ECAI 2023
K. Gal et al. (Eds.)

doi:10.3233/FAIA230267

Investigating Neural Fit Approaches for Sentence Embedding Model Paradigms

Helena Balabin[a, b;*], Antonietta Gabriella Liuzzi[a], Jingyuan Sun[b], Patrick Dupont[a], Rik Vanderberghe[a] and Marie-Francine Moens[b]

[a]Laboratory for Cognitive Neurology, Department of Neurosciences, KU Leuven
[b]Language Intelligence and Information Retrieval Lab, Department of Computer Science, KU Leuven

Abstract. In recent years, representations from brain activity patterns and pre-trained language models have been linked to each other based on neural fits to validate hypotheses about language processing. Nonetheless, open questions remain about what intrinsic properties of language processing these neural fits reflect and whether they differ across neural fit approaches, brain networks, and models. In this study, we use parallel sentence and functional magnetic resonance imaging data to perform a comprehensive analysis of four paradigms (masked language modeling, pragmatic coherence, semantic comparison, and contrastive learning) representing linguistic hypotheses about sentence processing. We include three sentence embedding models for each paradigm, resulting in a total of 12 models, and examine differences in their neural fit to four different brain networks using regression-based *neural encoding* and *Representational Similarity Analysis* (RSA). Among the different models tested, GPT-2, SkipThoughts, and S-RoBERTa yielded the strongest correlations with language network patterns, whereas contrastive learning-based models resulted in overall low neural fits. Our findings demonstrate that neural fits vary across brain networks and models representing the same linguistic hypothesis (e.g., GPT-2 and GPT-3). More importantly, we show the need for both neural encoding and RSA as complementary methods to provide full understanding of neural fits. All code used in the analysis is publicly available: `https://github.com/lcn-kul/sentencefmricomparison`.

1 Introduction

Linking representations from pre-trained language models (PLMs) and the human brain has proven to be a promising approach to gain insights about both language understanding in the human brain as well as the biological plausibility of PLMs. More specifically, PLMs can serve as hypotheses to explain language understanding and how it is realized in the human brain. Linking the two requires parallel datasets of text and brain activity patterns, which are often based on functional magnetic resonance imaging (fMRI) experiments. A recent development regarding the design of such experiments has been to move away from highly controlled settings, in which single words are shown towards using more naturalistic stimuli such as sentences, paragraphs [34], or even continuous stories [6], to study language processing in a more realistic setting.

To exploit the rich information in such datasets, PLMs such as BERT [13] or GPT-2 [36] are leveraged to derive contextualized representations (i.e., embeddings) from the stimuli [42, 9]. These embeddings are then used to define a neural fit between PLMs and features obtained from brain activity patterns to link the representations, typically using either (i) *Representational Similarity Analysis (RSA)* [24] or (ii) *neural encoding*. A series of recent studies on neural encoding has indicated that PLMs trained on next-word prediction objectives result in a high neural fit between word embeddings and brain activity patterns (e.g., [38, 9]), which is seen as evidence for the hypothesis of predictive processing playing a key role in language understanding (for a more detailed discussion see [3]).

However, it remains challenging to pinpoint the specific properties that determine a models' neural fit, and whether it can be distinctly attributed to such linguistic hypotheses, specifically at sentence level, mainly for three reasons. First, little is known about the generalizability of previous findings from word level to sentence level understanding and across different models representing the same linguistic hypothesis, neural fit methods (i.e., neural encoding and RSA), and brain networks. Second, the relation between performances on computational tasks, namely benchmarks used in natural language processing (NLP), and neural fits has not been investigated at sentence level yet. Third, the impact of inter-sentence context on the resulting representations that determine a neural fit remains unclear.

In this study, we address these challenges through a comprehensive analysis of different linguistic hypotheses about sentence processing, allowing for a more accurate investigation of brain regions involved in human language processing and explanation of the inner workings of PLMs. Based on four sentence embedding paradigms, we categorize a total of 12 sentence embedding models. By linking the respective sentence embeddings to representations derived from fMRI using both neural encoding and RSA as neural fits, we address the following three research questions:

- **R1** What properties can neural encoding and RSA reveal about the neural fit of linguistic hypotheses to different brain networks?
- **R2** Does the performance of sentence embedding models on current sentence embedding benchmarks correlate with its neural fit?
- **R3** Is the neural fit sensitive to sentence context, i.e., the surrounding sentences of a given sentence?

2 Related Work

2.1 Representational Similarity Analysis (RSA)

RSA [24] is a popular approach for linking models for various modalities in neuroscience, such as artificial neural network (ANN)

* Corresponding Author. Email: helena.balabin@kuleuven.be

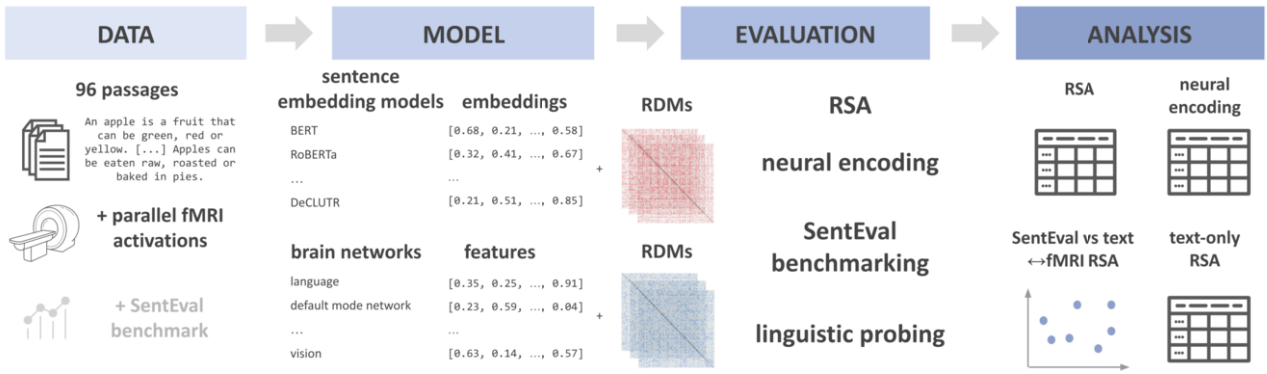


Figure 1: Methodology overview. This figure illustrates the pipeline used to perform several analysis on the parallel sentence and functional MRI (fMRI) data. First, sentence embeddings and fMRI features (i.e., response amplitudes) are derived using 12 sentence embedding models and four brain networks. Additionally, for both modalities, a representational dissimilarity matrix (RDM, depicted by the red and blue matrices) is calculated for each sentence embedding model/brain network using pairwise cosine distances across all inputs (i.e., embeddings or fMRI features). Afterwards, text-based evaluations are performed in the form of linguistic probing and SentEval benchmarking, and neural fits are determined through Representational Similarity Analysis (RSA) as well as neural encoding.

representations and features derived from the human brain. It consists of first calculating a representational dissimilarity matrix (RDM) for each model based on pairwise distances of its representations for a set of stimuli (e.g., embeddings for a list of sentences shown in an fMRI experiment). Then, in order to determine how closely related two given models are, a correlation between the two RDMs is calculated, resulting in a second-order isomorphism, which compares pairwise similarities rather than the individual vectors in the two given vector spaces. Recently, RSA has been applied to the domain of NLP [1, 26] for probing and explaining various hypotheses about word or sentence embeddings. However, it is insufficiently explored whether the identified links between PLMs and brain activations discovered by RSA differ from those revealed by neural encoding.

2.2 Neural Encoding

Neural encoding is a long-established approach for linking representations from ANNs with representations derived from brain activations. These ANN representations can cover various modalities such as text [22, 21, 42, 38, 9, 43], audio [30] or video [44]. Overall, the goal of neural encoding is to learn a mapping model $y = f(x)$ to predict brain activations y from ANN representations x, and evaluate the model on a held-out test set. Similarly, neural decoding is the inversion of this mapping, i.e., learning a mapping model from brain activations to predict ANN representations [17, 40, 41, 33]. In most studies, the basis for neural encoding consists of either single words (e.g., [21, 43]) or aggregated word embeddings of sentences (e.g., [38, 9]). In particular, the latter was reported to be problematic for creating informative sentence representations [37]. There are notable exceptions using sentence embedding models [2, 40, 41]. However, these approaches compare single sentence embedding models rather than paradigms to each other, and do not include any semantic comparison- and contrastive learning-based models (see Section 3.2), which we cover in our analyses. Further, previous linguistic probing analyses such as in [41] do not incorporate inter-sentence context. As a result, most of the evidence supporting the role of predictive processing in language originated from comparisons between word-level PLMs. Therefore, it remains unclear whether findings regarding the role of predictive processing generalize to equivalent paradigms in sentence-level PLMs.

3 Methods

To answer the research questions proposed in Section 1, we use a combination of neural fits, benchmarking, and linguistic probing methods (see Figure 1). First, we outline the fMRI and sentence data, as well as the respective brain regions and sentence embedding models serving as the starting point for the first research question (**R1**) in Sections 3.1 and 3.2. Then, we define the two methods to determine the neural fit in Section 3.3, and explain how we address research question **R2** in Section 3.4. Finally, to address question **R3**, we describe our RSA-based linguistic analysis and link it to the previously determined neural fits in Section 3.5. All preprocessing steps and analyses are implemented in Python based on sklearn, scipy, skipthoughts and Hugging Face's transformers library, and all code is publicly available[1]. Further, all experiments are conducted on a 24GB NVIDIA RTX A5000.

3.1 Data

We use the dataset from [34] consisting of parallel sentence and preprocessed fMRI data (see Figure 1). More specifically, we choose this dataset because it is (i) widely used (e.g., [41, 38]) and publicly available[2], (ii) reliable due to repeated measurements and (iii) sufficiently long in terms of sentence context for each measured brain response. To optimally utilize the available sentence context, we focus on experiment #2, which consists of 96 passages with four sentences each. Each passage is providing a Wikipedia-style definition of a concrete object (e.g., an apple). To obtain the respective brain activations, each passage was presented three times to a total of eight subjects. Here, we focus on subsets of the original set of brain voxels using four networks employed in the original approach as well as in [32], namely the **language** network, default mode network (**DMN**), **task**-positive network and **vision** network shown in Figure 2. More specifically, the frontotemporal language-selective network used in this approach is taken from [14] and includes the inferior temporal gyrus as well as the anterior temporal lobe, while the vision network is based on [8, 35] and mostly covers the occipital cortex. Moreover, the task-positive network is a cognitive control network that is activated whenever specific tasks are carried out, whereas the DMN is

[1] https://github.com/lcn-kul/sentencefmricomparison
[2] https://evlab.mit.edu/sites/default/files/documents/index2.html

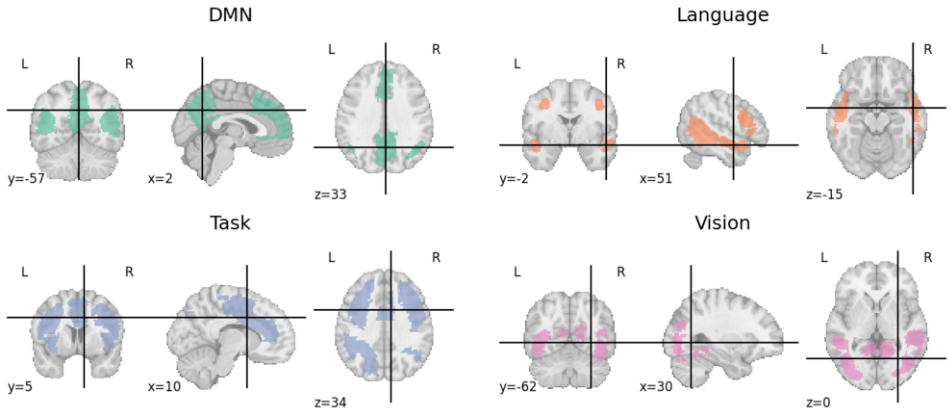


Figure 2: Binary maps of the brain networks adopted from [34, 32]. The four networks used in this analysis comprise regions of interest in the (i) default mode network (DMN), (ii) language network as defined by [14], (iii) task-positive network and (iv) vision network. The x, y and z values indicate the cut position for the sagittal, coronal and axial planes that show the largest connected component of each brain network.

a circuit that is deactivated during active versus control conditions, which has been linked to semantic processing, set shifting, and introspection. Additionally, we examine the neural fits for the language network separated into the left and right hemisphere and at the level of the whole brain in Appendix A[3].

3.2 Sentence Embedding Models

We choose a total of 12 sentence embedding models based on pre-trained language models with the aim of maximizing the variety of paradigms (i.e., training objectives) used to train the models while still including several examples per paradigm. We group the models based on their training objectives into four sentence embedding model paradigms (based on [27]).

Masked Language Modeling A straightforward approach to form sentence representations is to average word embeddings obtained from a model trained on a word-level task. To test this approach, we average the word embeddings from **BERT** [13], **RoBERTa** [28] and **DeBERTa** [19] and average their word embeddings to form sentence representations for each input. We chose averaging the word embeddings rather than using the `CLS` token as a sentence representation, since averaging has been shown to improve the quality of word embedding-based sentence representations [37]. These three models were trained using the masked language modeling (MLM) objective, which aims to predict masked tokens (i.e., sub-words) using the bidirectional context of a given input sequence.

Pragmatic Coherence To include wider sentence context, this paradigm focuses on the incorporation of coherence in terms of capturing the transition of meaning in longer contexts. Typically, such models are trained by predicting the correct subsequent input (i.e., word sequences or sentences), inspired by the concept of predictive coding [15, 4]. In this paradigm, we employ **SkipThoughts** [23], **GPT-2** [36] as well as **GPT-3** [7][4], as these approaches are based on next word or sentence prediction training objectives. Here, we include both GPT-2 and GPT-3 in order to examine the possible effect of the extended input length used during the pre-training procedure in GPT-3 (4096 tokens, compared to 1024 tokens for GPT-2) on the resulting neural fits. For these two models, we again derive aggregated representations by averaging all word embeddings for a given input.

Semantic Comparison This group of sentence embedding models is based on learning the semantic relationship between sentences, which is related to semantic processing in the human brain [25, 29]. The most common strategy for incorporating semantic sentence context is to fine-tune a sentence embedding model using Natural Language Inference (NLI), Question Answering (QA) or Semantic Textual Similarity (STS) tasks. For this paradigm, we choose the Sentence-RoBERTa (**S-RoBERTa**) `S-RoBERTa-NLI-STSb-large` [37], supervised SimCSE (**supSimCSE**) [16] and Sentence-T5 (**S-T5**) [31] models.

Contrastive Learning Given the recent success of several contrastive learning-based models on the SentEval [12] benchmark (see e.g., [16, 10]), we decided to incorporate these models as well. More specifically, we test the unsupervised SimCSE model (**unsupSimCSE**) [16], **DiffCSE** [10] and **DeCLUTR** [18]. To the best of our knowledge, there is no clear relationship between contrastive learning and language processing in the human brain. However, there is existing work on the links between contrastive learning-based models in computer vision and representations derived from the visual cortex [5]. Therefore, we test whether this finding generalizes to language processing in the human brain.

3.3 Neural Fit

For the first research question (**R1**), we determine the neural fit for each combination of sentence embedding model and brain network (resulting in $12 \times 4 = 48$ comparisons) using both RSA and regression-based neural encoding.

RSA The rationale behind Representational Similarity Analysis (RSA) is to compare two representational spaces (i.e., the embedding space of a sentence embedding model and the representational space of the fMRI features) based on pairwise distances across a set of inputs, independently from factors such as their dimensionality. More specifically, the first step of the RSA is to derive the

[3] All appendices can be accessed at https://github.com/lcn-kul/sentencefmricomparison

[4] The GPT-3 embeddings for all text inputs in this analysis are based on OpenAI's embedding API using the `text-embedding-ada-002` model

Paradigm	Model	Language	DMN	Task	Vision
Masked language modeling	**BERT**	-0.019	-0.006	-0.003	-0.018
	RoBERTa	-0.029	0.073***	0.080***	-0.035
	DeBERTa	0.032*	0.045**	0.065***	0.033*
	Mean	**-0.005**	**0.038**	**0.047**	**-0.007**
Pragmatic coherence	**SkipThoughts**	0.082***	0.043**	0.048***	0.142***
	GPT-2	0.067***	0.008	0.022	0.066***
	GPT-3	0.003	-0.037	-0.019	0.030*
	Mean	**0.051**	**0.005**	**0.017**	**0.080**
Semantic comparison	**S-RoBERTa**	0.051***	0.053***	0.047***	0.091***
	sup-SimCSE	-0.007	-0.004	0.024	0.017
	S-T5	0.048***	0.038**	0.048**	0.087***
	Mean	**0.031**	**0.029**	**0.040**	**0.065**
Contrastive learning	**unsup-SimCSE**	0.030*	0.030*	0.039**	0.048**
	DiffCSE	-0.025	-0.026	-0.007	-0.015
	DeCLUTR	-0.008	-0.004	-0.008	0.028*
	Mean	**-0.001**	**0.000**	**0.008**	**0.020**

Table 1: RSA-based neural fit results (**R1**). This table lists all Spearman rank correlations obtained through RSA for each pairing of sentence embedding model and brain network, grouped by the previously introduced sentence embedding paradigms. The **Mean** rows in each paradigm indicate the mean correlation for a given brain network for all models that are assigned to the respective paradigm. For each result (excluding means), the significance level is indicated using *** ($p \leq 0.001$), ** ($p \leq 0.01$), * ($p \leq 0.05$) or no asterisk ($p > 0.05$). For each brain network, the best performing individual model as well as the paradigm with the highest average correlation are underlined.

RDMs for the sentence embedding models and brain networks, respectively. For that, we calculate the pairwise cosine distances for both modalities (see Figure 1) based on either embedded paragraphs or fMRI features. For the fMRI RDMs, we first calculate an RDM for each subject, and derive a final RDM using the element-wise average of all subject-specific RDMs. Then, for each RDM, we select and vectorize the upper triangular matrix (excluding the diagonal) to calculate Spearman's rank correlation ρ between the vectorized RDMs for each sentence embedding model and brain network pair. We choose Spearman's rank correlation rather than the Pearson correlation coefficient to measure monotonic rather than linear associations. These correlation coefficients ρ are the resulting neural fits based on the RSA. Lastly, since the observed correlation values tend to be relatively small, we calculate the significance of each model-brain network correlation pair using permutation testing [24] with $n = 10,000$ repetitions. More specifically, for a given model-brain network pair, permutation testing determines the significance of an observed correlation by comparing it to a distribution of correlations between permuted sentence embedding model RDMs and the brain network RDM.

Neural Encoding To determine the neural fit based on neural encoding, we apply a standard regression-based neural encoding procedure (e.g., see [38, 41, 32, 39]). We choose neural encoding rather than decoding as we aim to investigate differences in the ability of sentence embedding models to predict brain activation features rather than the predictive ability of the brain activation features. For each possible mapping from a sentence embedding model onto a brain network, we construct a neural encoder by adding a linear mapping model (i.e., Ridge regression) on top of the output of the frozen sentence embedding model (i.e., in a feature extraction rather than a fine-tuning setting) to predict the fMRI features. More precisely, based on a training set of size M consisting of sentence embeddings $X \in \mathbb{R}^{M \times d_s}$ of dimensionality d_s and fMRI features $Y \in \mathbb{R}^{M \times d_f}$ of dimensionality d_f, the linear mapping model determines regression coefficients W to minimize the following objective function:

$$||Y - XW||_2^2 + ||W||_2^2$$

Based on an overall five-fold cross-validation procedure, we then train and test the encoder on the 96 paragraphs. Given a test set of size N, we evaluate each model's performance using pairwise accuracy, a commonly used metric to evaluate neural encoding [34, 41] as it focuses on binary comparisons rather than absolute distance values across paired representations. Based on each possible paragraph pair (i, j) (i.e., $\frac{N(N-1)}{2}$ pairs in total), its predicted and true fMRI vectors $\hat{y_i}, \hat{y_j}$ and y_i, y_j, and the cosine distance function D, we calculate the pairwise accuracy as follows:

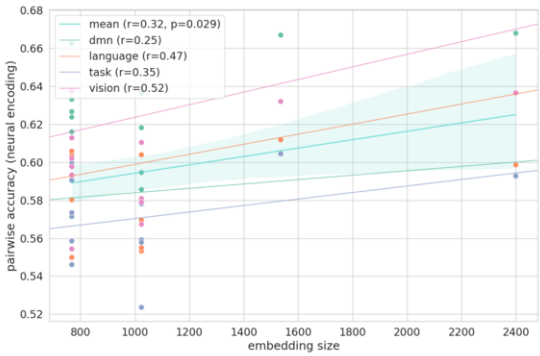
$$\text{pairwise accuracy} = \frac{2}{N(N-1)} \sum_{i=1}^{N-1} \sum_{j=i+1}^{N} \text{acc}(i,j)$$

$$\text{acc}(i,j) = \begin{cases} 1 & D(\hat{y}_i, y_i) + D(\hat{y}_j, y_j) < D(\hat{y}_i, y_j) + D(\hat{y}_j, y_i) \\ 0 & \text{otherwise} \end{cases}$$

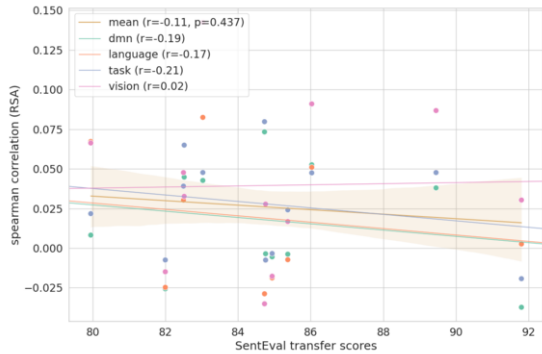
Intuitively speaking, the pairwise accuracy metric is measuring whether a predicted fMRI feature vector for a given sentence is closer to the underlying true fMRI feature vector from that same sentence rather than the one from another sentence. We use the average of the model's pairwise accuracy across all folds as its final neural fit based on neural encoding. Further, we perform an exemplary ablation study to investigate the effect of the hyperparameters of the neural encoder on the pairwise accuracy based on GPT-2 and the language network (see Appendix B). Moreover, to assess the role of the pre-training procedure on the neural fit, we compare GPT-2 to a randomly initialized baseline in Appendix C. Finally, we investigate the role of the topic of the passages on the resulting pairwise accuracies in Appendix D.

3.4 Sentence Embedding Model Properties

In addition to the RSA correlations from the previous section, we obtain sentence embedding benchmark scores using the SentEval [12] benchmark (or use previously reported SentEval results) for all sentence embedding models or to answer research question **R2**. The SentEval benchmark consists of 14 sentence classification tasks such as semantic textual similarity (STS), natural language inference (NLI) and sentiment analysis. We only use the transfer tasks (adopting the evaluation setup from [16]), mainly consisting of sentiment analysis tasks, and leave out the STS tasks, since some of the sentence embedding models have already been fine-tuned on such data (e.g., S-RoBERTa), which would result in an unfair comparison.



(a) Embedding size vs neural encoding-based neural fits



(b) SentEval benchmark scores vs RSA-based neural fits

Figure 3: Analysis of the sentence embedding model properties (**R2**). Figure 3a shows the correlation between sentence embedding sizes on the x-axis and neural encoding-based pairwise accuracy scores (i.e., multiple scores for multiple brain networks against a single embedding size per model) on the y-axis. Figure 3b is illustrating pairings of all RSA scores (y-axis) between fMRI and sentence embedding model representations against the average SentEval score (x-axis) for all sentence embedding models. The scores for the four brain networks as well as the mean are indicated in different colors, resulting in five data points for each sentence embedding model. The data points are arranged in one line due to having one associated embedding size or SentEval score per model.

Next, we average the classification accuracies on all seven transfer tasks to obtain an average score for each model. Then, we calculate the Pearson correlation coefficient between each pair of average SentEval and RSA-based neural fit score (using each brain network as an individual data point rather than an average across all brain networks) as in the previous section, and repeat the analyses for each brain network. Moreover, we investigate whether there is a link between the size (i.e., dimensionality) of the embeddings from the 12 sentence embedding models and their respective neural encoding performances. For that, we extract the embedding size for each model and correlate it to the previously obtained pairwise accuracy scores, again using each of the four brain networks as a separate data point. Lastly, to examine how representations from different sentence embedding models are correlated to each other, we create a correlogram based on the Pearson correlations between the RDMs of the sentence embedding models (see Appendix E).

3.5 Linguistic Probing

To better understand what linguistic properties a neural fit is capturing, we aim to gain deeper insights into how the previously determined neural fits are affected by sentence context (**R3**). For that, we use text-based RSA and an evaluation of neural fits based on different sentence contexts to examine the changes in the respective pairwise accuracies and correlations obtained with RSA caused by variations in textual inputs.

Text-Based RSA To compare the impact of sentence context on the resulting RDMs between the different sentence embedding models, we perform linguistic probing in the form of text-based RSA [26]. The general goal is to test different hypotheses about sentence processing based on correlating respective RDMs. Analogous to the method used in [26], we define a reference model, which functions as a starting point for comparing two different hypothesis models:

- **Reference model**: For each passage in the dataset (see Section 3.1), we extract the middle two sentences
- **Hypothesis model**: We extract the full passages from the dataset, i.e., the middle two sentences with the original preceding and subsequent sentences
- **Alternative hypothesis model**: We concatenate the middle two sentences with the first and last sentences that are randomly chosen from other passages

We argue that if a sentence embedding model is correctly encoding sentence context, there should be a high correlation ρ_{hyp} between the reference and hypothesis models. On the contrary, adding two random sentences to the original should significantly affect the representations in the alternative hypothesis and therefore lead to a decrease in the correlation ρ_{alt} between the reference and alternative hypothesis model. Again, we calculate RDMs for all sentence embedding models and all three linguistic models using cosine distance, and determine the Spearman's rank correlation. Finally, we calculate the difference $\Delta_\rho = \rho_{\text{hyp}} - \rho_{\text{alt}}$ between the two hypotheses for each sentence embedding model, which indicates whether it is encoding original and random contexts distinctly (high Δ_ρ) or not (low Δ_ρ).

Sentence Context Against Neural Fit Next, to better understand how sentence context is affecting the neural fits, we re-calculate the pairwise accuracies for the neural encoding approach by varying the sentence input based on the three models, resulting in the following three text inputs: (i) full paragraphs, (ii) middle two sentences and (iii) random first and last sentences appended to the middle two sentences. Importantly, we use the same fMRI features Y as in the previous sections in order to attribute the changes in the neural fit to the variations in text input X (namely $X_{\text{full}}, X_{\text{middle}}, X_{\text{random}}$) rather than variations in brain activation patterns. To assess how removing the first and last sentences as well as replacing first and last sentences with random sentences affect the resulting neural fits, we create two conditions, respectively: only-middle and random-first-last. More specifically, we report the differences in performance in the only-middle and the random-first-last conditions based on the difference in the average pairwise accuracy of each sentence embedding model across all four brain networks for the respective two inputs.

4 Results

4.1 Neural Fit

The neural fits of all 12 sentence embedding models for RSA and neural encoding are presented in Table 1 and Table 2, respectively.

Paradigm	Model	Language	DMN	Task	Vision
Masked language modeling	**BERT**	0.598	0.593	0.571	0.624
	RoBERTa	0.610	0.604	0.578	0.636
	DeBERTa	0.613	0.604	0.590	0.633
	Mean	**0.607**	**0.600**	**0.580**	**0.631**
Pragmatic coherence	**SkipThoughts**	0.636	0.599	0.593	0.668
	GPT-2	0.602	0.591	0.558	0.627
	GPT-3	0.632	0.612	0.604	0.667
	Mean	**0.623**	**0.600**	**0.585**	**0.654**
Semantic comparison	**S-RoBERTa**	0.581	0.569	0.559	0.618
	sup-SimCSE	0.567	0.553	0.524	0.586
	S-T5	0.637	0.606	0.599	0.662
	Mean	**0.595**	**0.576**	**0.561**	**0.622**
Contrastive learning	**unsup-SimCSE**	0.579	0.555	0.558	0.595
	DiffCSE	0.554	0.550	0.546	0.573
	DeCLUTR	0.593	0.580	0.573	0.616
	Mean	**0.575**	**0.562**	**0.559**	**0.595**

Table 2: Neural encoding-based neural fit results (**R1**). This table lists all pairwise accuracy scores for each possible pair of brain network and sentence embedding model, again grouped by the sentence embedding paradigm. The **Mean** rows in each paradigm indicate the mean accuracy for a given brain network for all models for that paradigm. Best results are indicated in the same manner as in Table 1.

Overall, with regard to research question **R1**, the average of the pragmatic coherence models leads to the highest RSA correlations across the language and vision networks, whereas the masked language modeling-based models lead to the highest correlations for the DMN and task network. For neural encoding, the pragmatic coherence paradigm results in the highest pairwise accuracy scores across all four networks. In both approaches, on average, the contrastive learning-based models result in the lowest performances across all networks. With respect to the individual models, SkipThoughts performs best for the language and vision networks for both neural fit approaches, whereas RoBERTa and GPT-3 perform best for the DMN and task network for RSA and neural encoding, respectively. Moreover, the RSA correlations were significant across all four brain networks for S-RoBERTa ($p \leq 0.001$), S-T5 ($p \leq 0.01$), SkipThoughts ($p \leq 0.01$), DeBERTa ($p \leq 0.05$) and unsup-SimCSE ($p \leq 0.05$). Finally, there are only minimal variations across neural encoding results obtained with different hyperparameters (see Appendix B).

We observe that some models show similar correlations across all brain networks, whereas others are more specific to a subset of networks. For instance, while models such as BERT and DiffCSE result in rather comparable (albeit low or even negative) correlations across all brain networks, there are large variations for models like GPT-2 and SkipThoughts, which are more specific to the vision and language networks than the DMN and task network. Conversely, the correlations observed for RoBERTa are highly specific to the DMN and task network. Furthermore, in both Table 1 and Table 2, the vision network is resulting in better performances in most cases.

Moreover, there are notable differences across models belonging to the same paradigm, specifically for the observed RSA correlations. For instance, while the GPT-2-based sentence representations lead to significant correlations for the language and vision brain networks, GPT-3 does not significantly correlate to any brain network apart from the vision network (see Table 1). However, for neural encoding, GPT-3 performs better than GPT-2 across all networks, presenting a complementary finding to the RSA results.

4.2 Sentence Embedding Model Properties

To examine whether the neural fits of the sentence embedding models can be explained by their inherent properties, we derived correlations between the sentence embedding models (Figure 3a) as well as SentEval benchmark and RSA scores (Figure 3b) (**R2**). We observe a negative, non-significant correlation ($r = -0.11, p \geq 0.05$) between the RSA and SentEval transfer scores in Figure 3b, suggesting that the benchmark scores of a sentence embedding model are not necessarily indicative of its neural fit. However, the correlations differ across the networks, including negative correlations for the DMN, language and task networks and a slightly positive correlation for the vision network. Furthermore, as shown in Figure 3a, there is a significantly positive correlation between the embedding size of a sentence embedding model and its pairwise accuracies, and the vision and language networks yield higher correlations than the DMN and task network. In addition, the correlogram across all sentence embedding models (see Appendix E) shows that while there are moderate to high correlations ($r \geq 0.5$ for all model pairs) across models within the contrastive learning and semantic comparison paradigms, the correlations within the masked language modeling and pragmatic coherence paradigms tend to be lower ($r \leq 0.5$ for some model pairs).

4.3 Linguistic Probing

To gain a deeper understanding about how the neural fit depends on the sentence context (**R3**), we apply linguistic probing in the form of text-based RSA. We test the ability of the models to produce distinct representations for matching versus random context sentences that are added to a shared reference, resulting in a hypothesis and an alternative hypothesis model (see Section 3.5). Table 3 summarizes the observed correlations between reference and hypothesis models (ρ_{hyp}), reference and alternative hypothesis models (ρ_{alt}) and their respective differences (Δ_ρ). Moreover, Table 4 shows how the neural encoding results are affected by removing the first and last sentences from the paragraphs or by replacing them with random sentences.

On average, the contrastive learning models result in the largest differences between the two hypotheses, followed by the semantic comparison paradigm, indicating that the two types of surrounding sentence context most distinctly alter the representations for these models. In addition, there are substantial differences across the paradigm-specific models. For instance, GPT-3 led to a much larger Δ_ρ than GPT-2 and SkipThoughts, and the reported difference for BERT was more than four times as large as the difference observed for RoBERTa. Furthermore, Table 4 indicates that while removing the first and last sentences only slightly affects the performances (except for RoBERTa), the insertion of random sentences results in a more drastic decrease in pairwise accuracy scores for all models, in particular GPT-3.

Paradigm	Model	ρ_{hyp}	ρ_{alt}	Δ_ρ
Masked language modeling	**BERT**	0.508***	-0.012	**0.520**
	RoBERTa	0.142***	0.018	**0.124**
	DeBERTa	0.458***	0.048***	**0.410**
	Mean	**0.369**	**0.018**	**0.351**
Pragmatic coherence	**SkipThoughts**	0.511***	0.297***	**0.214**
	GPT-2	0.263***	0.001	**0.262**
	GPT-3	0.808***	0.313***	**0.495**
	Mean	**0.527**	**0.204**	**0.324**
Semantic comparison	**S-RoBERTa**	0.81***	0.448***	**0.362**
	sup-SimCSE	0.621***	0.098***	**0.388**
	S-T5	0.722***	0.319***	**0.403**
	Mean	**0.718**	**0.288**	**0.384**
Contrastive learning	**unsup-SimCSE**	0.409***	0.022	**0.387**
	DiffCSE	0.353***	0.047***	**0.306**
	DeCLUTR	0.530***	0.044**	**0.486**
	Mean	**0.431**	**0.038**	**0.393**

Table 3: Text-based RSA (**R3**). This table contains the Spearman's rank correlations between the reference and hypothesis model (ρ_{hyp}) as well as between the reference and alternative hypothesis model (ρ_{alt}) for each sentence embedding model, grouped by the respective paradigms. The **Mean** rows indicate the mean of the two hypotheses across all models belonging to a given paradigm. The column Δ_ρ indicates the difference $\rho_{\text{hyp}} - \rho_{\text{alt}}$. Significance levels and best results are indicated in the same manner as in Table 1.

Paradigm	Model	only-middle	random-first-last
Masked language modeling	**BERT**	-0.013	-0.065
	RoBERTa	-0.039	-0.085
	DeBERTa	-0.024	-0.078
	Mean	**-0.025**	**-0.076**
Pragmatic coherence	**SkipThoughts**	0.011	-0.069
	GPT-2	-0.001	-0.074
	GPT-3	0.002	-0.090
	Mean	**0.004**	**-0.078**
Semantic comparison	**S-RoBERTa**	-0.013	-0.028
	sup-SimCSE	0.004	-0.019
	S-T5	-0.019	-0.089
	Mean	**-0.009**	**-0.045**
Contrastive learning	**unsup-SimCSE**	0.005	-0.036
	DiffCSE	0.020	-0.036
	DeCLUTR	-0.013	-0.044
	Mean	**0.004**	**-0.039**

Table 4: Sentence context against neural fit (**R3**). The **only-middle** column reports the difference in performance when replacing the full paragraphs with the middle two sentences, whereas the **random-first-last** column reports the difference when replacing the original first and last sentences with random sentences. Both conditions are based on the difference in the average pairwise accuracy of each sentence embedding model across all brain networks for the respective text inputs. The **Mean** rows indicate the mean of the differences across all models belonging to a given paradigm. The model and paradigm with the largest difference are underlined.

5 Discussion

The objective of this study was to gain a deeper understanding of linguistic hypotheses (i.e., sentence embedding model paradigms) in terms of their neural fit to several brain networks (**R1**). We found that models from the pragmatic coherence, semantic comparison and masked language modeling paradigms result in larger correlations for RSA as well as in higher pairwise accuracy scores for neural encoding compared to contrastive learning-based models. SkipThoughts, GPT-2 (for RSA) and GPT-3 (for neural encoding) proved to be the models with the best individual neural fits specifically for the language network, potentially providing further evidence for predictive processing playing a key role in language processing [38].

Furthermore, RSA and neural encoding seem to reveal complementary properties about the tested sentence embedding models, as indicated by variations in the respective neural fits. This observation becomes particularly apparent for GPT-3: Although it demonstrates superior performance in neural encoding, it does not yield significant correlations with any of the brain networks as measured by RSA. While neural encoding results are reflecting the predictive performance of the sentence embeddings as such, RSA is focusing on comparing distances between the sentences for fMRI and sentence embedding model representations. For neural encoding, we show that the embedding size and pairwise accuracies of the models are correlated to each other (see Figure 3a). A possible explanation for this result might be that embeddings of higher dimensionality are less compressed, and thus more informative. Taken together, these findings indicate that the high neural fit of GPT-3 based on neural encoding may be driven by its embedding size rather than the inherent properties of its representational space, as reflected by its comparatively low neural fit based on RSA. These results show the importance of using both neural fit approaches. Neural encoding and RSA complement each other, as they are driven by the predictive ability and the (dis)similarities across the sentence embeddings, respectively.

The variations across models belonging to the same paradigm highlight that caution is advised when drawing conclusions about language processing based on a high neural fit of a single model representing a particular linguistic hypothesis. For instance, closely related models (e.g., BERT/RoBERTa and GPT-2/GPT-3) lead to different neural fits, specifically for RSA. This result may be explained by the large difference in the linguistic probing results across the models (**R3**, see Table 3). More specifically, these differences indicate that the specific approach and pre-training setup (such as the aforementioned embedding size) for modeling the same paradigm substantially influence the sentence representations and thereby the neural fits. Moreover, the distinct encoding of correct versus random sentence context (i.e., large Δ_ρ values, see Table 3) for contrastive learning-based models can be attributed to their training objective, which aims to maximize the distance between embeddings of unrelated sentences. However, this characteristic does not seem to lead to higher neural fits (see the smaller differences shown for contrastive learning-based models in the rightmost column in Table 4 compared to pragmatic coherence models), implying that predictive processing may be a more plausible hypothesis for language processing.

Next, with regard to differences in the brain networks, we observe a general tendency of certain sentence embedding models being significantly correlated to all four brain networks versus others that are more specifically correlated to only some of the networks. Interestingly, models trained on inter-sentence context such as S-RoBERTa, S-T5 and SkipThoughts result in strongly significant ($p \leq 0.01$) correlations across all four brain networks, which can be seen as evidence for globally distributed semantic processing across the entire brain (cf. [2]). Conversely, word-level models such as GPT-2 or RoBERTa are specifically linked to a subset of the examined brain networks, which could be interpreted as models of language processing that resemble localized processing in these networks in the human brain. Moreover, the consistently high correlation of many sentence embedding models to the vision network can likely be explained by the experimental design of the dataset used in this study, which is focused on presenting definitions of mostly concrete concepts in written form. It can be argued that a biologically plausible sentence embedding model should only result in correlations specific to the language network and possibly the vision network in the case of visual imagination of concrete objects [11, 20].

Concerning question **R2**, our results are in line with previous literature. We found that neural fits do not correlate with SentEval benchmark scores, indicating a generalization of the lacking correlation of neural fits to GLUE benchmark scores reported in [38]. This suggests that benchmark performances used in NLP might not accurately reflect how well PLMs resemble language processing in the human brain. The recent success of several contrastive learning-based sentence embedding models, which is not reflected in their neural fit, provides evidence for this hypothesis. For SimCSE and DiffCSE, we hypothesize that using dropout-based positive examples and in-batch negative examples might explain the low neural fit of these models, especially because dropout is modifying (part of) the sentence representations at random, which is not necessarily cognitively plausible.

Overall, there remain many open questions regarding the qualities of language models that present the best match to language processing in the brain. Future work should perform controlled experiments for model comparisons based on neural fits to account for factors such as training data or model size. Finally, additional studies should examine how well findings regarding certain hypotheses tested with neural fits generalize to multimodal or multilingual settings.

6 Conclusion

In this work, we examined how linguistic hypotheses in the form of sentence embedding paradigms and their properties are related to their neural fit. We were able to confirm the important role of predictive processing at the level of sentences and the lacking coherence to sentence benchmark performances, indicating that sentence embedding models with high benchmark scores do not necessarily resemble language processing in the human brain. Moreover, we have shown the need for both neural encoding and RSA as complementary methods to provide a complete understanding of the neural fit of sentence embedding models. Finally, the high correlations between the sentence representations and the vision network point at their interconnectedness, inspiring further work on hypothesized brain regions in human language processing, the inner workings and future improvements of pre-trained language models, and brain-computer interfaces in general.

Ethics Statement

We used a publicly available dataset [34] in this work, which does not contain any data that can be directly linked to the participants' identities, since the authors shared preprocessed MATLAB files rather than raw NIfTI, PAR/REC or DICOM files.

Acknowledgments

This research was supported by funding from the Research Foundation - Flanders (Fonds Wetenschappelijk Onderzoek, FWO) grants 1154623N, 1247821N and from the European Research Council (ERC) under Grant Agreement No. 788506. The authors would like to thank Victor Milewski for his helpful comments.

References

[1] Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema, 'Blackbox Meets Blackbox: Representational Similarity & Stability Analysis of Neural Language Models and Brains', in *ACL-BlackboxNLP*, pp. 191–203, Florence, Italy, (August 2019).

[2] Andrew James Anderson, Douwe Kiela, Jeffrey R. Binder, Leonardo Fernandino, Colin J. Humphries, Lisa L. Conant, Rajeev D. S. Raizada, Scott Grimm, and Edmund C. Lalor, 'Deep Artificial Neural Networks Reveal a Distributed Cortical Network Encoding Propositional Sentence-Level Meaning', *Journal of Neuroscience*, **41**(18), 4100–4119, (May 2021).

[3] Richard Antonello and Alexander Huth, 'Predictive Coding or Just Feature Discovery? An Alternative Account of Why Language Models Fit Brain Data', *Neurobiology of Language*, 1–16, (November 2022).

[4] Vladimir Araujo, Andrés Villa, Marcelo Mendoza, Marie-Francine Moens, and Alvaro Soto, 'Augmenting BERT-style Models with Predictive Coding to Improve Discourse-level Representations', in *EMNLP*, pp. 3016–3022, Punta Cana, Dominican Republic, (November 2021).

[5] Shahab Bakhtiari, Patrick Mineault, Timothy Lillicrap, Christopher Pack, and Blake Richards, 'The functional specialization of visual cortex emerges from training parallel pathways with self-supervised predictive learning', in *NeurIPS*, (2021).

[6] Idan Blank, Nancy Kanwisher, and Evelina Fedorenko, 'A functional dissociation between language and multiple-demand systems revealed in patterns of BOLD signal fluctuations', *Journal of Neurophysiology*, **112**(5), 1105–1118, (September 2014).

[7] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei, 'Language Models are Few-Shot Learners', in *NeurIPS*, (2020).

[8] Randy L. Buckner, Jessica R. Andrews-Hanna, and Daniel L. Schacter, 'The brain's default network: anatomy, function, and relevance to disease', *Annals of the New York Academy of Sciences*, **1124**, 1–38, (March 2008).

[9] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King, 'Deep language algorithms predict semantic comprehension from brain activity', *Scientific Reports*, **12**(1), 16327, (September 2022).

[10] Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljačić, Shang-Wen Li, Wen-tau Yih, Yoon Kim, and James Glass, 'DiffCSE: Difference-based Contrastive Learning for Sentence Embeddings', *arXiv:2204.10298 [cs]*, (April 2022).

[11] Guillem Collell, Ted Zhang, and Marie-Francine Moens, 'Imagined Visual Representations as Multimodal Embeddings', *AAAI*, **31**(1), (February 2017).

[12] Alexis Conneau and Douwe Kiela, 'SentEval: An Evaluation Toolkit for Universal Sentence Representations', in *LREC*, Miyazaki, Japan, (May 2018).

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova, 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding', in *NAACL-HLT*, (June 2019).

[14] 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, (September 2011).

[15] Karl Friston, 'A theory of cortical responses', *Philosophical Transactions of the Royal Society B: Biological Sciences*, **360**(1456), 815–836, (April 2005).

[16] Tianyu Gao, Xingcheng Yao, and Danqi Chen, 'SimCSE: Simple Contrastive Learning of Sentence Embeddings', in *EMNLP*, pp. 6894–6910, Online and Punta Cana, Dominican Republic, (November 2021).

[17] Jon Gauthier and Roger Levy, 'Linking artificial and human neural representations of language', in *EMNLP-IJCNLP*, pp. 529–539, Hong Kong, China, (November 2019).

[18] John Giorgi, Osvald Nitski, Bo Wang, and Gary Bader, 'DeCLUTR: Deep Contrastive Learning for Unsupervised Textual Representations', in *ACL-IJCNLP*, pp. 879–895, Online, (August 2021).

[19] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen, 'DeBERTa: Decoding-enhanced BERT with Disentangled Attention', *arXiv:2006.03654*, (June 2020).

[20] Ingo Hertrich, Susanne Dietrich, and Hermann Ackermann, 'The Margins of the Language Network in the Brain', *Frontiers in Communication*, **5**, (2020).

[21] Nora Hollenstein, Antonio de la Torre, Nicolas Langer, and Ce Zhang,

'CogniVal: A Framework for Cognitive Word Embedding Evaluation', in *CoNLL*, (October 2019).
[22] Shailee Jain and Alexander G. Huth, 'Incorporating context into language encoding models for fMRI', in *NeurIPS*, pp. 6629–6638, (December 2018).
[23] Ryan Kiros, Yukun Zhu, Russ R Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler, 'Skip-Thought Vectors', in *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., (2015).
[24] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini, 'Representational similarity analysis - connecting the branches of systems neuroscience', *Frontiers in Systems Neuroscience*, **2**, 4, (2008).
[25] Matthew A. Lambon Ralph, Elizabeth Jefferies, Karalyn Patterson, and Timothy T. Rogers, 'The neural and computational bases of semantic cognition', *Nature Reviews Neuroscience*, **18**(1), (January 2017).
[26] Michael Lepori and R. Thomas McCoy, 'Picking BERT's Brain: Probing for Linguistic Dependencies in Contextualized Embeddings Using Representational Similarity Analysis', in *COLING*, pp. 3637–3651, Barcelona, Spain (Online), (December 2020).
[27] Ruiqi Li, Xiang Zhao, and Marie-Francine Moens, 'A Brief Overview of Universal Sentence Representation Methods: A Linguistic View', *Journal of the ACM*, 42, (May 2022).
[28] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, 'RoBERTa: A Robustly Optimized BERT Pretraining Approach', *arXiv:1907.11692*, (July 2019).
[29] Antonietta Gabriella Liuzzi, Patrick Dupont, Ronald Peeters, Rose Bruffaerts, Simon De Deyne, Gert Storms, and Rik Vandenberghe, 'Left perirhinal cortex codes for semantic similarity between written words defined from cued word association', *NeuroImage*, **191**, 127–139, (May 2019).
[30] Juliette Millet, Charlotte Caucheteux, Pierre Orhan, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, and Jean-Remi King. Toward a realistic model of speech processing in the brain with self-supervised learning, June 2022. arXiv:2206.01685 [cs, q-bio].
[31] Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang, 'Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models', in *Findings of the Association for Computational Linguistics: ACL 2022*, pp. 1864–1874, Dublin, Ireland, (May 2022).
[32] Subba Reddy Oota, Jashn Arora, Veeral Agarwal, Mounika Marreddy, Manish Gupta, and Bapi Surampudi, 'Neural Language Taskonomy: Which NLP Tasks are the most Predictive of fMRI Brain Activity?', in *NAACL-HLT*, pp. 3220–3237, Seattle, United States, (July 2022).
[33] Subba Reddy Oota, Jashn Arora, Manish Gupta, and Raju S. Bapi, 'Multi-view and Cross-view Brain Decoding', in *COLING*, pp. 105–115, Gyeongju, Republic of Korea, (October 2022).
[34] Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J. Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko, 'Toward a universal decoder of linguistic meaning from brain activation', *Nature Communications*, **9**(1), 963, (March 2018).
[35] Jonathan D Power, Alexander L Cohen, Steven M Nelson, Gagan S Wig, Kelly Anne Barnes, Jessica A Church, Alecia C Vogel, Timothy O Laumann, Fran M Miezin, Bradley L Schlaggar, and Steven E Petersen, 'Functional network organization of the human brain', *Neuron*, **72**(4), 665–678, (November 2011).
[36] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever, 'Language Models are Unsupervised Multitask Learners', Technical report, OpenAI, (2018).
[37] Nils Reimers and Iryna Gurevych, 'Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks', in *EMNLP-IJCNLP*, pp. 3982–3992, Hong Kong, China, (November 2019).
[38] 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), (November 2021).
[39] Jingyuan Sun and Marie-Francine Moens, 'Fine-tuned vs. prompt-tuned supervised representations: Which better account for brain language representations?', in *IJCAI*, Macau, China, (2023).
[40] Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong, 'Towards Sentence-Level Brain Decoding with Distributed Representations', *AAAI*, 7047–7054, (July 2019).
[41] Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong, 'Neural Encoding and Decoding With Distributed Sentence Representations', *IEEE Transactions on Neural Networks and Learning Systems*, **32**(2), 589–603, (February 2020).
[42] Mariya Toneva and Leila Wehbe, 'Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain)', in *NeurIPS*, Vancouver, Canada, (November 2019).
[43] Shaonan Wang, Yunhao Zhang, Xiaohan Zhang, Jingyuan Sun, Nan Lin, Jiajun Zhang, and Chengqing Zong, 'An fMRI Dataset for Concept Representation with Semantic Feature Annotations', *Scientific Data*, **9**(1), 721, (November 2022).
[44] Haiguang Wen, Junxing Shi, Yizhen Zhang, Kun-Han Lu, Jiayue Cao, and Zhongming Liu, 'Neural Encoding and Decoding with Deep Learning for Dynamic Natural Vision', *Cerebral Cortex*, **28**(12), 4136–4160, (December 2018).