



HAL
open science

Brains and algorithms partially converge in natural language processing

Charlotte Caucheteux, Jean-Rémi King

► **To cite this version:**

Charlotte Caucheteux, Jean-Rémi King. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 2022, 10.1038/s42003-022-03036-1 . hal-03361439

HAL Id: hal-03361439

<https://hal.science/hal-03361439v1>

Submitted on 1 Oct 2021

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

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

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

15 **Keywords** Natural Language Processing | Encoding | Functional Magnetic Resonance Imaging | Magneto-
16 encephalography

17 **1 Introduction**

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

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

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

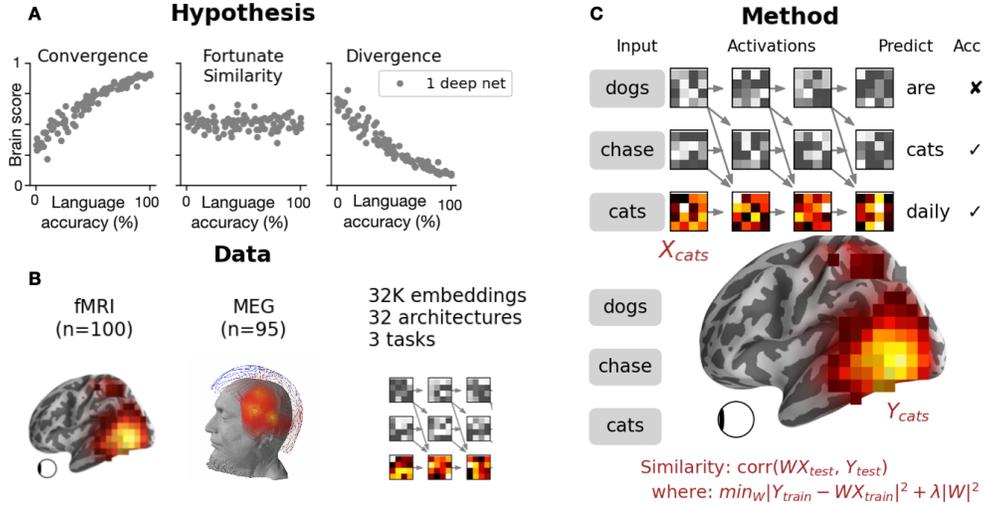


Figure 1: **Hypotheses and Methods.** **A.** The three panels represent three hypotheses on the link between language transformers and the brain. Each dot represents one hypothetical language transformer. Language transformers can be considered to converge to brain-like computations if their language 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.

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

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

38 To address these issues, we analyze the brain responses of 102 healthy adults, scanned with both functional magnetic
 39 resonance imaging (fMRI) and source-localized magneto-encephalography (MEG) by Schoffelen et al (3) during
 40 two 1 h-long sessions, during which they read isolated Dutch sentences composed of 9 to 15 words. We then extract
 41 32,400 embeddings from a variety of language transformers to compare their ability to linearly map onto these brain
 42 responses (‘brain score’ (32)). Finally, we assess how differences in training, architectures, and language performance
 43 independently contribute to these brain mappings (Figure 1).

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

46 2 Results

47 2.1 Intermediate layers predict brain responses best.

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

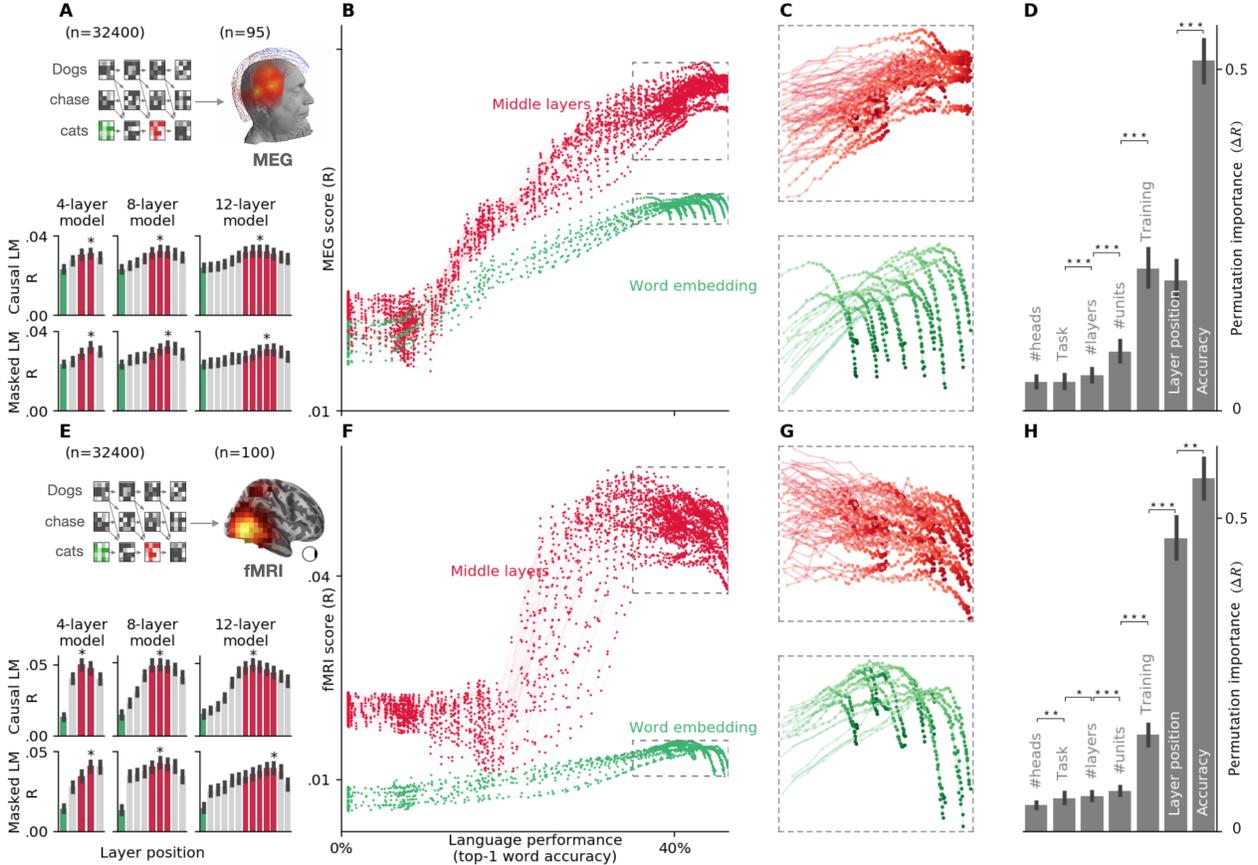


Figure 2: Language transformers tend to converge towards brain-like representations. **A.** Bar plots display the average MEG score (across time and channels) of six representative transformers varying in tasks (causal vs masked language modeling) and depth (4-12 layers). The green and red bars correspond 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.

56 We evaluate the brain scores of 32 transformer architectures (varying from 4 to 12 layers, each ranging from 128 to
 57 512 dimensions, and each benefiting from 4 to 8 attention heads), trained on the same Wikipedia dataset either with a
 58 ‘causal’ language modeling (CLM) or a ‘masked’ language modeling task (MLM). For each architecture, we input the
 59 model with the sentence read by the subjects in the MEG or fMRI scanner, extract the activations from every layer, and,
 60 finally, compute the corresponding fMRI and MEG scores.

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

66 Overall, the brain scores vary as a function of the relative depth of the embedding within the transformer. Specifically,
 67 both MEG and fMRI scores follow an inverted U-shaped pattern across layers for all architectures (Figure 2A and E):
 68 middle layers¹ systematically outperform output (fMRI: $\Delta R = .011 \pm .001$, $p < 10^{-18}$, MEG: $\Delta R = .003 \pm .0005$,

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

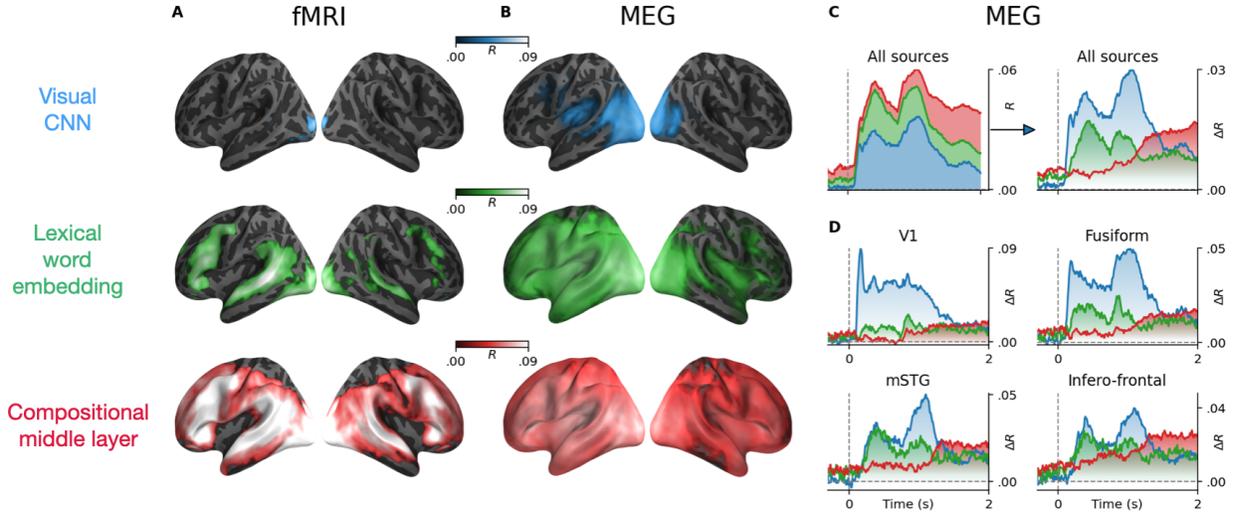


Figure 3: **The similarity between language transformers and the brain reveals the spatio-temporal hierarchy of language representations.** Lexical and compositional representations can be isolated from the word embedding (green) and one middle layer (red) of a typical language transformer (here, a 12-layer 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.

69 $p < 10^{-13}$) and input layers (fMRI: $\Delta R = .031 \pm .001$, $p < 10^{-18}$, MEG: $\Delta R = .009 \pm .001$, $p < 10^{-17}$). This
 70 result confirms that the intermediary representations of deep language transformers are more "brain-like" than input and
 71 output layers (26).

72 2.2 The emergence of brain-like representations predominantly depends on the network's ability to predict 73 missing words.

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

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

84 Second, brain scores strongly correlate with language accuracy in both MEG ($R = .77$ Pearson's correlation on average
 85 $\pm .01$ across subjects) and fMRI ($R = .57 \pm .02$, Figure 2B and C). The correlation is higher for middle (for fMRI:
 86 $R = .81 \pm .02$ and MEG: $R = .91 \pm .01$) than input ($R = .39 \pm .03$) and output layers ($R = .63 \pm .03$). Beta
 87 coefficients for each particular layer and architecture are displayed in Figure S5A and B. Furthermore, single-voxel
 88 analyses show that this correlation between brain score and language performance is driven mainly by the superior
 89 temporal sulcus and gyrus for the embedding layer (mean $R = .52 \pm .06$) and is widespread for the middle layers,
 90 exceeding $R = .85$ correlation in the superior temporal sulcus, infero-frontal, fusiform and angular gyri (Figure S5C).

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

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

97 2.3 Architectural and training factors impact brain scores too.

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

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

111 2.4 The mapping between the brain and language models helps to automatically decompose the cortical 112 hierarchy of language.

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

123 In fMRI, the brain scores of the visual embedding peak in the early visual cortex (V1) (mean brain scores across
124 voxels: $R = .022 \pm .003$, $p < 10^{-11}$). By contrast, the brain scores of lexical embedding peak in the left superior
125 temporal gyrus ($R = .052 \pm .004$, $p < 10^{-13}$) as well as in the inferior temporal cortex and middle frontal gyrus
126 ($R = .053 \pm .003$, $p < 10^{-15}$) and are significant across the entire language and reading network (Figure 3B). Finally,
127 the brain scores of the compositional embedding are significantly higher than those of lexical of embeddings in the
128 superior temporal gyrus ($\Delta R = .012 \pm .001$, $p < 10^{-16}$), the angular gyrus ($\Delta R = .010 \pm .001$, $p < 10^{-16}$), the
129 infero-frontal cortex ($\Delta R = .016 \pm .001$, $p < 10^{-16}$) and the dorsolateral prefrontal cortex ($\Delta R = .012 \pm .001$,
130 $p < 10^{-13}$). While these effects are lateralized (left hemisphere versus right hemisphere: $\Delta R = .010 \pm .001$,
131 $p < 10^{-14}$), they are significant across a remarkably large number of bilateral areas (Figure 3B).

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

134 2.5 The model-to-brain mapping reveals the unfolding of language representations over both time and space.

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

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

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

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

151 3 Discussion

152 Do deep language models and the human brain process sentences in the same way? Following a recent methodology
153 (34; 43; 44; 45; 46; 47; 48; 49; 31; 41; 26; 31), we address this issue by evaluating whether the activations of a large
154 variety of deep language models linearly map onto those of 102 human brains. Each subject was recorded by Schoffelen
155 and colleagues (3) with both MEG and fMRI, while they read isolated Dutch sentences composed of 9 to 15 words. Our
156 study provides two main contributions.

157 **Language performance is the primary factor that drives brain-mapping.** First, not only do language transformers
158 linearly map onto brain responses (24; 25; 26; 23; 27; 30; 32), but this property primarily depends on their language
159 performance: i.e. whether these models accurately predict words from their context. Our analysis indicates that
160 language performance is the most contributing factor explaining the variability of brain scores across models, spanning
161 32 architectures, two tasks, and 100 training steps (Figure 2D and H). Overall, deep language transformers thus appear
162 to mainly converge to brain-like representations during their training.

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

174 Finally, compositional embeddings peak in the brain regions associated with high-level language processing such
175 as the infero-frontal and anterior temporal cortices as well as the superior temporal cortex and the temporal-parietal
176 junction (58; 52; 28). We confirm that these left-lateralized representations are significant in both hemispheres (59; 60).
177 Critically, MEG suggests that these compositional effects become dominant and clearly bilateral long after word onset
178 (>800 ms). This surprisingly late response may be due to the nature of the sentences, whose complex syntactic structure
179 may slow down compositional computations.

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

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

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

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

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

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

226 4 Methods

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

230 4.1 Deep Neural Networks

231 4.1.1 Language Transformers

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

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

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

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

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

254 4.1.2 Visual Convolutional Neural Network

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

259 4.2 Neuroimaging

260 4.2.1 Protocol

261 For all the analyses, we used the open-source dataset released by Schoffelen and colleagues (3), gathering the functional
 262 magnetic resonance imaging (fMRI) and magneto-encephalography (MEG) recordings of 204 native Dutch speakers
 263 (100 males), aged from 18 to 33 years. Here, we focused on the 102 right-handed speakers who performed a *reading*
 264 task while being recorded by a CTF magneto-encephalography (MEG) and, in a separate session, with a SIEMENS
 265 Trio 3T Magnetic Resonance scanner (3).

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

275 4.2.2 Magnetic Resonance Imaging (MRI)

276 Structural images were acquired with a T1-weighted magnetization-prepared rapid gradient-echo (MP-RAGE) pulse
 277 sequence. The full acquisition details, available in (3), are summarized here for simplicity: TR=2,300 ms, TE=3.03
 278 ms, 8 degree flip-angle, 1 slab, slice-matrix size=256x256, slice thickness=1 mm, field of view=256 mm, isotropic
 279 voxel-size=1.0x1.0x1.0 mm. Structural images were defaced by Schoffelen and colleagues. Preprocessing of the
 280 structural MRI was performed with Freesurfer (67), using the `recon-all` pipeline and a manual inspection of the
 281 cortical segmentations, realigned to 'fsaverage'. Region-of-interest analyses were selected from the PALS Brodmann's
 282 Area atlas (68) and the Destrieux atlas (69).

283 Functional images were acquired with a T2*-weighted functional echo-planar blood oxygenation level-dependent
 284 (EPI-BOLD) sequence. The full acquisition details, available in (3), are summarized here for simplicity: TR=2.0
 285 seconds, TE=35ms, flip angle=90 degrees, anisotropic voxel size=3.5x3.5x3.0 mm extracted from 29 oblique slices.
 286 fMRI was preprocessed with fMRIPrep with default parameters (70). The resulting BOLD times series were detrended
 287 and de-confounded from 18 variables (the 6 estimated head-motion parameters ($\text{trans}_{x,y,z}$, $\text{rot}_{x,y,z}$) and the first 6
 288 noise components calculated using anatomical CompCorr (71) and 6 DCT-basis regressors using Nilearn's `clean_img`

²Each 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>

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

295 4.2.3 Magneto-encephalography (MEG)

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

309 4.3 Noise Ceiling: Brain \rightarrow Brain mapping

310 To estimate the amount of explainable signal in each MEG and fMRI recording, we trained and evaluated, through
 311 cross-validation, a linear mapping model W to predict the brain responses of a given subject to each sentence Y from the
 312 aggregated brain responses of all other subjects who read the same sentence X . Specifically, five cross-validation splits
 313 were implemented across 5-sentence blocks with scikit-learn 'GroupKFold' (36). For each word of each sentence i , all
 314 but one subject who read the corresponding sentence were averaged with one another to form a template brain response:
 315 $x_i \in \mathbb{R}^n$ with n the number of MEG channels or fMRI voxels, as well as a target brain response $y_i \in \mathbb{R}^n$ corresponding
 316 to the remaining subject. X and Y were normalized (mean=0, std=1) across sentences for each spatio-temporal
 317 dimension, using a robust scaler clipping below and above the 0.01st and 99.99th percentiles, respectively. A linear
 318 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:

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

319 with λ the l_2 regularization parameter, chosen amongst 20 values log-spaced between 10^{-3} and 10^8 with nested
 320 leave-one-out cross-validation for each dimension separately (as implemented in (36)). Brain predictions $\hat{Y} = WX$
 321 were evaluated with a Pearson correlation on the test set:

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

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

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

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

326 4.4 Brain score and similarity: Network \rightarrow Brain mapping

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

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

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

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

$$\Delta R_{M_1} = R_{M_1} - R_{M_2} \quad (4)$$

346 4.4.1 Convergence analysis

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

356 4.4.2 Permutation feature importance

357 To systematically quantify how the architecture, language accuracy, and training of the language transformers impacted
 358 their ability to linearly map onto brain activity, we fitted, for each subject separately, a Random Forest across the
 359 models' properties to predict their brain scores, using scikit-learn's `RandomForest` (35; 36). Specifically, we input
 360 the following features to the random forest: the training task (causal language modeling "CLM" vs. masked language
 361 modeling "MLM"), the number of attention heads $\in [4, 8]$, the total number of layers $\in [4, 8, 12]$, dimensionality
 362 $\in [128, 256, 512]$, training step (number of gradient updates, $\in [0, 4.5M]$), language modeling accuracy (top-1 accuracy
 363 at predicting a masked word) and the relative position of the representation (a.k.a 'layer position', between 0 for the
 364 word-embedding layer, and 1 for the last layer). The performance of the Random Forest was evaluated for each subject
 365 separately with a Pearson correlation R using five-split cross-validation across models.

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

371 4.5 Population statistics

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

377 4.6 Brain parcellation

378 In Section 2.2, Section 2.5 and Figure 3, we focus on particular regions of interest using the Brodmann's areas from the
 379 PALS parcellation of freesurfer⁴. The superior temporal gyrus (BA22) is split into its anterior, middle and posterior
 380 parts to increase granularity. For clarity, we rename certain areas as specified in the table below.

⁴https://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 Nee-lakantan, 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.

- 418 [14] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword
419 information. *arXiv preprint arXiv:1607.04606*, 2016.
- 420 [15] T. M. Mitchell, S. V. Shinkareva, A. Carlson, K.-M. Chang, V. L. Malave, R. A. Mason, and M. A. Just. Predicting
421 human brain activity associated with the meanings of nouns. 320(5880):1191–1195, 2008.
- 422 [16] Andrew James Anderson, Edmund C. Lalor, Feng Lin, Jeffrey R. Binder, Leonardo Fernandino, Colin J. Humphries,
423 Lisa L. Conant, Rajeev D. S. Raizada, Scott Grimm, and Xixi Wang. Multiple regions of a cortical network
424 commonly encode the meaning of words in multiple grammatical positions of read sentences. 29(6):2396–2411,
425 2019. Publisher: Oxford Academic.
- 426 [17] Jona Sassenhagen and Christian J. Fiebach. Traces of meaning itself: Encoding distributional word vectors in
427 brain activity. *bioRxiv*, 2019.
- 428 [18] Subba Reddy Oota, Naresh Manwani, and Bapi Raju S. fMRI Semantic Category Decoding using Linguistic
429 Encoding of Word Embeddings. *arXiv e-prints*, page arXiv:1806.05177, Jun 2018.
- 430 [19] Samira Abnar, Rasyan Ahmed, Max Mijnheer, and Willem H. Zuidema. Experiential, distributional and
431 dependency-based word embeddings have complementary roles in decoding brain activity. *CoRR*, abs/1711.09285,
432 2017.
- 433 [20] Yu-Ping Ruan, Zhen-Hua Ling, and Yu Hu. Exploring semantic representation in brain activity using word
434 embeddings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*,
435 pages 669–679, Austin, Texas, November 2016. Association for Computational Linguistics.
- 436 [21] Christian Brodbeck, L Elliot Hong, and Jonathan Z Simon. Rapid transformation from auditory to linguistic
437 representations of continuous speech. *Current Biology*, 28(24):3976–3983, 2018.
- 438 [22] Jon Gauthier and Anna Ivanova. Does the brain represent words? an evaluation of brain decoding studies of
439 language understanding. *arXiv preprint arXiv:1806.00591*, 2018.
- 440 [23] Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. Aligning context-based statistical models of
441 language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Nat-
442 ural Language Processing (EMNLP)*, pages 233–243, Doha, Qatar, October 2014. Association for Computational
443 Linguistics.
- 444 [24] Shailee Jain and Alexander Huth. Incorporating context into language encoding models for fmri. In S. Bengio,
445 H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information
446 Processing Systems 31*, pages 6628–6637. Curran Associates, Inc., 2018.
- 447 [25] Nikos Athanasiou, Elias Iosif, and Alexandros Potamianos. Neural activation semantic models: Computational
448 lexical semantic models of localized neural activations. In *Proceedings of the 27th International Conference
449 on Computational Linguistics*, pages 2867–2878, Santa Fe, New Mexico, USA, August 2018. Association for
450 Computational Linguistics.
- 451 [26] Mariya Toneva and Leila Wehbe. Interpreting and improving natural-language processing (in machines) with
452 natural language-processing (in the brain). *CoRR*, abs/1905.11833, 2019.
- 453 [27] Micha Heilbron, Kristijan Armeni, Jan-Mathijs Schoffelen, Peter Hagoort, and Floris P de Lange. A hierarchy of
454 linguistic predictions during natural language comprehension. *bioRxiv*, 2020.
- 455 [28] Jonathan R Brennan and Liina Pykkänen. Meg evidence for incremental sentence composition in the anterior
456 temporal lobe. *Cognitive science*, 41:1515–1531, 2017.
- 457 [29] John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan R Brennan. Finding syntax in human encephalography
458 with beam search. *arXiv preprint arXiv:1806.04127*, 2018.
- 459 [30] Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A Nastase, Amir
460 Feder, Dotan Emanuel, Alon Cohen, et al. Thinking ahead: prediction in context as a keystone of language in
461 humans and machines. *bioRxiv*, 2020.
- 462 [31] Alexander Kell, Daniel Yamins, Erica Shook, Sam Norman-Haignere, and Josh McDermott. A task-optimized
463 neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing
464 hierarchy. *Neuron*, 98, 04 2018.
- 465 [32] Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua
466 Tenenbaum, and Evelina Fedorenko. Artificial neural networks accurately predict language processing in the
467 brain. *BioRxiv*, 2020.
- 468 [33] Juliette Millet and Jean-Rémi King. Inductive biases, pretraining and fine-tuning jointly account for brain
469 responses to speech. February 2021.

- 470 [34] Daniel LK Yamins, Ha Hong, Charles F Cadieu, Ethan A Solomon, Darren Seibert, and James J DiCarlo.
471 Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the*
472 *National Academy of Sciences*, 111(23):8619–8624, 2014.
- 473 [35] Leo Breiman. Random forests. 45(1):5–32, 2001.
- 474 [36] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel,
475 Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python.
476 *Journal of machine learning research*, 12(Oct):2825–2830, 2011.
- 477 [37] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words
478 and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q.
479 Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 3111–3119. Curran Associates,
480 Inc., 2013.
- 481 [38] Christopher D. Manning, Kevin Clark, John Hewitt, Urvashi Khandelwal, and Omer Levy. Emergent linguistic
482 structure in artificial neural networks trained by self-supervision. page 201907367, 2020.
- 483 [39] Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. What does BERT learn about the structure of language? In
484 *ACL 2019 - 57th Annual Meeting of the Association for Computational Linguistics*, Florence, Italy, July 2019.
- 485 [40] Jeonghun Baek, Geewook Kim, Junyeop Lee, Sungrae Park, Dongyoon Han, Sangdoon Yun, Seong Joon Oh, and
486 Hwalsuk Lee. What is wrong with scene text recognition model comparisons? dataset and model analysis. In
487 *Proceedings of the IEEE International Conference on Computer Vision*, pages 4715–4723, 2019.
- 488 [41] Alexander G. Huth, Wendy A. de Heer, Thomas L. Griffiths, Frédéric E. Theunissen, and Jack L. Gallant. Natural
489 speech reveals the semantic maps that tile human cerebral cortex. 532(7600):453–458, 2016.
- 490 [42] Anna Seydell-Greenwald, Xiaoying Wang, Elissa Newport, Yanchao Bi, and Ella Striem-Amit. Spoken language
491 comprehension activates the primary visual cortex. *bioRxiv*, 2020.
- 492 [43] Hanlin Tang, Martin Schrimpf, William Lotter, Charlotte Moerman, Ana Paredes, Josue Ortega Caro, Walter
493 Hardesty, David Cox, and Gabriel Kreiman. Recurrent computations for visual pattern completion. *Proceedings*
494 *of the National Academy of Sciences*, 115(35):8835–8840, 2018.
- 495 [44] Seyed-Mahdi Khaligh-Razavi and Nikolaus Kriegeskorte. Deep supervised, but not unsupervised, models may
496 explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.
- 497 [45] Nikolaus Kriegeskorte. Deep neural networks: a new framework for modeling biological vision and brain
498 information processing. *Annual review of vision science*, 1:417–446, 2015.
- 499 [46] Umut Güçlü and Marcel AJ van Gerven. Deep neural networks reveal a gradient in the complexity of neural
500 representations across the ventral stream. *Journal of Neuroscience*, 35(27):10005–10014, 2015.
- 501 [47] Michael Eickenberg, Alexandre Gramfort, Gaël Varoquaux, and Bertrand Thirion. Seeing it all: Convolutional
502 network layers map the function of the human visual system. *NeuroImage*, 152:184–194, 2017.
- 503 [48] Daniel LK Yamins and James J DiCarlo. Using goal-driven deep learning models to understand sensory cortex.
504 *Nature neuroscience*, 19(3):356, 2016.
- 505 [49] Andrew Saxe, Stephanie Nelli, and Christopher Summerfield. If deep learning is the answer, what is the question?
506 *Nature Reviews Neuroscience*, pages 1–13, 2020.
- 507 [50] Cathy J Price. The anatomy of language: a review of 100 fmri studies published in 2009. *Annals of the new York*
508 *Academy of Sciences*, 1191(1):62–88, 2010.
- 509 [51] Stanislas Dehaene and Laurent Cohen. The unique role of the visual word form area in reading. *Trends in cognitive*
510 *sciences*, 15(6):254–262, 2011.
- 511 [52] Gregory Hickok and David Poeppel. The cortical organization of speech processing. 8(5):393–402, 2007. Number:
512 5 Publisher: Nature Publishing Group.
- 513 [53] Aakash Agrawal, KVS Hari, and SP Arun. A compositional neural code in high-level visual cortex can explain
514 jumbled word reading. *Elife*, 9:e54846, 2020.
- 515 [54] Evelina Fedorenko, Idan Blank, Matthew Siegelman, and Zachary Mineroff. Lack of selectivity for syntax relative
516 to word meanings throughout the language network. *bioRxiv*, page 477851, 2020.
- 517 [55] Stanislas Dehaene, Laurent Cohen, Mariano Sigman, and Fabien Vinckier. The neural code for written words: a
518 proposal. 9(7):335–341.
- 519 [56] Dora Hermes, Vinitha Rangarajan, Brett L Foster, Jean-Remi King, Itir Kasikci, Kai J Miller, and Josef Parvizi.
520 Electrophysiological responses in the ventral temporal cortex during reading of numerals and calculation. *Cerebral*
521 *cortex*, 27(1):567–575, 2017.

- 522 [57] Oscar Woolnough, Cristian Donos, Patrick S Rollo, Kiefer James Forseth, Yair Lakretz, Nathan E Crone, Simon
523 Fischer-Baum, Stanislas Dehaene, and Nitin Tandon. Spatiotemporal dynamics of orthographic and lexical
524 processing in the ventral visual pathway. *bioRxiv*, 2020.
- 525 [58] Christophe Pallier, Anne-Dominique Devauchelle, and Stanislas Dehaene. Cortical representation of the constituent
526 structure of sentences. *Proceedings of the National Academy of Sciences*, 108(6):2522–2527, 2011.
- 527 [59] Evelina Fedorenko, Po-Jang Hsieh, Alfonso Nieto-Castañón, Susan Whitfield-Gabrieli, and Nancy Kanwisher.
528 New method for fmri investigations of language: defining rois functionally in individual subjects. *Journal of*
529 *neurophysiology*, 104(2):1177–1194, 2010.
- 530 [60] Gregory B Cogan, Thomas Thesen, Chad Carlson, Werner Doyle, Orrin Devinsky, and Bijan Pesaran. Sensory–
531 motor transformations for speech occur bilaterally. *Nature*, 507(7490):94–98, 2014.
- 532 [61] Karl Friston. The free-energy principle: a unified brain theory? *Nature reviews neuroscience*, 11(2):127–138,
533 2010.
- 534 [62] João Loula, Marco Baroni, and Brenden M. Lake. Rearranging the familiar: Testing compositional generalization
535 in recurrent networks. *CoRR*, abs/1807.07545, 2018.
- 536 [63] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. Hellaswag: Can a machine really finish
537 your sentence? *arXiv preprint arXiv:1905.07830*, 2019.
- 538 [64] Andy Coenen, Emily Reif, Ann Yuan, Been Kim, Adam Pearce, Fernanda Viégas, and Martin Wattenberg.
539 Visualizing and measuring the geometry of BERT.
- 540 [65] Liberty S Hamilton and Alexander G Huth. The revolution will not be controlled: natural stimuli in speech
541 neuroscience. *Language, Cognition and Neuroscience*, 35(5):573–582, 2020.
- 542 [66] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke
543 Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan
544 Herbst. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual*
545 *Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster*
546 *Sessions*, pages 177–180. Association for Computational Linguistics.
- 547 [67] Bruce Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.
- 548 [68] David C Van Essen. A population-average, landmark-and surface-based (pals) atlas of human cerebral cortex.
549 *Neuroimage*, 28(3):635–662, 2005.
- 550 [69] Christophe Destrieux, Bruce Fischl, Anders Dale, and Eric Halgren. Automatic parcellation of human cortical
551 gyri and sulci using standard anatomical nomenclature. *Neuroimage*, 53(1):1–15, 2010.
- 552 [70] Oscar Esteban, Christopher J Markiewicz, Ross W Blair, Craig A Moodie, A Ilkay Isik, Asier Erramuzpe, James D
553 Kent, Mathias Goncalves, Elizabeth DuPre, Madeleine Snyder, et al. fmriprep: a robust preprocessing pipeline for
554 functional mri. *Nature methods*, 16(1):111–116, 2019.
- 555 [71] Yashar Behzadi, Khaled Restom, Joy Liau, and Thomas T Liu. A component based noise correction method
556 (compcor) for bold and perfusion based fmri. *Neuroimage*, 37(1):90–101, 2007.
- 557 [72] Alexandre Abraham, Fabian Pedregosa, Michael Eickenberg, Philippe Gervais, Andreas Mueller, Jean Kossaifi,
558 Alexandre Gramfort, Bertrand Thirion, and Gaël Varoquaux. Machine learning for neuroimaging with scikit-learn.
559 *Frontiers in neuroinformatics*, 8:14, 2014.
- 560 [73] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck,
561 Lauri Parkkonen, and Matti S. Hämäläinen. Mne software for processing meg and eeg data. *NeuroImage*, 86:446 –
562 460, 2014.
- 563 [74] Peter Hagoort. The neurobiology of language beyond single-word processing. *Science*, 366(6461):55–58, 2019.
- 564 [75] Evelina Fedorenko, Terri L Scott, Peter Brunner, William G Coon, Brianna Pritchett, Gerwin Schalk, and Nancy
565 Kanwisher. Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of*
566 *Sciences*, 113(41):E6256–E6262, 2016.
- 567 [76] Matthew J. Nelson, Imen El Karoui, Kristof Giber, Xiaofang Yang, Laurent Cohen, Hilda Koopman, Sydney S.
568 Cash, Lionel Naccache, John T. Hale, Christophe Pallier, and Stanislas Dehaene. Neurophysiological dynamics
569 of phrase-structure building during sentence processing. *Proceedings of the National Academy of Sciences*,
570 114(18):E3669–E3678, May 2017. Publisher: National Academy of Sciences Section: PNAS Plus.

571 6 Supplementary information

572 6.1 Anatomical and temporal characteristics of average brain responses to reading.

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

581 6.2 Noise ceilings.

582 To compare the brain responses to the activations of deep language models, it is important to estimate the level of
 583 signal-to-noise ratio (Figure S4). For this, we fit, for each subject separately, a "noise-ceiling" model across subjects:
 584 for each recording of each subject and each sentence Y_{train} , we fit a linear model W from the recordings of all other
 585 subjects who read the same sentence X_{train} to predict each voxel and each MEG sensor at each time sample, separately.
 586 Using a cross-validation scheme across sentences, we then evaluate the Pearson correlation R between (1) the true
 587 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
 588 sensor separately. This procedure can be thought of as approximating an optimal black box: i.e. evaluating a one-hot
 589 encoder of brain responses is trained and evaluated on each element of a unique sentence. Noise ceiling peaks within the
 590 expected language network (75) (Figure 1 F-H). These estimates are relatively low: for example, fMRI noise ceilings
 591 reach, on average, $R = 0.129 (\pm 0.004 \text{ SEM across subjects})$ in the superior temporal gyrus, whereas MEG noise
 592 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	$p < 10^{-8}$
593	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	$p < 10^{-8}$

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

596 6.3 Probe analysis of the language transformer.

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

608 6.4 Video materials

609 Below the captions for the two videos provided in supplementary.

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

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

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

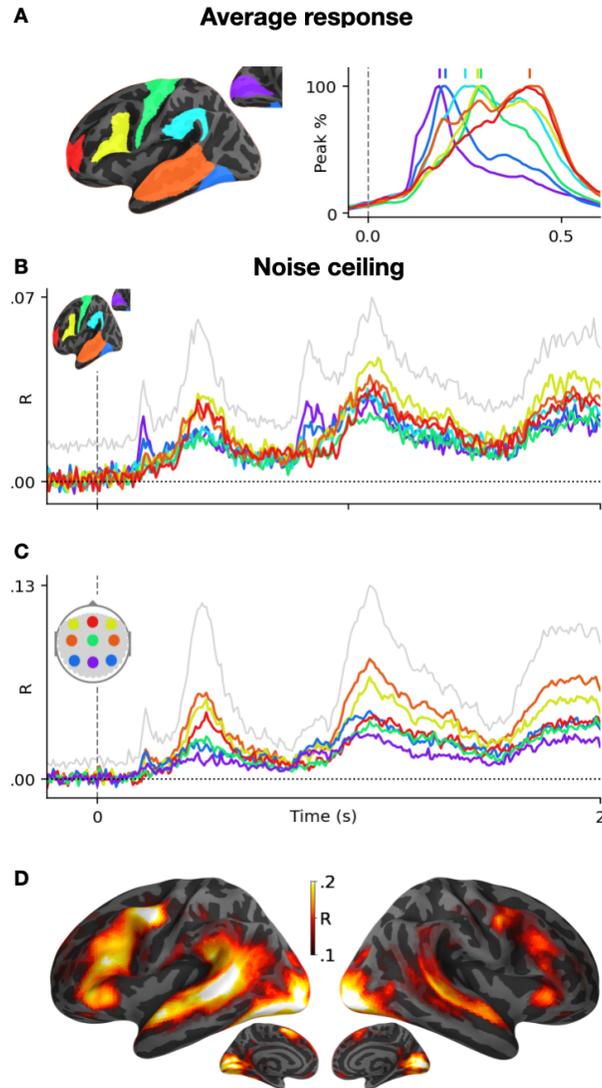


Figure S4: **Noise ceilings.** **A.** Grand average MEG source estimates to word onset ($t=0$) for 7 regions typically associated with reading (V1: purple, M1: green, fusiform gyrus: dark blue, supramarginal gyrus: light blue, superior temporal gyrus: orange, infero-frontal gyrus: yellow and fronto-polar 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 (B) in sensor space. **D.** Noise ceiling estimates of fMRI recordings.

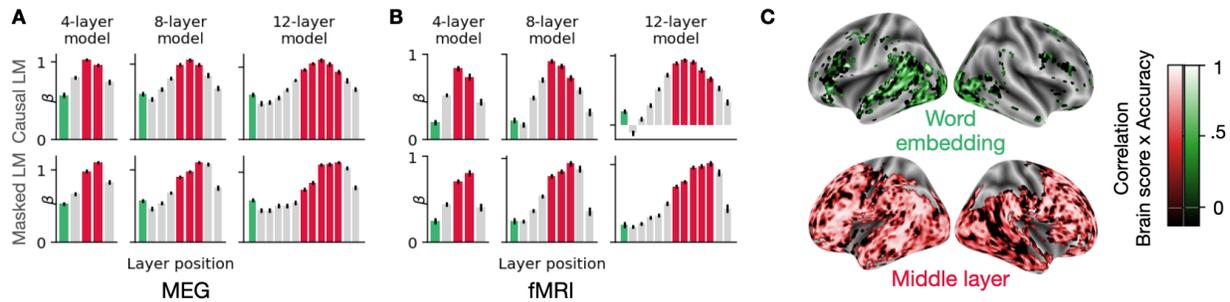


Figure S5: **Correlation between the network's performance and brain score.** A-B. Standardized beta coefficients between the language modeling performance of the network and its MEG (A) or fMRI (B) scores. For each subject, the brain scores are first scaled (0-mean, 1-std). Then, a linear regression 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).

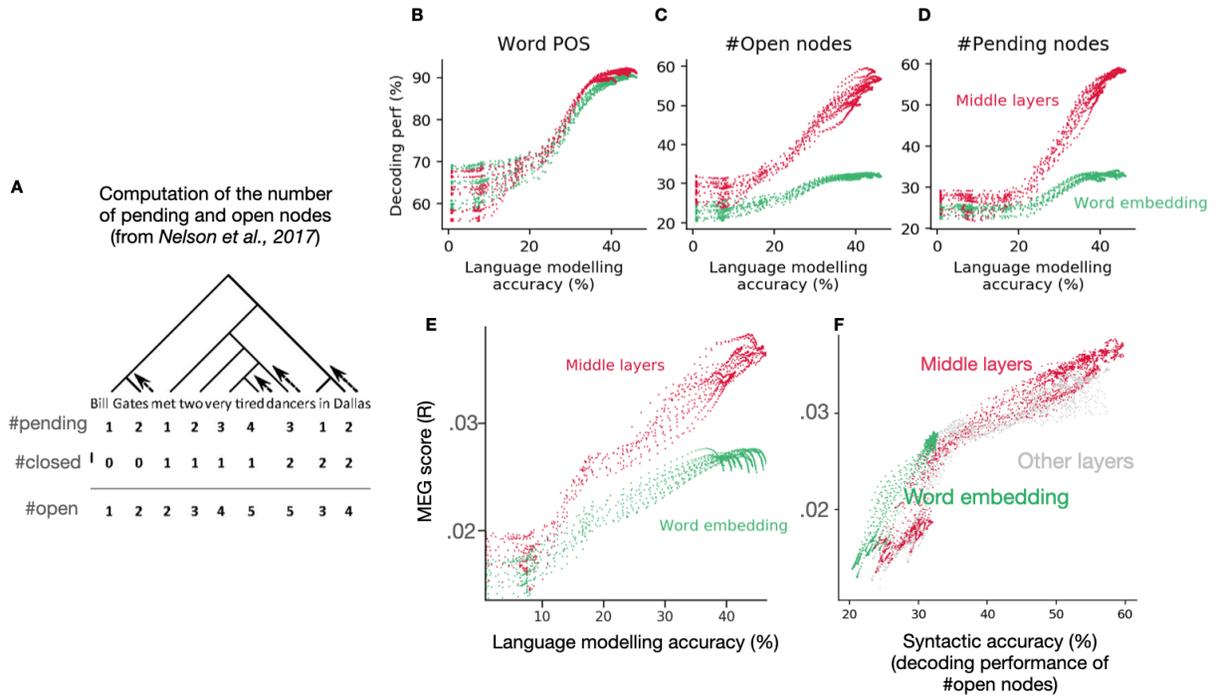


Figure S6: **What linguistic information drives the brain score?** **A**. From the stimulus, we compute three linguistic features: the part-of-speech of the words (i) (as given by Spacy), and two higher-level syntactic features: the number of pending nodes (ii) and open nodes (iii). These two syntactic features are derived from the constituency trees of the sentences, following (76). **B-D**. A ℓ_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.