Appendix

A Left and Right Language Network and Whole Brain Results

Here, we list the neural encoding results obtained by splitting the language network into voxels belonging to the right and left hemisphere (rather than using the combined language network). For that, we use the left/right language network indications from the original dataset [2]. Also, we include the neural encoding results based on using all available voxels together (i.e., at the level of the whole brain).

Paradigm	Model	Left	Right	Whole
Masked language modeling	**BERT**	0.623	0.572	0.574
	RoBERTa	0.640	0.581	0.588
	DeBERTa	0.639	0.587	0.598
	Mean	**0.634**	**0.580**	**0.587**
Pragmatic coherence	**SkipThoughts**	0.672	0.600	0.608
	GPT-2	0.636	0.568	0.588
	GPT-3	0.673	0.591	0.611
	Mean	**0.660**	**0.586**	**0.602**
Semantic comparison	**S-RoBERTa**	0.620	0.542	0.551
	sup-SimCSE	0.589	0.545	0.539
	S-T5	0.677	0.597	0.601
	Mean	**0.629**	**0.561**	**0.564**
Contrastive learning	**unsup-SimCSE**	0.603	0.555	0.561
	DiffCSE	0.578	0.530	0.551
	DeCLUTR	0.621	0.566	0.590
	Mean	**0.601**	**0.550**	**0.567**

Table 1: Neural encoding-based neural fits for the language network split into left and right hemispheres and for the whole brain.Neural encoding-based neural fit results (R1). Best results are indicated in the same manner as in Table 1 (main text).

As shown in Table 1, there are substantial differences for the encoding performances for the language networks derived from the left and right brain hemispheres (0.631 and 0.570 for left and right, averaged across the 12 models). Higher performance for the left hemisphere is to be expected, since language processing is typically associated with the left brain hemisphere in right-handed individuals. Moreover, at whole-brain level, GPT-3 (0.611), SkipThoughts (0.608) and S-T5 (0.601) yield the highest scores, which is consistent with the performances split across the four networks (see Table 2, main text).

B Neural Encoder Hyperparameter Ablation Study

To assess the effect of the hyperparameter choices on the resulting pairwise accuracy scores, we performed an exemplary ablation study using GPT-2 [3] and the language network [1]. We focused on the hyperparameters of the neural encoder (i.e., the Ridge regression layer on top of the output of the frozen sentence embedding model) rather than the pre-trained models themselves, for which we used the respective default implementations and hyperparameters. We varied the following three hyperparameters of the ridge regression layer based on its implementation in the sklearn library in Python and tested all possible combinations ($\alpha \in [0, 0.5, 1, 2]$, tol $\in$ [1e-3, 1e-4, 1e-5], max_iter $\in [100, 500, 1000]$). This results in 36 measured pairwise accuracies (see Table 2) with a mean of 0.601 and a standard deviation of $s = 0.003$. Not only is this reported result nearly identical to the pairwise accuracy reported for GPT-2 and the language network (0.602, see Table 2 in the main text), but the standard deviation is negligible compared to the standard deviation across the 12 models ($s = 0.027$).

	pairwise accuracy	α	max_iter	tol
	0.597	0	100	1e-03
	0.597	0	500	1e-04
	0.600	0.5	1000	1e-04
	0.606	2	100	1e-03
	0.600	0.5	500	1e-04
	0.606	2	1000	1e-04
	0.597	0	100	1e-05
	0.600	0.5	100	1e-04
	0.597	0	500	1e-05
	0.600	0.5	500	1e-05
	0.597	0	100	1e-04
	0.602	1	1000	1e-03
	0.600	0.5	1000	1e-03
	0.606	2	100	1e-05
	0.602	1	500	1e-03
	0.597	0	1000	1e-05
	0.606	2	1000	1e-05
	0.602	1	100	1e-05
	0.602	1	100	1e-04
	0.597	0	500	1e-03
	0.597	0	1000	1e-04
	0.606	2	1000	1e-03
	0.602	1	500	1e-05
	0.600	0.5	100	1e-05
	0.597	0	1000	1e-03
	0.602	1	100	1e-03
	0.600	0.5	500	1e-03
	0.606	2	100	1e-04
	0.606	2	500	1e-04
	0.606	2	500	1e-05
	0.600	0.5	1000	1e-05
	0.600	0.5	100	1e-03
	0.606	2	500	1e-03
	0.602	1	500	1e-04
	0.602	1	1000	1e-05
	0.602	1	1000	1e-04
mean	**0.601**			
std	**0.003**			

Table 2: Pairwise accruacy scores for each combination of the hyperparameters α, **max_iter** and **tol**

.

C Random Baseline

Model	Language	DMN	Task	Vision
GPT-2	0.602	0.591	0.558	0.627
GPT-2-random	0.519	0.508	0.509	0.496

Table 3: Pairwise accuracy scores for all four brain networks for GPT-2 and GPT-2-random, a GPT-2-based model that has been initialized with random weights (i.e., that has not been pre-trained)

To examine whether the pre-training procedure indeed contributes to the neural fit of a language model, we compare a pre-trained model with a random baseline in the exemplary case of GPT-2. More specifically, we introduce a baseline version of GPT-2 called GPT-2-random, based on using a newly initialized (i.e., not pre-trained) model with the exact same architecture as GPT-2 with randomly initialized weights, for which we then determine its nerual fit. The ran-

dom baseline shown in Table 3 results in an average pairwise accuracy of 0.508 across all four networks, which is close to chance level (i.e., 0.5) for the pairwise accuracy metric. Moreover, the substantially higher performances of the pre-trained GPT-2 model prove the effectiveness of the pre-training procedure for generating representations that lead to a higher neural fit.

D Topic-Passage Variations

To better understand the role of the topics of the 96 passages on the neural encoding results, we aim to replicate different pairwise accuracy evaluation settings (taken from [4]) for the exemplary case of GPT-2 in this section. For that, we calculate pairwise accuracies for pairs either from the same or different topics:

- *same topic different passage (STDP)*: For a given passage, we only calculate the pairwise accuracies for pairings in which the other passage relates to the same topic
- *different topic different passage (DTDP)*: Here, the pairwise accuracies are based on passages from other topics for a given passage

Setting	Language	DMN	Task	Vision
STDP	0.548	0.573	0.494	0.553
DTDP	0.605	0.591	0.560	0.630

Table 4: Pairwise accuracy scores for two different evaluation settings, *same topic different passage (STDP)* and *different topic different passage (DTDP)*, based on the topic categories provided in [2]

Note that that the *same topic same passage* setting is not applicable in our case as we are using whole passages instead of single sentences. The results shown in Table 4 indicate that the pairwise accuracies are higher for passages from different topics than within the same topic, which is expected, as it is easier to distinguish less semantically similar paragraph pairs.

E Correlogram for Sentence Embedding Model Representational Dissimilarity Matrices (RDMs)

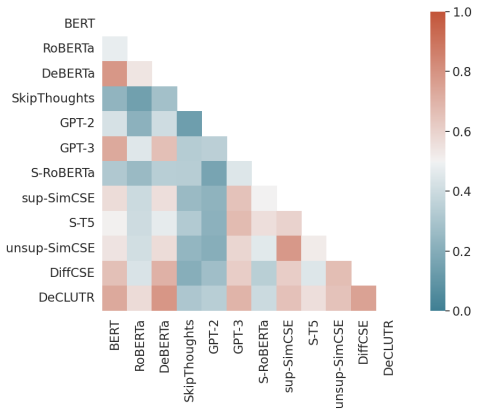

Figure 1: Correlogram for the sentence embedding model RDMs

Here, we create a correlogram based on the Pearson correlations between the RDMs of the sentence embedding models, to examine how representations from different sentence embedding models are correlated to each other. As shown in Figure 1, there are moderate to high correlations ($r \geq 0.5$ for all model pairs) across models within the contrastive learning and semantic comparison paradigms. In contrast, the correlations within the masked language modeling and pragmatic coherence paradigms are lower ($r \leq 0.5$ for some model pairs).

References

[1] Evelina Fedorenko, Michael K. Behr, and Nancy Kanwisher. "Functional specificity for high-level linguistic processing in the human brain". In: *Proceedings of the National Academy of Sciences* 108.39 (Sept. 2011), pp. 16428–16433.

[2] Francisco Pereira et al. "Toward a universal decoder of linguistic meaning from brain activation". In: *Nature Communications* 9.1 (Mar. 2018), p. 963.

[3] Alec Radford et al. *Language Models are Unsupervised Multitask Learners*. Tech. rep. OpenAI, 2018, p. 24.

[4] Jingyuan Sun et al. "Towards Sentence-Level Brain Decoding with Distributed Representations". en. In: *AAAI* (July 2019), pp. 7047–7054.