LANGUAGE PROCESSING IN BRAINS AND DEEP NEURAL NETWORKS: COMPUTATIONAL CONVERGENCE AND ITS LIMITS

A PREPRINT

Charlotte Caucheteux
Facebook AI Research

Jean-Rémi King
PSL University, CNRS
Facebook AI Research

June 29, 2020

ABSTRACT

Deep learning has recently allowed substantial progress in language tasks such as translation and completion. Do such models process language similarly to humans, and is this similarity driven by systematic structural, functional and learning principles? To address these issues, we tested whether the activations of 7,400 artificial neural networks trained on image, word and sentence processing linearly map onto the hierarchy of human brain responses elicited during a reading task, using source-localized magneto-encephalography (MEG) recordings of one hundred and four subjects. Our results confirm that visual, word and language models sequentially correlate with distinct areas of the left-lateralized cortical hierarchy of reading. However, only specific subsets of these models converge towards brain-like representations during their training. Specifically, when the algorithms are trained on language modeling, their middle layers become increasingly similar to the late responses of the language network in the brain. By contrast, input and output word embedding layers often diverge away from brain activity during training. These differences are primarily rooted in the sustained and bilateral responses of the temporal and frontal cortices. Together, these results suggest that the compositional - but not the lexical - representations of modern language models converge to a brain-like solution.

***K*eywords** Natural Language Processing | Neurobiology of Language | Magneto-encephalography

1 Introduction

Deep neural networks trained on "language modeling" (i.e. guessing masked words in a given text) have recently led to substantial progress (Vaswani et al., 2017; Devlin et al., 2018; Lample and Conneau, 2019) on tasks traditionally associated with human intelligence (Turing, 2009; Chomsky, 2006). While still limited (Loula et al., 2018), the improvements in translation, dialogue and summarization allowed by these models, lead to a simple question: do these artificial neural networks tend to process language similarly to humans, or do they just superficially mimic our behavior?

This question is all-the-more challenging that the neurobiology of language remains in its infancy: the transformation of sensory input into phonemic and orthographic representations are becoming increasingly understood (Mesgarani et al., 2014; Dehaene and Cohen, 2011; Hickok and Poeppel, 2007; Di Liberto et al., 2015; Tang et al., 2017; Donhauser and Baillet, 2019). However, the large inter-individual variability and the distributed nature of high-level representations (Fedorenko et al., 2010; Price, 2010; Fedorenko et al., 2020) have limited the study of lexical (how individual words are represented) and compositional representations (i.e. how words are combined with one another) to piecemeal investigations (Binder et al., 1997; Pylkkänen and Marantz, 2003; Binder et al., 2009; Price, 2010; Fedorenko et al., 2010; Pallier et al., 2011; Fedorenko et al., 2016; Blank et al., 2016; Brennan et al., 2016; Brennan and Pylkkänen, 2017; Pylkkanen and Brennan, 2019; Hagoort, 2019).

Reading, for example, depends on a cortical hierarchy that originates in the primary visual cortex (V1), continues within the visual word form area (in the fusiform gyrus - where letters and strings are identified, and reaches the angular gyrus, the anterior temporal lobe and the middle temporal gyrus - associated with word meaning (Price, 2010; Dehaene

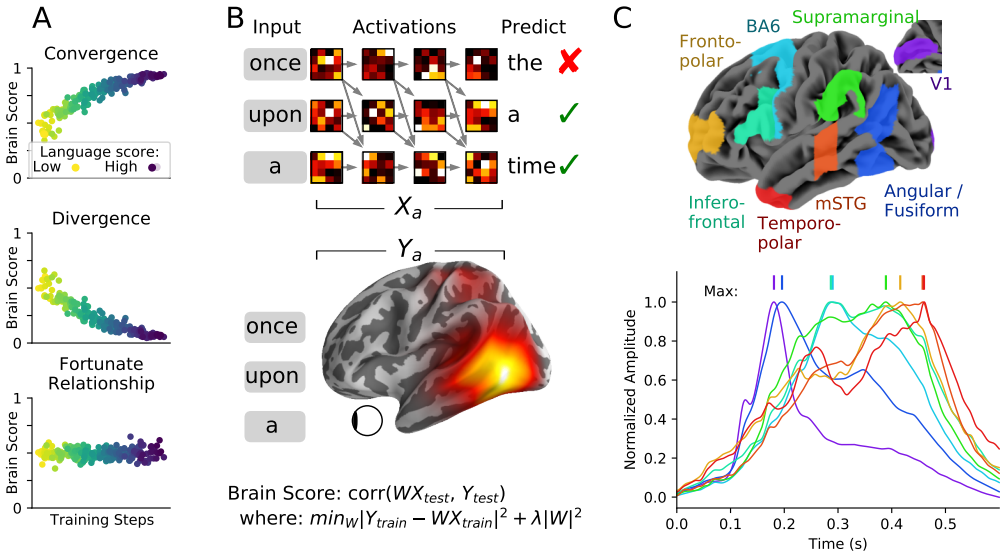

Figure 1: A. Hypotheses. An artificial neural network is said to *converge* to brain-like representations if training makes its activations become increasingly correlated with those of the brain, and *vice versa* if it diverges. Because artificial neural networks are high dimensional, part of a random network could significantly correlate with brain activity, and thus lead to a "fortunate relationship" between the brain and the algorithm. In this schema, each dot represents one artificial neural network frozen at a given training step. B. To quantify the similarity between an artificial neural network and the brain, a linear regression can be fit from the model's activations (X) to the brain response (Y) to the same stimulus sequence (here: 'once upon a'). The resulting "brain score" Yamins et al. (2014) is independent of the training objective of the model (e.g. predicting the next word). C. Average (absolute) MEG responses to word onset in various regions of the cortical network associated with reading, normalized by their maximum amplitudes to highlight their relative onsets and peaks (top ticks). See Movie 1 for additional results.

and Cohen, 2011; Hickok and Poeppel, 2007; Huth et al., 2016). This hierarchical pathway and a parallel motor route - responsible for articulatory codes (Hickok and Poeppel, 2007) together connect with the inferior frontal gyrus, repeatedly associated with the compositional processes of language (Dehaene and Cohen, 2011; Hickok and Poeppel, 2007; Hagoort, 2005; Mollica et al.; Fedorenko et al., 2020)). However, the precise nature, format and dynamics of each of these language regions remain largely unknown (Fedorenko et al., 2020; Dehaene and Cohen, 2011; Hagoort, 2019; Hickok and Poeppel, 2007).

Recently, several neuroimaging studies have directly linked the representations of language models to those of the brain. In this view, the word embeddings developed for natural language processing (Bengio et al., 2003; Mikolov et al., 2013; Pennington et al., 2014; Bojanowski et al., 2016) have been shown to linearly correlate with the brain responses elicited by words presented in isolation (Mitchell et al., 2008; Anderson et al., 2019; Sassenhagen and Fiebach, 2019) or within narratives (Reddy Oota et al., 2018; Abnar et al., 2017; Ruan et al., 2016; Brodbeck et al., 2018; Gauthier and Ivanova, 2018). Furthermore, *contextualized* word embeddings significantly improve such correlations in the prefrontal, temporal and parietal cortices (Jain and Huth, 2018; Athanasiou et al., 2018; Toneva and Wehbe, 2019). However, these important findings call for a principled investigation: do modern language models systematically (1) converge to, (2) anecdotally correlate with, or (3) even diverge away from the sequence of computations the brain implements to process language (Figure 1 A-B)? On the contrary, do some language models better correlate with brain activity than simply because (i) they are higher dimensional, (ii) they are trained on different texts, (iii) they naturally express a large variety of non-linear dynamics etc?

To address this issue, we assessed whether the activations of 7,400 artificial neural networks varying in their architectures, objectives, training and performances, linearly correlate with source-localized magneto-encephalography (MEG) recordings of 104 subjects reading words sequentially presented in random word lists or arranged into meaningful sentences. The results reveal the functional organization of the reading cortical network with a remarkable spatio-temporal clarity (Videos 1 and 2). Critically, the extensive comparison across models reveal that, during training, the

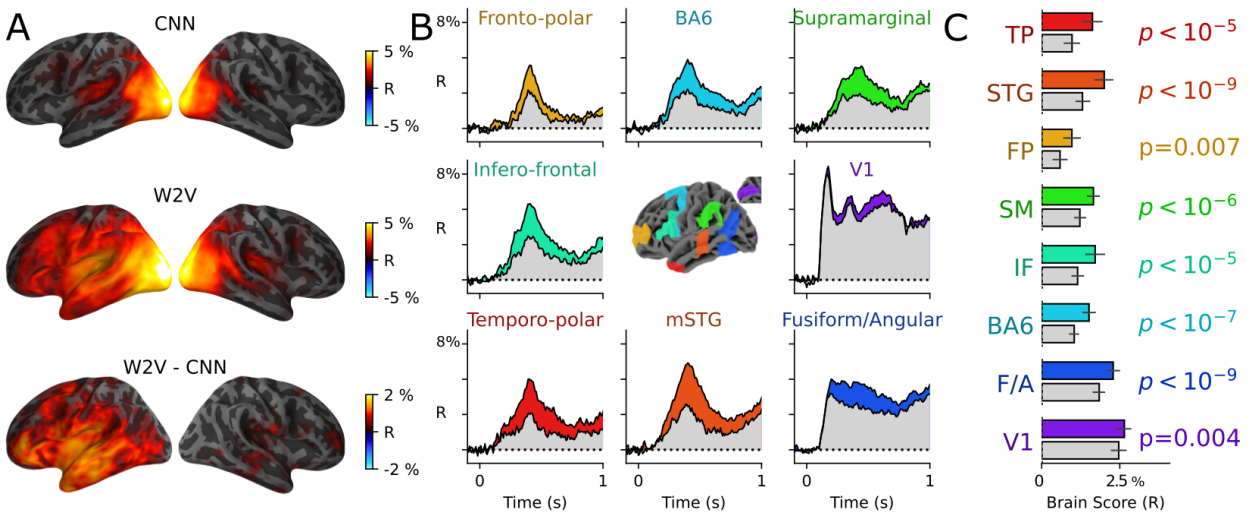

Figure 2: A. Average brain scores across time (0 - 1 sec after word onset) and subjects for the deep CNN trained on character recognition (top), Word2Vec (middle) and the difference between the two (bottom) in response to words presented in random word lists. B. Average brain scores within each region-of-interest (panels) obtained with the CNN (gray) and with Word2Vec (W2V). The coloured area indicate when W2V is higher than CNN. C. Second-level statistical comparison across subjects of the brain scores obtained within each region-of-interest (averaged from 0 - 1 s) with the CNN (gray) and W2V (color), resulting from a two-sided Wilcoxon signed-rank test. Error bars are the 95% confidence intervals of the scores' distribution across subjects.

middle - but not the outer - layers of deep language models systematically converge to the sustained representations of the bilateral frontal and temporal cortices.

2 Results

Towards a spatio-temporal decomposition of the reading network

In spite of substantial individual variability, our MEG source reconstructions were consistent with the anatomy of the reading network (Dehaene and Cohen, 2011): on average, written words elicited a sequence 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 1 C, Video 1).

Some of these brain responses likely represent non-linguistic features (e.g. visual onsets, width of the words etc). Following previous works (Kriegeskorte, 2015; Güçlü and van Gerven, 2015; Eickenberg et al., 2017; Yamins and DiCarlo, 2016), we thus tested whether single-trial MEG responses linearly correlated with the activations of the last layer of a deep convolutional neural network (CNN) trained on character recognition (Baek et al., 2019). Specifically, we input the CNN with 100x32 pixels images of each the words presented to the subjects in order to convert them into 888-dimensional visual embedding vectors. We then fitted a ridge-regularized linear regression across these visual embeddings to predict the MEG responses to the corresponding words presented within random lists. Finally, we estimated the precision of this model-to-brain mapping on out-of-vocabulary predictions with a Pearson R correlation score, hereafter referred to as "brain score".

A spatio-temporal summary of the brain scores obtained with the CNN are displayed in Figure 2. Overall, our results confirm that single-trial brain responses can be linearly predicted by the activations of a deep CNN, with a peak brain score of R=8.4% $\pm$ .46% (standard error of the mean across subjects) in the early visual cortex at 150 ms. Note that the present MEG brain scores are substantially lower than those reported with analogies fMRI and electrophysiology studies (Yamins et al., 2014; Jain and Huth, 2018; Huth et al., 2016), partly because we here predict individual recording samples rather than averaged responses or noise-corrected repetitions. These effects are nonetheless robust across subjects (Supplementary Figure 1).

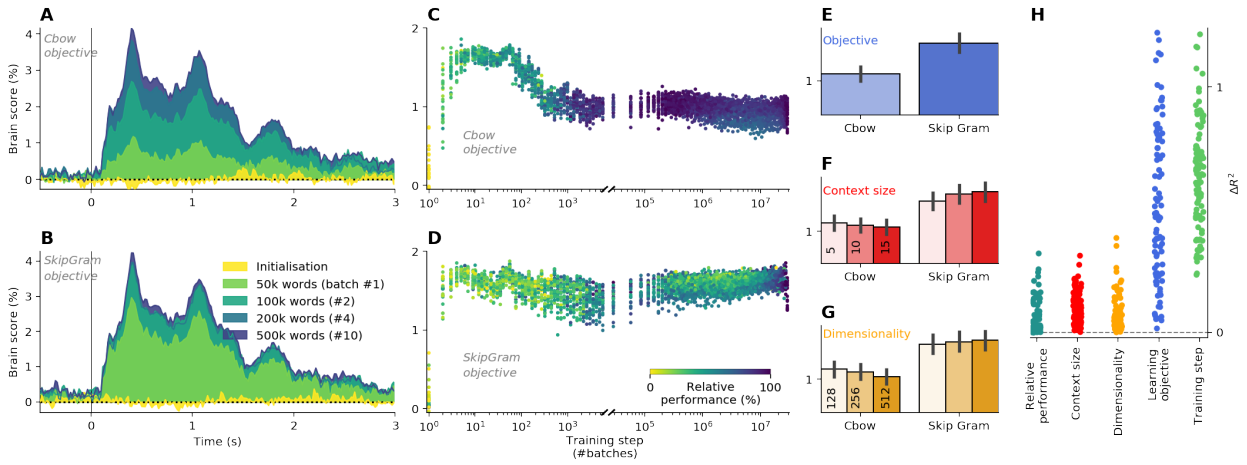

Figure 3: A. Brain scores of a 256-dimensional CBOW Word2Vec embedding, trained with a context size of 5 words at different training stages (full corpus size: 278M words from Wikipedia), as a function of time (averaged across MEG senors). B Same as (A) using a Skip-gram Word2Vec. C. Averaged brain scores obtained across time, space (all MEG channels) and subjects (y-axis), for each of the word embedding of CBOW Word2Vec embeddings (dots, n=6,889 models in total) as a function of their training (x-axis) and performance on the training task (color) (see methods for details). D. Same as (C) using a Skip-gram Word2Vec. E-G. Averaged Brain scores obtained as a function of the objective (CBOW vs Skipgram), context size and dimensionality, irrespective of other properties (e.g. training, performance etc). Error bars indicate the 95% confidence interval. H. Feature importance estimated from a random forest regressor fitted for each subject separately (dots, n=95) to predict the brain scores (averaged across time and space) of an embedding model given its loss, training step, objective, context size and dimensionality. Overall, the performance of word embedding has a modest and non-monotonic impact on brain scores.

Word embeddings specifically correlate with late, distributed and lateralized brain responses to words

Visual embeddings cannot capture the arbitrary meanings of written words. Word semantics, however, is partially captured by algorithms trained to predict word vectors from their context (CBOW) or vice versa (Skipgram, (Mikolov et al., 2013)). We thus applied the above brain score analysis with Word2Vec embeddings, in order to identify where and when semantic representations may account for brain responses above and beyond low-level visual features.

Word2Vec (Mikolov et al., 2013) trained with a CBOW objective better predicted brain responses than the visual embeddings of the CNN from $\approx$ 200ms, and with a peak difference around 400ms (mean $\Delta R = 1.6\%$ across sources and subjects) especially in the left-lateralized temporal and prefrontal responses (statistics summarized in Figure 2C). These results confirm that the corresponding brain regions specifically generate lexical representations.

Word embeddings do not systematically converge towards brain-like representations

To test whether the similarity between word embeddings and cortical responses is fortunate or principled, we trained 18 word embeddings varying in random seeds , dimensionality ($n \in 128, 256, 512$), context sizes ($n \in 2, 4, 6$) and learning objectives (Skipgram vs CBOW). We then estimated for each of them, how their brain score (averaged over time and independently of brain location) varied as a function of their training steps (cf. Methods) and performance on the training task (log-likelihood on a test set, cf.SI-2-4). Note that, because losses between architectures are hardly comparable because their training tasks differ, we here focus on their *relative* performance (%): i.e. the loss at a current step divided by the final loss. Figure 3 summarizes how the average brain scores vary with each structural and functional properties word embeddings'.

We then used the feature-importance analysis of random forests (Breiman, 2001), in order to quantify the independent impact of each model property (relative performance, context size, dimensionality, CBOW or Skipgram objective and training step) on the final brain scores (cf. Methods). The dimensionality ($\Delta R^2 = 7\% \pm 0.8\%, p < 10^{-16}$, as estimated with a second-level Wilcoxon test across subjects), the context size ($10\% \pm 0.6\%, p < 10^{-16}$) and the performance of the model ($\Delta R^2 = 6\% \pm 0.7\%, p < 10^{-16}$) significantly but modestly varied the extent to which a word embedding linearly correlates with brain responses to words. By contrast, the amount of training ($\Delta R^2 = 64\% \pm 2\%, p < 10^{-16}$)

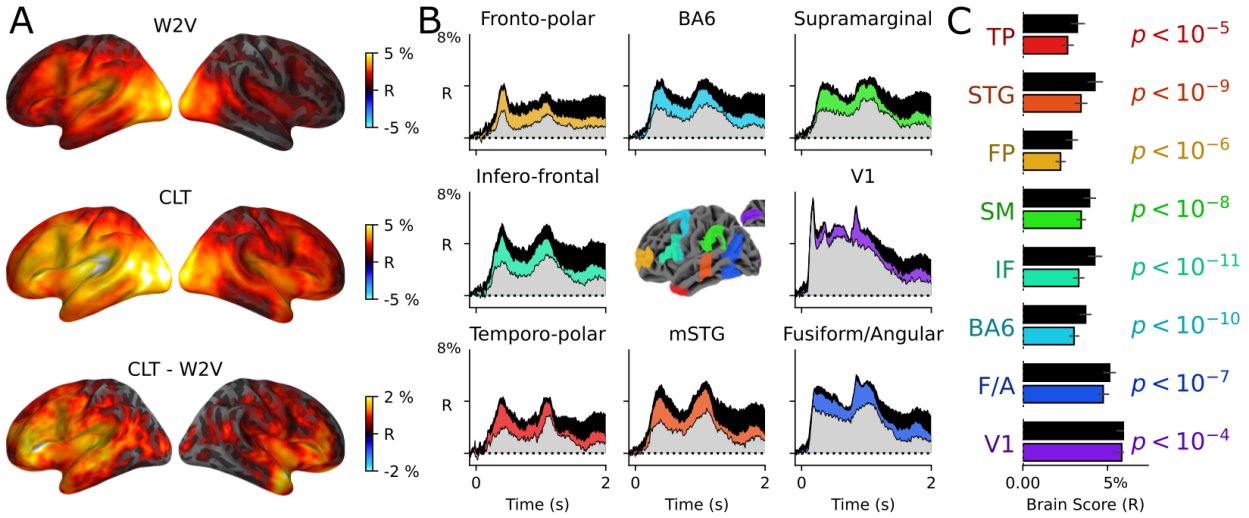

Figure 4: A. Average brain scores across time (0 - 2 sec after word onset) and subjects for Word2Vec (top), the ninth layer of a 13-layer Causal Language Transformer (CLT, middle) and the difference between the two (bottom) in response to words presented within sentences. B. Average brain scores within each region-of-interest (panels) obtained with the CNN (gray), Word2Vec (color) and the CLT (black). C. Second-level statistical comparison across subjects of the brain scores obtained within each region-of-interest (averaged from 0 - 2 s) with Word2Vec (color) and the CLT (black), resulting from a two-sided Wilcoxon signed-rank test. Error bars are the 95% confidence intervals of the scores' distribution across subjects. See Movie 2 for additional results.

and the learning objective (CBOW vs Skipgram) ($\Delta R^2 = 50\% \pm 3\%, p < 10^{-16}$) appeared to be important predictors of whether an embedding would linearly correlate with brain responses to words.

To our surprise, however, brain scores did not vary monotonically with the training and performance of word embeddings. For example, the brain scores of CBOW-trained Word2Vec steeply increased at the beginning of training, decreased after 60 training steps (i.e. after having been exposed to 3M words and 155,000 distinct words), and finally reached a plateau after 3,000 training steps (150M words and 750,000 distinct words). (Fig. 3 C). Similarly, the Word2Vec embeddings trained with a skip-gram objective reached a plateau around only five training steps (i.e. ≈250,000 words).

Together, these results suggest that training word embedding algorithms does not make them systematically converge to brain-like representations.

Contextualized word embeddings specifically correlate with the delayed responses of the bilateral prefrontal and temporal cortices

Word embeddings are not contextualized: i.e. words are associated with a unique vector, independently of the context in which they are used. Yet, the meaning of a sentence depends not only on the meaning of its words, but also on the rules used to combine them (Dummett, 1981). To test whether modern language models systematically *combine* words into representations that linearly correlate with those of the human brain, we applied the above brain score analyses with the *contextualized* word representations generated by a variety of transformers (Devlin et al., 2018; Vaswani et al., 2017; Radford et al., 2019) - state-of-the art deep neural networks trained to predict words given a context. To make sure that the models, like the brain, process and combine words sequentially, we restrict our analyses to *causal* transformers (i.e. unidirectional, from left to right) (CLT), and now focus on the brain responses to words presented within isolated but meaningful sentences.

Figure 4 summarizes the brain scores obtained with visual, word and contextualized word embeddings, generated by the visual CNN, a representative Word2Vec embedding and the ninth layer of a representative 13-layer CLT, respectively. Video 2 displays the brain scores obtained with the visual CNN (blue), as well as the *gain* in brain scores obtained with the Word2Vec embedding (green) and the *gain* in brain scores obtained with the ninth layer of a CLT (red). Overall, the spatio-temporal brain scores obtained with visual and word embeddings in the context of sentences were largely similar to those obtained in the context of random word lists: the former peaked around 100 ms in the primary visual cortex and rapidly propagated across the reading network, whereas the latter peaked around 400 ms and were primarily distributed over the left temporal and frontal cortices (Figure 4, Video 2). The contextualized word embeddings led to significantly

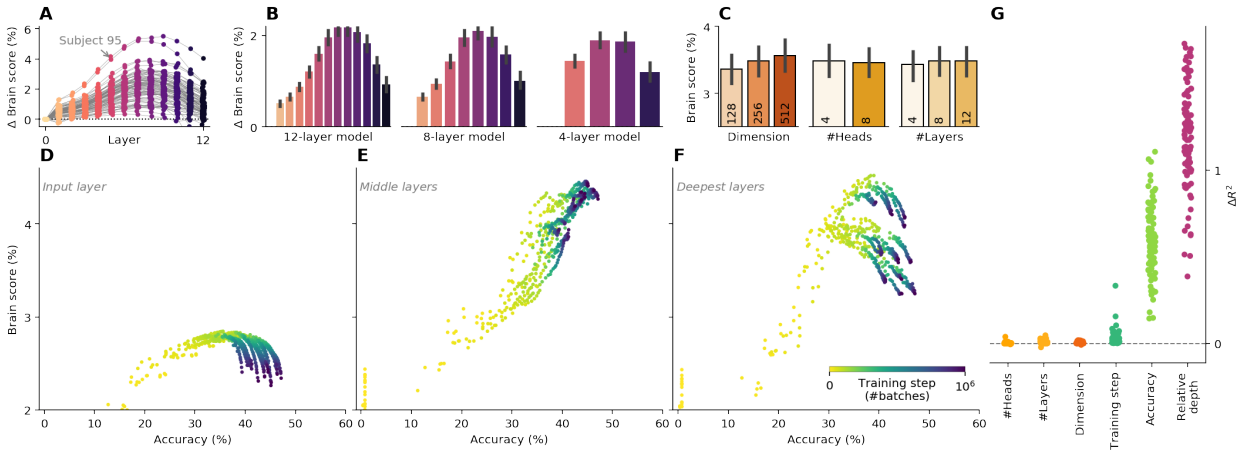

Figure 5: A. Gain in brain score (y-axis) between the representation extracted from each layer of a 13-layer CLT (x-axis) and its input layer (average across all time samples). Each line depicts a subject. B. Same as (A) showing the average effect across subjects for clarity (left), as well as the average gain in brain scores obtained with 8-layer and 4-layer CLT. C. Brain scores averaged over time and models as a function of their dimensionality (left), number attention heads (middle) and total number of layers (right) D. Brain scores (averaged time and subjects, y-axis) for the first (non-contextualized) layer of 594 CLTs (18 architectures x 33 color-coded training steps). Each dot corresponds to the averaged brain score of a frozen model. E. Same as D using the middle layers of the CLTs ($\frac{n_{layers}}{3} < layer \leq 2\frac{n_{layers}}{3}$). F. Same as D using the deepest layers of the CLTs ($layer > 2\frac{n_{layers}}{3}$). G. Feature importance estimated, for each subject separately (dots, n=95), with a random forest fitted to predict the average brain score from the model's hyperparameters (number of attention heads, total number of layers, and dimensionality of its layers), training step and performance, as well as from the layer relative depth of the layer used to compute this brain score. Error bars are the 95% confidence intervals of the scores' distribution across subjects.

higher brain scores than word embeddings especially one second after word onset. These improvements in brain scores peaked in the bilateral temporal and prefrontal cortices, especially around the infero-frontal cortex, anterior temporal lobe and middle and superior temporal gyrus. Overall these results show that we can precisely track the formation of visual, lexical and compositional representations in the expected brain regions (Price, 2010; Dehaene and Cohen, 2011; Hagoort, 2019).

Middle layers converge towards brain responses while outer layers diverge away from them

Is the gain in brain score obtained with the middle layer of a 13-layer CLT fortunate or, on the contrary, does it reflect a convergence of these modern language models towards brain-like representations? To address this issue, we trained 18 CLT varying in dimensionality ($n \in 128, 256, 512$), number of attention heads ($n \in 4, 8$) and total number of layers ($n \in 4, 8, 12$) but all trained on the same Wikipedia corpus. We froze each of these networks at different training stages (cf. Methods), input them with the word sequences that were presented to the subjects, and extracted their representations from each of their layers in order to estimate the corresponding brain scores.

First, we observed that, at the end of training, the middle layers weakly but systematically outperformed outer layers (e.g. first and last): e.g. the difference in brain scores between the middle and output layers of a 13-layer CLT was above chance ($\Delta R = 1\%$, ± 0.06 across subjects, $p < 10^{-29}$) and this difference holds across architectures ($\Delta R = 0.75\% \pm 0.15$ across architectures, $p < 10^{-7}$ as assessed with a second-level Wilcoxon test across subjects, Figure 5A-B).

Second, we investigated how the brain score of each layer varied with the training and the accuracy of the artificial neural network, to predict the word following a context (cf. Methods, Supplementary Figure 3). On average, the brain scores of the input layer increased within the 400k first training steps ($\approx 80M$ processed words), but ultimately presented a steady decrease over training steps (Figure 3): Pearson correlation between training steps and brain scores: R=-63% $\pm 3.1\% p < 10^{-34}$, Figure 5 D). The brain scores of the deepest layers (layer above the 66^{th} percentile of the model's depth) also rapidly started to decrease with training: $-35\% \pm 2.6\%$ correlation after 2 epochs, $p < 10^{-21}$ (Figure 5 F). By contrast, the brain scores of the middle layers (layer in between the 33^{th} and the 66^{th} percentile of the model's depth) systematically increased with the training of the model ($R = 33\% \pm 1.5\%, p < 10^{-35}$, Figure 5 E).

To summarize how brain scores are independently modulated by the models architecture, their training and the relative depth of their representations, we implemented a feature analysis across models based on a random forest (Figure 5G). These results confirmed that the relative depth of the extracted representation ($\Delta R^2 = 120\% \pm 3\%, p < 10^{-16}$), the performance of the models on language modeling ($\Delta R^2 = 59\% \pm 2\%, p < 10^{-16}$) and the network's amount of training ($\Delta R^2 = 2\% \pm 0.4\%, p < 10^{-16}$) had each a major impact on brain scores (Figure 5H). By contrast, the dimensionality of the layers and total number of layers modestly influenced the brain scores ($\Delta R^2 < 0.3\%$).

Overall, these results suggest that, beyond the marginal effects of the models' architectures, the middle - but not the outer - layers of deep language models systematically converge towards brain-like representations.

3 Discussion

Whether modern A.I. algorithms converge to the computational solutions of the human brain, or whether they simply find far-out tricks to superficially mimic human behavior is largely unknown. Following recent achievements in electrophysiology (Yamins et al., 2014; Tang et al., 2018) and neuroimaging (Khaligh-Razavi and Kriegeskorte, 2014; Kriegeskorte, 2015; Güçlü and van Gerven, 2015; Eickenberg et al., 2017; Huth et al., 2016; Toneva and Wehbe, 2019; Kell et al., 2018), we here tackled this challenge on the restricted issue of word and sentence processing by assessing how brain responses to random word lists (Figure 2, 3) and meaningful sentences (Figures 4 and 5, Video 2) variably correlated with different (parts of) artificial neural networks depending on their architectures, training and language performances.

Our results reveal how and when the reading network transforms written words into visual, lexical and compositional representations. In particular, the sequential recruitment of the early visual cortex and of the fusiform gyrus match and extend previous fMRI and electro-cortigography studies (Dehaene and Cohen, 2011; Hermes et al., 2017). Furthermore, the brain scores obtained with word embeddings delineate a neural code of word semantics distributed over the frontal, temporal and parietal cortices similarly to what has been reported in recent fMRI studies (Mitchell et al., 2008; Jain and Huth, 2018; Toneva and Wehbe, 2019). Finally, the compositional representations of deep language models peaked precisely in the brain regions traditionally associated with high-level sentence processing (Pallier et al., 2011; Hickok and Poeppel, 2007; Brennan and Pylkkänen, 2017). As expected (Fedorenko et al., 2010; Cogan et al., 2014), most of these effects appeared left-lateralized, but significantly recruited both hemispheres. As usual, such MEG source estimates, however, should be considered with parsimony (Baillet, 2017). In particular, the ventro-medial prefrontal and anterior cingulate cortices presently observed (Video 2) are specific to language processing. Their responses could thus relate to cognitive processes, that are not specific to language, such as emotional evaluation and working memory load.

Critically, we found that the middle - but not the outer - layers of modern deep language models become increasingly similar to brain activity as they learn to accurately predict the word that should follow a given context. This suggests that the way language models learn to *combine* - as opposed to *represent* - words converges to a brain-like solution. We should stress, however, that this convergence is almost certainly partial. First, current language models are still far from human-level performance on a variety of tasks such as dialogue, summarization, and systematic generalization (Loula et al., 2018; Zellers et al., 2019). In addition, they can be disrupted by adversarial examples (Nie et al., 2019). Finally, the architecture of the popular transformer network (Vaswani et al., 2017) is in many ways *not* biologically plausible: while the brain has been repeatedly associated with a predictive coding architecture, where prediction errors are computed at each level of an interconnected hierarchy of recurrent networks (Friston, 2010), transformers access an unreasonably large buffer of words through an attentionally-gated set of "duplicated" feedforward networks, which together only compute prediction errors at their final layer. In light of these major architectural differences, it is thus remarkable to see that the brain and the middle layers of these models find a partially common solution to language processing.

The reason why outer layers fail to converge to brain-like representations surprised us, as word embedding have been repeatedly used to model brain activity (Mitchell et al., 2008; Sassenhagen and Fiebach, 2019; Huth et al., 2016). We can speculate that this phenomenon results from the way language models are trained. Specifically, language models aim to predict words from other words. Presumably, this artificial task is best achieved by learning syntax, pragmatics as well as general knowledge, which, together can help transform a series of words into a meaningful construct - meaningful construct which can then be used to constrain the probability of subsequent words. By contrast, the human brain may use a different objective (Richards et al.; Friston, 2010): e.g. maximizing the information exchange during dialogue, predicting the trajectory of a narrative or minimizing surprise at each the level of representational hierarchy - as opposed to minimizing lexical surprise. These alternative objectives may happen to also necessitate words to be combined into meaningful constructs: i.e. to necessitate going through representations that are similar to the networks'. While speculative, this issue highlights the necessity, both for A.I. and cognitive neuroscience, to explore further *what*

the brain aims to achieve during language processing (i.e. finding the learning objective), rather than *how* it achieves it (i.e. finding the representations necessary to achieve this goal).

Finally, the present study only *starts* to elucidate the precise nature of linguistic representations in the brains and artificial neural network. Similarly, it only *starts* to unravel the complex interplay between the regions of the language network. How the mind builds and organizes its lexicon and how it parses and manipulate sentences thus remain open questions. We hope that the present workwill serve as a stepping stone to progress on these historical questions.

4 Methods

4.1 Models

We aimed to compare brain activity to three families of models, targeting visual, lexical and compositional representations, respectively.

Visual Network

To model visual representations, every word presented to the subjects were 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 (Baek et al., 2019), resulting in 888-dimensional embeddings. These embeddings aim to replicate previous work on the similarity between visual neural networks and brain activity in response to images (e.g. (Yamins et al., 2014; Kriegeskorte, 2015; Güçlü and van Gerven, 2015)) while ensuring that the stimuli input to the network are similar to those used for their training.

Word embedding networks

To model lexical representations, every word presented to the subjects was lower-cased and input to Word2Vec models (Mikolov et al., 2013). Word2Vec consists of a one-hidden-layer neural network stacked on a look-up-table (also called embedding layer). Depending on the learning objective, they were trained to predict the current word w_t given its context $w_{t-k}, ...w_{t-2}, w_{t-1}, w_{t+1}, w_{t+2}, ...w_{t+k}$ (Continuous Bag-Of-Words, CBOW), or the context given the current word (a.k.a Skip-gram). The models were trained using Gensim implementation (Řehůřek and Sojka, 2010) (hyper-parameters: batch size of 50,000 words, 0.001 sampling, hierarchical soft-max, i.e without negative sampling), on 288,450,116 lower-cased, tokenized (regexp tokenizer from Gensim) words from Wikipedia, without punctuation. We restricted the vocabulary to the stimuli words used in the present MEG study and to the words that appeared at least 5 times in the training corpus (resulting in $\approx$ 820K distinct vocabulary words in total).

To evaluate the networks' performance, we computed their loss (negative-log-likelihood) on a test dataset of 458,622 words from Wikipedia. Losses between CBOW and Skip-Gram with various context sizes are hardly comparable because the training task differ. We thus reported (3), for each network, its performance stage (%), from worst (0%) to the best performance (100%).

In total, we investigated 18 Word2Vec architectures varying in learning objective (Skipgram vs CBOW), dimensionality ($\in [128, 256, 512]$) and context size during training ($k \in [2, 4, 6]$) for three different random seeds. We froze each network at different training stages: 130 steps between initialisation and epoch 1,000. Networks were stopped after 3 days of training on 6 CPUs independently of their training stages, resulting in 6,889 models of non-contextualized word embeddings in total (see Supplementary Information for model details).

Causal Language Transformer

To model compositional representations (in the sense of non-linear interactions between words), we extracted the contextualized word representation of a variety of transformers trained on language modeling (Vaswani et al., 2017). We restricted our analyses to *causal* networks (i.e unidirectional, processing words from left to right) to facilitate the interpretation of the model-to-brain mapping. For example, given the sentence 'THE CAT IS ON THE MAT', the brain responses to 'ON' are solely compared to the activations of CLT input with 'THE CAT IS ON'. CLT representations were generated for every word, by generating word sequences from the three previous sentences. We did not observe qualitatively different results when using longer contexts. Note that sentences were isolated, and were not part of a narrative.

Algorithms were trained on a language modelling task using XLM implementation (see Supplementary Information for details on the hyper-parameters) (Lample and Conneau, 2019), on the same Wikipedia corpus of 278,386,651 Wikipedia

words extracted using WikiExtractor [1] and pre-processed using Moses tokenizer (Koehn et al.), with punctuation. We restricted the vocab to the 50,000 most frequent words, concatenated with all of the words used in the study (50,341 vocabulary words in total). These design choices enforces that the difference in brain scores observed across models cannot be explained by differences in corpora and text preprocessing.

To evaluate the networks' performance on language modeling, we computed their perplexity (exponential of the entropy) and their accuracy (accuracy at predicting the current word given previous words) on a test dataset of 180,883 words from Wikipedia. Note that we had to use a larger test dataset to evaluate Word2Vec networks because of the loss was not stable.

In total, we investigated 18 distinct CLT architectures varying in dimensionality ($\in [128, 256, 512]$), number of layers ($\in [4, 8, 12]$) and attention heads ($\in [4, 8]$). We froze the networks at 33 training stages between initialisation and epoch 200. The CLTs were stopped after 3 days of training on 8 GPUs, resulting in 579 CLT models in total, and 5,223 contextualized word representations (one per layer).

4.2 Magneto-encephalography

One-hundred and two subjects performed a one-hour-long reading task while being recorded with a 257 CTF magneto-encephalography (MEG) scanner by Schoffolen and colleagues (Schoffelen et al., 2019). Words were flashed one at a time to the subjects, and grouped into sequences of 9 - 15 words ("word sequences"), for a total of approximately 2,700 words per subject. The inter-stimulus interval varied between 300 ms and 1,400 ms, and the inter-sequence interval consisted of 5s-long blank screen. Word sequences were either meaningful sentences, or random word lists. Sentences and word lists were blocked into five sequences. Twenty percents of the sequences were followed by a yes/no question about the content of the previous sentences (e.g. "Did grandma give a cookie to the girl?) and word lists (e.g. Was the word 'grandma' mentioned?) to ensure that subjects were paying attention.

The 257 MEG time series were pre-processed using the Python library MNE (Gramfort et al., 2014). Signals were band-passed filter between 0.1 and 40 Hz filtered using MNE default parameters and segmented between -0.5s to 2s relative to word onset.

Seven subjects were excluded from the analyses because of difficulties processing metadata.

4.2.1 Brain score

To assess the linear relationship between the MEG signals of one subject and the activations of one artificial neural network (e.g a word embedding, or the activations of the first layer of a CLT), we split the subject's MEG samples into K train-test folds (test folds=20% of words, K=5). For the analyses of word lists (section 2, 2, 2), train and test sets contained separate words (out-of-vocabulary cross-validation). For the analyses of 'sentences' (section 2, 2), train and test sets contained separate sentences. For each train fold, $|T|$ cross-validated ridge regressions were fitted to predict the MEG signal elicited t seconds after the word onset, given the corresponding word embedding as input ($t \in T$, $|T|$ varies between 6 and 180 depending on the experiment, cf. 4.3.1). We used the RidgeCV regressor from scikit-learn (Pedregosa et al., 2011), with penalization parameter varying between 10^{-3} and 10^{8} (20 values, logarithmically scaled) to reliably optimize hyperparameters on each training set.

We then evaluated the performance of the $K * |T|$ ridge regressions fits by computing the Pearson's correlation between predicted and actual MEG on test folds. We obtained $K * |T|$ correlation scores, corresponding to the ability the network's activations to correlate with brain activity at time step t, for fold k. We call "brain scores" the correlation scores averaged across folds.

4.2.2 Source reconstruction

To evaluate the anatomical location of these effects, we performed the brain scores on source-reconstructed MEG signals, by correlating the single-trial source estimates ("true" source) with the single-trial source predictions generated by the model-to-brain mapping regressions.

To this aim, Freesurfer (Fischl, 2012) was used to automatically segment the T1-weighted anatomical MRI of individual subjects, using the 'recon-all' pipeline. We then manually co-referenced the subjects' segmented skull with the head-markers digitized prior to the MEG acquisition. A single-layer forward model was made using MNE-Python's default parameters (Gramfort et al., 2014). Because of lack of empty-room recordings, the noise covariance matrix used for the inverse operator was estimated from the zero-centered 200ms of baseline MEG activity preceding word onset.

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

The average brain responses of Figure 1 C were computed as the square of the source-reconstruction of the average evoked related field across all words for each subject separately, then averaged across subjects and finally divided by their respective maxima, in order to highlight temporal differences. Region-of-interest analyses were selected from the PALS Brodmann' area atlas (Van Essen, 2005) and the Destrieux Atlas segmentation (Destrieux et al., 2010) to summarize these large-dimensional effects. Movie 1 displays the non-normalized sources. The model-to-brain mapping regressions fitted across the baseline-corrected sensors $Y_{True}^{sensors}$ were used to generated single-trial predictions at the sensors levels $Y_{Pred}^{sensors}$. Then, both $Y_{True}^{sensors}$ and $Y_{Pred}^{sensors}$ were projected onto the subjects' source space using a dSPM inverse operator with default parameters (e.g. fixed orientation, depth=0.8 and lambda2=1). Finally, the brain-score was computed from the correlation between $Y_{True}^{sources}$ and $Y_{Pred}^{sources}$.

4.3 Statistics

4.3.1 Convergence analysis

All neural networks but the visual CNN were trained from scratch on the same corpus (cf 4.1 and 4.1) and systematically computed the brain scores of their activations on each subject, sensor and time sample independently. For computational reasons, we restricted ourselves to six representative time samples regularly distributed between $\in [-400, 1,600]$ms. Brain scores were then averaged across channels, time samples and subjects to obtain the results in Figure 3 and 5. To evaluate the convergence of a model, we computed, for each subject, the Pearson's correlation between the brain scores of the network and its performance and/or its training step.

4.3.2 Feature importance

To systematically quantify how the architecture, the performance and the learning of the artificial neural networks impacted their ability to linearly correlate with brain activity, we fitted, for each subject separately, a random forest across the models' properties (e.g. dimensionality, training stage) to predict their brain scores, using scikit-learn's RandomForest (Breiman, 2001; Pedregosa et al., 2011). The performance of the random forests (R^2) was evaluated with a 5-fold cross-validation across models for each subject separately.

For Word2Vec embeddings (Figure 3), we used the learning objective (skip-gram versus continuous-bag-of-words), the context size during training ($\in \{2, 4, 6\}$, the dimensionality ($\in \{128, 256, 512\}$), the performance stage the model ($\in [0, 1]$, cf. Word embedding networks) and its training stage (epochs), as the input data to the random forest.

For CLT embeddings (Figure 5), we used the number of attention heads ($\in [4, 8]$), total number of layers ($\in [4, 8, 12]$), dimensionality ($\in [128, 256, 512]$), training step (epochs, $\in [0, 200]$), accuracy and the relative depth of the representation (between 0 and 1, $depth = \frac{layer}{n_{layers}}$), as the input data to the random forest.

In both case, we implemented a "feature importance" analyses to assess the contribution of each characteristic on brain-score. Feature importance measures ΔR^2: the decrease in R^2 when shuffling one feature repeatedly (here, $n_{repeats} = 100$). For each subject, we reported the average decrease across the cross-validation folds (Figure 3, 5). The resulting importance score (ΔR^2) observed per feature, are expected to be centered around 0 if the model property does not impact on brain-score, and positive otherwise. Note that the decrease in R2 can be greater than one when the model performs worse than chance (with negative R2) when shuffling a feature.

4.3.3 Population statistics

For clarity, most figures report average effect sizes obtained across subjects. To estimate the robustness of these estimates, 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 chance level.

Error bars and $\pm$ refer to standard of the mean intervals.

4.4 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

S. Abnar, R. Ahmed, M. Mijnheer, and W. H. Zuidema. Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity. *CoRR*, abs/1711.09285, 2017. URL http://arxiv.org/abs/1711.09285.

A. J. Anderson, E. C. Lalor, F. Lin, J. R. Binder, L. Fernandino, C. J. Humphries, L. L. Conant, R. D. S. Raizada, S. Grimm, and X. 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. ISSN 1047-3211. doi: 10.1093/cercor/bhy110. URL https://academic.oup.com/cercor/article/29/6/2396/4996559. Publisher: Oxford Academic.

N. Athanasiou, E. Iosif, and A. 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, Aug. 2018. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/C18-1243.

J. Baek, G. Kim, J. Lee, S. Park, D. Han, S. Yun, S. J. Oh, and H. 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.

S. Baillet. Magnetoencephalography for brain electrophysiology and imaging. *Nature neuroscience*, 20(3):327–339, 2017.

Y. Bengio, R. Ducharme, and P. 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. URL http://papers.nips.cc/paper/1839-a-neural-probabilistic-language-model.pdf.

J. R. Binder, J. A. Frost, T. A. Hammeke, R. W. Cox, S. M. Rao, and T. Prieto. Human brain language areas identified by functional magnetic resonance imaging. *Journal of Neuroscience*, 17(1):353–362, 1997.

J. R. Binder, R. H. Desai, W. W. Graves, and L. L. Conant. Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. 19(12):2767–2796, 2009. ISSN 1047-3211. doi: 10.1093/cercor/bhp055. URL https://academic.oup.com/cercor/article/19/12/2767/376348. Publisher: Oxford Academic.

I. Blank, Z. Balewski, K. Mahowald, and E. Fedorenko. Syntactic processing is distributed across the language system. *Neuroimage*, 127:307–323, 2016.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

L. Breiman. Random forests. 45(1):5–32, 2001. ISSN 1573-0565. doi: 10.1023/A:1010933404324. URL https://doi.org/10.1023/A:1010933404324.

J. R. Brennan and L. Pylkkänen. Meg evidence for incremental sentence composition in the anterior temporal lobe. *Cognitive science*, 41:1515–1531, 2017.

J. R. Brennan, E. P. Stabler, S. E. Van Wagenen, W.-M. Luh, and J. T. Hale. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and language*, 157:81–94, 2016.

C. Brodbeck, L. E. Hong, and J. Z. Simon. Rapid transformation from auditory to linguistic representations of continuous speech. *Current Biology*, 28(24):3976–3983, 2018.

N. Chomsky. *Language and mind*. Cambridge University Press, 2006.

G. B. Cogan, T. Thesen, C. Carlson, W. Doyle, O. Devinsky, and B. Pesaran. Sensory–motor transformations for speech occur bilaterally. *Nature*, 507(7490):94–98, 2014.

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

C. Destrieux, B. Fischl, A. Dale, and E. Halgren. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *Neuroimage*, 53(1):1–15, 2010.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805, 2018. URL http://arxiv.org/abs/1810.04805.

G. M. Di Liberto, J. A. O'Sullivan, and E. C. Lalor. Low-frequency cortical entrainment to speech reflects phoneme-level processing. *Current Biology*, 25(19):2457–2465, 2015.

P. W. Donhauser and S. Baillet. Two distinct neural timescales for predictive speech processing. *Neuron*, 2019.

M. Dummett. *Frege: Philosophy of language*. Harvard University Press, 1981.

M. Eickenberg, A. Gramfort, G. Varoquaux, and B. Thirion. Seeing it all: Convolutional network layers map the function of the human visual system. *NeuroImage*, 152:184–194, 2017.

E. Fedorenko, P.-J. Hsieh, A. Nieto-Castañón, S. Whitfield-Gabrieli, and N. Kanwisher. New method for fmri investigations of language: defining rois functionally in individual subjects. *Journal of neurophysiology*, 104(2): 1177–1194, 2010.

E. Fedorenko, T. L. Scott, P. Brunner, W. G. Coon, B. Pritchett, G. Schalk, and N. Kanwisher. Neural correlate of the construction of sentence meaning. *Proceedings of the National Academy of Sciences*, 113(41):E6256–E6262, 2016.

E. Fedorenko, I. Blank, M. Siegelman, and Z. Mineroff. Lack of selectivity for syntax relative to word meanings throughout the language network. *bioRxiv*, page 477851, 2020.

B. Fischl. Freesurfer. *Neuroimage*, 62(2):774–781, 2012.

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

J. Gauthier and A. Ivanova. Does the brain represent words? an evaluation of brain decoding studies of language understanding. *arXiv preprint arXiv:1806.00591*, 2018.

A. Gramfort, M. Luessi, E. Larson, D. A. Engemann, D. Strohmeier, C. Brodbeck, L. Parkkonen, and M. S. Hämäläinen. Mne software for processing meg and eeg data. *NeuroImage*, 86:446 – 460, 2014. ISSN 1053-8119. doi: https://doi.org/10.1016/j.neuroimage.2013.10.027. URL `http://www.sciencedirect.com/science/article/pii/S1053811913010501`.

U. Güçlü and M. A. 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.

P. Hagoort. On broca, brain, and binding: a new framework. *Trends in Cognitive Sciences*, 9:416–423, 2005.

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

D. Hermes, V. Rangarajan, B. L. Foster, J.-R. King, I. Kasikci, K. J. Miller, and J. Parvizi. Electrophysiological responses in the ventral temporal cortex during reading of numerals and calculation. *Cerebral cortex*, 27(1):567–575, 2017.

G. Hickok and D. Poeppel. The cortical organization of speech processing. 8(5):393–402, 2007. ISSN 1471-0048. doi: 10.1038/nrn2113. URL `https://www.nature.com/articles/nrn2113`. Number: 5 Publisher: Nature Publishing Group.

A. G. Huth, W. A. de Heer, T. L. Griffiths, F. E. Theunissen, and J. L. Gallant. Natural speech reveals the semantic maps that tile human cerebral cortex. 532(7600):453–458, 2016. ISSN 0028-0836, 1476-4687. doi: 10.1038/nature17637. URL `http://www.nature.com/articles/nature17637`.

S. Jain and A. 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. URL `http://papers.nips.cc/paper/7897-incorporating-context-into-language-encoding-models-for-fmri.pdf`.

A. Kell, D. Yamins, E. Shook, S. Norman-Haignere, and J. McDermott. A task-optimized neural network replicates human auditory behavior, predicts brain responses, and reveals a cortical processing hierarchy. *Neuron*, 98, 04 2018. doi: 10.1016/j.neuron.2018.03.044.

S.-M. Khaligh-Razavi and N. Kriegeskorte. Deep supervised, but not unsupervised, models may explain it cortical representation. *PLoS computational biology*, 10(11):e1003915, 2014.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. 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. URL `https://www.aclweb.org/anthology/P07-2045`.

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

G. Lample and A. Conneau. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

J. Loula, M. Baroni, and B. M. Lake. Rearranging the familiar: Testing compositional generalization in recurrent networks. *CoRR*, abs/1807.07545, 2018. URL http://arxiv.org/abs/1807.07545.

N. Mesgarani, C. Cheung, K. Johnson, and E. F. Chang. Phonetic feature encoding in human superior temporal gyrus. 343(6174):1006–1010, 2014. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1245994. URL https://science.sciencemag.org/content/343/6174/1006. Publisher: American Association for the Advancement of Science Section: Report.

T. Mikolov, K. Chen, G. Corrado, and J. Dean. Efficient estimation of word representations in vector space. 2013. URL http://arxiv.org/abs/1301.3781.

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. ISSN 0036-8075, 1095-9203. doi: 10.1126/science.1152876. URL https://www.sciencemag.org/lookup/doi/10.1126/science.1152876.

F. Mollica, M. Siegelman, E. Diachek, S. T. Piantadosi, Z. Mineroff, R. Futrell, H. Kean, P. Qian, and E. Fedorenko. Composition is the core driver of the language-selective network. 1(1):104–134. doi: 10.1162/nol_a_00005. URL https://doi.org/10.1162/nol_a_00005. Publisher: MIT Press.

Y. Nie, A. Williams, E. Dinan, M. Bansal, J. Weston, and D. Kiela. Adversarial nli: A new benchmark for natural language understanding. *arXiv preprint arXiv:1910.14599*, 2019.

C. Pallier, A.-D. Devauchelle, and S. Dehaene. Cortical representation of the constituent structure of sentences. *Proceedings of the National Academy of Sciences*, 108(6):2522–2527, 2011. ISSN 0027-8424. doi: 10.1073/pnas.1018711108. URL https://www.pnas.org/content/108/6/2522.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, et al. Scikit-learn: Machine learning in python. *Journal of machine learning research*, 12(Oct): 2825–2830, 2011.

J. Pennington, R. Socher, and C. D. Manning. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, 2014. URL http://www.aclweb.org/anthology/D14-1162.

C. 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.

L. Pylkkanen and J. R. Brennan. Composition: The neurobiology of syntactic and semantic structure building. 2019.

L. Pylkkänen and A. Marantz. Tracking the time course of word recognition with meg. *Trends in cognitive sciences*, 7 (5):187–189, 2003.

A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

S. Reddy Oota, N. Manwani, and B. Raju S. fMRI Semantic Category Decoding using Linguistic Encoding of Word Embeddings. *arXiv e-prints*, art. arXiv:1806.05177, Jun 2018.

R. Řehůřek and P. Sojka. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta, May 2010. ELRA. http://is.muni.cz/publication/884893/en.

B. A. Richards, T. P. Lillicrap, P. Beaudoin, Y. Bengio, R. Bogacz, A. Christensen, C. Clopath, R. P. Costa, A. de Berker, S. Ganguli, C. J. Gillon, D. Hafner, A. Kepecs, N. Kriegeskorte, P. Latham, G. W. Lindsay, K. D. Miller, R. Naud, C. C. Pack, P. Poirazi, P. Roelfsema, J. Sacramento, A. Saxe, B. Scellier, A. C. Schapiro, W. Senn, G. Wayne, D. Yamins, F. Zenke, J. Zylberberg, D. Therien, and K. P. Kording. A deep learning framework for neuroscience. 22(11):1761–1770. ISSN 1097-6256, 1546-1726. doi: 10.1038/s41593-019-0520-2. URL http://www.nature.com/articles/s41593-019-0520-2.

Y.-P. Ruan, Z.-H. Ling, and Y. 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, Nov. 2016. Association for Computational Linguistics. doi: 10.18653/v1/D16-1064. URL https://www.aclweb.org/anthology/D16-1064.

J. Sassenhagen and C. J. Fiebach. Traces of meaning itself: Encoding distributional word vectors in brain activity. *bioRxiv*, 2019. doi: 10.1101/603837. URL https://www.biorxiv.org/content/early/2019/04/09/603837.

J.-M. Schoffelen, R. Oostenveld, N. Lam, J. Udden, A. Hultén, and P. Hagoort. A 204-subject multimodal neuroimaging dataset to study language processing. *Scientific Data*, 6, 12 2019. doi: 10.1038/s41597-019-0020-y.

C. Tang, L. Hamilton, and E. Chang. Intonational speech prosody encoding in the human auditory cortex. *Science*, 357 (6353):797–801, 2017.

H. Tang, M. Schrimpf, W. Lotter, C. Moerman, A. Paredes, J. O. Caro, W. Hardesty, D. Cox, and G. Kreiman. Recurrent computations for visual pattern completion. *Proceedings of the National Academy of Sciences*, 115(35):8835–8840, 2018.

M. Toneva and L. Wehbe. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). *CoRR*, abs/1905.11833, 2019. URL `http://arxiv.org/abs/1905.11833`.

A. M. Turing. Computing machinery and intelligence. In *Parsing the Turing Test*, pages 23–65. Springer, 2009.

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

A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin. Attention is all you need. In *NIPS*, 2017.

D. L. Yamins and J. J. DiCarlo. Using goal-driven deep learning models to understand sensory cortex. *Nature neuroscience*, 19(3):356, 2016.

D. L. Yamins, H. Hong, C. F. Cadieu, E. A. Solomon, D. Seibert, and J. 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.

R. Zellers, A. Holtzman, Y. Bisk, A. Farhadi, and Y. Choi. Hellaswag: Can a machine really finish your sentence? *arXiv preprint arXiv:1905.07830*, 2019.