences* **118.** ISSN: 0027-8424 (2021).
16. Antonello, R., Turek, J. S., Vo, V. & Huth, A. Low-dimensional structure in the space of language representations is reflected in brain responses. *Advances in neural information processing systems* **34,** 8332–8344 (2021).
17. Pasquiou, A., Lakretz, Y., Hale, J. T., Thirion, B. & Pallier, C. *Neural Language Models are not Born Equal to Fit Brain Data, but Training Helps* in *International Conference on Machine Learning* (PMLR, 2022), 17499–17516.
18. Aliko, S., Huang, J., Gheorghiu, F., Meliss, S. & Skipper, J. I. A naturalistic neuroimaging database for understanding the brain using ecological stimuli. *Scientific Data* **7,** 1–21. ISSN: 2052-4463 (2020).
19. Hamilton, L. S. & Huth, A. G. The revolution will not be controlled: natural stimuli in speech neuroscience. *Language, cognition and neuroscience* **35,** 573–582. ISSN: 2327-3798 (2020).
20. Nastase, S. A. *et al.* The "Narratives" fMRI dataset for evaluating models of naturalistic language comprehension. *Scientific data* **8,** 250 (2021).
21. Zhang, Y., Kim, J.-H., Brang, D. & Liu, Z. Naturalistic stimuli: A paradigm for multiscale functional characterization of the human brain. *Current opinion in biomedical engineering* **19,** 100298. ISSN: 2468-4511 (2021).
22. Li, J., Lai, M. & Pylkkänen, L. Semantic composition in experimental and naturalistic paradigms. *Imaging Neuroscience* **2,** 1–17 (2024).
23. Kriegeskorte, N. & Douglas, P. K. Interpreting encoding and decoding models. *Current opinion in neurobiology* **55,** 167–179 (2019).
24. Bruffaerts, R. *et al.* Redefining the resolution of semantic knowledge in the brain: advances made by the introduction of models of semantics in neuroimaging. *Neuroscience & Biobehavioral Reviews* **103,** 3–13. ISSN: 0149-7634 (2019).

25. Hagoort, P. The meaning-making mechanism(s) behind the eyes and between the ears. *Philosophical Transactions of the Royal Society B* **375,** 20190301. ISSN: 0962-8436 (2020).

26. Wang, X. *et al.* Organizational principles of abstract words in the human brain. *Cerebral Cortex* **28,** 4305–4318. ISSN: 1047-3211 (2018).

27. Fernandino, L., Tong, J.-Q., Conant, L. L., Humphries, C. J. & Binder, J. R. Decoding the information structure underlying the neural representation of concepts. *Proceedings of the National Academy of Sciences* **119,** e2108091119. ISSN: 0027-8424 (2022).

28. Tong, J. *et al.* A distributed network for multimodal experiential representation of concepts. *Journal of Neuroscience* **42,** 7121–7130 (2022).

29. Acunzo, D. J., Low, D. M. & Fairhall, S. L. Deep neural networks reveal topic-level representations of sentences in medial prefrontal cortex, lateral anterior temporal lobe, precuneus, and angular gyrus. *NeuroImage* **251,** 119005. ISSN: 1053-8119 (2022).

30. Fairhall, S. L. Sentence-level embeddings reveal dissociable word-and sentence-level cortical representation across coarse-and fine-grained levels of meaning. *Brain and Language* **250,** 105389 (2024).

31. Fodor, J., De Deyne, S. & Suzuki, S. Compositionality and Sentence Meaning: Comparing Semantic Parsing and Transformers on a Challenging Sentence Similarity Dataset. *Computational Linguistics,* 1–52 (2024).

32. Lake, B. M. & Murphy, G. L. Word meaning in minds and machines. *Psychological Review* **130,** 401–31. ISSN: 1939-1471 (2021).

33. Žabokrtskỳ, Z., Zeman, D. & Ševčíková, M. Sentence meaning representations across languages: what can we learn from existing frameworks? *Computational Linguistics* **46,** 605–665 (2020).

34. Anderson, A., Kiela, D., Clark, S. & Poesio, M. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association for Computational Linguistics* **5,** 17–30 (2017).

35. Just, M., Wang, J. & Cherkassky, V. L. Neural representations of the concepts in simple sentences: Concept activation prediction and context effects. *Neuroimage* **157,** 511–520. ISSN: 1053-8119 (2017).

36. Fedorenko, E., Ivanova, A. A. & Regev, T. I. The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience* **25,** 289–324 (2024).

37. Handjaras, G. *et al.* How concepts are encoded in the human brain: a modality independent, category-based cortical organization of semantic knowledge. *Neuroimage* **135,** 232–242. ISSN: 1053-8119 (2016).

38. Shain, C. *et al.* Distributed sensitivity to syntax and semantics throughout the language network. *Journal of Cognitive Neuroscience* **36,** 1427–1471 (2024).

39. Fedorenko, E., Blank, I. A., Siegelman, M. & Mineroff, Z. Lack of selectivity for syntax relative to word meanings throughout the language network. *Cognition* **203,** 104348 (2020).

40. Marelli, M. *et al. A SICK cure for the evaluation of compositional distributional semantic models* in *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)* (2014), 216–223.

41. Goucha, T. & Friederici, A. D. The language skeleton after dissecting meaning: A functional segregation within Broca's area. *NeuroImage* **114,** 294–302 (2015).

42. Mollica, F. *et al.* Composition is the core driver of the language-selective network. *Neurobiology of Language* **1,** 104–134. ISSN: 2641-4368 (2020).

43. Fedorenko, E., Nieto-Castanon, A. & Kanwisher, N. Lexical and syntactic representations in the brain: an fMRI investigation with multi-voxel pattern analyses. *Neuropsychologia* **50,** 499–513 (2012).

44. Matchin, W., Hammerly, C. & Lau, E. The role of the IFG and pSTS in syntactic prediction: Evidence from a parametric study of hierarchical structure in fMRI. *cortex* **88,** 106–123 (2017).

45. Călinescu, L., Ramchand, G. & Baggio, G. How (not) to look for meaning composition in the brain: A reassessment of current experimental paradigms. *Frontiers in Language Sciences* **2,** 1096110 (2023).

46. Oota, S. R., Gupta, M. & Toneva, M. *Joint processing of linguistic properties in brains and language models* in *Advances in Neural Information Processing Systems* **36** (2024).

47. Feghhi, E., Hadidi, N., Song, B., Blank, I. A. & Kao, J. C. What Are Large Language Models Mapping to in the Brain? a Case Against Over-Reliance on Brain Scores. *arXiv preprint arXiv:2406.01538* (2024).

48. Hosseini, E. A. *et al.* Artificial Neural Network Language Models Predict Human Brain Responses to Language Even After a Developmentally Realistic Amount of Training. *Neurobiology of Language* **5,** 43–63. ISSN: 2641-4368. `https://doi.org/10.1162/nol%5C_a%5C_00137` (Apr. 2024).

49. Caucheteux, C., Gramfort, A. & King, J.-R. *Disentangling syntax and semantics in the brain with deep networks* in *International conference on machine learning* (2021), 1336–1348.

50. Kauf, C., Tuckute, G., Levy, R., Andreas, J. & Fedorenko, E. 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,** 7–42. ISSN: 2641-4368 (2024).

51. Hadidi, N., Feghhi, E., Song, B. H., Blank, I. A. & Kao, J. C. Illusions of Alignment Between Large Language Models and Brains Emerge From Fragile Methods and Overlooked Confounds. *bioRxiv,* 2025–03 (2025).

52. Nelson, M. J. *et al.* Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences* **114,** E3669–E3678. ISSN: 0027-8424 (2017).

53. Schuster, S., Hawelka, S., Himmelstoss, N. A., Richlan, F. & Hutzler, F. The neural correlates of word position and lexical predictability during sentence reading: Evidence from fixation-related fMRI. *Language, Cognition and Neuroscience* **35,** 613–624 (2020).

54. Woolnough, O. *et al.* Spatiotemporally distributed frontotemporal networks for sentence reading. *Proceedings of the National Academy of Sciences* **120,** e2300252120. ISSN: 0027-8424 (2023).

55. Deniz, F., Tseng, C., Wehbe, L., la Tour, T. D. & Gallant, J. L. Semantic representations during language comprehension are affected by context. *Journal of Neuroscience* **43,** 3144–3158. ISSN: 0270-6474 (2023).

56. Baron, S. G., Thompson-Schill, S. L., Weber, M. & Osherson, D. An early stage of conceptual combination: Superimposition of constituent concepts in left anterolateral temporal lobe. *Cognitive neuroscience* **1,** 44–51 (2010).

57. Frankland, S. M. & Greene, J. D. Concepts and compositionality: in search of the brain's language of thought. *Annual review of psychology* **71,** 273–303. ISSN: 0066-4308 (2020).

58. Desbordes, T., King, J.-R. & Dehaene, S. Tracking the neural codes for words and phrases during semantic composition, working-memory storage, and retrieval. *Cell Reports* **43** (2024).

59. Stanojević, M., Brennan, J. R., Dunagan, D., Steedman, M. & Hale, J. T. Modeling structure-building in the brain with CCG parsing and large language models. *Cognitive science* **47,** e13312 (2023).

60. Fresen, A. J., Choenni, R., Heilbron, M., Zuidema, W. & de Heer Kloots, M. *Language Models That Accurately Represent Syntactic Structure Exhibit Higher Representational Similarity To Brain Activity* in *Proceedings of the Annual Meeting of the Cognitive Science Society* **46** (2024).

61. Chang, T. A. & Bergen, B. K. Language model behavior: A comprehensive survey. *Computational Linguistics* **50,** 293–350 (2024).

62. Clark, K., Khandelwal, U., Levy, O. & Manning, C. D. *What Does BERT Look at? an Analysis of BERT's Attention* in *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (2019), 276–286.

63. Manning, C. D., Clark, K., Hewitt, J., Khandelwal, U. & Levy, O. Emergent linguistic structure in artificial neural networks trained by self-supervision. *Proceedings of the National Academy of Sciences* **117,** 30046–30054 (2020).

64. Timkey, W. & van Schijndel, M. *All Bark and No Bite: Rogue Dimensions in Transformer Language Models Obscure Representational Quality* in *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (2021), 4527–4546.

65. Cai, X., Huang, J., Bian, Y. & Church, K. *Isotropy in the contextual embedding space: Clusters and manifolds* in *International Conference on Learning Representations* (2021).

66. Conneau, A., Kiela, D., Schwenk, H., Barrault, L. & Bordes, A. *Supervised Learning of Universal Sentence Representations from Natural Language Inference Data* in *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Copenhagen, Denmark, Sept. 2017), 670–680. `https://aclanthology.org/D17-1070`.

67. Cer, D. *et al. Universal sentence encoder for English* in *Proceedings of the 2018 conference on empirical methods in natural language processing: system demonstrations* (2018), 169–174.

68. Sun, Y. *et al.* Ernie 2.0: A continual pre-training framework for language understanding. *Proceedings of the AAAI conference on artificial intelligence* **34,** 8968–8975 (2020).

69. Reimers, N. & Gurevych, I. *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks* in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (2019), 3982–3992.

70. Tsukagoshi, H., Sasano, R. & Takeda, K. *DefSent: Sentence Embeddings using Definition Sentences* in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)* (2021), 411–418.

71. Bai, X., Chen, Y. & Zhang, Y. *Graph Pre-training for AMR Parsing and Generation* in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (2022), 6001–6015.

72. Opitz, J. & Frank, A. *SBERT studies Meaning Representations: Decomposing Sentence Embeddings into Explainable Semantic Features* in *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)* (2022), 625–638.

73. Bevilacqua, M., Blloshmi, R. & Navigli, R. *One SPRING to Rule Them Both: Symmetric AMR Semantic Parsing and Generation without a Complex Pipeline* in *Proceedings of AAAI* (2021), 12564–12573.

74. Goyal, P. & Ferrara, E. Graph embedding techniques, applications, and performance: A survey. *Knowledge-Based Systems* **151,** 78–94 (2018).

75. Rossi, A., Barbosa, D., Firmani, D., Matinata, A. & Merialdo, P. Knowledge graph embedding for link prediction: A comparative analysis. *ACM Transactions on Knowledge Discovery from Data (TKDD)* **15,** 1–49 (2021).

76. Cai, S. & Knight, K. *Smatch: an evaluation metric for semantic feature structures* in *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)* (2013), 748–752.

77. Opitz, J., Daza, A. & Frank, A. Weisfeiler-leman in the bamboo: Novel AMR graph metrics and a benchmark for AMR graph similarity. *Transactions of the Association for Computational Linguistics* **9,** 1425–1441 (2021).

78. Esteban, O. *et al.* fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods* **16,** 111–116 (2019).

79. Prince, J. S. *et al.* Improving the accuracy of single-trial fMRI response estimates using GLMsingle. *Elife* **11,** e77599. ISSN: 2050-084X (2022).

80. Abe, K., Yokoi, S., Kajiwara, T. & Inui, K. *Why is sentence similarity benchmark not predictive of application-oriented task performance?* in *Proceedings of the 3rd Workshop on Evaluation and Comparison of NLP Systems* (2022), 70–87.

81. Abdalla, M., Vishnubhotla, K. & Mohammad, S. *What Makes Sentences Semantically Related? A Textual Relatedness Dataset and Empirical Study* in *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics* (2023), 782–796.

82. Glasser, M. F. *et al.* A multi-modal parcellation of human cerebral cortex. *Nature* **536,** 171–178 (2016).

83. Nili, H. *et al.* A toolbox for representational similarity analysis. *PLoS computational biology* **10,** e1003553 (2014).

3 Methods

3.1 Stimuli and computational models

3.1.1 Word embedding models

In this study we compared four different approaches for representing sentence meaning. The baseline for all comparisons was the Mean model in which sentence embeddings are constructed by elementwise averaging of word embeddings. We also evaluated two alternative models for combining word embeddings into sentence embeddings. Multiplicative (Mult) embeddings were constructed by adding one to each element of the word embeddings (to avoid negative numbers), then performing elementwise multiplication of all word embeddings. Convolutional (Conv) embeddings were constructed by adding one to each element of the word embeddings, then iteratively performing circular discrete convolution of each word embedding with the convolution of all previous word embeddings. For all three models based on word embeddings, sentence embeddings were constructed after removing a list of stop words containing words with little semantic content such as pronouns, modal verbs, conjunctions, and common prepositions. Cosine similarity was used to compute the similarity of each pair of sentence embeddings.

3.1.2 Transformer models

We computed the representations for a range of transformer architectures, along with the older InferSent LSTM model for comparison, as summarised in Table 1. As per our preregistration, for the statistical analysis we averaged the RSA correlation with brain representations over five different transformer architectures: ERNIE 2.0, AMRbart, SentBERT, DefSent, and OpenAI. For all transformers, sentence embeddings were normalised by subtracting the mean and dividing by the standard deviation of each feature. This is motivated by research indicating that without normalisation, transformers tend to learn very anisotropic embeddings with a few dimensions dominating over all the others[64,65]. Sentence similarities were computed using cosine similarity.

3.1.3 Graph models

We adopted AMR as a representative graph-based approach for representing sentence meaning. We used the SapienzaNLP (Spring) AMR parser[73] to parse all sentences, as it is among the best-performing AMR parses with freely available and easily implementable code.

Table 1: Summary of models of sentence meaning analysed in this study.

Model name	Explanation	Citation
InferSent	A bi-directional LSTM trained on a variety of natural language inference tasks.	66
USE	Universal Sentence Encodings, based on a transformer architecture trained on a range of language tasks. Sentence embedding constructed by averaging all token embeddings from the final layer.	67
ERNIE 2.0	A transformer based on the BERT architecture trained using multi-task learning. Sentence embedding constructed by averaging all token embeddings from the final layer.	68
SentBERT	A sentence transformer model with additional training to predict paired sentences from a large dataset. We used the mpnet embeddings.	69
DefSent-cls	A RoBERTa-large transformer adapted to produce sentence embeddings with fine-tuning on about 100,000 words paired with their dictionary definitions.	70
AMRbart	A sentence transformer architecture trained to encode AMR graphs.	71
S3BERT	A sentence transformer based on SentBERT with extra training to use AMR graph-based metrics to construct an overall similarity score.	72
OpenAI	Embeddings provided from the OpenAI API model text-embedding-ada-002, based on a large transformer with additional fine-tuning from human feedback.	8
Llama 3	A decoder-only transformer architecture trained using multi-task learning with additional fine-tuning from human feedback. Embeddings extracted from the 8B model. Sentence embedding constructed by averaging all token embeddings from the final layer.	9

Evaluating syntax-based models using STS datasets requires a method for computing the similarity between the graphs for each sentence. While various techniques have been developed for converting graphs into vector embeddings, these have typically focused on knowledge databanks rather than natural language[74,75]. Furthermore, we are interested in testing graph-based models of representing sentences more directly, rather than the embeddings produced from these graphs. As such, we analyse the similarity of AMR-embeddings using two existing methods for comparing graph similarity directly: SMATCH[76] and WWLK[77]. In the main manuscript we report the results for the more widely-used SMATCH metric, as it achieved much higher correlations than the WWLK metric.

3.1.4 Hybrid models

To compute representations for the VerbNet-CN hybrid model, we used the GPT-4 model of the OpenAI Chat Completions API to parse each of the 108 sentences by assigning parts of the sentence to one of eight semantic roles: Verb, Agent, Patient, Theme, Time, Manner, Location, Trajectory. After parsing by semantic role, embeddings for each semantic role as before, by averaging the static ConceptNet embeddings of each constituent word after the removal of stop words. Words that are not associated with any semantic role were discarded. As before, the result is a set of role embeddings which constitutes the representation of the meaning of the sentence in terms of vector representations of each major semantic role.

To compute the similarity between two sentences, we first aligned the two sentences based on the semantic roles. Matching semantic roles were then accorded a similarity of 0.5 plus the computed cosine similarity between the rolewise embeddings. In cases where the semantic role was present in one sentence but not the other, a rolewise similarity of zero was used. Overall sentence similarity was computed as the weighted average of these eight rolewise similarities. We used fixed weights of 3 for the Verb, and 2 for Agent, Patient, and Theme, and 0.5 for Time, Manner, Location, and Trajectory, adopted from our previous study[31].

To compute representations for the AMR-CN hybrid model[31], we first parsed sentences using the SapienzaNLP (Spring) AMR parser[73]. Each token in the sentence was then assigned an 'AMR role' in accordance with its location in the parse tree by concatenating all nested parse labels. Role similarities were computed as the cosine similarity between the averaged ConceptNet word embeddings for all tokens with the same AMR role in each sentence of a sentence pair. Finally, the overall sentence similarity was computed as average role similarity over all roles found.

3.1.5 Sentence stimuli

A set of 108 sentences was hand-crafted specifically for this study. The aim was to include sentence pairs rang-

ing from very similar to very dissimilar, while also providing for many pairwise sentence comparisons in which lexical similarity was high while overall meaning was different, owing to interchanging of semantic roles in the sentence. This allows for better model discrimination by ensuring that only models sensitive to sentence structure are able to accurately differentiate the meaning of such interchanged sentences.

The process by which sentences were constructed is summarised in Table 2. All sentences consisted of a single clause written in the active voice describing a specific event. Pronouns, proper nouns, and subordinate clauses were excluded for simplicity and to limit sources of syntactic variation. Sentences were produced by constructing systematic variations of an initial 'base' sentence by altering elements such as the subject, verb, and object, or adding modifiers like adjectives, location, or time. Several different categories of modified sentences were constructed. A small number of 'same' sentences were constructed by adding a single adjective with only minimal effect on sentence meaning, for example 'the equipment' becomes 'the new equipment'. 'Modified' sentences were constructed by adding two or three modifier elements such as location, manner, or time when the event occurred. 'Substituted' sentences were designed to investigate the effect of altering key sentence elements, such as changing the subject, object, or verb of the sentence.

Critical to the study design was construction of 'swapped' sentences, in which one or more pairs of words interchanged roles in the sentence. For example, if in the initial sentence the subject is 'the cameraman', the direct object 'the equipment', and the indirect object 'the director', then in the interchanged sentence the subject is now 'the director', the direct object is 'the cameraman', and the indirect object is 'the equipment'. As with the 'base' sentence, the swapped sentences were also systematically varied through substitutions and addition of modifiers. The aim of this procedure was to develop a set of sentence pairs with gradations of similarity while approximately controlling for lexical similarity. Differences in meaning in these sets of sentences are therefore mostly attributable to sentence structure and semantic roles, not simply use of different words. The complete set of stimuli are provided in Supplementary Information.

Using the methods described above, six distinct subsets each consisting of 18 related sentences were developed. This resulted in 5,778 pairwise comparisons across all sentences, of which 4,860 were 'different' sentence pairs and 918 were block diagonal sentence pairs of primary interest in this study. The RSA design matrix for all 108 sentences is shown in Figure 1a.

In our study preregistration (see https://osf.io/jme7x), we predicted that over the block diagonal set of sentence pairs, the Hybrid model would have a higher correlation with brain representations than the average over five specified Transformers, which in turn would have a higher correlation than the Mean model. We did not make predictions for any other models.

3.2 fMRI data collection

3.2.1 Participants

Thirty-nine participants (23 women, 14 men, 2 other) between the ages of 18 and 40 (mean=22.2) were recruited from our university campus (The University of Melbourne) for the study. All self-identified as native speakers of English, and all but one (a last-minute replacement) identified as right-handed. Participants received $70 as compensation for their time, which corresponds to about $23 per hour for a three-hour session. Nine participants were excluded from the main analysis: seven for scoring below 70% on the attention task (see details below), and two for head motion exceeding 4mm maximum framewise deviation averaged over eight runs, leaving data from 30 participants for subsequent analysis. Note that owing to somewhat poorer performance of participants compared to those in our pilot, we lowered the cutoff slightly from the 75% stated in the preregistration, which led to the inclusion of a single additional participant who scored 73%. In Supplementary Information Figure S11 we show that accuracy on attention check questions had a strong association with model correlations.

The study protocol was approved by the University of Melbourne Human Research Ethics Committee (Reference Number: 2023-28035-47583-3).

3.2.2 Experimental task

While undergoing scanning, participants were presented with a set of 108 sentences, each shown one at a time. They were instructed simply to read each sentence and think about its meaning. Sentence timing was varied with the length of the sentence, to allow sufficient time for reading longer sentences while avoiding leaving time for participants to engage in mind wandering after reading the shorter sentences. The time for each sentence was computed using a quadratic formula in the number of characters, with parameters chosen based on feedback from pilot participants. Presentation time ranged from 2-7 seconds, with an average of 4.29 seconds per sentence. The inter-stimulus interval was selected from a uniform random distribution between 2-7s, with an average of 4.5s. The order of sentences was randomised separately for each participant, with 54 sentences presented during each 508s run. The entire set of 108 sentences was presented every two runs, such that upon completion of all eight runs participants had seen each sentence four times. For five participants, only six runs were included, either because the participant did not complete the full scan or due to excessive head motion on the remaining two runs.

3.2.3 Attention task

To check attention and task engagement, participants were presented with four questions randomly distributed throughout each of the eight runs (40 questions total).

Table 2: Explanation of the process of constructing sentences used in the study. Added or altered elements in the second sentence in each pair are italicised.

Type	Example sentence pair
Same	The cameraman brought the equipment to the director. The cameraman brought the *new* equipment to the director.
Modified	The cameraman brought the equipment to the director. The cameraman *begrudgingly* brought the *awkward* equipment *up from the basement* to the director.
Substituted	The cameraman brought the equipment to the director. The cameraman *sold* the equipment to the director.
Modified & substituted	The cameraman *sold* the equipment to the director. The cameraman *begrudgingly* brought the *awkward* equipment *up from the basement* to the director.
Swapped (simple case)	The cameraman brought the equipment to the director. The *director* brought the *cameraman* to the *equipment*.
Swapped (with substitution)	The cameraman brought the equipment to the director. The *director sold* the *cameraman's equipment*.
Swapped (with modifiers)	The cameraman brought the equipment to the director. The *renowned director* brought the *impatient cameraman* to the *equipment from his last project*.
Different	The cameraman brought the equipment to the director. The psychologist warned the secretary about the patient.

All questions were four-option multiple choice questions relating to the meaning of the immediately preceding sentence. Each question, along with its potential answers, was displayed on screen for 5 seconds. Participants selected the answer using one of the two-button boxes held in each hand.

3.2.4 Image acquisition

The fMRI data was acquired using a 7 Tesla Siemens MAGNETOM scanner at the Melbourne Brain Centre (Parkville, Victoria) with a 32-channel radio frequency coil. The BOLD signal was measured using a multiband echo-planar imaging sequence (TR = 800 ms, TE= 22.2 ms, FA = 45°). We acquired 636 volumes on each of the eight runs, each with 84 interleaved slices (thickness = 1.6 mm, gap = 0 mm, FOV = 208mm, matrix = 130x130, multi-band factor = 6, voxel size=$1.6\times1.6\times1.6\text{mm}^3$). Cardiac and respiratory traces were also recorded.

3.2.5 Preprocessing

Preprocessing was performed using fMRIprep with default parameters[78]. First, the T1-weighted (T1w) structural image was skull-stripped and normalized to the MNI152NLin2009-cAsym standard space. Second, each of the 8 BOLD runs was slice-time corrected and the volumes were motion-corrected by registering them to the single-band reference (SBRef) for each run. Distortion correction was applied by mapping field coefficients onto the reference image. All BOLD runs were then co-registered to the T1w reference, and resampled into the standard 1.6mm MNI152NLin2009cAsym space. Full details of this process are given in Supplementary Information.

3.2.6 GLM Model

To model the brain activity pattern resulting from each sentence, a general linear model (GLM) was fitted using a boxcar function for each separate sentence convolved with the canonical haemodynamic response (HRF). This approach yields beta coefficients for each voxel and each distinct sentence stimulus. GLMs were fitted using GLMsingle[79], a sophisticated software package able to automatically detect and remove sources of noise, and also fit an appropriate HRF for each voxel.

A constant stimulus duration of 3s was used for all stimuli for two reasons. First, GLMsingle does not support variable stimuli lengths. Second, participants will not form a full mental representation of a sentence until they finish reading it, so it is appropriate to only include the final portion of the stimulus for longer sentences.

In our preregistration we stated we would extract the representation over the final 3s for each stimulus. However, during the course of the study it became clear from participant feedback that the time provided for reading longer sentences was more than necessary, particularly for repeated trials. As such, in the main manuscript we instead report results for the middle 3s of each stimulus. For example, for a 7s sentence representations are evaluated during the window 2-5s. We show in Supplementary Information Figure S9 that our results are similar when

using the final 3s but with lower absolute magnitudes, presumably because participants begin to disengage with the task at the end of longer sentence presentations.

Three regressors of no interest were included in the GLM. The first was the number of characters displayed to the participant at any given time, as a control for the optical size of the visual stimulus. The final two regressors specified the timing of button presses for question responses, with one regressor each for left-hand and right-hand presses.

Regressions were run for each subject using the default parameters. Beta coefficients for each presentation of all 108 sentences were then extracted from the final 'TYPED_FITHRF_GLMDENOISE' output of GLMSingle, and averaged over all four presentations of each sentence.

3.3 Behavioural data collection

3.3.1 Participants

A total of 502 participants (267 male, 223 female, and 17 other; age range, 18-45; mean age ± SD, 29.80 ± 6.0) were recruited using the Prolific platform (`https://www.prolific.com/`). Participants were paid £4.50 for completing the task, which took an average of 22.5 minutes, amounting to an hourly rate of £11.96. All participants were self-declared native English speakers in Australia or the United States.

The study protocol was approved by the University of Melbourne Human Research Ethics Committee (Reference Number: 2023-23559-36378-6).

3.3.2 Survey task

Each participant provided similarity judgements on a 7-point Likert scale (1-7) of 102 sentence pairs randomly selected from the pool of all 5,778 sentence pairs. As our primary interest was in the block diagonal sentences, we over-sampled from these sentence pairs relative to the other sentence pairs. As such, each participant rated 42 block diagonal sentence pairs and 60 other sentence pairs.

Given the inherent vagueness of the similarity judgement task, previous studies have noted that lengthy instructions on how to make similarity judgements are often unclear, or may bias participant responses[80,81]. Because our goal was to elicit intuitive judgements without imposing any particular framework which might influence results, we did not provide participants with any special training or instructions about how to assign ratings. Participants were simply instructed to "consider both the similarity in meaning of the individual words contained in the sentences, as well as the similarity of the overall idea or meaning expressed by the sentences". The full instructions given to participants can be found in the Supplementary Information.

In addition to the sentence pairs derived from the 108 experimental sentences, participants were also presented with additional 10 sentence pairs that served as an attention check. These stimuli consisted of either pairs of identical sentences (high similarity) or one simple sentence paired with a grammatically correct but nonsensical sentence (low similarity).

3.3.3 Preprocessing

We excluded all participants who failed more than two of the ten attention check items, resulting in 486 of 502 participants being retained. This amounted to 49,572 judgements, providing an average of 22 ratings for each block diagonal sentence pair and 6 for each of the other sentence pairs. Similarity judgements were averaged over participants and normalised between 0 and 1.

3.4 Representational Similarity Analysis

3.4.1 Voxel Selection

Voxel selection was performed in two different ways. To provide an overall brain representation, we extracted all voxels within the cortical mask from the MNI152NLin2009cAsym template. To eliminate potential confound from visual regions, we also constructed a cortical mask excluding voxels in visual cortical regions V1-V4 from the cortical mask. In our preregistration we stated that we would remove any voxels having an absolute correlation with sentence length greater than 0.5. However during our analysis we found this to be infeasible given the large number of voxels sensitive to sentence length. We subsequently became aware that several previous studies have found similar length effects in the language network[52–54]. As such, we instead directly remove the visual cortex regions V1-V4 from analysis. As an additional check, we also performed all analyses controlling for the minimum sentence length, with the results shown in Supplementary Information Figure S4. In addition, we also analysed voxels within a language region of interest (ROI) mask. The mask contains 26,000 voxels found to be primarily sensitive to linguistic stimuli in a series of previous experiments involving contrasting sentence stimuli with pseudowords[15].

To identify voxels sensitive to sentence stimuli, the stability score was computed for each voxel as the average correlation between its time series of activity on different presentations of the stimuli[2]. All voxels within the mask with stability scores above a threshold of 0.07 were selected for computing RSA matrices. We show in Supplementary Information Figure S10 that alternative stability thresholds yield similar results, though with higher magnitudes when higher thresholds are used.

Masks for cortical regions of interest were constructed using the Glasser atlas[82]. Parcel indices included in each region were as follows. Dorsolateral prefrontal cortex: 67,68,71,73,83,84,85,86,87; dorsomedial prefrontal cortex: 26,43,63,69; precuneus: 15,27,29,30,31,45,121,142; posterior cingulate: 14,32,33,34,35,38,161,162; primary

visual cortex: 1,4,5,6; primary somatosensory cortex: 9,51,52,53.

3.4.2 Computing RSA matrices

For fMRI data, RSA matrices were computed by first normalising GLMSingle beta coefficients by subtracting the mean and dividing by the standard deviation for each voxel. Cosine similarities were then computed between the voxel representations of each sentence (using only the subset of included voxels) for each distinct pair of sentences, yielding an RSA matrix for each participant.

RSA matrices for computational models were computed differently depending on the model in question. For all vector-based models (including Mean and Transformers) sentence embeddings were extracted for each sentence, and then normalised by subtracting the mean and dividing by the standard deviation for each dimension. Pairwise sentence similarities were then computed using cosine similarity between the corresponding embeddings. For models not entirely based on vector representations (Smatch-AMR, and VerbNet-ConceptNet), we compute pairwise similarities as specified in subsection 3.1.

3.4.3 Data-model RSA correlations

RSA matrices for brain representations were compared with those of the computational models by calculating for each participant the partial Spearman correlation controlling for the difference in sentence lengths, then averaging over participants. We use the pingouin 0.5.4 python package, which utilises the inverse covariance matrix for computing partial Spearman correlations. This has been proven more reliable than the alternative regression residuals technique when a subset of variables are discrete (see discussion at `https://github.com/raphaelvallat/pingouin/issues/147`). This is especially relevant for the Graph model using the SMATCH metric, which outputs discrete similarity scores.

In addition to the simple average across participants, we implemented an alternative method adapted from several previous studies[26,27,37], in which a group-averaged RSA matrix was first constructed by averaging pairwise sentence similarities over participants, and then the correlation computed between each model RSA and this group-averaged RSA matrix.

For the simple average method, confidence intervals and statistical testing was performed using simple two-sided t-tests computed over participants. For the group average method, confidence intervals were estimated by bootstrapping over participants, performed 100 times. In the preregistration we planned to perform bootstrapping over stimuli as well as over participants, however in retrospect we judged this to be inappropriate since our sentences were not a random sampling from some corpus, but were specially constructed to provide specified semantic and syntactic variation. For both methods, the Bonferroni correction was used to adjust for three independent model comparisons (Mean to Transformer, Transformer to Hybrid, and Hybrid to Syntax), yielding a significance level of α=0.05/3 = 0.0167.

We also computed the correlation between human-rated similarities and the brain RSA similarities, though we did not perform a statistical test as we had no prior hypothesis about this correlation.

3.5 Searchlight RSA

To visualise the location of the cortical regions responsible for encoding sentence information, we implemented RSA-searchlight[83]. Using the mne-rsa package (see `https://users.aalto.fi/~vanvlm1/mne-rsa/index.html`), we performed an 8mm searchlight analysis over all voxels within the cortical mask. Images were smoothed with 5mm FWHM and thresholded at z=2 using threshold-free cluster enhancement (tfce) correction for display.