Brains and algorithms partially converge in natural language processing

Charlotte Caucheteux, Jean-Rémi King

THE MAPPING OF DEEP LANGUAGE MODELS ON BRAIN RESPONSES PRIMARILY DEPENDS ON THEIR PERFORMANCE

A PREPRINT

Charlotte Caucheteux
Facebook AI Research

Jean-Rémi King
PSL University, CNRS
Facebook AI Research

October 1, 2021

ABSTRACT

Recent deep networks like transformers not only excel in several language tasks, but their activations linearly map onto the human brain during language processing. Is this functional similarity caused by specific factors, such as the language abilities and the architecture of the algorithms? To address this issue, we analyze the brain responses to isolated sentences in a large cohort of 102 subjects, each recorded with both functional magnetic resonance imaging (fMRI) and magnetoencephalography (MEG). We then compare the ability of 32,400 transformer embeddings to linearly map onto these brain responses. Finally, we evaluate how the architecture, training, and performance of the models independently account for this brain mapping. Our analyses reveal two main findings. First, the similarity between brain responses and the activations of language models primarily depends on their ability to predict words from the context. Second, this similarity allows us to decompose and precisely track the rise and maintenance of perceptual, lexical, and compositional representations within each cortical region. Overall, this study evidences a partial convergence of language transformers to brain-like solutions, and shows how this phenomenon helps unravel the brain bases of natural language processing.

***K*eywords** Natural Language Processing | Encoding | Functional Magnetic Resonance Imaging | Magneto-encephalography

1 Introduction

Deep learning has recently made remarkable progress in harnessing capabilities hitherto considered unique to the human species (4; 5; 6). In particular, language transformers demonstrate unprecedented completion, translation, and summarization abilities (7; 8; 9; 10). Do these algorithms process words and sentences like the human brain?

Preliminary evidence suggests that they might. First, word embeddings – high dimensional dense vectors trained to predict lexical neighborhood (11; 12; 13; 14) – have been shown to linearly map onto the brain responses elicited by words presented either in isolation (15; 16; 17) or within narratives (18; 19; 20; 21; 22; 23). Second, the "contextualized" activations of language transformers improve the precision of this mapping, especially in the prefrontal, temporal and parietal cortices (24; 25; 26). Third, specific computations of deep language models, such as the estimations of word surprisal (i.e. the probability of a word given its context) and the parsing of syntactic constituents have been shown to correlate with evoked related potentials (27; 28; 29; 30).

However, the comparison between deep language models and the brain is fragmentary. First, most studies map the high-dimensional activations of deep language models onto fMRI: yet, these slow brain signals are unable to determine the sequence of brain representations elicited as the sentence unfolds. Second, past studies are based on i) a small number of subjects and ii) on a small set of language models varying in dimensionality, architecture, training objective, and training corpus. These computational differences prevent a formal comparison between the brain and these algorithms. Third, the mapping between brains and algorithms could be driven by factors largely independent from language

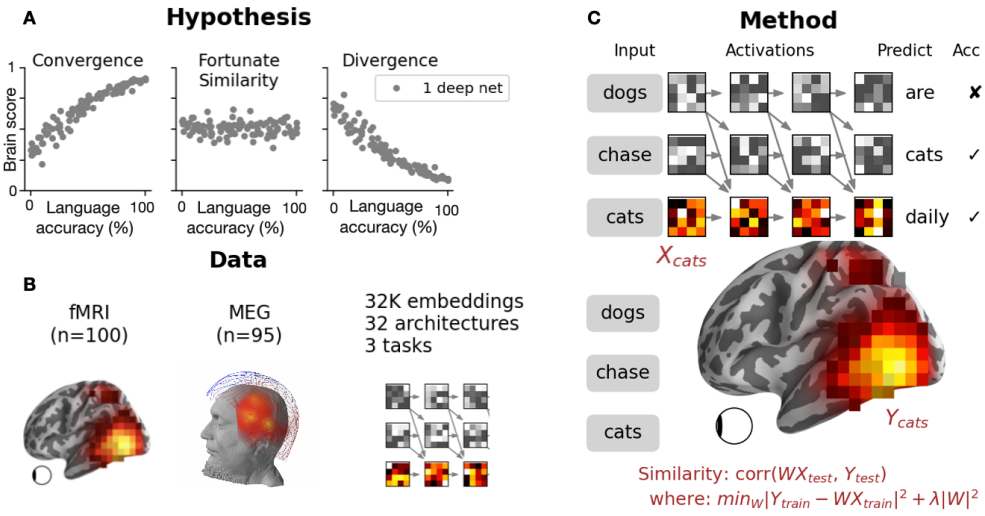

Figure 1: **Hypotheses and Methods**. **A**. The three panels represent three hypotheses on the link between language transformers and the brain. Each dot represents one hypothetical language transformer. Language transformers can be considered to converge to brain-like computations if their language performance (x-axis: i.e. top-1 accuracy at predicting a word from its previous context) correlates with their ability to map onto brain responses to the same stimuli (i.e. y-axis: brain score), and *vice versa* for a divergence hypothesis. High-dimensional neural networks can, in principle, capture relevant information (1; 2), and thus lead to a fortunate similarity with brain responses. **B**. Using fMRI and MEG recordings in the same subjects (3), we compare both the language performance and the brain-mapping of the layer-wise activations ('embeddings') extracted from a large variety of language transformers. **C**. To compute the brain scores, we (1) fit a linear regression W from the model's activations X to predict brain responses Y and (2) evaluate this mapping with a correlation between the predicted and true brain responses to held-out sentences Y_{test}. MEG scores and fMRI scores are computed independently.

processes: for instance, networks with random weights have been shown to significantly predict brain responses to sound, speech and language stimuli (31; 32; 33).

Thus, we ask two questions: is the similarity between language algorithms and the brain driven by specific factors? If so, can these algorithms help reveal the spatiotemporal hierarchy of language computations in the human brain?

To address these issues, we analyze the brain responses of 102 healthy adults, scanned with both functional magnetic resonance imaging (fMRI) and source-localized magneto-encephalography (MEG) by Schoffelen et al (3) during two 1 h-long sessions, during which they read isolated Dutch sentences composed of 9 to 15 words. We then extract 32,400 embeddings from a variety of language transformers to compare their ability to linearly map onto these brain responses ('brain score' (32)). Finally, we assess how differences in training, architectures, and language performance independently contribute to these brain mappings (Figure 1).

The results demonstrate that the similarity between language models and the brain primarily depends on their language ability (i.e. accurately predicting words from the context).

2 Results

2.1 Intermediate layers predict brain responses best.

To evaluate how the activations (X) of a deep language model map onto the brain (y) in response to the same sentences, we fit, within each subject, an ℓ_2-penalized linear regression (W) to predict single-sample fMRI and MEG responses for each voxel/sensor independently. We assess the accuracy of this mapping with a Pearson R correlation (hereafter referred to as 'brain score' (34)) between true and predicted brain activations on held-out recordings of distinct sentences, using a five-fold cross-validation. Except if stated otherwise, we report the average brain scores across all voxels (or across MEG sensors and time samples) in the text, and refer the reader to the figures for a more complete description. Finally, we assess the statistical significance of these (average or single-voxel/channel) brain scores with a two-sided Wilcoxon test across subjects.

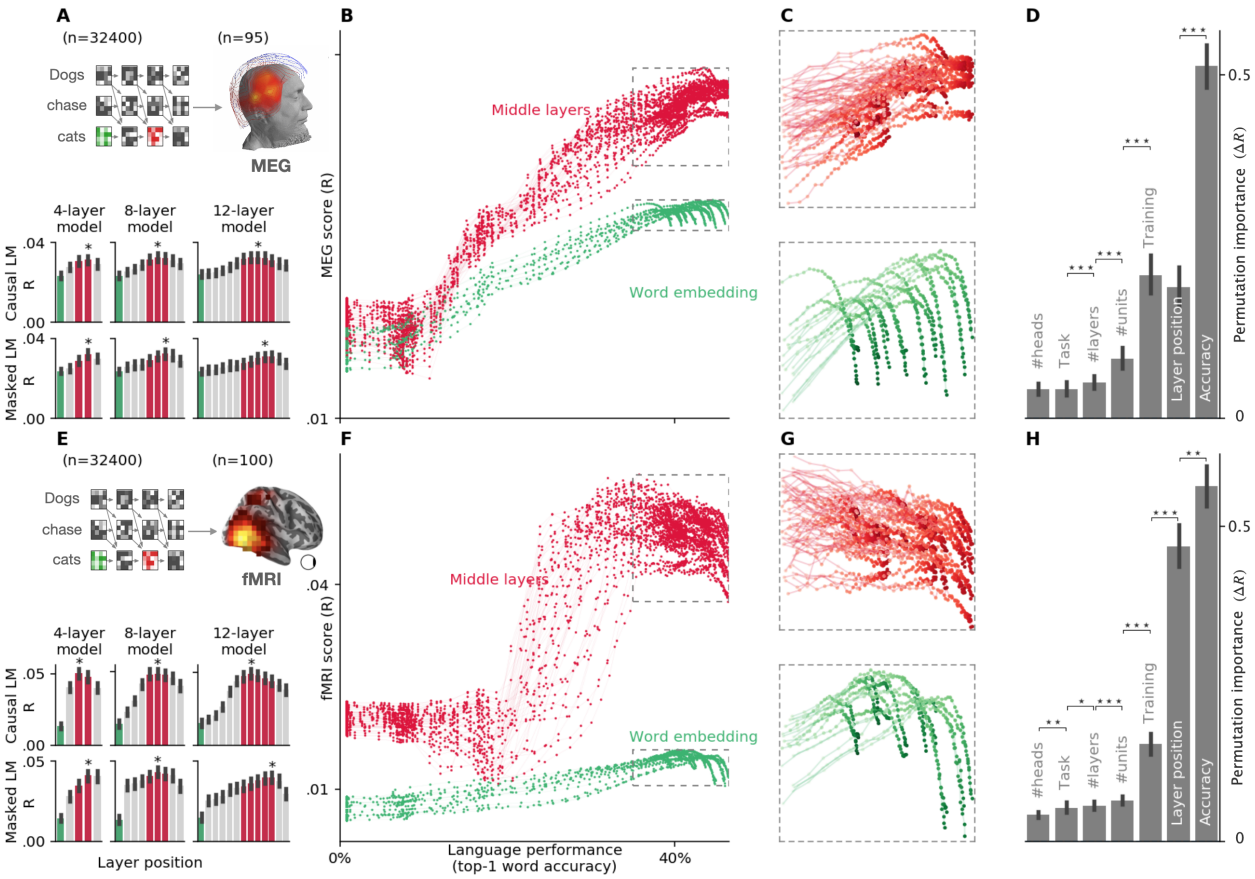

Figure 2: **Language transformers tend to converge towards brain-like representations. A**. Bar plots display the average MEG score (across time and channels) of six representative transformers varying in tasks (causal vs masked language modeling) and depth (4-12 layers). The green and red bars correspond to the word-embedding and middle layers, respectively. The star indicates the layer with the highest MEG score. **B**. Average MEG scores (across subjects, time, and channels) of each of the embeddings (dots) extracted from 18 *causal* architectures, separately for the input layer (word embedding, green) and the middle layers (red). **C**. Zoom of (B), focusing on the best neural networks (i.e. accuracy >35 %), revealing a slight plateau and divergence of the middle and input layers, respectively. **D**. Permutation importance reveals how each property of the language transformers specifically contribute to the brain scores (ΔR). All properties significantly contribute to the brain scores ($\Delta R > 0$, all $p < 0.0001$ across subjects). Ordered pairwise comparisons of the permutation scores are marked with a star ('*': $p < .05$, '**': $p < .01$, '***': $p < .001$). **E-H**. Same as A-D, but evaluated on fMRI recordings. All error bars are the 95% confidence intervals across subjects.

We evaluate the brain scores of 32 transformer architectures (varying from 4 to 12 layers, each ranging from 128 to 512 dimensions, and each benefiting from 4 to 8 attention heads), trained on the same Wikipedia dataset either with a 'causal' language modeling (CLM) or a 'masked' language modeling task (MLM). For each architecture, we input the model with the sentence read by the subjects in the MEG or fMRI scanner, extract the activations from every layer, and, finally, compute the corresponding fMRI and MEG scores.

The brain scores of all these trained language models are significantly above chance (all $p < 10^{-9}$, Figure 2A and E). As detailed in supplementary analyses, the modest values of these brain scores reflect the notoriously high level of noise in single-sample single-voxel/channel neuroimaging data. Indeed, fMRI and MEG scores reach $R = .048$ and $R = .041$, respectively, for the best layer of a typical 12-layer CLM, which is close to and even exceeds the noise ceiling (fMRI: $R = .060$, MEG: $R = .020$, Figure S4).

Overall, the brain scores vary as a function of the relative depth of the embedding within the transformer. Specifically, both MEG and fMRI scores follow an inverted U-shaped pattern across layers for all architectures (Figure 2A and E): middle layers [1] systematically outperform output (fMRI: $\Delta R = .011 \pm .001$, $p < 10^{-18}$, MEG: $\Delta R = .003 \pm .0005$,

[1] For simplicity, we refer to 'middle layers' as the layer $l \in [n_{\text{layers}}/2, 3n_{\text{layers}}/4]$, Figure 2A and E

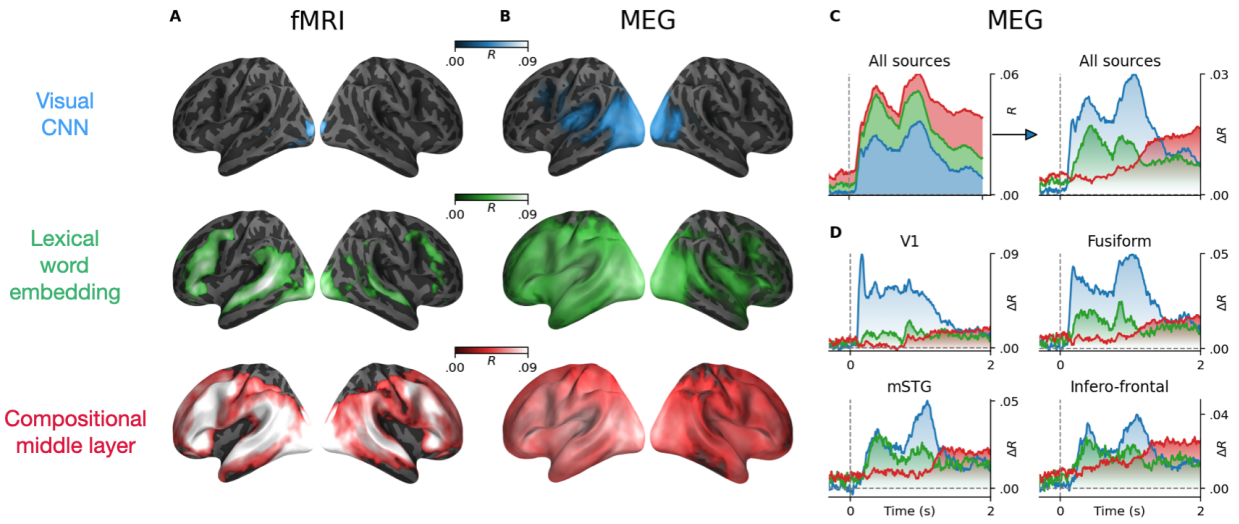

Figure 3: **The similarity between language transformers and the brain reveals the spatio-temporal hierarchy of language representations.** Lexical and compositional representations can be isolated from the word embedding (green) and one middle layer (red) of a typical language transformer (here, a 12-layer causal transformer). To account for low-level visual representations, we also compute the brain scores of a convolutional neural network trained on character recognition (blue). **A**. Mean (across subjects) fMRI scores obtained with the convolutional neural network (blue), the word embedding layer (green), and the ninth layer of a 12-layer transformer (red). All colored regions display significant fMRI scores across subjects after FDR correction for multiple comparisons. **B**. The temporal resolution of MEG allows to precisely track the unfolding of the three types of representation over time. In B, the mean MEG scores averaged across all time samples and subjects. **C**. Left: mean MEG scores averaged across all sensors. Right: mean MEG gains averaged across all sensors: i.e. green: [word embedding] - [visual embedding]; red: [compositional embedding] - [word embedding]). **D**. Mean MEG gains in four regions of interest. For a whole-brain depiction of the MEG gains, see Video 2.

$p < 10^{-13}$) and input layers (fMRI: $\Delta R = .031 \pm .001$, $p < 10^{-18}$, MEG: $\Delta R = .009 \pm .001$, $p < 10^{-17}$). This result confirms that the intermediary representations of deep language transformers are more "brain-like" than input and output layers (26).

2.2 The emergence of brain-like representations predominantly depends on the network's ability to predict missing words.

The above findings result from *trained* neural networks. However, recent studies suggest that random (i.e. untrained) networks can significantly map onto brain responses (31; 32; 33). To test whether brain mapping specifically depends on the language proficiency of the model, we assess the brain scores of each of the 32 architectures trained with 100 distinct amounts of data. For each of these training steps, we compute the top-1 accuracy of the model at predicting masked or incoming words from their contexts. This analysis thus results in 32,400 embeddings, whose brain scores can be evaluated as a function of language performance (Figure 2B and F).

We observe three main findings. First, random embeddings systematically lead to significant brain scores across subjects and architectures. The mean fMRI score across voxels is $R = .19 \pm .01$, $p < 10^{-16}$. The mean MEG score across channels and time sample is $R = .18 \pm .008$, $p < 10^{-16}$). This result suggests that language transformers partially map onto brain responses independent of their language abilities.

Second, brain scores strongly correlate with language accuracy in both MEG ($R = .77$ Pearson's correlation on average $\pm .01$ across subjects) and fMRI ($R = .57 \pm .02$, Figure 2B and C). The correlation is higher for middle (for fMRI: $R = .81 \pm 02$ and MEG: $R = .91 \pm 01$) than input ($R = .39 \pm 03$) and output layers ($R = .63 \pm 03$). Beta coefficients for each particular layer and architecture are displayed in Figure S5A and B. Furthermore, single-voxel analyses show that this correlation between brain score and language performance is driven mainly by the superior temporal sulcus and gyrus for the embedding layer (mean $R = .52 \pm .06$) and is widespread for the middle layers, exceeding $R = .85$ correlation in the superior temporal sulcus, infero-frontal, fusiform and angular gyri (Figure S5C).

Overall, this result suggests that the better language models are at predicting words from the context, the more their activations linearly map onto those of the brain.

Third, the highest brain scores are *not* achieved by the very best language transformers (Figure 2C and G). For instance, CLM transformers best map onto MEG ($R = .039$) and fMRI ($R = .056$) when they reach a language performance of 43% and 32%, respectively. By contrast, the very best transformers reach a language accuracy of 46%, but have significantly smaller brain scores (Figure 2C and G).

2.3 Architectural and training factors impact brain scores too.

Language proficiency co-varies with the amount of training as well as with several architectural variables. To disentangle the contribution of each of these variables to the brain scores, we perform a permutation feature importance analysis. Specifically, we train a Random Forest estimator (35) to predict the average brain scores (across voxels or MEG sensors) of each subject independently, given the layer of the representation, the architectural properties (number of layers, dimensionality, attention head), task (CLM, MLM), amount of training (number of steps) and language performance (top-1 accuracy) of the transformer. Permutation feature importance then estimates the unique contribution of each feature in explaining the variability of brain scores across models (36; 35). The results confirm that language performance is the most important factor that drives the brain scores (Figure 2 D,H). This factor supersedes other covarying factors such as the amount of training, and the relative position of the embedding with regard to the architecture ('layer position'): $\Delta R = .56 \pm .01$ for fMRI, $\Delta R = .51 \pm .02$ for MEG. Nevertheless, these other factors contribute significantly to the prediction of brain scores ($p < 10^{-16}$across subjects for all variables).

Overall, these results show that the ability of deep language models to map onto the brain *primarily* depends on their ability to predict words from the context, and is best supported by the representations of their middle layers.

2.4 The mapping between the brain and language models helps to automatically decompose the cortical hierarchy of language.

Where and when are the language representations of the brain similar to those of deep language models? To address this question, we extract the activations of the first (X_{CLM_1},a.k.a "word embedding") and ninth layers (X_{CLM_9}) of a representative 12-layer transformer trained on causal language modeling (CLM). Unlike the 9$^{\text{th}}$ layer, the word embedding layer represents each word as a unique vector independent of its context (37). On the contrary, deeper layers are compositional: they combine word representations to best predict incoming words and can thus capture sentence-level properties like syntax (38; 39). To control for sub-lexical features, we also extract the activations X_{CNN} of the last layer of a convolutional neural network (CNN) trained on character recognition (40) and input with the image of each word. By definition, such CNN is devoid of context and is thus unable to capture the meaning of words. Consequently, we hereafter refer to these different sets of activations as *visual*, *lexical* and *compositional* embeddings (Figure 3A).

In fMRI, the brain scores of the visual embedding peak in the early visual cortex (V1) (mean brain scores across voxels: $R = .022 \pm .003, p < 10^{-11}$). By contrast, the brain scores of lexical embedding peak in the left superior temporal gyrus ($R = .052 \pm .004, p < 10^{-13}$) as well as in the inferior temporal cortex and middle frontal gyrus ($R = .053 \pm .003, p < 10^{-15}$) and are significant across the entire language and reading network (Figure 3B). Finally, the brain scores of the compositional embedding are significantly higher than those of lexical of embeddings in the superior temporal gyrus ($\Delta R = .012 \pm .001$, $p < 10^{-16}$), the angular gyrus ($\Delta R = .010 \pm .001$, $p < 10^{-16}$), the infero-frontal cortex ($\Delta R = .016 \pm .001$, $p < 10^{-16}$) and the dorsolateral prefrontal cortex ($\Delta R = .012 \pm .001$, $p < 10^{-13}$). While these effects are lateralized (left hemisphere versus right hemisphere: $\Delta R = .010 \pm .001$, $p < 10^{-14}$), they are significant across a remarkably large number of bilateral areas (Figure 3B).

Overall, these results confirm that trained deep neural networks linearly map onto the brain (41; 26). However, the *sequence* of representations underlying these shared representations remains unknown.

2.5 The model-to-brain mapping reveals the unfolding of language representations over both time and space.

To characterize the unfolding of brain responses over both time and space, we perform the same analysis using source-localized MEG recordings. The resulting brain scores are consistent with – although less spatially precise than – the fMRI results (Figure 3C, average brain score between 0 and 2 s). For clarity, Figure 3D and Video 2 (SI.6.4) plot the *gain* in MEG scores: i.e. the difference of prediction performance between i) word and visual embeddings (green) and ii) the difference between compositional and word embedding (red). The brain scores of the visual embedding peak around 100 ms in V1 ($R = .008 \pm .002, p < 10^{-3}$), and rapidly propagate to higher-level areas (Figure 3D, Video 2, SI.6.4). The *gain* achieved by the word embedding can be observed in the left posterior fusiform gyrus around 200 ms

and peaks around 400 ms and in the left temporal and frontal cortices. Finally, the *gain* achieved by the compositional embedding is observed in a large number of bilateral brain regions, and peaks after $\approx$ 1s (Figure 3C and D).

After that period, brain areas outside the language network, such as area V1, appear to be better predicted by word and compositional embeddings than by visual ones (e.g between visual and word in V1: $\Delta R = .016 \pm .002$, $p < 10^{-10}$). These effects could thus reflect feedback activity (42) and explain why the corresponding fMRI responses are better accounted for by word and compositional embeddings than by visual ones.

Together with Supplementary Figure S5, these results show with unprecedented spatio-temporal precision, that the brain-mapping of our three representative embeddings automatically recovers the hierarchy of visual, lexical, and compositional representations of language in each cortical region.

3 Discussion

Do deep language models and the human brain process sentences in the same way? Following a recent methodology (34; 43; 44; 45; 46; 47; 48; 49; 31; 41; 26; 31), we address this issue by evaluating whether the activations of a large variety of deep language models linearly map onto those of 102 human brains. Each subject was recorded by Schoffelen and colleagues (3) with both MEG and fMRI, while they read isolated Dutch sentences composed of 9 to 15 words. Our study provides two main contributions.

Language performance is the primary factor that drives brain-mapping. First, not only do language transformers linearly map onto brain responses (24; 25; 26; 23; 27; 30; 32), but this property primarily depends on their language performance: i.e. whether these models accurately predict words from their context. Our analysis indicates that language performance is the most contributing factor explaining the variability of brain scores across models, spanning 32 architectures, two tasks, and 100 training steps (Figure 2D and H). Overall, deep language transformers thus appear to mainly converge to brain-like representations during their training.

Language transformers help decompose the cortical hierarchy of language in space and time. Second, our model comparison decomposes the visual, lexical, and compositional representations in the cortex. Whereas the areas involved in language processing are well known (50; 51; 52; 41; 53), the precise nature, format, and dynamics of their lexical and compositional representations remain largely unknown (54; 55; 52). Here, we track these hierarchical representations with unprecedented spatio-temporal precision. Early visual responses (<150 ms) are quasi-entirely accounted for by visual embeddings, and then transmitted to the posterior fusiform gyrus, which switches from visual to lexical representations around 200 ms (Video 2 and SI.6.4). This finding strengthens the claim that this area is responsible for orthographic and morphemic computations (51; 56; 57). Then, around 400 ms, lexical embeddings predict a large fronto-temporo-parietal network which peaks in the left temporal gyrus; these representations are then maintained for several seconds (15; 24; 26; 17). This result not only confirms the wide spread distribution of meaning in the brain (41), but also reveals its remarkably long-lasting nature.

Finally, compositional embeddings peak in the brain regions associated with high-level language processing such as the infero-frontal and anterior temporal cortices as well as the superior temporal cortex and the temporal-parietal junction (58; 52; 28). We confirm that these left-lateralized representations are significant in both hemispheres (59; 60). Critically, MEG suggests that these compositional effects become dominant and clearly bilateral long after word onset (>800 ms). This surprisingly late response may be due to the nature of the sentences, whose complex syntactic structure may slow down compositional computations.

Overall, our results suggest a convergence between the brain and language transformers. However, several factors qualify this conclusion.

Brain scores are limited by the signal-to-noise ratio. The mapping between these models and brain recordings is low. This phenomenon is expected: i) neuroimaging is notoriously noisy and ii) we analyze and model here single-sample responses of single-voxel/sensor. However, the resulting brain scores are i) highly significant (all $p < 10^{-9}$ on average across both all fMRI voxels and MEG sensors), and ii) in the same order of magnitude to our noise ceilings (Figure S4) as well as previous reports (e.g. (41), before correcting for the noise ceiling). Besides, we generally report brain scores *averaged* across *all* voxels or MEG channels, even though many brain areas do not strongly respond to language (S4). Yet, brain scores often reach $R > .10$ in the brain areas associated with language (Figure 3). Critically, the core of our study is the *link* between brain scores and language performance. This effect is very strong: the correlation between the language performance and brain scores is above R= .90 for MEG and R= .80 for fMRI (Figure S5).

Language performance is not the only variable modulating brain scores. Permutation feature importance shows that several factors such as the amount of training and the architecture significantly impact brain scores. This finding contributes to a growing list of variables that lead deep language models to behave more-or-less similarly to the brain. For example, Hale et al. (29) show that the amount and the type of corpus impact the ability of deep language parsers to linearly correlate with EEG responses. The present work complements this finding by evaluating the full set of activations of deep language models. It further demonstrates that the key ingredient to make a model more brain-like is, for now, to improve its language performance. This conclusion, however, should be qualified, because the brain scores of some layers of the very best models tend to ultimately decrease with language performance, especially in fMRI (Figure 2G). We speculate that this unexpected phenomenon rises because transformers may over-specialize in an objective that differs from the human brain's: predicting a word from its context, as opposed to generating the meaning of a sentence.

Structure of language is different from that of the brain. The training and the architecture of transformers (7) are in many ways *not* biologically plausible. On the one hand, the brain i) is a *recurrent* architecture ii) is trained on a relatively small amount of grounded sentences, and iii) presumably computes prediction errors at each level of the language hierarchy (61). On the other hand, transformers are i) *feedforward* neural networks, ii) trained on huge but strictly textual corpora (10), iii) can memorize and access a very large number of words, and iv) only minimize prediction errors at their final layer. Besides, language transformers are still far from human-level performance in a variety of tasks such as dialogue, summarization, and systematic generalization (62; 63). Thus, it is all-the-more remarkable to see that such algorithms partially map onto brain responses.

Input and output layers show a limited convergence. The input and output layers converge less than the middle layers (Figure S5). Why is there such a difference? We speculate that syntactic representations may drive the convergence of the middle layers. Indeed, unlike word embeddings, middle layers have been shown to encode syntactic trees (38) and co-references (39; 64). Our supplementary analyses support this possibility: middle layers best encode syntactic features and this information varies with language performance similar to brain scores (Figure S5). Studying the precise *nature* of the shared representations between brains and transformers is an exciting direction for future work.

Overall, this study provides evidence of shared language representations between the adult human brain and language transformers, suggesting a *partial* convergence between the two systems. This result is important for three reasons. First, it suggests that there is a limited number of – and perhaps unique – solutions to process language. Second, this convergence provides a concrete framework to understand the computational bases of language: deep language networks can be used as meaningful models of language processing only if they process language like our brain. Similarly, these models allow the community to move away from factorial design, and capitalize on the incremental properties of uncontrolled settings (65). Third, current language models remain relatively poor at general understanding, and zero-shot generalization (although see (10). Our results thus provide a stepping stone to unravel the cognitive operations specific to the human species, and, ultimately, implement them in machine learning algorithms.

4 Methods

We assess the similarity between (i) the activations of deep neural networks and (ii) those of the brain of 102 subjects, recorded with magneto-encephalography (MEG) and functional magnetic resonance imaging (fMRI) by Schoffelen et al. (3), when these two sets of systems are input with the same 400 isolated sentences.

4.1 Deep Neural Networks

4.1.1 Language Transformers

To model word and sentence representations, we trained a variety of transformers (7), input them with the same sentences that the subject read, and extracted the corresponding activations from each layer. We always extract activation in a "causal" way: for example, given the sentence 'THE CAT IS ON THE MAT', the brain response to 'ON' would be solely compared to the activations of the transformer input with 'THE CAT IS ON', and extracted from the 'ON' contextualized embeddings. Word embeddings and contextualized embeddings were generated for every word, by generating word sequences from the three previous sentences. We did not observe qualitatively different results when using shorter or longer contexts. It is to be noted that the sentences were isolated, and were not part of a narrative.

In total, we investigated 32 distinct architectures varying in their dimensionality ($\in [128, 256, 512]$), number of layers ($\in [4, 8, 12]$), attention heads ($\in [4, 8]$), and training task ("causal" language modeling and "masked" language modeling). While "causal" language transformers are trained to predict a word from its previous context, "masked" language transformers predict randomly masked words from a surrounding context. We froze the networks at ≈ 100

training stages (log distributed between 0 and 4,5M gradient updates, which corresponds to $\approx$ 35 passes over the full corpus), resulting in 3,600 networks in total, and 32,400 word representations (one per layer). The training was early-stopped when the networks' performance did not improve after 5 epochs on a validation set. Therefore, the number of frozen steps varied between 96 and 103 depending on the training length.

The algorithms were trained using XLM implementation [2] (9), on the same Wikipedia corpus of 278,386,651 words (in Dutch) extracted using WikiExtractor [3] and pre-processed using Moses tokenizer (66), with punctuation. We restricted the vocabulary to the 50,000 most frequent words, concatenated with all words used in the study (50,341 vocabulary words in total). These design choices enforce that the difference in brain scores observed across models cannot be explained by differences in corpora and text preprocessing.

To evaluate the language processing performance of the networks, we computed their performance (top-1 accuracy on word prediction given the context) using a test dataset of 180,883 words from Dutch Wikipedia.

4.1.2 Visual Convolutional Neural Network

To model visual representations, every word presented to the subjects was rendered on a gray 100 x 32 pixel background with a centered black Arial font, and input to a VGG network pretrained to recognize words from images (40), resulting in an 888-dimensional embedding. This embedding was used to replicate and extend previous work on the similarity between visual neural network activations and brain responses to the same images (e.g. (34; 45; 46)).

4.2 Neuroimaging

4.2.1 Protocol

For all the analyses, we used the open-source dataset released by Schoffelen and colleagues (3), gathering the functional magnetic resonance imaging (fMRI) and magneto-encephalography (MEG) recordings of 204 native Dutch speakers (100 males), aged from 18 to 33 years. Here, we focused on the 102 right-handed speakers who performed a *reading* task while being recorded by a CTF magneto-encephalography (MEG) and, in a separate session, with a SIEMENS Trio 3T Magnetic Resonance scanner (3).

Words (in Dutch) were flashed one at a time with a mean duration of 351 ms (ranging from 300 to 1400 ms), separated with a 300 ms blank screen, and grouped into sequences of 9 - 15 words, for a total of approximately 2,700 words per subject. Sequences were separated by a 5 s-long blank screen. We restricted our study to meaningful sentences (400 distinct sentences in total, 120 per subject). The exact syntactic structures of sentences varied across all sentences. Roughly, sentences were either composed of a main clause and a simple subordinate clause, or contained a relative clause. Twenty percent of the sentences were followed by a yes/no question (e.g. "Did grandma give a cookie to the girl?") to ensure that subjects were paying attention. Questions were not included in the dataset, and thus excluded from our analyses. Sentences were grouped into blocks of five sequences. This grouping was used for cross-validation to avoid information leakage between the train and test sets.

4.2.2 Magnetic Resonance Imaging (MRI)

Structural images were acquired with a T1-weighted magnetization-prepared rapid gradient-echo (MP-RAGE) pulse sequence. The full acquisition details, available in (3), are summarized here simplicity: TR=2,300 ms, TE=3.03 ms, 8 degree flip-angle, 1 slab, slice-matrix size=256×256, slice thickness=1 mm, field of view=256 mm, isotropic voxel-size=1.0×1.0×1.0 mm. Structural images were defaced by Schoffelen and colleagues. Preprocessing of the structural MRI was performed with Freesurfer (67), using the `recon-all` pipeline and a manual inspection of the cortical segmentations, realigned to 'fsaverage'. Region-of-interest analyses were selected from the PALS Brodmann's Area atlas (68) and the Destrieux atlas (69).

Functional images were acquired with a T2*-weighted functional echo-planar blood oxygenation level-dependent (EPI-BOLD) sequence. The full acquisition details, available in (3), are summarized here for simplicity: TR=2.0 seconds, TE=35ms, flip angle=90 degrees, anisotropic voxel size=3.5×3.5×3.0 mm extracted from 29 oblique slices. fMRI was preprocessed with fMRIPrep with default parameters (70). The resulting BOLD times series were detrended and de-confounded from 18 variables (the 6 estimated head-motion parameters ($\texttt{trans}_{x,y,z}$, $\texttt{rot}_{x,y,z}$) and the first 6 noise components calculated using anatomical CompCorr (71) and 6 DCT-basis regressors using nilearn's `clean_img`

[2] Each algorithm was trained each on 8 GPUs using early stopping with training perplexity criteria, 16 streams per batch, 128 words per stream, epoch size of 200 000 streams, 0.1 dropout, 0.1 attention dropout, gelu activation, inverse (sqrt) adam optimizer with learning rate 0.0001, 0.01 weight decay.

[3] https://github.com/attardi/wikiextractor

pipeline and otherwise default parameters (72). The resulting volumetric data lying along a 3mm "line" orthogonal to the mid-thickness surface were linearly projected to the corresponding vertices. The resulting surface projections were spatially decimated by 10, and are hereafter referred to as voxels, for simplicity. Finally, each group of 5 sentences was separately and linearly detrended. It is noteworthy that our cross-validation never splits such groups of five consecutive sentences between the train and test sets. Two subjects were excluded from the fMRI analyses because of difficulties in processing the metadata, resulting in 100 fMRI subjects.

4.2.3 Magneto-encephalography (MEG)

The MEG time series were preprocessed using MNE-Python and its default parameters except when specified (73). Signals were band-passed filtered between 0.1 and 40 Hz filtered, spatially corrected with a Maxwell Filter, clipped between the 0.01^{st} and 99.99^{th} percentiles, segmented between -500 ms to +2,000 ms relative to word onset and baseline-corrected before t=0. Reference channels and non-MEG channels were excluded from subsequent analyses, leading to 273 MEG channels per subject. We manually co-referenced (i) the skull segmentation of subjects' anatomical MRI with (ii) the head markers digitized before MEG acquisition. A single-layer forward model was generated with the Freesurfer-wrapper implemented in MNE-Python (73). Due to the lack of empty-room recordings, the noise covariance matrix used for the inverse operator was estimated from the zero-centered 200 ms of baseline MEG activity preceding word onset. Subjects' source space inverse operators were computed using a dSPRM. The average brain responses displayed in Figure 1D were computed as the square of the average evoked related field across all words for each subject separately, averaged across subjects, and finally divided by their respective maxima, to highlight temporal differences. Video 1 displays the average sources without normalization (SI.6.4). Seven subjects were excluded from the MEG analyses because of difficulties in processing the metadata, resulting in 92 usable MEG recordings.

4.3 Noise Ceiling: Brain $\rightarrow$ Brain mapping

To estimate the amount of explainable signal in each MEG and fMRI recording, we trained and evaluated, through cross-validation, a linear mapping model W to predict the brain responses of a given subject to each sentence Y from the aggregated brain responses of all other subjects who read the same sentence X. Specifically, five cross-validation splits were implemented across 5-sentence blocks with scikit-learn 'GroupKFold' (36). For each word of each sentence i, all but one subject who read the corresponding sentence were averaged with one another to form a template brain response: $x_i \in \mathbb{R}^n$ with n the number of MEG channels or fMRI voxels, as well as a target brain response $y_i \in \mathbb{R}^n$ corresponding to the remaining subject. X and Y were normalized (mean=0, std=1) across sentences for each spatio-temporal dimension, using a robust scaler clipping below and above the 0.01^{st} and 99.99^{th} percentiles, respectively. A linear mapping $W \in \mathbb{R}^{n \times n}$ was then fit with a ridge regression to best predict Y from X on the train set:

$$W = (X_{\text{train}}^T X_{\text{train}} + \lambda I)^{-1} X_{\text{train}}^T Y_{\text{train}} \tag{1}$$

with λ the $l2$ regularization parameter, chosen amongst 20 values log-spaced between 10^{-3} and 10^8 with nested leave-one-out cross-validation for each dimension separately (as implemented in (36)). Brain predictions $\hat{Y} = WX$ were evaluated with a Pearson correlation on the test set:

$$R = \text{Corr}(Y_{\text{test}}, \hat{Y}_{\text{test}}) \tag{2}$$

For the MEG source noise estimate, the correlation was also performed after source projection:

$$R = \text{Corr}(KY_{\text{test}}, K\hat{Y}_{\text{test}}) \tag{3}$$

with $K \in \mathbb{R}^{n \times m}$ the inverse operator projecting the n MEG sensors onto m sources. Correlation scores were finally averaged across cross-validation splits for each subject, resulting in one correlation score ('brain score') per voxel (or per MEG sensor/time sample) per subject.

4.4 Brain score and similarity: Network $\rightarrow$ Brain mapping

To estimate the functional similarity between each artificial neural network and each brain, we followed the same analytical pipeline used for noise ceiling, but replaced X with the activations of the deep learning models. Specifically, using the same cross-validation, and for each subject separately, we trained a linear mapping $W \in \mathbb{R}^{o,n}$ with o the number of activations, to predict brain responses Y from the network activations X. X was normalized across words (mean=0, std=1).

To account for the hemodynamic delay between word onset and the BOLD response recorded in fMRI, we used a finite impulse response (FIR) model with five delays (from 2 to 10 seconds) to build X^* from X. W was found using the

same ridge regression described above, and evaluated with the same correlation scoring procedure. The resulting brain correlation scores measure the linear relationship between the brain signals of one subject (measured either by MEG or fMRI) and the activations of one artificial neural network (e.g a word embedding). For MEG, we simply fit and evaluated the model activations X at each time sample independently.

In principle, one may orthogonalize low-level representations (e.g. visual features) from high-level network models (e.g. language model), to separate the specific contribution of each type of model. This is because middle layers have access to the word-embedding layer, and can, in principle, simply copy some of its activations. Similarly, word embedding can implicitly contain visual information: e.g. frequent words tend to be visually smaller than rare ones. In our case, however, the middle layers of transformers were much better than word embeddings, which were much better than visual embeddings. To quantify the gain ΔR achieved by a higher-level model M_1 (e.g. the middle layers of a transformer) and a lower level model M_2 (e.g. a word embedding) we thus simply compared the difference of their encoding scores:

$$\Delta R_{M_1} = R_{M_1} - R_{M_2} \tag{4}$$

4.4.1 Convergence analysis

All neural networks but the visual CNN were trained from scratch on the same corpus (cf. section 4.1.1). We systematically computed the brain scores of their activations on each subject, sensor (and time sample in the case of MEG) independently. For computational reasons, we restricted model comparison on MEG encoding scores to ten time samples regularly distributed between $[0, 2]$s. Brain scores were then averaged across spatial dimensions (i.e. MEG channels or fMRI surface voxels), time samples, and subjects to obtain the results in Figure 2. To evaluate the convergence of a model, we computed, for each subject separately, the correlation between (1) the average brain score of each network and (2) its performance or its training step (Figure 2 and S5). Positive and negative correlations indicate convergence and divergence, respectively. Brain scores above 0 before training indicate a fortuitous relationship between the activations of the brain and those of the networks.

4.4.2 Permutation feature importance

To systematically quantify how the architecture, language accuracy, and training of the language transformers impacted their ability to linearly map onto brain activity, we fitted, for each subject separately, a Random Forest across the models' properties to predict their brain scores, using scikit-learn's `RandomForest` (35; 36). Specifically, we input the following features to the random forest: the training task (causal language modeling "CLM" vs. masked language modeling "MLM"), the number of attention heads $\in [4, 8]$, the total number of layers $\in [4, 8, 12]$, dimensionality $\in [128, 256, 512]$, training step (number of gradient updates, $\in [0, 4.5M]$), language modeling accuracy (top-1 accuracy at predicting a masked word) and the relative position of the representation (a.k.a 'layer position', between 0 for the word-embedding layer, and 1 for the last layer). The performance of the Random Forest was evaluated for each subject separately with a Pearson correlation R using five-split cross-validation across models.

"Permutation feature importance" summarizes how each of the covarying properties of the models (their task, architecture, etc.) specifically impacts the brain scores (35). Permutation feature importance was implemented with scikit-learn (36) and is summarized with ΔR: the decrease in R when shuffling one feature (using 50 repetitions). For each subject, we reported the average decrease across the cross-validation splits (Figure 2). The resulting scores (ΔR) are expected to be centered around 0 if the corresponding feature does not impact the brain scores , and positive otherwise.

4.5 Population statistics

To estimate the robustness of our results, we systematically performed second-level analyses across subjects. Specifically, we applied Wilcoxon signed-rank tests across subjects' estimates to evaluate whether the effect under consideration was systematically different from the chance level. The p-values of individual voxel/source/time samples were corrected for multiple comparisons, using a False Discovery Rate (Benjamini/Hochberg) as implemented in MNE-Python ((73)). Error bars and $\pm$ refer to the standard error of the mean (SEM) interval across subjects.

4.6 Brain parcellation

In Section 2.2, Section 2.5 and Figure 3, we focus on particular regions of interest using the Brodmann's areas from the PALS parcellation of freesurfer[4]. The superior temporal gyrus (BA22) is split into its anterior, middle and posterior parts to increase granularity. For clarity, we rename certain areas as specified in the table below.

[4]https://surfer.nmr.mgh.harvard.edu/fswiki/PALS_B12

Label	Corresponding Brodmann's areas
V1	BA17
Fusiform	BA37
Angular	BA39
aSTG	BA22-anterior
mSTG	BA22-middle
pSTG	BA22-posterior
Supramarginal	BA40
Infero-frontal	BA44 / BA45 / BA47
Fronto-polar	BA10
Temporo-polar	BA38

4.7 Ethics

This study was conducted in compliance with the Helsinki Declaration. No experiments on living beings were performed for this study. These data were provided (in part) by the Donders Institute for Brain, Cognition, and Behaviour after having been approved by the local ethics committee (CMO – the local "Committee on Research Involving Human Subjects" in the Arnhem-Nijmegen region).

5 Acknowledgement

This work was supported by ANR-17-EURE-0017, the Fyssen Foundation, and the Bettencourt Foundation to JRK for his work at PSL.

References

[1] Ella Bingham and Heikki Mannila. Random projection in dimensionality reduction: applications to image and text data. In *Proceedings of the seventh ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 245–250, 2001.

[2] Jonathan Frankle and Michael Carbin. The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*, 2018.

[3] Jan-Mathijs Schoffelen, Robert Oostenveld, Nietzsche Lam, Julia Udden, Annika Hultén, and Peter Hagoort. A 204-subject multimodal neuroimaging dataset to study language processing. *Scientific Data*, 6, 12 2019.

[4] Alan M Turing. Computing machinery and intelligence. In *Parsing the Turing Test*, pages 23–65. Springer, 2009.

[5] Noam Chomsky. *Language and mind*. Cambridge University Press, 2006.

[6] Stanislas Dehaene, LE Yann, and Jacques Girardon. *La plus belle histoire de l'intelligence: des origines aux neurones artificiels: vers une nouvelle étape de l'évolution*. Robert Laffont, 2018.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, 2017.

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018.

[9] Guillaume Lample and Alexis Conneau. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[10] Tom B Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *arXiv preprint arXiv:2005.14165*, 2020.

[11] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 932–938. MIT Press, 2003.

[12] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. 2013.

[13] Jeffrey Pennington, Richard Socher, and Christopher D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014.

[14] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

[15] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just. Predicting human brain activity associated with the meanings of nouns. 320(5880):1191–1195, 2008.

[16] Andrew James Anderson, Edmund C. Lalor, Feng Lin, Jeffrey R. Binder, Leonardo Fernandino, Colin J. Humphries, Lisa L. Conant, Rajeev D. S. Raizada, Scott Grimm, and Xixi Wang. Multiple regions of a cortical network commonly encode the meaning of words in multiple grammatical positions of read sentences. 29(6):2396–2411, 2019. Publisher: Oxford Academic.

[17] Jona Sassenhagen and Christian J. Fiebach. Traces of meaning itself: Encoding distributional word vectors in brain activity. *bioRxiv*, 2019.

[18] Subba Reddy Oota, Naresh Manwani, and Bapi Raju S. fMRI Semantic Category Decoding using Linguistic Encoding of Word Embeddings. *arXiv e-prints*, page arXiv:1806.05177, Jun 2018.

[19] Samira Abnar, Rasyan Ahmed, Max Mijnheer, and Willem H. Zuidema. Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity. *CoRR*, abs/1711.09285, 2017.

[20] Yu-Ping Ruan, Zhen-Hua Ling, and Yu Hu. Exploring semantic representation in brain activity using word embeddings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 669–679, Austin, Texas, November 2016. Association for Computational Linguistics.

[21] Christian Brodbeck, L Elliot Hong, and Jonathan Z Simon. Rapid transformation from auditory to linguistic representations of continuous speech. *Current Biology*, 28(24):3976–3983, 2018.

[22] Jon Gauthier and Anna Ivanova. Does the brain represent words? an evaluation of brain decoding studies of language understanding. *arXiv preprint arXiv:1806.00591*, 2018.

[23] Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 233–243, Doha, Qatar, October 2014. Association for Computational Linguistics.

[24] Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems 31*, pages 6628–6637. Curran Associates, Inc., 2018.

[25] Nikos Athanasiou, Elias Iosif, and Alexandros Potamianos. Neural activation semantic models: Computational lexical semantic models of localized neural activations. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2867–2878, Santa Fe, New Mexico, USA, August 2018. Association for Computational Linguistics.

[26] Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *CoRR*, abs/1905.11833, 2019.

[27] Micha Heilbron, Kristijan Armeni, Jan-Mathijs Schoffelen, Peter Hagoort, and Floris P de Lange. A hierarchy of linguistic predictions during natural language comprehension. *bioRxiv*, 2020.

[28] Jonathan R Brennan and Liina Pylkkänen. Meg evidence for incremental sentence composition in the anterior temporal lobe. *Cognitive science*, 41:1515–1531, 2017.

[29] John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan R Brennan. Finding syntax in human encephalography with beam search. *arXiv preprint arXiv:1806.04127*, 2018.

[30] Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, et al. Thinking ahead: prediction in context as a keystone of language in humans and machines. *bioRxiv*, 2020.

[31] Alexander Kell, Daniel Yamins, Erica Shook, Sam Norman-Haignere, and Josh McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98, 04 2018.

[32] 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*, 2020.

[33] Juliette Millet and Jean-Rémi King. Inductive biases, pretraining and fine-tuning jointly account for brain responses to speech. February 2021.

[34] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, 111(23):8619–8624, 2014.

[35] Leo Breiman. Random forests. 45(1):5–32, 2001.

[36] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct):2825–2830, 2011.

[37] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates, Inc., 2013.

[38] Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. Emergent linguistic structure in artificial neural networks trained by self-supervision. page 201907367, 2020.

[39] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 2019.

[40] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoo Yun, Seong Joon Oh, and Hwalsuk Lee. What is wrong with scene text recognition model comparisons? dataset and model analysis. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4715–4723, 2019.

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

[42] Anna Seydell-Greenwald, Xiaoying Wang, Elissa Newport, Yanchao Bi, and Ella Striem-Amit. Spoken language comprehension activates the primary visual cortex. *bioRxiv*, 2020.

[43] Hanlin Tang, Martin Schrimpf, William Lotter, Charlotte Moerman, Ana Paredes, Josue Ortega Caro, Walter Hardesty, David Cox, and Gabriel Kreiman. Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences*, 115(35):8835–8840, 2018.

[44] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

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

[46] Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.

[47] Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194, 2017.

[48] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356, 2016.

[49] Andrew Saxe, Stephanie Nelli, and Christopher Summerfield. If deep learning is the answer, what is the question? *Nature Reviews Neuroscience*, pages 1–13, 2020.

[50] Cathy J Price. The anatomy of language: a review of 100 fmri studies published in 2009. *Annals of the new York Academy of Sciences*, 1191(1):62–88, 2010.

[51] Stanislas Dehaene and Laurent Cohen. The unique role of the visual word form area in reading. *Trends in cognitive sciences*, 15(6):254–262, 2011.

[52] Gregory Hickok and David Poeppel. The cortical organization of speech processing. 8(5):393–402, 2007. Number: 5 Publisher: Nature Publishing Group.

[53] Aakash Agrawal, KVS Hari, and SP Arun. A compositional neural code in high-level visual cortex can explain jumbled word reading. *Elife*, 9:e54846, 2020.

[54] Evelina Fedorenko, Idan Blank, Matthew Siegelman, and Zachary Mineroff. Lack of selectivity for syntax relative to word meanings throughout the language network. *bioRxiv*, page 477851, 2020.

[55] Stanislas Dehaene, Laurent Cohen, Mariano Sigman, and Fabien Vinckier. The neural code for written words: a proposal. 9(7):335–341.

[56] Dora Hermes, Vinitha Rangarajan, Brett L Foster, Jean-Remi King, Itir Kasikci, Kai J Miller, and Josef Parvizi. Electrophysiological responses in the ventral temporal cortex during reading of numerals and calculation. *Cerebral cortex*, 27(1):567–575, 2017.

[57] Oscar Woolnough, Cristian Donos, Patrick S Rollo, Kiefer James Forseth, Yair Lakretz, Nathan E Crone, Simon Fischer-Baum, Stanislas Dehaene, and Nitin Tandon. Spatiotemporal dynamics of orthographic and lexical processing in the ventral visual pathway. *bioRxiv*, 2020.

[58] Christophe Pallier, Anne-Dominique Devauchelle, and Stanislas Dehaene. Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, 108(6):2522–2527, 2011.

[59] Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, Susan Whitfield-Gabrieli, and Nancy Kanwisher. New method for fmri investigations of language: defining rois functionally in individual subjects. *Journal of neurophysiology*, 104(2):1177–1194, 2010.

[60] Gregory B Cogan, Thomas Thesen, Chad Carlson, Werner Doyle, Orrin Devinsky, and Bijan Pesaran. Sensory–motor transformations for speech occur bilaterally. *Nature*, 507(7490):94–98, 2014.

[61] Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138, 2010.

[62] João Loula, Marco Baroni, and Brenden M. Lake. Rearranging the familiar: Testing compositional generalization in recurrent networks. *CoRR*, abs/1807.07545, 2018.

[63] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.

[64] Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg. Visualizing and measuring the geometry of BERT.

[65] Liberty S Hamilton and Alexander G Huth. The revolution will not be controlled: natural stimuli in speech neuroscience. *Language, Cognition and Neuroscience*, 35(5):573–582, 2020.

[66] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180. Association for Computational Linguistics.

[67] Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

[68] David C Van Essen. A population-average, landmark-and surface-based (pals) atlas of human cerebral cortex. *Neuroimage*, 28(3):635–662, 2005.

[69] 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, 2010.

[70] Oscar Esteban, Christopher J Markiewicz, Ross W Blair, Craig A Moodie, A Ilkay Isik, Asier Erramuzpe, James D Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, et al. fmriprep: a robust preprocessing pipeline for functional mri. *Nature methods*, 16(1):111–116, 2019.

[71] Yashar Behzadi, Khaled Restom, Joy Liau, and Thomas T Liu. A component based noise correction method (compcor) for bold and perfusion based fmri. *Neuroimage*, 37(1):90–101, 2007.

[72] Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi, Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Machine learning for neuroimaging with scikit-learn. *Frontiers in neuroinformatics*, 8:14, 2014.

[73] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Lauri Parkkonen, and Matti S. Hämäläinen. Mne software for processing meg and eeg data. *NeuroImage*, 86:446 – 460, 2014.

[74] Peter Hagoort. The neurobiology of language beyond single-word processing. *Science*, 366(6461):55–58, 2019.

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

[76] Matthew J. Nelson, Imen El Karoui, Kristof Giber, Xiaofang Yang, Laurent Cohen, Hilda Koopman, Sydney S. Cash, Lionel Naccache, John T. Hale, Christophe Pallier, and Stanislas Dehaene. Neurophysiological dynamics of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*, 114(18):E3669–E3678, May 2017. Publisher: National Academy of Sciences Section: PNAS Plus.

6 Supplementary information

6.1 Anatomical and temporal characteristics of average brain responses to reading.

When and where do textual sentences elicit brain activity? As expected (54; 51; 74; 52), average fMRI and MEG responses to written words peak in a distributed and bilateral cortical network, including the primary visual cortex, the left fusiform gyrus, the supra-marginal, and the superior temporal cortices, as well as the motor, premotor and infero-frontal areas (Figure S4A). MEG source reconstruction, based on structural MRI and minimum norm estimates, further clarifies the dynamics of this cortical network: on average, word onset elicits a series of brain responses originating in V1 around ≈100 ms and continuing within the left posterior fusiform gyrus around 200 ms, the superior and middle temporal gyri, as well as the pre-motor and infero-frontal cortices between 150 and 500 ms after word onset (Figure S4A, Video 1).

6.2 Noise ceilings.

To compare the brain responses to the activations of deep language models, it is important to estimate the level of signal-to-noise ratio (Figure S4). For this, we fit, for each subject separately, a "noise-ceiling" model across subjects: for each recording of each subject and each sentence Y_{train}, we fit a linear model W from the recordings of all other subjects who read the same sentence X_{train} to predict each voxel and each MEG sensor at each time sample, separately. Using a cross-validation scheme across sentences, we then evaluate the Pearson correlation R between (1) the true brain responses of subject Y_{test} and (2) the predicted brain responses $\hat{Y}_{\text{test}} = W \cdot X_{\text{test}}$ for each voxel and each MEG sensor separately. This procedure can be thought of as approximating an optimal black box: i.e. evaluating a one-hot encoder of brain responses is trained and evaluated on each element of a unique sentence. Noise ceiling peaks within the expected language network (75) (Figure 1 F-H). These estimates are relatively low: for example, fMRI noise ceilings reach, on average, $R = 0.129$ (± 0.004 SEM across subjects) in the superior temporal gyrus, whereas MEG noise ceilings peak at $R = 0.069 \pm 0.001$.

Fronto-polar cortex:	0.054 ± 0.003	$p < 10^{-8}$
Fusiform:	0.120 ± 0.004	$p < 10^{-8}$
Infero-frontal:	0.139 ± 0.005	$p < 10^{-8}$
M1:	0.042 ± 0.003	$p < 10^{-8}$
STG:	0.129 ± 0.004	$p < 10^{-8}$
Supramarginal:	0.078 ± 0.003	$p < 10^{-8}$
V1:	0.150 ± 0.006	$p < 10^{-8}$

Supplementary Table 1. Average noise ceiling within each region-of-interest. Mean, standard error of the mean and p-values across subjects.

6.3 Probe analysis of the language transformer.

Middle layers better map onto brain responses than input and output layers. Why is there such a difference between layers? To tackle the question, we measure the level to which the 32,400 transformer embeddings linearly predict two types of linguistic features: part-of-speech (i.e a lexical feature), and the number of open and pending nodes (i.e compositional syntactic features (76)). More precisely, we fit and evaluate an ℓ_2-penalized linear model to predict each of these features given the transformer's embedding and plot this decoding performance as a function of the language performance of the model (Figure S6). While the word embedding and middle layers similarly predict word-level features (word length and part-of-speech of the word), the two high-level syntactic features (number of open and pending nodes) are better predicted by the middle layers of transformers. Finally, the decoding performance of the two syntactic features varies with the layer and the performance, in a manner strikingly similar to the brain score. These analyses suggest that middle layers are more "brain-like" than extremity layers because they learn to encode abstract linguistic properties like syntax.

6.4 Video materials

Below the captions for the two videos provided in supplementary.

Video 1. Anatomical and temporal hierarchy of reading. Average brain responses elicited by the onset of visual words (≈2,700 words were presented to each of the 95 subjects), as estimated with minimum source estimates (MNE) of the single-trial responses constrained by the individual subjects' anatomy (cortical surface extracted from T1 scans).

These results correspond to the ones summarized in Figure S4A. Overall, these results confirm that we can track the sequential recruitment of the cortical hierarchy of reading, starting from early visual cortex, moving up through the expected location of the visual word form area, and then igniting the temporal, prefrontal and parietal areas typically associated with language processing. Although these effects are bilateral, the typical left-lateralization associated with language processing can be observed.

Video 2. The main levels of the hierarchy of language revealed by deep neural networks. Single-trial encoding scores obtained for three representative embeddings reveal the types of representations that are generated within each region and at each time instant. Blue, green and red colors indicates when and where brain responses to words are specifically predicted by visual, word and compositional embeddings, respectively (a.k.a gain in brain scores). The animated legend illustrates the same data without the anatomy: each dot corresponds to a brain source, radius corresponds to effect size (center: no effect, circle: maximum effect), and angle corresponds to the type of representation (visual, lexical or compositional). Overall, these results show when and where the brain transforms visual representations into lexical and compositional representations.

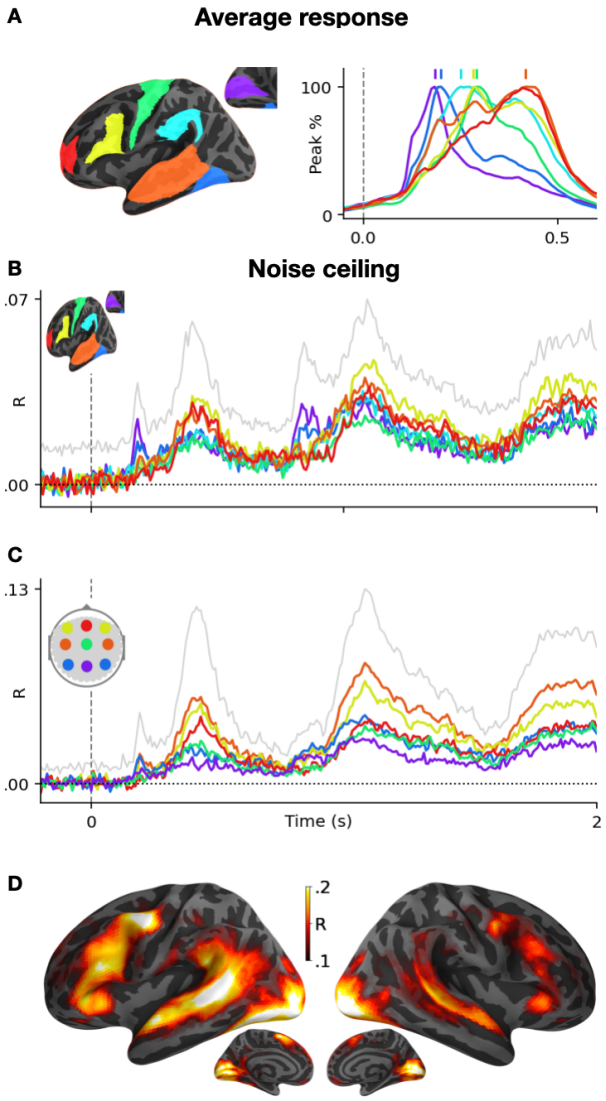

Figure S4: **Noise ceilings.** **A**. Grand average MEG source estimates to word onset (t=0) for 7 regions typically associated with reading (V1: purple, M1: green, fusiform gyrus: dark blue, supramarginal gyrus: light blue, superior temporal gyrus: orange, infero-frontal gyrus: yellow and fronto-polar gyrus: red), normalized to their peak response. Vertical bars indicate the peak time of each region. The full (not normalized) data is displayed in Video 1. **B**. MEG noise ceilings, approximated by predicting brain responses of a given subject from those of all other subjects. Colored lines depict the mean noise ceiling in each region of interest. The grey line depicts the best noise ceiling across sources. **C**. Same as (D) in sensor space. **D**. Noise ceiling estimates of fMRI recordings.

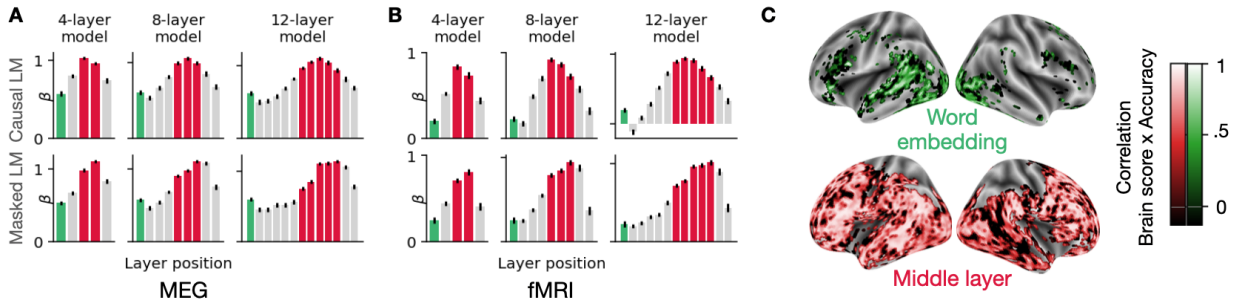

Figure S5: **Correlation between the network's performance and brain score. A-B.** Standardized beta coefficients between the language modeling performance of the network and its MEG (A) or fMRI (B) scores. For each subject, the brain scores are first scaled (0-mean, 1-std). Then, a linear regression is fit to predict the brain score (averaged across channels and time for MEG, across voxels for fMRI) of each layer of 100 networks (all 512-dimensional, with 12 layers and 8 heads) given their language performance (top-1 accuracy). The beta coefficients of the language performance are reported (y-axis). Results are consistent across 4-, 8-, and 12-layer transformers, trained on a causal (top) or masked (bottom) language modeling task. Error bars are the standard error of the mean beta coefficients across subjects. **C.** Pearson correlation between the performance of the 100 transformers (all 512-dimensional, with 12 layers and 8 heads) and the brain score of their word embedding (top) and ninth layer (bottom), for each voxel. Correlation scores are computed for each (subject, voxel) pair, then averaged across subjects. Only significant voxels are displayed, as assessed with a two-sided Wilcoxon test across subjects and corrected for multiple comparison using false discovery rate across voxels (threshold: .001).

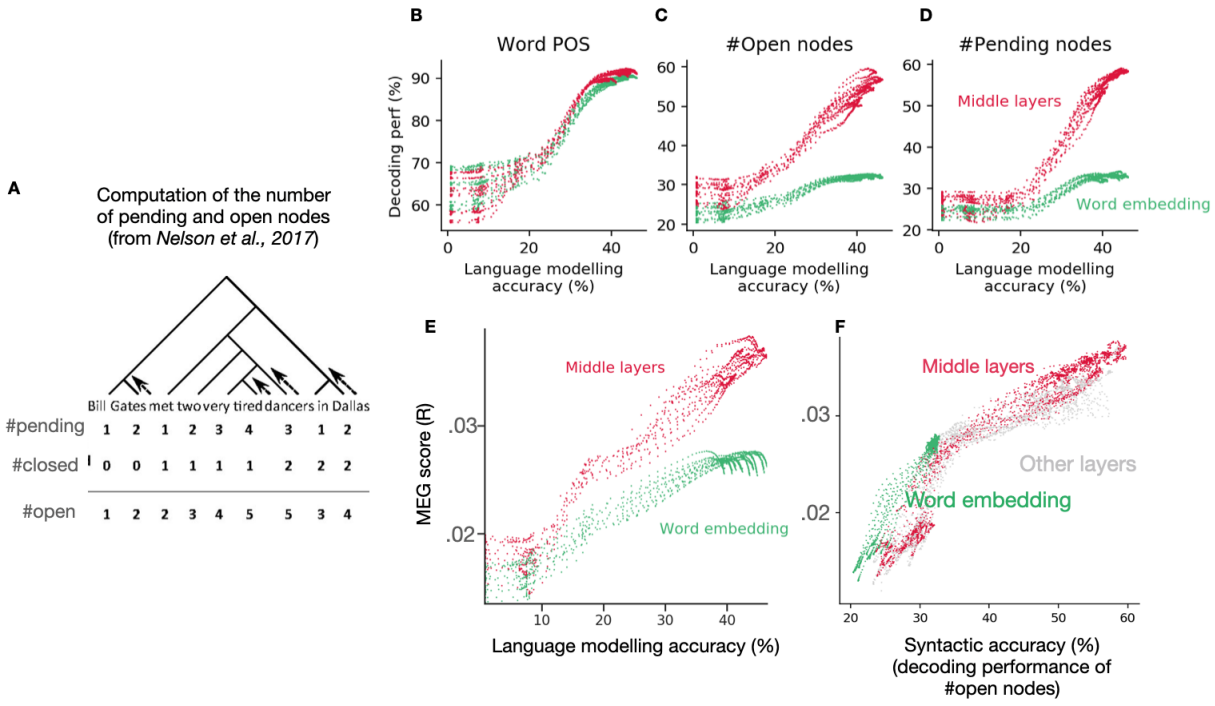

Figure S6: **What linguistic information drives the brain score? A.** From the stimulus, we compute three linguistic features: the part-of-speech of the words (i) (as given by Spacy), and two higher-level syntactic features: the number of pending nodes (ii) and open nodes (iii). These two syntactic features are derived from the constituency trees of the sentences, following (76). **B-D**. A $\ell 2$-penalized linear regression is fit to predict the three linguistic features from the word embeddings (green), and middle layers (red) of the causal models studied in Figure 2B. The decoding performance is reported on the y-axis (accuracy at predicting the part-of-speech for B, r-squared for C, D and E). **E.** MEG scores (averaged across sensors and time) of the embeddings given their language modeling performance (top-1 accuracy at predicting the next word, Figure 2B). **F.** MEG scores of the embeddings given their ability to predict the number of open nodes.

Task	Dim	#Layers	#Heads	Best perplexity	Best accuracy
mlm	512	12	8	4.75	67.25
mlm	512	12	4	4.86	66.71
mlm	512	8	8	5.10	65.86
mlm	512	8	4	5.10	66.03
mlm	512	4	8	5.62	64.21
mlm	512	4	4	5.90	63.61
mlm	256	12	8	6.21	63.08
mlm	256	12	4	6.26	63.00
mlm	256	8	8	6.68	62.03
mlm	256	8	4	6.82	61.53
mlm	256	4	8	7.83	59.51
mlm	256	4	4	8.11	58.95
mlm	128	12	8	9.32	57.15
mlm	128	12	4	9.95	56.35
mlm	128	8	8	10.25	55.99
mlm	128	8	4	10.48	55.76
mlm	128	4	8	12.17	53.58
mlm	128	4	4	12.96	52.82
clm	512	12	8	15.16	46.30
clm	512	12	4	15.23	46.21
clm	512	8	8	15.78	45.78
clm	512	8	4	15.74	45.72
clm	512	4	8	16.97	44.77
clm	512	4	4	17.09	44.64
clm	256	12	8	17.95	44.17
clm	256	12	4	18.20	44.12
clm	256	8	8	19.09	43.47
clm	256	8	4	19.06	43.51
clm	256	4	8	20.93	42.32
clm	256	4	4	20.98	42.37
clm	128	12	8	23.79	41.15
clm	128	12	4	23.72	41.22
clm	128	8	8	24.87	40.62
clm	128	8	4	24.92	40.64
clm	128	4	8	27.33	39.49
clm	128	4	4	27.69	39.48

Figure S7: **Performance of the 32 transformer architectures.** Best perplexity (the lower the better) and top-1 accuracy (the higher the better) of 32 transformer architectures, evaluated on a test test of 180K words from Wikipedia. Transformers are trained with a masked ('mlm') or causal ('clm') language modeling objective. They vary in their dimensionality ("Dim"), number of layers ('Layers') and number of attention heads ('Heads'). The models are trained on a set of 280K words from Wikipedia (in Dutch). The training is stopped when the perplexity on a validation set does not decrease for 5 epochs.