GPT-2's activations predict the degree of semantic comprehension in the human brain

Charlotte Caucheteux[1,2,*], **Alexandre Gramfort**[2], **and Jean-Rémi King**[1,3]

[1]Facebook AI Research, Paris, France; [2]Université Paris-Saclay, Inria, CEA, Palaiseau, France; [3]École normale supérieure, PSL University, CNRS, Paris, France

Language transformers, like GPT-2, have demonstrated remarkable abilities to process text, and now constitute the backbone of deep translation, summarization and dialogue algorithms. However, whether these models actually understand language is highly controversial. Here, we show that the representations of GPT-2 not only map onto the brain responses to spoken stories, but also predict the extent to which subjects understand the narratives. To this end, we analyze 101 subjects recorded with functional Magnetic Resonance Imaging while listening to 70 min of short stories. We then fit a linear model to predict brain activity from GPT-2 activations, and correlate this mapping with subjects' comprehension scores as assessed for each story. The results show that GPT-2's brain predictions significantly correlate with semantic comprehension. These effects are bilaterally distributed in the language network and peak with a correlation above 30% in the infero-frontal and medio-temporal gyri as well as in the superior frontal cortex, the planum temporale and the precuneus. Overall, this study provides an empirical framework to probe and dissect semantic comprehension in brains and deep learning algorithms.

Neuroscience of language | Deep Neural Networks

In less than two years, language transformers like GPT-2 have revolutionized the field of natural language processing (NLP). These deep learning architectures are typically trained on very large corpora to complete partially-masked texts, and provide a one-fit-all solution to translation, summarization, and question-answering tasks and algorithms (1).

Critically, their hidden representations have been shown to – at least partially – correspond to those of the brain: single-sample fMRI (2–4), MEG (2, 4), and intracranial responses to spoken and written texts (3, 5) can be significantly predicted from a linear combination of the hidden vectors generated by these deep networks. Furthermore, the quality of these predictions directly depends on the models' ability to complete text (3, 4).

In spite of these achievements, strong doubts subsist on whether language transformers actually generate meaningful constructs (6). When asked to complete "I had \$20 and gave \$10 away. Now, I thus have \$", GPT-2 predicts "20"*. Similar trivial errors can be observed for geographical locations, temporal ordering, pronoun attribution and causal reasoning. These results have thus led some to argue that "the system has no idea what it is talking about" (7).

Here, we propose to investigate semantic comprehension, not by assessing whether GPT-2's predictions match human behavior, but by comparing its hidden representations to those of the brain. First, we compare GPT-2's activations with the functional Magnetic Resonance Imaging of 101 subjects listening to 70 min of seven short stories. Second, we evaluate how these representations – shared between GPT-2 and the brain – vary with semantic comprehension, as individually assessed by a questionnaire at the end of each story.

GPT-2's activations linearly map onto fMRI responses to spoken narratives. To assess whether GPT-2 generates similar representations to those of the brain, we first evaluate, for each voxel, subject and narrative independently, whether the fMRI responses can be predicted from a linear combination of GPT-2's activations (Figure 1A). To mitigate fMRI spatial resolution and the necessity to correct each observation by the number of statistical comparisons, we here report either 1) the average scores across voxels or 2) the average score within 314 regions of interest (following a subdivision of Destrieux atlas (8), cf. SI.1). Consistent with previous findings (2, 4, 9), these brain scores are significant over a distributed and bilateral cortical network, and peak in middle- and superior-temporal gyri and sulci, as well as in the supra-marginal and the infero-frontal cortex (2, 4, 9) (Figure 1B.).

By extracting GPT-2 activations from multiple processing stages, called "layers" (from layer one to layer twelve), we confirmed that middle layers best aligned with the brain (Figure 1C), as seen in previous studies (2, 4, 9). The following analyses focus on the activations extracted from the *ninth* layer of GPT-2.

GPT-2's brain predictions correlate with speech comprehension. Does the linear mapping between GPT-2 and the brain reflect a fortunate correspondence (4)? Or, on the contrary, does it reflect similar representations of high-level semantics? To address this issue, we correlate these brain scores to comprehension scores, assessed for each subject-story pair. On average across all voxels, this correlation reaches $\mathcal{R} = 0.33$ ($p < 10^{-6}$, Figure 1D, as assessed with Pearson's p-value provided by SciPy). This correlation is significant across a wide variety of the bilateral temporal, parietal and prefrontal cortices typically linked to language processing (Figure 1E). Together, these results suggest that the shared representations between GPT-2 and the brain encode relate to semantic comprehension.

Low-level processing only partially accounts for the correlation between comprehension and GPT-2's mapping Is this correlation driven by a modulation of low-level processing by uncontrolled factors, such as subjects' attention? To test this issue, we evaluate the predictability of fMRI given low-level phonological features: the word rate, phoneme rate, phonemes, stress and tone of the narrative. The corresponding brain scores correlate with the subjects' understanding ($\mathcal{R} = 0.17, p < 10^{-2}$)

*as assessed using Huggingface interface (https://github.com/huggingface/transformers) and GPT-2 pretrained model with temperature=0.

DRAFT

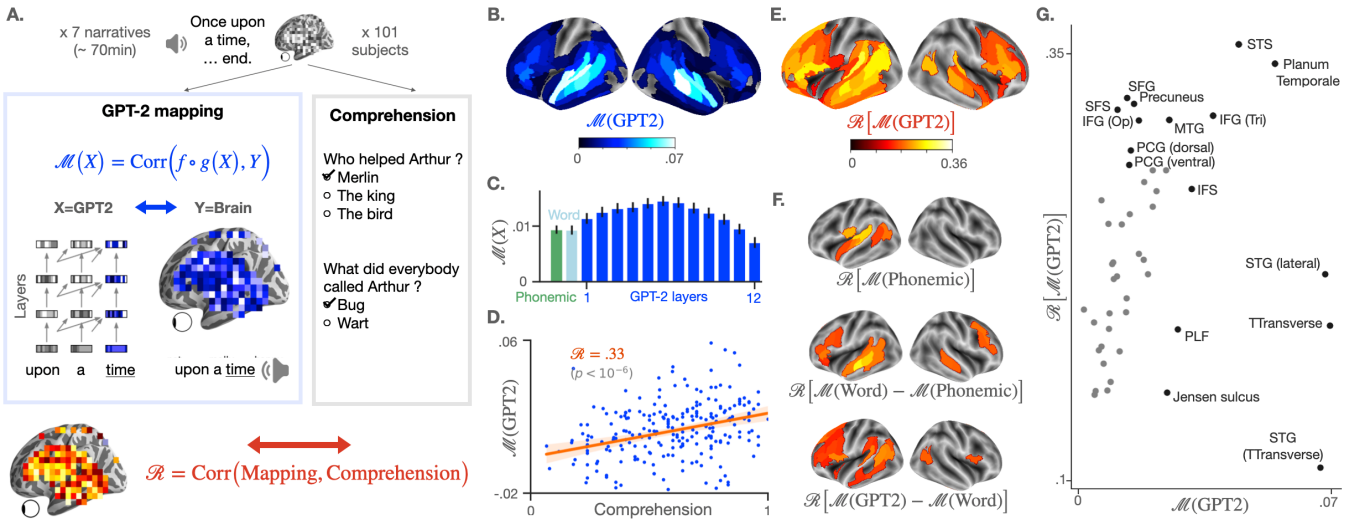

Fig. 1. A. 101 subjects listen to narratives (70 min of unique audio stimulus in total) while their brain signal is recorded using functional MRI. At the end of each story, a questionnaire is submitted to each subject to assess their understanding, and the answers are summarized into a comprehension score specific to each (narrative, subject) pair (grey box). In parallel (blue box on the left), we measure the mapping between the subject's brain activations and the activations of GPT-2, a deep network trained to predict a word given its past context, both elicited by the same narrative. To this end, a linear spatio-temporal model ($f \circ g$) is fitted to predict the brain activity of one voxel Y, given GPT-2 activations X as input. The degree of mapping, called "brain score" is defined for each voxel as the Pearson correlation between predicted and actual brain activity on held-out data (blue equation, cf. Methods). Finally, we test the correlation between the comprehension scores of the subjects and their corresponding brain scores using Pearson's correlation (red equation). A positive correlation means that the representations shared across the brain and GPT-2 are key for the subjects to understand a narrative. **B.** Brain scores (fMRI predictability) of the activations of the ninth layer of GPT-2. Scores are averaged across subjects, narratives, and voxels within brain regions (157 regions in each hemisphere, following a subdivision of Destrieux Atlas (8), cf. SI.1). Only significant regions are displayed, as assessed with a two-sided Wilcoxon test across (subject, narrative) pairs, testing whether the brain score is significantly different from zero (threshold: .05). **C.** Brain scores, averaged across fMRI voxels, for different activation spaces: phonological features (word rate, phoneme rate, phonemes, tone and stress, in green), the non-contextualized word embedding of GPT-2 ("Word", light blue) and the activations of the contextualized layers of GPT-2 (from layer one to layer twelve, in blue). The error bars refer to the standard error of the mean across (subject, narrative) pairs (n=237). **D.** Comprehension and GPT-2 brain scores, averaged across voxels, for each (subject, narrative) pair. In red, Pearson's correlation between the two (denoted $\mathcal{R}$), the corresponding regression line and the 95% confidence interval of the regression coefficient. **E.** Correlations ($\mathcal{R}$) between comprehension and brain scores over regions of interest. Brain scores are first averaged across voxels within brain regions (similar to B.), then correlated to the subjects' comprehension scores. Only significant correlations are displayed (threshold: .05). **F.** Correlation scores ($\mathcal{R}$) between comprehension and the subjects' brain mapping with phonological features (M(Phonemic) (i), the share of the word-embedding mapping that is not accounted by phonological features $\mathcal{M}(\text{Word}) - \mathcal{M}(\text{Phonemic})$ (ii) and the share of the GPT-2 ninth layer's mapping not accounted by the word-embedding $\mathcal{M}(\text{GPT2}) - \mathcal{M}(\text{Word})$ (iii). **G.** Relationship between the average GPT-2-to-brain mapping (ninth layer) per region of interest (similar to B.), and the corresponding correlation with comprehension ($\mathcal{R}$, similar to D.). Only regions of the left hemisphere, significant in both B. and D. are displayed. For simplicity, we only use the 75 parcellation regions (in each hemisphere) of Destrieux Atlas. In black, the top ten regions in terms of brain and correlation scores (cf. SI.1 for the acronyms). Significance in D, E and F is assessed with Pearson's p-value provided by SciPy[†]. In B, E and F, p-values are corrected for multiple comparison using a False Discovery Rate (Benjamin/Hochberg) over the 2×157 regions of interest.

DRAFT

but less so than the brain scores of GPT-2 ($\Delta\mathcal{R} = 0.16$). These low-level correlations with comprehension peak in the left superior temporal cortex (Figure 1F). Overall, this result suggests that the link between comprehension and GPT-2's brain mapping may be partially explained by, but not reduced to a modulation of low-level processing.

High-level processing best explains the link between comprehension and GPT-2's predictions Is the correlation between comprehension and GPT-2's mapping driven by a *lexical* effect as opposed to a high-level ability to meaningfully combine words? To tackle this issue, we extract the non-contextualized word embeddings from GPT-2, compute their brain mapping, and, again, evaluate the correlation between comprehension and these lexical mappings. On average across voxels, the resulting $\mathcal{R}$ scores are higher than those of phonological features, and lower than those of GPT-2's ninth layer ($\Delta\mathcal{R} = 0.03$).

Comparisons between phonological, word-embedding and GPT-2 ninth layer's $\mathcal{R}$ scores are displayed in Figure 1F. Lexical effects (word-embedding *versus* phonological) peak in the superior-temporal lobe and in pars triangularis. In addition, higher-level effects (GPT-2 ninth layer *versus* word-embedding) peak in the superior-frontal, posterior superior-temporal gyrus, in the precuneus and in both the pars opercularis and pars triangularis – a network typically associated with high-level language comprehension (10).

Comprehension effects are mainly driven by individuals' variability The variability in comprehension scores could results from inter-stimulus variability – some stories may be harder to comprehend than others for GPT-2 and/or for subjects. In addition, longer narratives could be easier to understand, and the brain mapping be higher due to a larger amount of data. To control for the this effect, we fit a linear mixed model to predict comprehension scores given brain scores, specifying the narrative as a random effect (cf. SI.1). The fixed effect of brain score (shared across narratives) is highly significant: $\beta = 5.9, p < 10^{-6}$, cf. SI.1). However, the random effect (slope specific to each single narrative) is not ($\beta = 1.4$, $p = .88$). Overall, these analyses confirm that the link between GPT-2 and speech comprehension is mainly driven by subjects' individual differences in their ability to make sense of the narratives.

Discussion Our analyses reveal a positive correlation between semantic comprehension and the degree to which GPT-2's activations map onto those of the brain. Thus, GPT-2's activations

– at least partially – relate to the semantic representations our brain generates when we process a spoken narrative.

Our fMRI results strengthen and complete prior work on the neural correlates of semantic comprehension. In particular, previous studies have used inter-subject brain correlation to reveal the brain regions associated with understanding (11). For example, Lerner et al have presented subjects with normal texts or texts scrambled at the word, sentence or paragraph level, in order to parametrically manipulate their level of comprehension (10). The corresponding fMRI signals proved to correlate across subjects in the primary and second auditory areas when the input was scrambled below the lexical level. By contrast, they only became correlated in the bilateral infero-frontal and temporo-parietal cortex when the scrambling was performed at the level of sentences and paragraphs. Our model-based results are remarkably consistent with this hierarchical organization. In sum, the present work illustrates that language processes can be directly investigated with uncontrolled stimuli and may not necessitate artificial – and costly – experimental conditions (12).

The present results do not mean that GPT-2 fully understands what it reads. First, the correlation between speech comprehension and GPT-2's ability to map onto the brain reaches, at best 36% (in STS) (i.e. 13% of explained variance), indicating that the variation in comprehension remains, mostly unaccounted for. Second, language comprehension is here operationally defined as subjects' ability to correctly answer a set of questions after each text. This measure may thus be noisy because of mnesic and attentional processes. Third, fMRI's signal-to-noise ratio and temporal resolution limits our ability to determine whether GPT-2's representations of specific events, goals, and mental states match those of the brain.

Finally, the present study strengthens and clarifies the similarity between the brain and deep language models, repeatedly observed in the past three years (2–4, 9, 13). Together, these findings reinforce the relevance of deep language models for understanding the neural bases of narrative comprehension.

Materials and Methods

Our analyses rely on the "Narratives" dataset (14), composed of the brain signals, recorded using fMRI, of 345 subjects listening to 27 narratives.

Narratives and comprehension score Among the 27 stories of the dataset, we selected the seven stories for which subjects were asked to answer a comprehension questionnaire at the end, and for which the answers varied across subjects (more than ten different comprehension scores across subjects), resulting in 70 minutes of audio stimuli in total, from four to 19 minutes per story (Figure 2). Questionnaires were either multiple-choice, fill-in-the blank, or open questions (answered with free text) rated by humans (14). Here, we used the comprehension score computed in the original dataset which was either a proportion of correct answers or the sum of the human ratings, scaled between 0 and 1 (14). It summarizes the comprehension of one subject for one narrative (specific to each (narrative, subject) pair).

Brain activations The brain activations of the 101 subject who listened to the selected narratives were recorded using fMRI, as described in (14). As suggested in the original paper, pairs of (subject, narrative) were excluded because of noisy recordings, resulting in 237 pairs in total.

GPT-2 activations GPT-2 (1) is a high-performing neural language model trained to predict a word given its previous context (it does not have access to succeeding words), given millions of examples (e.g Wikipedia texts). It consists of multiple Transformer modules (twelve, each of them called "layer") stacked on a non-contextual word embedding (a look-up table that outputs a single vector per vocabulary word) (1). Each layer l can be seen as a nonlinear system that takes a sequence of w words as input, and outputs a contextual vector of dimension (w, d), called the "activations" of layer l ($d = 768$). Intermediate layers were shown to better encode syntactic and semantic information than input and output layers (15), and to better map onto brain activity (2, 4). Thus, for our analyses, we selected the activations of the *ninth* layer of GPT-2, elicited by the same narratives as the subjects.

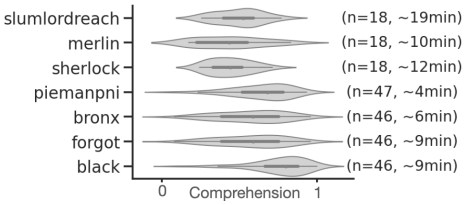

Fig. 2. For each of the seven narratives: number of subjects (n), distribution of comprehension scores across subjects and length of the narrative.

In practice, the narratives' transcripts were formatted (replacing special punctuation marks such as "–" and duplicated marks "?." by dots), tokenized using GPT-2 tokenizer and input to the GPT-2 pretrained model provided by Huggingface [‡]. The activation vectors between successive fMRI measurement were summed to obtain one vector per measurement. To match the fMRI measurements and the GPT-2 vectors over time, we used the speech-to-text correspondences provided in the fMRI dataset (14).

Linear mapping between GPT-2 and the brain For each (subject, narrative) pair, we measure the mapping between i) the fMRI activations elicited by the narrative and ii) the activations of GPT-2 (layer nine) elicited by the same narrative. To this end, a linear spatiotemporal model is fitted on a train set to predict the fMRI scans given the GPT-2 activations as input. Then, the mapping is evaluated by computing the Pearson correlation between predicted and actual fMRI scans on a held out set I:

$$\mathcal{M}^{(s,w)} : I \mapsto \mathcal{L}\left(f \circ g(X^{(w)})_{i \in I}, (Y_i^{(s,w)})_{i \in I}\right)$$

With $f \circ g$ the fitted estimator (g: temporal and f: spatial mappings), $\mathcal{L}$ Pearson's correlation, $X^{(w)}$ the activations of GPT-2 and $Y^{(s,w)}$ the fMRI scans of subjects s, both elicited by the narrative w.

In practice, f is a ℓ_2-penalized linear regression (following scikit-learn implementation[§]) and g is a finite impulse response (FIR) model with 5 delays, where each delay sums the activations of GPT-2 input with the words presented between two TRs. For each subject, this procedure is repeated across five train (80% of the fMRI scans) and disjoint test folds (20% of the fMRI scans), and the Pearson correlations are averaged across folds to obtain a single score per (subject, narrative) pair. This score, denoted $\mathcal{M}(X)$ in Figure 1A, measures the mapping between the activations space X and the brain of one subject, elicited by one narrative.

Significance Significance was either assessed by using (i) a second-level Wilcoxon test (two-sided) across subject-narrative pairs, testing whether the mapping (one value per pair) was significantly different from zero (Figure 1B), (ii) either by using the first-level Pearson p-value provided by SciPy[¶] (Figure 1D-G). In Figure 1B, E, F, p-values were corrected for multiple comparison (2×157 ROIs) using False Discovery Rate (Benjamin/Hochberg)[‖].

Supporting Information (SI).

Brain parcellation. In Figure 1B, E, and F, we used a subdivision of the parcellation from Destrieux Atlas (8). Regions with

[‡] https://github.com/huggingface/transformers
[§] https://scikit-learn.org/
[¶] https://www.scipy.org/
[‖] https://mne.tools/

DRAFT

more than 400 vertices were split into smaller regions (so that each regions contains less than 400 vertices). The original parcellation consists of 75 regions per hemisphere. Our custom parcellation consists in 157 regions per hemisphere.

In Figure 1G, we use the original parcellation for simplicity, and the following acronyms:

- STG/STS (Superior temporal gyrus/sulcus)
- MTG/MTS (Medial temporal gyrus/sulcus)
- SFG/SFS (Superior frontal gyrus/sulcus)
- IFG/IFS (Inferior frontal gyrus/sulcus)
- Tri/Op (Pars triangularis/opercularis)
- IPG/IPS (Inferior parietal gyrus/sulcus)
- TTransverse (Temporal transverse sulcus)
- PCG (Posterior cingulate gyrus)
- PLF (Posterior lateral fissure)

Mixed-effect model. Not all subjects listened to the same stories. To check that the $\mathcal{R}$ scores (correlation between comprehension and brain mapping) were not driven by the narratives and questionnaires' variability, a linear mixed-effect model was fit to predict the comprehension of a subject given its brain mapping scores, specifying the narrative as a random effect. More precisely, if $_{w_i} \in \mathbb{R}$ corresponds to the mapping scores of the ith subject that listened to the story w, and $C_{w_i} \in \mathbb{R}$ refers to the comprehension scores, we estimate the fixed effect parameters $\widetilde{\beta} \in \mathbb{R}$ and $\widetilde{\eta} \in \mathbb{R}$ (shared across narratives), and the random effect parameter $\beta_w \in \mathbb{R}$ and $\eta_w \in \mathbb{R}$ (specific to the narrative w) such that:

$$C_{w_i} = (\widetilde{\beta} + \beta_w) \times_{w_i} + (\widetilde{\eta} + \eta_w) + \epsilon_{w_i}$$

with ϵ_{w_i} a vector of i.i.d normal errors with mean 0 and variance σ^2. In practice, we use the statsmodels[**] implementation of linear mixed-effect models. Significance of the coefficients were assessed with a t-test, as implemented in statsmodels.

References

1. Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. page 24, 2018.
2. Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *arXiv:1905.11833 [cs, q-bio]*, November 2019. arXiv: 1905.11833.
3. Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua Tenenbaum, and Evelina Fedorenko. Artificial Neural Networks Accurately Predict Language Processing in the Brain. *bioRxiv*, page 2020.06.26.174482, June 2020. . Publisher: Cold Spring Harbor Laboratory Section: New Results.
4. Charlotte Caucheteux and Jean-Rémi King. Language processing in brains and deep neural networks: computational convergence and its limits. *bioRxiv*, page 2020.07.03.186288, July 2020. . Publisher: Cold Spring Harbor Laboratory Section: New Results.
5. Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Catherine Kim, Colton Casto, Fanda Lora, Adeen Flinker, Sasha Devore, Werner Doyle, Patricia Dugan, Daniel Friedman, Avinatan Hassidim, Michael Brenner, Yossi Matias, Ken A. Norman, Orrin Devinsky, and Uri Hasson. Thinking ahead: prediction in context as a keystone of language in humans and machines. *bioRxiv*, page 2020.12.02.403477, January 2021. . Publisher: Cold Spring Harbor Laboratory Section: New Results.
6. Yixin Nie, Adina Williams, Emily Dinan, Mohit Bansal, Jason Weston, and Douwe Kiela. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*, 2019.
7. Gary Marcus. Gpt-2 and the nature of intelligence. *The Gradient*, 2020.
8. Christophe Destrieux, Bruce Fischl, Anders Dale, and Eric Halgren. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, 53 (1):1–15, October 2010. ISSN 1053-8119. .
9. Shailee Jain and Alexander G Huth. Incorporating Context into Language Encoding Models for fMRI. preprint, Neuroscience, May 2018.
10. Y. Lerner, C. J. Honey, L. J. Silbert, and U. Hasson. Topographic Mapping of a Hierarchy of Temporal Receptive Windows Using a Narrated Story. *Journal of Neuroscience*, 31(8): 2906–2915, February 2011. ISSN 0270-6474, 1529-2401. .
11. Evelina Fedorenko, Terri Scott, Peter Brunner, William Coon, Brianna Pritchett, Gerwin Schalk, and Nancy Kanwisher. Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences of the United States of America*, 113, September 2016. .
12. Liberty Hamilton and Alexander Huth. The revolution will not be controlled: natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, 35:1–10, July 2018. .
13. Jon Gauthier and Anna Ivanova. Does the brain represent words? An evaluation of brain decoding studies of language understanding. *arXiv:1806.00591 [cs]*, June 2018. arXiv: 1806.00591.
14. Samuel A. Nastase, Yun-Fei Liu, Hanna Hillman, Asieh Zadbood, Liat Hasenfratz, Neggin Keshavarzian, Janice Chen, Christopher J. Honey, Yaara Yeshurun, Mor Regev, Mai Nguyen, Claire H. C. Chang, Christopher Baldassano, Olga Lositsky, Erez Simony, Michael A. Chow, Yuan Chang Leong, Paula P. Brooks, Emily Micciche, Gina Choe, Ariel Goldstein, Tamara Vanderwal, Yaroslav O. Halchenko, Kenneth A. Norman, and Uri Hasson. Narratives: fMRI data for evaluating models of naturalistic language comprehension. preprint, Neuroscience, December 2020.
15. Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What Does BERT Learn about the Structure of Language? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3651–3657, Florence, Italy, 2019. Association for Computational Linguistics. .

[**] https://www.statsmodels.org/

DRAFT