Pasquiou, A., Lakretz, Y., Thirion, B. & Pallier, C. (2023). Information-Restricted Neural Language Models Reveal Different Brain Regions' Sensitivity to Semantics, Syntax and Contex. Neurobiology of Language, Advance publication. https://doi.org/10.1162/nol_a_00125.

Information-Restricted Neural Language Models Reveal Different Brain Regions' Sensitivity to Semantics, Syntax and Context

Alexandre Pasquiou[1, 2], Yair Lakretz[1], Bertrand Thirion[2], and Christophe Pallier[1]

[1]UNICOG, Cognitive Neuroimaging Unit, INSERM, CEA, Neurospin, Gif-sur-Yvette, France

[2]MIND, INRIA, CEA, Neurospin, Gif-sur-Yvette, France

Author Note

Alexandre Pasquiou https://orcid.org/0000-0002-7966-8083

Correspondence concerning this article should be addressed to Alexandre Pasquiou, MIND, INRIA, CEA, Neurospin, Gif-sur-Yvette, France. E-mail: alexandre.pasquiou@inria.fr

Abstract

A fundamental question in neurolinguistics concerns the brain regions involved in syntactic and semantic processing during speech comprehension, both at the lexical (word processing) and supra-lexical levels (sentence and discourse processing). To what extent are these regions separated or intertwined? To address this question, we introduce a novel approach exploiting neural language models to generate high-dimensional feature sets that separately encode semantic and syntactic information. More precisely, we train a lexical language model, Glove, and a supra-lexical language model, GPT-2, on a text corpus from which we selectively removed either syntactic or semantic information. We then assess to what extent the features derived from these information-restricted models are still able to predict the fMRI time-courses of humans listening to naturalistic text. Furthermore, to determine the windows of integration of brain regions involved in supra-lexical processing, we manipulate the size of contextual information provided to GPT-2. The analyses show that, while most brain regions involved in language comprehension are sensitive to both syntactic and semantic features, the relative magnitudes of these effects vary across these regions. Moreover, regions that are best fitted by semantic or syntactic features are more spatially dissociated in the left hemisphere than in the right one, and the right hemisphere shows sensitivity to longer contexts than the left. The novelty of our approach lies in the ability to control for the information encoded in the models' embeddings by manipulating the training set. These "information-restricted" models complement previous studies that used language models to probe the neural bases of language, and shed new light on its spatial organization.

Keywords: fMRI; encoding models; syntax; context; semantics; LLM

Information-Restricted Neural Language Models Reveal Different Brain Regions' Sensitivity to Semantics, Syntax and Context

Author summary

The study of how the brain processes language, and in particular sentence meaning (semantics) and structure (syntax), is crucial for improving language therapies and technologies. Yet, identifying the specific brain regions involved in these processes is challenging because they are interconnected. To solve this problem, we developed an approach that produces Artificial Intelligence (AI) models of different linguistic aspects of a story, like semantic or syntactic information. We first trained AI models on texts from which we removed either syntactic or semantic information. We then compared the fMRI brain activity of individuals listening to a story with the syntactic or semantic AI-derived models on the same story. If these models align with a specific brain region, then we assume that the model and the region likely encode the same information. By doing this, we identified brain regions involved in processing word meaning, sentence structure, and contextual information. The novelty of our method lies in the ability to control for the information inside models. Our approach complements previous studies that used AI models to find the brain regions processing language, and sheds new light on their spatial organization.

Introduction

Understanding the neural bases of language processing has been one of the main research efforts in the neuroimaging community for the past decades (see, e.g., Binder et al., 2009; Friederici, 2011, for reviews). However, the complex nature of language makes it difficult to discern how the various processes underlying language processing are topographically and dynamically organized in the human brain, and therefore many questions remain open to this date.

One central open question is whether semantic and syntactic information are encoded and processed jointly or separately in the human brain. Language comprehension

requires to access word meanings (lexical semantics), but also to compose these meanings to construct the meaning of entire sentences. In languages such as English, the meaning of a sentence strongly depends on word order – for example, 'The boy kissed the girl' has a different meaning than 'The girl kissed the boy' although both sentences contain the exact same words (in other languages, inflectional cues rather than word order signal the roles of words in the sentence). Importantly, meaning construction of new sentences would be roughly done in the same way if only the structure of the sentences remains the same ('The X kissed the Y'), independently of the lexical meanings of the single nouns in the sentences ('boy' and 'girl'). This combinatorial property of language allows to construct meanings of sentences that we have never heard before and suggests that it might be computationally advantageous for the brain to have developed neural mechanisms for composition that are separate from those dedicated to the processing of lexico-semantic content. Such neural mechanisms for composition would be sensitive to only the abstract structure of sentences and would implement the syntactic rules according to which sentence parts should be composed.

Following related considerations, the dominant view over the past decades claimed that syntactic information is represented and processed in specialized brain regions, akin to the classic modular view (Chomsky, 1984; Fodor, 1983). Neuronal modularity of language processing gained support from early lesion studies suggesting that syntactic processing takes place in localized and specialized brain regions such as Broca's area, showing double dissociations between syntactic and semantic processing (Caramazza & Zurif, 1976; Goodglass, 1993). Neuroimaging studies (Embick, 2000; Friederici et al., 2017; Friederici et al., 2006; Garrard et al., 2004; Grodzinsky & Santi, 2008; Hagoort, 2014; Hashimoto & Sakai, 2002; Matchin & Hickok, 2020; Pallier et al., 2011; Shetreet & Friedmann, 2014; Vigliocco, 2000) as well as simulation work on language acquisition and processing in artificial neural language models (Lakretz et al., 2021; Lakretz et al., 2019; O'Reilly & Frank, 2006; Russin et al., 2019; Ullman, 2004) have provided further support to this view

since then.

However, in parallel, an opposing view has argued that semantics and syntax are processed in a common distributed language processing system (Bates & Dick, 2002; Bates & MacWhinney, 1989; Dick et al., 2001). Recent work in support of this view has raised concerns regarding the replicability of some of the early results from the modular view (Siegelman et al., 2019) and provided evidence that semantic and syntactic processing in the language network might not be so easily dissociated from one another (Fedorenko et al., 2020; Mollica et al., 2018).

Neuroimaging studies, cited to defend one or the other view, have mainly relied on one of two methodological approaches: on the one hand, controlled experimental paradigms, which manipulate the words or sentences (Bottini et al., 1995; Caplan et al., 1998; Mazoyer et al., 1993; Pallier et al., 2011; Stromswold et al., 1996) and, on the other hand, naturalistic paradigms that make use of stimuli closer to what one could find in a daily environment. The former approach probes linguistic dimensions in one of the following ways: varying the presence or absence of syntactic or semantic information (Friederici, Kotz, et al., 2009; Friederici et al., 2003) or varying the syntactic structure difficulty or the semantic interpretation difficulty (e.g. Cooke et al., 2001; Friederici, Makuuchi, et al., 2009; Kinno et al., 2007; Newman et al., 2010; Santi & Grodzinsky, 2010). However, the conclusions from such studies may be bounded to the peculiarity of the task and setup used in the experiment (Nastase et al., 2020). To overcome these shortcomings, over the last years, researchers have become increasingly interested in data using "Ecological Paradigms", in which participants are engaged in more natural tasks, such as conversation or story listening (LeBel et al., 2022; Lerner et al., 2011; Nastase et al., 2021; Pasquiou et al., 2022; Regev et al., 2013; Wehbe et al., 2014). This avoids any task-induced bias and takes into consideration both lexical and supra-lexical levels of syntax and semantic processing. Integrating supra-lexical level information is essential for understanding language processing in the brain, because the lexical-semantic information

of a word and the resulting semantic compositions depend on its context.

More recently, following advances in natural language processing, neural language models have been increasingly employed in the analysis of data collected from ecological paradigms. Neural language models are models based on neural networks, which are trained to capture joint probability distributions of words in sentences using next-word, or masked-word prediction tasks Devlin et al. (e.g. 2019), Elman (1991), Pennington et al. (2014), and Radford et al. (2019). By doing so, the models have to learn semantic and syntactic relations among word tokens in the language. To study brain data collected from ecological paradigms, neural language models are presented with the same sentence stimuli, then, their activations (aka, embeddings) are extracted and used to fit and predict the brain data (Caucheteux & King, 2022; Huth et al., 2016; Pasquiou et al., 2022; Wehbe et al., 2014). This approach has led to several discoveries, such as wide networks associated with semantic processing uncovered by Huth et al. (2016) using word embeddings (see also Pereira et al. (2018a)), or context-sensitivity maps discovered by Jain and Huth (2018) and Toneva and Wehbe (2019).

Despite these advances and extensive neuroscientific and cognitive explorations, the neural bases of semantics, syntax and the integration of contextual information still remain debated. In particular, a central puzzle remains in the field: on the one hand, studies investigating syntax and semantics found vastly distributed networks when using naturalistic stimuli (Caucheteux et al., 2021; Fedorenko et al., 2020) and others found more localized activations for syntax, typically in inferior frontal and posterior temporal regions, when using constrained experimental paradigms e.g., (Matchin et al., 2017; Pallier et al., 2011). Thus, whether there is a hierarchy of brain regions integrating contextual information or the extent to which syntactic information is independently processed from semantic information, in at least some brain regions, remains largely debated to date.

So far, insights from neural language models about this central puzzle were also rather limited. This is mostly due to the complexity of the models in terms of size, training

and architecture. This complexity makes it difficult to identify how and what information is encoded in their latent representations, and how to use their embeddings to study brain function.

Caucheteux et al. (2021) used a neural language model, GPT-2, in an novel way to separate semantic and syntactic processing in the brain. Specifically, using a pre-trained GPT-2 model, they built syntactic predictors by averaging the embeddings of words from sentences that shared syntactic but no semantic properties, and used them to identify syntactic-sensitive brain regions. They defined as semantic-sensitive brain regions, the regions that were better predicted by the GPT-2's embeddings computed on the original text, compared to the syntactic predictors. They observed that syntax and semantics, defined in this way, rely on a common set of distributed brain areas.

Jain and Huth (2018) used pre-trained LSTM models to study context integration. They varied the amount of context used to generate word embeddings, and obtained a map indicating brain regions' sensitivity to different sizes of context.

Here, we propose a new approach to tackle the questions of syntactic vs. semantic processing and contextual integration, by fitting brain activity with word embeddings derived from *information-restricted* models. By this, we mean that the models are trained on text corpora from which specific types of information (syntactic, semantic, or contextual) were removed. We then assess the ability of these information-restricted models to fit brain activations, and compared it to the predictive performance of a neural model trained on the integral dataset.

More precisely, we created a text corpus of novels from the Gutenberg Project (http://www.gutenberg.org) and used it to define three different sets of features: (i) *Integral features*, the full text from the corpus (ii) *Semantic features*, the content words from the corpus; (iii) *Syntactic features*, where each word and punctuation sign from the corpus is replaced by syntactic characteristics. We then trained two types of models on each feature space: a non-contextual model, Glove (Pennington et al., 2014), and a

contextual model, GPT-2 (Radford et al., 2019) (See Fig. 1A). The text transcription of the audio-book, that participants listened to in the scanner, was then presented to the neural language models from which we derived embedding vectors. After fitting these embedded representations to fMRI brain data with linear encoding models, we computed the cross-validated correlations between the encoding models' predicted time courses and the observed time-series. In a first set of analyses, this allowed us to quantify the sensitivity to syntactic and semantic information in each voxel (Fig. 1B). In a second set of analyses, we identified brain regions integrating information beyond the lexical level. We first compared the contextual model (GPT-2) and the non contextual model (Glove), before investigating the brain regions processing short (5 words), medium (15 words) and long (45 words) contexts, using a non-contextualized GloVe model as a 0-context baseline (See Fig. 1C.).

Methods and Materials

Creation of datasets; Semantic, Syntactic and Integral features

We selected a collection of English novels from Project Gutenberg (www.gutenberg.org; data retrieved on February 21, 2016). This *integral dataset* comprised 4.4GB of text for training purposes and 1.1GB for validation. From it, we created two information-restricted datasets: the *semantic dataset* and the *syntactic dataset*. In the *semantic dataset*, only content words were kept, while all grammatical, function words and punctuation signs were filtered out. In the *syntactic dataset*, each token (word or punctuation sign) was replaced by an identifier encoding a triplet (POS, Morph, NCN) where POS is the Part-of-speech computed using Spacy (Honnibal & Montani, 2017), Morph corresponds to the morphological features obtained from Spacy and NCN stands for the Number of Closing Nodes in the parse tree, at the current token, computed using the Berkeley Neural Parser (Kitaev & Klein, 2018) available with Spacy.

In this paper, we refer to the content of the integral dataset as *integral features*, the content of the semantic dataset as *semantic features*, and the content of the syntactic dataset as *syntactic features*. Examples of integral, semantic and syntactic features are

given in Appendix 1-Models training.

GloVe Training

GloVe (Global Vectors for Word Representation) relies on the co-occurence matrix of words in a given corpus to generate fixed embedding vectors that capture the distributional properties of the words (Pennington et al., 2014). Using the open-source code provided by Pennington and al. (https://nlp.stanford.edu/projects/glove/), we trained GloVe on the three datasets (integral, semantic and syntactic), setting the context window size to 15 words, the embedding vectors' size to 768, and the number of training epochs to 20, until no further improvement on the validation set could be observed; convergence assessments are provided in Appendix 1-Convergence of the language models during training (Appendix 1-Fig.D2).

GPT-2 Training

GPT-2 (Generative Pretrained Transformer 2) is a deep learning transformer-based language model. We trained the open-source implementation GPT2LMHeadModel, provided by HuggingFace (Wolf, 2020), on the three datasets (integral, semantic and syntactic).

The GPT2LMHeadModel architecture is trained on a next-token prediction task using a CrossEntropyLoss and the Pytorch Python package (Paszke et al., 2019). The training procedure can easily be extended to any feature type by adapting both vocabulary size and tokenizer to each vocabulary. Indeed, the inputs given to GPT2LMHeadModel are ids encoding vocabulary items. All the analyses reported in this paper were performed with 4-layer models having 768 units per layer and 12 attention heads. As shown in (Pasquiou et al., 2022), these 4-layer models fit brain data nearly as well as the usual 12-layer models. We presented the models with input sequences of 512 tokens, and let the training run for 5 epochs until no further improvement on the validation set could be observed; convergence assessments are provided in Appendix 1-Convergence of the language models during training (Appendix 1-Fig.D1).

For the GTP-2 trained on the semantic features, small modifications had to be made to the model architecture in order to remove all residual syntax. By default, GPT-2 encodes the absolute positions of tokens in sentences. When training GPT-2 on the semantic features, as word ordering might contain syntactic information, we had to make sure that position information could not be leveraged by means of its positional embeddings, yet keeping information about word proximity as it influences semantics. We modified the implementation so that the GPT-2 trained on semantic features follows these specifications (see Appendix 1-Removing absolute position information in GPT-2 trained on semantic features).

Stimulus: The Little Prince story

The stimulus used to obtain activations from humans and from NLP models was *The Little Prince* novella. Humans listened to an audio-book version, spliced into 9 tracks that lasted approximately 11 minutes each (see Li et al., 2022). In parallel, NLP models were provided with an exact transcription of this audio-book, enriched with punctuation signs from the written version of the Little Prince. The text comprised 15,426 words and 4,482 punctuation signs. The acoustic onsets and offsets of the spoken words were marked to align the audio recording with the *The Little Prince* text.

Computing Embeddings from the Little Prince text

The tokenized versions of the Little Prince (one for each feature type) were run through Glove and GPT-2 in order to generate embeddings that could be compared with fMRI data.

For GloVe, we simply retrieved the fixed embedding vector learnt during training for each token.

For GPT-2, we retrieved the contextualized third layer hidden-state (aka embedding) vector for each token, so that the dimension is comparable to the dimension of GloVe's embeddings (768 units). Layer 3 (out of 4) was selected because it has been demonstrated that late middle layers of recurrent language models are best able to predict

brain activity (Jain & Huth, 2018; Toneva & Wehbe, 2019) (see Appendix 1-Fig.L1).

The embedding built by GPT-2 for a given token rely on the past tokens (aka past context). The bigger the past context, the more reliable the token embedding will be. We designed the following procedure to ensure that the embedding of each token used similar past context size: the input sequence was limited to a maximum of 512 tokens. The text was scanned with a sliding window of size $N = 512$ tokens, and a step of 1 token. The embedding vector of the next to last token (in the sliding window) was then retrieved. For the context-constrained versions of GPT-2 (denoted GPT-$2_{Context-k}$), the input text was formatted as the training data (see Fig.1C) in batches of input sequences of length $(k + 5)$ tokens (see Appendix 1-Context-limited models for examples), and only the embedding vector of the current token was retrieved. Embedding matrices were built by concatenating words embeddings. More precisely, calling d the dimension of the embeddings retrieved from of a neural model, corresponding to the number of units in one layer in our case, and w the total number of tokens in the text, we obtained an embedding matrix $\boldsymbol{X} \in \mathbb{R}^{w \times d}$ after the presentation of the entire text to the model.

Decoding of syntax and semantics categories from embeddings

We designed two decoding tasks: a syntax decoding task in which we tried to predict the triplet (Part-of-speech, morphological information and number of closing nodes) of each word from its embedding vector (355 categories), and a semantic decoding task in which we tried to predict each *content word*'s semantic category (from *Wordnet*, https://wordnet.princeton.edu/) from its embedding vector (837 categories).

We used Logistic Classifiers and the text of *The Little Prince*, which was split using a 9-fold cross-validation on runs, training on 8 runs and evaluating on the remaining one for each split. Dummy classifiers were fitted and used as estimations of chance-level for each task and model. It is crucial to acknowledge that the baseline performance level varies based on both the decoding task and the specific model employed. Specifically, the models trained on semantic features were exclusively trained on content words. Consequently,

when assessing the syntactic decoding accuracy of these models, only content words were considered, resulting in an elevated baseline performance level. Conversely, for the models trained on syntactic/integral features, the syntactic decoding accuracy encompasses the evaluation of all tokens. All classifiers implementations were taken from Scikit-Learn (Pedregosa et al., 2011).

MRI data

We used the functional Magnetic Resonance Imaging (fMRI) data of 51 English speaking participants who listened to an entire audio-book of The Little Prince during about one hour and a half. These data, available at https://openneuro.org/datasets/ds003643/versions/1.0.2 are described in details by Li et al. (2022). In short, the acquisition used echo-planar imaging (TR=2s; resolution=3.75x3.75x3.75mm) with a multi-echo (3 echos) sequence to optimize signal-to-noise (Kundu et al., 2018). Preprocessing comprised multi-echo independent components analysis (ME-ICA) to denoise data for motion, physiology and scanner artifacts, correction for slice-timing differences, and nonlinear alignment to the MNI template brain.

For each participant, there were 9 runs of fMRI acquisition representing about 10 minutes of brain activations each. We re-sampled the preprocessed individual scans at 4x4x4 mm (to reduce computation load) and applied linear detrending and standardization (mean removal and scaling to unit variance) to each voxel's time-series.

Finally, we computed a global brain mask to only keep voxels containing useful signal (using nilearn's *compute_epi_mask* function, we find the least dense point of the total image histogram) across all runs for at least 50% of all participants. This global mask contained 26,164 voxels at 4x4x4mm resolution. All analyses reported in this paper were performed within this global mask.

Correlations between embeddings and individual fMRI data

The embeddings ($\mathbf{X}$) derived from neural language models were mapped to each subject's fMRI activations ($\mathbf{Y}_s, s = 1..S$) following the pipeline outlined in Fig.1B.

The process, using the standard model-based encoding approach to modelling fMRI signals (Huth et al., 2016; Naselaris et al., 2011; Pasquiou et al., 2022), is detailed in Appendix 1-Mapping NLM activations to brain data. In brief, each column of $\mathbf{X}$ was first aligned with the words' offsets in the audio stream and convolved with the default *SPM* haemodynamic kernel (using Nilearn's *compute_regressor* function from the *nilearn.glm.first_level* module). The resulting time-course was sub-sampled to match the sampling frequency of the scans $\mathbf{Y}_s$ (giving $\tilde{\mathbf{X}}$). Next, in each individual voxel, the time-course of brain activation was regressed on $\tilde{\mathbf{X}}$ using Ridge regression. The Ridge regression regularization was estimated using a nested-cross validation scheme (see Appendix 1-Mapping NLM activations to brain data for more details). Finally, the cross-validated Pearson correlation R between the encoding model's prediction and the fMRI signal for subject s in voxel v was computed. The output of this process is a map of correlations between the encoding models' predictions and the observed time series, for a given participant.

Baseline fMRI model

To obtain a more accurate evaluation of the specific impact of the embeddings on brain scores, we removed the contribution of three confounding variables from all maps presented in this paper. The three confounding variables were: a) *the acoustic energy* (root mean squared of the audio signal sampled every 10ms) b) *the word-rate* (one event at each word offset) c) *the log of the unigram lexical frequency* of each word (modulator of the word events). An fMRI Ridge linear model that only included these three regressors was used to compute a map of cross-validated correlations for each participant.

The R-maps presented in Fig.3 of this paper are corrected for the contribution of these variables, that is they display ΔR, the increase in R when adding a model to the

baseline model versus the baseline model by itself.

Appendix 1-Fig.G1 displays the significant correlations in the group-level R maps associated with the Baseline Model, corrected for multiple comparison using a FDR correction ($p < 0.005$).

Group-level Maps

The brain maps presented in this document display group average increase in R scores obtained from individuals correlation maps (relative to the baseline model or to another model). Only voxels showing statistically significant increase in R score are shown.

Significance was assessed through one-sample t-tests applied to the spatially smoothed correlation maps, with an isotropic Gaussian kernel with FWHM of 6mm. In each voxel, the test assessed whether the distribution of R_{test} values across participants was significantly larger than zero. To control for multiple comparisons, all maps were corrected using a False Discovery Rate (FDR) correction with $p < 0.005$ (Benjamini & Hochberg, 1995). On each corrected figure, the FDR threshold on the z-scores, named z_{FDR}, is indicated at the bottom, that is, values reported on these maps (e.g. R scores) are shown only for voxels whose z-score survived this threshold ($z_{voxel} > z_{FDR}$).

While all analyses were done on volume data, all brain maps were projected onto brain surface for visualization purposes, using '*fsaverage5*' (from Nilearn's *datasets.fetch_surf_fsaverage*) mesh and the '*vol_to_surf*' function (from Nilearn's *surface* module).

Syntax and Semantics peak regions

We decided to also report brain maps' *peak regions*, i.e. the 10% of the voxels having the highest R score in a brain map. The motivation is that two different language processes might elicit lots of brain regions in common, while the regions that are better fitted by the representations derived from each process might differ. The peak regions of the neural correlates of semantic and syntactic representations are displayed on surface brain maps. The proportions of voxels belonging to each peak region as well as the Jaccard

score between syntax and semantics are displayed for each model and hemisphere. Subject-level maps were added in the Appendix to complement our group-level analysis.

Jaccard index

The Jaccard index (computed using scikit-learn *jaccard_score* function from the *metrics* module) for two sets X and Y is defined in the following manner: $J(X,Y) = |X \cap Y|/|X \cup Y|$. It behaves as a similarity coefficient: when the two sets completely overlap, J=1; when their intersection is nil, J=0.

Specificity index

To quantify how much each voxel v is influenced by semantic and syntactic embeddings, we defined a *specificity index* in the following manner:

$$x_{specificity}(v) = \log_{10}\left(\frac{r_{Semantic}(v)}{r_{Syntax}(v)}\right)$$

r_{Syntax} is the R score increase relative to the baseline model for the syntactic embeddings. $r_{Semantic}$ is the R score increase relative to the baseline model for the semantic embeddings.

In Fig.4, the higher x is, the more sensitive it is to semantic embeddings compared to syntactic embeddings. The lower x is, the more sensitive it is to syntactic embeddings compared to semantic embeddings. x close to 0, indicates an equal sensitivity to syntactic and semantic embeddings.

Group average specificity index maps were computed from each subject's map and significance was assessed through one-sample t-tests applied to the spatially smoothed specificity maps, with an isotropic Gaussian kernel with FWHM of 6mm. A FDR correction ($p < 0.005$) was used to correct for multiple comparisons.

Results

Dissociation of syntactic and semantic information in embeddings

We first assessed the amount of syntactic and semantic information contained in the embedding vectors derived from GloVe and GPT-2 trained on the different sets of features. In order to do so, we trained logistic classifiers to decode either the semantic category or the syntactic category from the embeddings generated from the text of *The Little Prince*.

The decoding performances of the logistic classifiers are displayed in Fig.2. The models trained directly on the integral features, that is, the intact texts, have relatively high performance on the two tasks (75% in average for both GloVe and GPT-2). The models trained on the syntactic features performed well on the syntax decoding task (decoding accuracy >95%), but are near chance-level on the semantic decoding task (decoding accuracy around 25% with a chance-level at 16%). Similarly, the models trained on the semantic features display good performance on the semantic decoding task (decoding accuracy greater than 80%), but a relatively poorer decoding accuracy on the syntax decoding task (45%, chance level: 16%). These results validate the experimental manipulation by showing that syntactic embeddings essentially encode syntactic information and semantic embeddings essentially encode semantic information. The high decoding accuracy of GloVe models is to be expected as we are decoding fixed categories associated with each word. Most of the information contained in the syntactic label (POS+Morph) and the semantic label is independent of the context, thus, GloVe performs well because it ignores contextual information. On the other hand, GPT-2 may be slightly affected by contextual cues. Despite this, the decoding task remains useful in demonstrating the presence of specific information within a model's embeddings.

In the Appendix, we present the decoding accuracy of the models when independently decoding the Part-of-speech (POS), syntactic morphological features (Morph), and the Number of Closing Nodes (NCN). These findings reveal that models trained on semantic features perform at chance-level when predicting the NCN, while

surpassing chance-level accuracy when predicting the Morph and POS. This improved performance in Morph prediction can be attributed to the retention of certain features such as gender, plural, or tense, which were preserved to maintain semantic integrity. POS is well decoded because of the small number of POS labels compared to the vocabulary size (number of content words).

Correlations of fMRI data with syntactic and semantic embeddings

Our objective was to evaluate how well the embeddings computed from GloVe and GPT-2 on the syntactic and semantic features fit the fMRI signal in various parts of the brain. For each model/features combination, we computed the increase in R score when the resulting embeddings were appended to a baseline model that comprised low-level variables (acoustic energy, word onsets and lexical frequency). This was done separately for each voxel. The resulting maps are displayed in Fig.3A.

The maps reveal that semantic and syntactic feature-derived embeddings from GloVe or GPT-2 significantly explain the signal in a set of bilateral brain regions including frontal and temporal regions, as well as the Temporo-parietal junction, the Precuneus and Dorso-Medial Prefrontal Cortex (dMPC). The classical left-lateralized language network, which includes the Inferior Frontal Gyrus (IFG) and the Superior Temporal Sulcus (STS), is entirely covered. Overall, a vast network of regions is modulated by both semantic and syntactic information.

Nevertheless, detailed inspection of the maps shows different R score distribution profiles (see Appendix 1-R Scores Distribution for GloVe and GPT-2 Trained on Semantic or Syntactic Features Appendix 1-Fig.I1). For example, syntactic embeddings yield the highest fits in the Superior Temporal Lobe, extending from the Temporal Pole (TP) to the Temporo-Parietal Junction (TPJ), as well as the Inferior Frontal Gyrus (IFG, BA-44 and 47), the Superior Frontal Gyrus (SFG), the Dorso-Medial Prefrontal Cortex (dMPC) and the posterior Cingulate cortext (pCC). Semantic embeddings, on the other hand, show peaks in the posterior Middle Temporal Gyrus (pMTG), the Angular Gyrus (AG), the

Inferior Frontal Sulcus (IFS), the dMPC and the Precuneus/pCC.

Regions best fitted by semantic or syntactic embeddings

As noticed above, despite the fact that the regions fitted by semantic and syntactic embeddings essentially overlap (Fig.3A), the areas where each model has the highest R scores differ. To better visualize the maxima from these maps, we selected, for each of them, the 10% of voxels having the highest R scores. Thresholding at the 90-th percentile of the distributions (threshold values displayed in Appendix 1-Fig.I1) produces the maps presented in Fig.3B.

A first observation is that the number of supra-threshold voxels is quite similar in the left (19%) and right (21%) hemispheres, whether GPT-2 or Glove is considered, showing that during the processing of natural speech, both syntactic and semantic features modulate activations in both hemispheres to a similar extent. The regions involved include, bilaterally, the TP, the STS, the IFG and IFS, the DMPC, the pMTG, the TPJ, the Precuneus and pCC.

One noticeable difference between the two hemispheres, apparent in Fig.3B, concerns the *overlap* between the semantic and syntactic peak regions: it is stronger in the right than in the left hemisphere. To assess this overlap, we computed the Jaccard indices (see Jaccard index) between voxels modulated by syntax and voxels modulated by semantics. The Jaccard indices were much larger in the right hemisphere ($J^{right}_{GloVe} = 0.52$ and $J^{right}_{GPT-2} = 0.60$) than in the left ($J^{left}_{GloVe} = 0.14$ and $J^{left}_{GPT-2} = 0.20$).

The left hemisphere displayed distinct peak regions for semantics and syntax; syntax involving the STS, the pSTG, the anterior TP, the IFG (BA-44/45/47) and the MFG, while semantics involves the pMTG, AG, the TPJ and the IFS. We only observe overlap in the upper IFG (BA-44), AG and posterior STS. On medial faces, semantics and syntax share peak regions in the Precuneus, the pCC and the DMPC. In the right hemisphere, syntax and semantics share the STS, pMTG and most frontal regions, with only syntax-specific peak regions in the TP and SFG and semantics-specific peak regions in the TPJ.

In addition to the group-level analysis, we conducted subject-level analyses that yielded consistent findings (see Appendix 1-Fig.K,K2,K3). Our results demonstrate the following patterns:

- We observed higher Jaccard scores in the right hemisphere compared to the left.
- Syntactic peak regions were identified in the Temporal regions, the IFG (Inferior Frontal Gyrus), and dmPFC (dorsomedial Prefrontal Cortex).
- Semantic peak regions were found near the IFS (Inferior Frontal Sulcus)/pMTG (posterior Middle Temporal Gyrus)/TPJ (TemporoParietal Junction).

These subject-level analyses further support and reinforce the patterns observed at the group level.

Overall, this shows that the neural correlates of syntactic and semantic features appear more separable in the left than in the right hemisphere .

Gradient of sensitivity to syntax or semantics

The analyses presented above revealed a large distributed network of brain regions sensitive to both syntax and semantics but with varying local sensitivity to both conditions.

We further investigated these differences by defining a *specificity index* that reflects, for each voxel, the logarithm of the ratio between the R scores derived from the semantic and the syntactic embeddings (see Specificity index). A score of x indicates that the voxel is 10^x-times more sensitive to semantics compare to syntax if $x > 0$ (green), and conversely, the voxel is 10^x-times more sensitive to syntax compare to semantics if $x < 0$ (red). Voxels with specificity indexes close to 0, are colored in yellow and show equal sensitivity to both conditions. Specificity indexes are plotted on surface maps in Fig.4. The top row shows the specificity index of voxels where there was a significant effect for syntactic or for semantic embeddings in Fig.3A, while the bottom row shows group specificity indexes corrected for multiple comparison using an FDR-correction of 0.005 (N=51).

The top row of Fig.4 shows that voxels that are more sensitive to Syntax include, bilaterally, the anterior Temporal Lobes (aTL), the STG, the Supplementary Motor Area

(SMA), the MFG and sub-parts of the IFG. Voxels more sensitive to Semantics are located in the pMTG, the TPJ/AG, the IFS, SFS and the Precuneus. Voxels sensitive to both types of features are located in the posterior STG, the STS, the dMPC, the CC, the MFG and in the IFG.

More specifically, in Fig.4 bottom, one can observe significantly low ratios (in favor of the syntactic embeddings) in the STG, aTL and pre-SMA, and significantly large ratios (in favor of the semantic embeddings) in the pMTG, the AG and the IFS. Specificity index maps are consistent with the maps of R score differences between semantic and syntactic embeddings for Glove and GPT-2 (see Appendix 1-Fig.J1), but provide more insights into the relative sensitivity to syntax and semantics. These maps highlight that some brain regions show stronger responses to the semantic or to the syntactic condition even when they show sensitivity to both.

Unique contributions of syntax and semantics

The previous analyses allowed us to quantify the amounts of brain signal explained by the information encoded in various embeddings. Yet, when two embeddings explain the same amount of signal, that is, have similar R score, it remains to be clarified whether they hinge on information represented redundantly in the embeddings or information specific to each embedding. To address this issue, we analyzed the additional information brought by each embedding on top of the other one. To this end, we evaluated correlations that are uniquely explained by the semantic embeddings compared to the syntactic embeddings, and conversely.

To quantify the unique contribution of each feature space to the prediction of the fMRI signal, we first estimated the Pearson correlation explained by the embeddings learned from the individual feature space - e.g., using only syntactic embeddings or semantic embeddings. We then assessed the correlation explained by the concatenation of embeddings derived from different feature spaces - e.g., concatenating syntactic and semantic embedding vectors (de Heer et al., 2017).

Because it can identify single voxels whose responses can be partly explained by different feature spaces, this approach provides more information than simple subtractive analyses that estimate the R score difference per voxel (see Appendix 1-Fig.J1).

Syntactic embeddings (Fig.5A) uniquely explained brain data in localized brain regions: the STG, the TP, the pre-SMA and in the IFG, with R scores increases of about 5%.

Semantic embeddings (Fig.5B) uniquely explained signal bilaterally in the same wide network of brain regions as the one highlighted in Fig.3A, including frontal and temporo-parietal regions bilaterally as well as the Precuneus and pCC medially, with similar R scores increases around 5%.

This suggests that even if most of the brain is sensitive to both syntactic and semantic conditions, syntax is preferentially processed in more localized regions than semantics which is widely distributed.

Synergy between syntax and semantics

To probe regions where the joint effect of syntax and semantics is greater than the sum of the contributions of these features, we compared the R scores of the embeddings derived from the integral features with the R scores of the encoding models concatenating the semantic and syntactic embeddings (see Fig.5C).

For the embeddings obtained with GloVe, this analysis did not reveal any significant effect. For the embeddings obtained with GPT-2, significant effects were observed in most of the brain, but with higher effects in the semantic peak regions: pMTG, TPJ, AG and in frontal regions.

Integration of contextual information

To further examine the effect of context, we compared GPT-2, the supra-lexical model which takes context into account, to GloVe, a purely lexical model. The differences in R scores between the two models, trained on each of the three datasets are presented in Fig.6.

GPT-2 embeddings elicit stronger R scores than GloVe. The difference spreads over wider regions when the models were trained on syntax compared to semantics (see Fig.6 top left and right). The comparison for syntax led to significant differences bilaterally in the STS/STG, from the Temporal Pole to the TPJ, in superior, middle and inferior frontal regions, and medially in the pCC and dMPC. For semantics, the comparison only led to significant differences in the Precuneus, the right STS and posterior STG. Fig.6 (bottom left) shows the comparison between GPT-2 and GloVe when trained on the Integral features. Given that both semantic and syntactic contextual information were available to GPT-2, these maps reflect the regions that benefit from context during story listening.

To show that context has an effect is one thing, but different brain regions are likely to have different integration window's size. To address this question, we developed a fixed-context window training protocol to control for the amount of contextual information used by GPT-2 (Fig.1C). We trained models with short (5 tokens), medium (15 tokens) and long (45 tokens) range windows sizes. This ensures that GPT-2 was not sampling out of the learnt distribution at inference, and not using more context than what was available in the context window.

Comparing GPT-2 with 5 tokens to GloVe (0-size context) highlighted a large network of frontal and temporo-parietal regions. Medially, it included the Precuneus, the pCC and the DMPC (Fig.6, short). Short context-sensitivity showed peak effects in the Supramarginal gyri, the pMTG and medially in the Precuneus and pCC. Counting the number of voxels showing significant short-context effects highlighted an asymmetry between the left and right hemisphere with 1.6 times more significant voxels in the left hemisphere compared to the right. Contrasting a GPT-2 model using 15 tokens of context (the average size of a sentence in *The Little Prince*) versus a GPT-2 model using only 5 tokens, yielded localized significant differences in the SFG/SFS, the TP, MFG and STG near Heschl's gyri and medially in the Precuneus and pCC (Fig.6, Medium). The biggest medium context effects included the left MFG, the right SFG and DMPC and bilaterally

the Precuneus and pCC. Finally, contrasting models using respectively 45 and 15 tokens of context revealed 2.8 times as many significant differences in the right hemisphere as in the left. Significant effects were the highest bilaterally and medially in the pCC, followed, in the right hemisphere, by the Precuneus, the DMPC, MFG, SFG, STS and TP (see Fig.6, bottom).

Taken together, our results show 1) that syntax dominantly determines the integration of contextual information, 2) that a bilateral network of frontal and temporo-parietal regions is modulated by short context, 3) that short-range context integration is preferentially located in the left hemisphere, 4) that the right hemisphere is involved in the processing of longer context sizes, and finally 5) that medial regions (Precuneus and pCC) are core regions of context integration, showing context effects at all scales.

Discussion

Language comprehension in humans is a complex process, which involves several interacting sub-components (word recognition, processing of syntactic and semantic information to construct sentence meaning, pragmatic and discourse inference, ...) (Jackendoff, 2002, e.g.). Discovering how the brain implements these processes is one of the major goals of neurolinguistics. A lot of attention has been devoted, in particular, to the syntactic and semantic components (Binder & Desai, 2011; Friederici, 2017, for reviews) and the extent to which they are implemented in (practically) distinct or identical regions is still debated (e.g. Fedorenko et al., 2020).

It must be noted that a fair proportion of these studies relied on controlled experimental paradigms with single words or sentences, based on the manipulation of complexity or violations of expectations. To study language processing in a more natural way, several recent studies have presented naturalistic texts to participants, and have analyzed their brain activations using Artificial Neural Language Models (e.g. Huth et al., 2016; Pasquiou et al., 2022; Pereira et al., 2018a; Schrimpf et al., 2020). These models are

known to encode some aspects of semantics and syntax (e.g. Hewitt & Manning, 2019; Lakretz et al., 2019; Pennington et al., 2014). In the current work, to further dissect brain activations into separate linguistic processes, we trained NLP models on a corpus from which we selectively removed syntactic, semantic or contextual information and examined how well these information-restricted models could explain fMRI signal recorded from participants who had listened to an audiobook. The rationale was to highlight brain regions representing syntactic and semantic information, at the lexical and supralexical levels (comparing a lexical model GloVe, and a contextual one, GPT-2). Additionaly, by varying the amount of context provided to the supralexical model, we sought to identify the brain regions sensitive to different context sizes (see Jain and Huth (2018) for a similar analysis).

Whether models were trained on syntactic features or on semantic features, they fit fMRI activations in a wide bilateral network which goes beyond the classic language network comprising the IFG and temporal regions: it also includes most of the dorso lateral and medial prefrontal cortex, the inferior parietal cortex, and on the internal face, the precuneus and posterior cingulate cortex (see Fig.3). Nevertheless, the regions *best* predicted by syntactic features on the one hand, and semantic features on the other hand, are not exactly the same. While they overlap quite a lot in the right hemisphere, they are more dissociated in the left hemisphere Fig.3, panel B and Appendix 1-Fig.K1). In addition, the relative sensitivity to syntax and semantics varies from region to region, with syntax predominating in the temporal lobe (see Fig.4). Elimination of shared variance between syntactic and semantic features confirmed that pure syntactic effects are restricted to STG/STS, bilaterally, IFG, and pre-SMA, while pure semantic effects occur throughout the network (Fig.5 A-B).

The comparison between the supralexical model (GPT-2) and the lexical one (GloVe), revealed brain regions involved in compositionality (Fig.6) and a synergy between syntax and semantics that arises only at the supralexical level (Fig.5C). Finally, analyses of the influence of the size of context provided to GPT-2 when computing word embeddings,

show that (1) a bilateral network of fronto-temporo-parietal regions is sensitive to short context, that (2) there is a dissociation between the left and right hemispheres, respectively associated with short-range and long-range context integration, and finally that (3) the medial Precuneus and posterior Cingulate gyri show the highest effects at every scale, hinting at an important role in large context integration (Fig.7).

In summary, this study shows that

- there is a difference between the right and left hemispheres with respect to the separation of syntactic and semantic processing. We found more segregation in the left compared to the right hemisphere. This provides support to classic theories on the functional difference between the left and right hemispheres (Beeman & Chiarello, 2013).
- the right hemisphere is sensitive to longer contexts than the left one (Beeman & Chiarello, 2013)
- neural language models are a beneficial tool in the study of brain function. Manipulating the training corpus or the size of the context window, possible only in simulations, was shown to lead to new findings about language processing in the human brain.

Models trained on semantic and syntactic features fit brain activity in a widely distributed network, but with varying relative degrees.

When trained on the integral corpus, that is on the integral features, both the lexical (GloVe) and contextual (GPT-2) models captured brain activity in a large *extended language network* (Appendix 1-Fig.H1). This large extended language network goes beyond the *core* language network, that is, the left IFG and temporal regions, encompassing homologous areas in the right hemisphere, the dorsal prefrontal regions, both on the lateral and medial surfaces, as well as in the inferior parietal, Precuneus and posterior Cingulate. The result is consistent with the ones from previous studies that have looked at brain responses to naturalistic text, whether analysed with NLP models (e.g. Caucheteux et al.,

2021; Huth et al., 2016; Jain & Huth, 2018; Pereira et al., 2018b) or not (Chang et al., 2022; Lerner et al., 2011).

The Precuneus/pCC, inferior parietal and dorsomedial prefrontal cortex are part of the Default Mode Network (DMN) (Raichle, 2015). The same areas are actually also relevant in language and high-level cognition. For example, early studies examining the role of coherence during text comprehension had pointed out the same regions (Ferstl & von Cramon, 2001; Xu et al., 2005): coherent discourses elicit stronger activations than incoherent ones. Recent work by (Chang et al., 2022) has revealed that the DMN is the last stage in a temporal hierarchy of processing naturalistic text, integrating information on the scale of paragraphs and narrative events, see also (Baldassano et al., 2017; Simony et al., 2016). These regions are not language-specific though, as they have been shown to be activated during various theory of mind tasks, relying on language or not, and have thus also been dubbed the "Mentalizing network" (Baetens et al., 2014; Mar, 2011).

Models trained on the information-restricted semantic and syntactic features fit signal in this widely distributed network (Fig.3A). This is in agreement with Caucheteux et al. (2021) and Fedorenko et al. (2020) who, using very different approaches, found that syntactic predictors modulated activity throughout the language network. Caucheteux et al. (2021) first constructed new texts that matched, as well as possible, the text presented to participants in terms of their syntactic properties. The lexical items being different, the semantics of the new texts bear little relation with the original text. Then, using a pre-trained version of GPT-2, the authors obtained embeddings from these new texts and averaged them to create syntactic predictors. They found that these syntactic embeddings fitted a network of regions (ibid. Fig5D) similar to the one we observed (Fig.3A). Further, defining the effect of semantics as the difference between the scores obtained from the embeddings from the original text, and the scores from the syntactic embeddings, Caucheteux et al. (2021) observed that semantics had a significant effect throughout the same network (ibid. Fig5G).

Should one conclude that syntax and semantics equally modulate the entire language network? Our results reveal a more complex picture. Figure 4 presents a semantics vs syntax specificity index map, showing higher sensitivity to syntax in the STG and anterior temporal lobe, whereas the parietal regions are more sensitive to semantics, consistent with Binder et al. (2009).

Our study helps to reconcile two apparently contradicting results in the literature. On the one hand, classic results on syntactic processing found a localized set of brain regions involved in syntactic processing (Friederici, 2016; Matchin et al., 2017; Pallier et al., 2011), whereas recent studies, using naturalistic (so-called 'ecological') paradigms, found a more widely spread, distributed, network of brain regions involved in syntactic processing (Caucheteux et al., 2021; Fedorenko et al., 2020). Our study reconciles these two apparently contradicting results by providing a more graded view of syntactic processing in the brain, showing that sensitivity to syntactic processing peaks at around the same set of localized brain regions identified in classic studies.

Another point to take in consideration is that syntactic and semantic features are not perfectly orthogonal. Indeed, the logistic decoder trained on the embeddings from the semantic dataset was better than chance at recovering both syntactic Morphological features and the Part-of-Speech (Fig.2 and Appendix 1-Fig.E1). This might be due, for example, to the fact that some features like gender or number are present in both datasets, explicitly in the syntactic dataset and implicitly in the semantic dataset. Part-of-speech can be easily decoded from semantic features because the number of POS labels is much smaller than the vocabulary size of the semantic features. To focus on the unique contributions of syntax and semantic, we remove the shared variance from the syntactic and semantic models using model comparisons (Fig.5).

"Pure" semantic but not "pure" syntactic features modulate activity in a wide set of brain regions.

The unique effect of semantics, when its shared component with syntax was removed, remains widespread (Fig.5B). This is consistent with the notion that semantic information is widely distributed over the cortex, an idea popularized by embodiment theories (Hauk et al., 2004; Pulvermüller, 2013), but which was already supported by the neuropsychological observations revealing domain-specific semantic deficits in patients (Damasio et al., 2004).

On the other hand the "pure" effect of syntax "shrinked" to the STG and aTL (bilaterally), the IFG (on the left) and the pre-SMA (Fig.5A). The left IFG and STG/STS have previously been implicated in syntactic processing Friederici (2011) and Friederici (2017, e.g.), and this is confirmed by the new approach employed here. Note that we are not claiming that these regions are specialized for syntactic processing only. Indeed they also appear to be sensitive to the "pure" semantic component (Fig.5B).

The contributions of the right hemisphere.

A striking feature of our results is the strong involvement of the right hemisphere. The notion that the right hemisphere has some linguistic abilities is supported by the studies on split-brains (Sperry, 1961) and by the patterns of recovery of aphasic patients after lesions in the left hemisphere (Dronkers et al., 2017). Moreover, a number of brain imaging studies have confirmed the right hemisphere involvement in higher-level language tasks, such as comprehending metaphors or jokes, generating the best endings to sentences, mentally repairing grammatical errors, detecting story inconsistencies (see Beeman and Chiarello (2013) and Jung-Beeman (2005)). All in all, this suggests that the right hemisphere is apt at recognizing distant relations between words. This conclusion is further reinforced by our observation of long-range (paragraph-level) context effects in the right hemisphere (Fig.7, Long).

The effects we observed in the right hemisphere are not simply the mirror image of

the left hemisphere. Spatially, syntax and semantics dissociate more in the left than the right. An observation that is consistent both at subject-level (see Appendix 1-Fig.K1) and group-level (see Fig.3, Panel B). Moreover, the regions of overlap correspond to the regions integrating long context (Fig.7C, bottom row), suggesting that the left hemisphere is relatively more involved in the processing of local semantic or syntactic information, whereas the right hemisphere integrates both information at a larger time-scale (supra sentential).

Syntax drives the integration of contextual information.

The comparison between the predictions of the integral model trained on the intact texts, and the predictions of the combined syntactic and semantic embeddings from the information-restricted models (Fig.5C), highlights a striking contrast between GloVe and GPT2. While the former, a purely lexical model, does not benefit from being trained on the integral text, GPT-2 shows clear synergetic effects of syntactic and semantic information. GPT-2's embeddings fit brain activation better when syntactic and semantic information can contribute together. The fact that the regions that benefit most from this synergetic effect are high-level integrative regions, at the end of the temporal processing hierarchy described by Chang et al. (2022), suggests that the availability of syntactic information drives the semantic interpretation at the sentence level.

These regions are quite similar to the semantic peak regions highlighted in Fig. 3B, and overlap with the regions showing context effects (Fig.7). This replicates, and extends, the results from Jain and Huth (2018) who, varying the amount of context fed to LSTM models, from 0 to 19 words, found shorter context effects in temporal regions (ibid. Fig 4).

It is crucial to clarify that the influence of syntactic information on semantic interpretation at the sentence level does not imply that syntactic information drives the alignment performance between artificial and biological neural networks. By examining the sensitivity index maps and comparing models trained on semantic features with those trained on syntactic features (see Appendix 1-Fig.J1), it becomes apparent that models

trained on semantic features account for a larger proportion of variance in most areas of the brain. The finding that semantic information accounts for a greater proportion of variance than syntactic information aligns with previous studies in the literature (Kauf et al., 2023; Mollica et al., 2019; Sinha et al., 2021).

Limitations of our study

Two limitations of our study must be acknowledged.

The dissociation between syntax and semantics is not perfect. The way we created the semantic dataset by removing function words clearly impacts supra-lexical semantics. For example, removing instances of *and* and *or* prevents the NLP model from distinguishing between the meaning of "A or B" and "A and B". In other words, the logical form of sentences can be perturbed. The decline in compositional semantics becomes apparent when examining the layer-wise encoding performance of the semantic model (see Appendix 1-Fig.L1). In contrast to the models trained on integral or syntactic features, which exhibit optimal encoding performance in the later layers, the semantic model demonstrates a decrease in performance. This observation indicates that the model struggles to effectively utilize the structural information necessary for composing the meanings of more extensive linguistic structures.

This may partly explain the synergetic effect of syntax and semantics described above. Removing pronouns is also problematic as this removed the arguments of some verbs. Ideally, one would like to find transformations of the sentences that keep the semantic information associated to the function words like conjunctions or pronouns, but it is not clear how to do that.

A second limitation concerns potential confounding effects of prosody. One cannot exclude that the embeddings of the models captured some prosodic variables correlated with syntax (Bennett & Elfner, 2019). For example, certain categories of words (e.g. determiners or pronouns) are shorter and less accented than others. Also, although the models are purely trained on written text, they acquire the capacity to predict the end of

sentences, which are more likely to be followed by pauses in the acoustic signal. We included acoustic energy and the words' offsets in the baseline models to try and diminish the impact of such factors, but such controls cannot be perfect. One way to address this issue would be to have participants *read* the text, presented at a fixed presentation rate. This would effectively remove all low-level effects of prosody.

Conclusion

State-of-the-art Natural Language Processing models, like transformers, trained with large enough corpora, can generate essentially flawless grammatical text, showing that they can acquire the grammar of the language. Using them to fit brain data has become a common endeavour, even if their architecture rules them out of plausible models of the brain. Yet, despite their low biological plausibility, their ability to build rich distributed representations can be exploited to study language processing in the brain. In this paper, we have demonstrated that restricting information provided to the model during training can be used to show which brain areas encode this information. Information-restricted models are powerful and flexible tools to probe the brain as they can be used to investigate whatever representational space chosen, such as semantics, syntax or context. Moreover, once they are trained, these models can be used directly on any dataset in order to generate information-restricted features for model-brain alignment. This approach is highly beneficial, both in term of richness of the features, and scalability, compared to classical approaches that use manually crafted features or focus on specific contrasts. In future experiments, more fine grained control of both the information given to the models as well as model's representations will permit more precise characterisation of the role of the various regions involved in language comprehension.

Data Availability

The Integral Dataset (train, test and dev) is available at: https://osf.io/jzcvu/. The semantic and syntactic datasets can be derived from the Integral Dataset using the scripts provided in https://github.com/AlexandrePsq/Information-Restricted-NLMs.

All analyses, as well as model training, features extraction and the fitting of encoding models were performed using Python 3.7.6 and can be replicated using the code provided in the same Github repository (https://github.com/AlexandrePsq/Information-Restricted-NLMs). The required packages are listed there. A non-exhaustive list includes numpy (Harris et al., 2020), scipy (Virtanen et al., 2020), scikit-learn (Pedregosa et al., 2011), matplotlib (Hunter, 2007), pandas (McKinney et al., 2010) and nilearn (https://nilearn.github.io/stable/index.html).

The fMRI dataset is publicly available at https://openneuro.org/datasets/ds003643/versions/1.0.2, and all details regarding the dataset are described in details by Li et al. (2022).

Acknowledgments

The authors acknowledge the use of the computing resources of NeuroSpin the Gipsi.

References

Baetens, K., Ma, N., Steen, J., & Van Overwalle, F. (2014). Involvement of the mentalizing network in social and non-social high construal. *Social Cognitive and Affective Neuroscience*, *9*(6), 817–824. https://doi.org/10.1093/scan/nst048

Baldassano, C., Chen, J., Zadbood, A., Pillow, J. W., Hasson, U., & Norman, K. A. (2017). Discovering event structure in continuous narrative perception and memory. *Neuron*, *95*(3), 709–721.e5. https://doi.org/https://doi.org/10.1016/j.neuron.2017.06.041

Bates, E., & Dick, F. (2002). Language, gesture, and the developing brain [Publisher: Wiley Online Library]. *Developmental Psychobiology: The Journal of the International Society for Developmental Psychobiology*, *40*(3), 293–310. https://pubmed.ncbi.nlm.nih.gov/11891640/

Bates, E., & MacWhinney, B. (1989). Functionalism and the competition model. In B. MacWhinney & E. Bates (Eds.), *The crosslinguistic study of sentence processing* (pp. 3–73). Cambridge University Press. https://www.researchgate.net/publication/230875840_Functionalism_and_the_Competition_Model/link/545a97170cf2c16efbbbc1d5/download

Beeman, M. J., & Chiarello, C. (2013). *Right hemisphere language comprehension: Perspectives from cognitive neuroscience*. Psychology Press.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300. Retrieved February 21, 2023, from http://www.jstor.org/stable/2346101

Bennett, R., & Elfner, E. (2019). The Syntax–Prosody Interface. *Annual Review of Linguistics*, *5*(1), 151–171. https://doi.org/10.1146/annurev-linguistics-011718-012503

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences*, *15*(11), 527–536. https://doi.org/10.1016/j.tics.2011.10.001

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies. *Cerebral Cortex*, *19*(12), 2767–2796. https://doi.org/10.1093/cercor/bhp055

Bottini, G., Corcoran, R., Sterzi, R., Paulesu, E., Schenone, P., Scarpa, P., Frackowiak, R., & Frith, C. (1995). The role of the right hemisphere in the interpretation of figurative aspects of language. a positron emission tomography activation study. *Brain : a journal of neurology*, *117 (Pt 6)*, 1241–53. https://doi.org/10.1093/brain/117.6.1241

Caplan, D., Alpert, N., & Waters, G. (1998). Effects of Syntactic Structure and Propositional Number on Patterns of Regional Cerebral Blood Flow [_eprint: https://direct.mit.edu/jocn/article-pdf/10/4/541/1931814/089892998562843.pdf]. *Journal of Cognitive Neuroscience*, *10*(4), 541–552. https://doi.org/10.1162/089892998562843

Caramazza, A., & Zurif, E. B. (1976). Dissociation of algorithmic and heuristic processes in language comprehension: Evidence from aphasia. *Brain and language*, *3*(4), 572–582. https://pubmed.ncbi.nlm.nih.gov/974731/

Caucheteux, C., Gramfort, A., & King, J.-R. (2021). Disentangling Syntax and Semantics in the Brain with Deep Networks. *ICML 2021 - 38th International Conference on Machine Learning*, 13.

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Communications Biology*. https://doi.org/10.1038/s42003-022-03036-1

Chang, C. H. C., Nastase, S. A., & Hasson, U. (2022). Information flow across the cortical timescale hierarchy during narrative construction [Publisher: Proceedings of the National Academy of Sciences]. *Proceedings of the National Academy of Sciences*, *119*(51), e2209307119. https://doi.org/10.1073/pnas.2209307119

Chomsky, N. (1984). *Modular Approaches to the Study of the Mind* (Vol. 1). San Diego State University Press San Diego. https://archive.org/details/modularapproache00noam/page/n9/mode/2up

Cooke, A., Zurif, E. B., DeVita, C., Alsop, D., Koenig, P., Detre, J., Gee, J., PinÃ£ngo, M., Balogh, J., & Grossman, M. (2001). Neural basis for sentence comprehension: Grammatical and short term memory components. *Human Brain Mapping*, *15*(2), 80–94. https://doi.org/10.1002/hbm.10006

Damasio, H., Tranel, D., Grabowski, T., Adolphs, R., & Damasio, A. (2004). Neural systems behind word and concept retrieval. *Cognition*, *92*(1-2), 179–229. https://doi.org/10.1016/j.cognition.2002.07.001

de Heer, W. A., Huth, A. G., Griffiths, T. L., Gallant, J. L., & Theunissen, F. E. (2017). The Hierarchical Cortical Organization of Human Speech Processing. *The Journal of Neuroscience*, *37*(27), 6539–6557. https://doi.org/10.1523/JNEUROSCI.3267-16.2017

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [arXiv: 1810.04805]. *arXiv:1810.04805 [cs]*. Retrieved December 10, 2020, from http://arxiv.org/abs/1810.04805

Dick, F., Bates, E., Wulfeck, B., Utman, J. A., Dronkers, N., & Gernsbacher, M. A. (2001). Language deficits, localization, and grammar: Evidence for a distributive model of language breakdown in aphasic patients and neurologically intact individuals. [Publisher: American Psychological Association]. *Psychological review*, *108*(4), 759. https://psycnet.apa.org/record/2001-18918-004

Dronkers, N. F., Ivanova, M. V., & Baldo, J. V. (2017). What Do Language Disorders Reveal about Brain–Language Relationships? From Classic Models to Network Approaches. *Journal of the International Neuropsychological Society : JINS*, *23*(9-10), 741–754. https://doi.org/10.1017/S1355617717001126

Elman, J. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, *7*, 195–225. https://link.springer.com/article/10.1007/BF00114844

Embick, D. (2000). Features, syntax, and categories in the latin perfect. *Linguistic Inquiry*, *31*(2), 185–230. Retrieved November 8, 2022, from http://www.jstor.org/stable/4179104

Fedorenko, E., Blank, I., Siegelman, M., & Mineroff, Z. (2020). Lack of selectivity for syntax relative to word meanings throughout the language network [Publisher: Cold Spring Harbor Laboratory]. *bioRxiv*, 477851.

Ferstl, E. C., & von Cramon, D. Y. (2001). The role of coherence and cohesion in text comprehension: An event-related fMRI study. *Cognitive Brain Research*, *11*(3), 325–340. https://doi.org/10.1016/S0926-6410(01)00007-6

Fodor, J. (1983). *The modularity of mind*. MIT press. https://mitpress.mit.edu/9780262560252/the-modularity-of-mind/

Friederici, A. D. (2011). The Brain Basis of Language Processing: From Structure to Function. *Physiol Rev*, *91*, 36. https://pubmed.ncbi.nlm.nih.gov/22013214/

Friederici, A. D. (2016). The neuroanatomical pathway model of Language: Syntactic and semantic networks. In *Neurobiology of Language* (pp. 349–356). Elsevier.

Friederici, A. D., Chomsky, N., Berwick, R. C., Moro, A., & Bolhuis, J. J. (2017). Language, mind and brain. *Nature human behaviour*, *1*(10), 713–722.

Friederici, A. D., Fiebach, C. J., Schlesewsky, M., Bornkessel, I. D., & von Cramon, D. Y. (2006). Processing Linguistic Complexity and Grammaticality in the Left Frontal Cortex. *Cerebral Cortex*, *16*(12), 1709–1717. https://doi.org/10.1093/cercor/bhj106

Friederici, A. D., Kotz, S. A., Scott, S. K., & Obleser, J. (2009). Disentangling syntax and intelligibility in auditory language comprehension. *Human Brain Mapping*, *31*(3), 448–457. https://doi.org/10.1002/hbm.20878

Friederici, A. D., Makuuchi, M., & Bahlmann, J. (2009). The role of the posterior superior temporal cortex in sentence comprehension. *NeuroReport*, *20*(6), 563–568. https://doi.org/10.1097/WNR.0b013e3283297dee

Friederici, A. D., RÃ¼schemeyer, S.-A., Hahne, A., & Fiebach, C. J. (2003). The Role of Left Inferior Frontal and Superior Temporal Cortex in Sentence Comprehension: Localizing Syntactic and Semantic Processes. *Cerebral Cortex*, *13*(2), 170–177. https://doi.org/10.1093/cercor/13.2.170

Friederici, A. D. (2017). Neurobiology of Syntax as the Core of Human Language. *BIOLINGUISTICS*, *11*. https://bioling.psychopen.eu/index.php/bioling/article/view/9093

Garrard, P., Carroll, E., Vinson, D., & Vigliocco, G. (2004). Dissociation of lexical syntax and semantics: Evidence from focal cortical degeneration [PMID: 15788273]. *Neurocase*, *10*(5), 353–362. https://doi.org/10.1080/13554790490892248

Goodglass, H. (1993). *Understanding aphasia.* Academic Press. https://www.jstor.org/stable/416147

Grodzinsky, Y., & Santi, A. (2008). The battle for broca's region. *Trends in Cognitive Sciences*, *12*(12), 474–480. https://doi.org/https://doi.org/10.1016/j.tics.2008.09.001

Hagoort, P. (2014). Nodes and networks in the neural architecture for language: Broca's region and beyond [Publisher: Elsevier]. *Current opinion in Neurobiology*, *28*, 136–141.

Harris, C. R., Millman, K. J., van der Walt, S. J., Gommers, R., Virtanen, P., Cournapeau, D., Wieser, E., Taylor, J., Berg, S., Smith, N. J., Kern, R., Picus, M., Hoyer, S., van Kerkwijk, M. H., Brett, M., Haldane, A., del Río, J. F., Wiebe, M., Peterson, P., ... Oliphant, T. E. (2020). Array programming with numpy. *Nature*, *585*(7825), 357–362. https://doi.org/10.1038/s41586-020-2649-2

Hashimoto, R., & Sakai, K. L. (2002). Specialization in the left prefrontal cortex for sentence comprehension. *Neuron*, *35*(3), 589–597. https://doi.org/https://doi.org/10.1016/S0896-6273(02)00788-2

Hauk, O., Johnsrude, I., & Pulvermüller, F. (2004). Somatotopic representation of action words in human motor and premotor cortex. *Neuron*, *41*(2), 301–307. Retrieved February 12, 2015, from http://www.sciencedirect.com/science/article/pii/S0896627303008389

Hewitt, J., & Manning, C. D. (2019). A Structural Probe for Finding Syntax in Word Representations. *North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL)*, 10. https://aclanthology.org/N19-1419/

Honnibal, M., & Montani, I. (2017). *spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing* [To appear]. https://spacy.io/usage

Hunter, J. D. (2007). Matplotlib: A 2d graphics environment. *Computing in Science & Engineering*, *9*(3), 90–95. https://doi.org/10.1109/MCSE.2007.55

Huth, A. G., de Heer, W. A., Griffiths, T. L., Theunissen, F. E., & Gallant, J. L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458. https://doi.org/10.1038/nature17637

Jackendoff, R. (2002). *Foundations of language: Brain, meaning, grammar, evolution*. Oxford University Press UK. https://academic.oup.com/book/32834

Jain, S., & Huth, A. (2018). Incorporating context into language encoding models for fmri. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, & R. Garnett (Eds.), *Advances in neural information processing systems* (p. 10). Curran Associates, Inc. https://proceedings.neurips.cc/paper/2018/file/f471223d1a1614b58a7dc45c9d01df19-Paper.pdf

Jung-Beeman, M. (2005). Bilateral brain processes for comprehending natural language. *Trends in Cognitive Sciences*, *9*(11), 512–518. https://doi.org/10.1016/j.tics.2005.09.009

Kauf, C., Tuckute, G., Levy, R., Andreas, J., & Fedorenko, E. (2023). Lexical semantic content, not syntactic structure, is the main contributor to ANN-brain similarity of fMRI responses in the language network. *bioRxivorg*.

Kinno, R., Kawamura, M., Shioda, S., & Sakai, K. L. (2007). Neural correlates of noncanonical syntactic processing revealed by a pictured sentence matching task. *Human Brain Mapping*, *29*(9), 1015–1027. https://doi.org/10.1002/hbm.20441

Kitaev, N., & Klein, D. (2018). Constituency parsing with a self-attentive encoder. *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 2676–2686. https://doi.org/10.18653/v1/P18-1249

Kundu, P., Voon, V., Balchandani, P., Lombardo, M. V., Poser, B. A., & Bandettini, P. A. (2018). Multi-echo fMRI: A review of applications in fMRI denoising and analysis of BOLD signals. *NeuroImage*, *154*. https://doi.org/10.1016/j.neuroimage.2017.03.033

Lakretz, Y., Hupkes, D., Vergallito, A., Marelli, M., Baroni, M., & Dehaene, S. (2021). Mechanisms for handling nested dependencies in neural-network language models and humans. *Cognition*, *213*, 104699. https://arxiv.org/abs/2006.11098

Lakretz, Y., Kruszewski, G., Desbordes, T., Hupkes, D., Dehaene, S., & Baroni, M. (2019). The emergence of number and syntax units in lstm language models. *NAACL-HLT (1)*, 11–20. https://arxiv.org/abs/1903.07435

LeBel, A., Wagner, L., Jain, S., Adhikari-Desai, A., Gupta, B., Morgenthal, A., Tang, J., Xu, L., & Huth, A. G. (2022). A natural language fmri dataset for voxelwise encoding models. *Biorxiv*. https://doi.org/10.1101/2022.09.22.509104

Lerner, Y., Honey, C. J., Silbert, L. J., & Hasson, U. (2011). Topographic Mapping of a Hierarchy of Temporal Receptive Windows Using a Narrated Story. *Journal of Neuroscience*, *31*(8), 2906–2915. https://doi.org/10.1523/JNEUROSCI.3684-10.2011

Li, J., Bhattasali, S., Zhang, S., Franzluebbers, B., Luh, W.-M., Spreng, N., Brennan, J. R., Yang, Y., Pallier, C., & Hale, J. (2022). Le petit prince multilingual naturalistic fmri corpus. *Scientific Data*, *9*. https://doi.org/10.1038/s41597-022-01625-7

Mar, R. A. (2011). The neural bases of social cognition and story comprehension. *Annual review of psychology*, *62*, 103–134. https://pubmed.ncbi.nlm.nih.gov/21126178/

Matchin, W., Hammerly, C., & Lau, E. (2017). The role of the IFG and pSTS in syntactic prediction: Evidence from a parametric study of hierarchical structure in fMRI [Publisher: Elsevier]. *cortex*, *88*, 106–123.

Matchin, W., & Hickok, G. (2020). The cortical organization of syntax. *Cerebral Cortex*, *30*(3), 1481–1498. https://doi.org/10.1093/cercor/bhz180

Mazoyer, B. M., Tzourio, N., Frak, V., Syrota, A., Murayama, N., Levrier, O., Salamon, G., Dehaene, S., Cohen, L., & Mehler, J. (1993). The Cortical Representation of Speech [_eprint: https://direct.mit.edu/jocn/article-pdf/5/4/467/1932303/jocn.1993.5.4.467.pdf]. *Journal of Cognitive Neuroscience*, *5*(4), 467–479. https://doi.org/10.1162/jocn.1993.5.4.467

McKinney, W., et al. (2010). Data structures for statistical computing in python. *Proceedings of the 9th Python in Science Conference*, *445*, 51–56. https://conference.scipy.org/proceedings/scipy2010/pdfs/mckinney.pdf

Mollica, F., Diachek, E., Mineroff, Z., Kean, H., Siegelman, M., Piantadosi, S. T., Futrell, R., Qian, P., & Fedorenko, E. (2019). Composition is the core driver of the language-selective network. *bioRxiv*. https://doi.org/10.1101/436204

Mollica, F., Siegelman, M., Diachek, E., Piantadosi, S. T., Mineroff, Z., Futrell, R., & Fedorenko, E. (2018). High local mutual information drives the response in the

human language network. *bioRxiv*, 436204. https://www.biorxiv.org/content/10.1101/436204v1.full

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *NeuroImage*, *56*(2), 400–410. https://doi.org/10.1016/j.neuroimage.2010.07.073

Nastase, S. A., Goldstein, A., & Hasson, U. (2020). Keep it real: Rethinking the primacy of experimental control in cognitive neuroscience, [Publisher: NeuroImage]. *NeuroImage*, *222*. https://doi.org/10.1016/j.neuroimage.2020.117254

Nastase, S. A., Liu, Y.-F., Hillman, H., Zadbood, A., Hasenfratz, L., Keshavarzian, N., Chen, J., Honey, C. J., Yeshurun, Y., Regev, M., & et al. (2021). The "narratives" fmri dataset for evaluating models of naturalistic language comprehension. *Scientific Data*, *8*(1). https://doi.org/10.1038/s41597-021-01033-3

Newman, S. D., Ikuta, T., & Burns, T. (2010). The effect of semantic relatedness on syntactic analysis: An fMRI study. *Brain and language*, *113*(2), 51–58. https://doi.org/10.1016/j.bandl.2010.02.001

O'Reilly, R. C., & Frank, M. J. (2006). Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural computation*, *18*(2), 283–328. https://pubmed.ncbi.nlm.nih.gov/16378516/

Pallier, C., Devauchelle, A.-D., & Dehaene, S. (2011). Cortical representation of the constituent structure of sentences [Publisher: National Acad Sciences]. *Proceedings of the National Academy of Sciences*, *108*(6), 2522–2527.

Pasquiou, A., Lakretz, Y., Hale, J. T., Thirion, B., & Pallier, C. (2022). Neural Language Models are not Born Equal to Fit Brain Data, but Training Helps. *Proceedings of the 39th International Conference on Machine Learning (ICML)*, *162*, 17499–17516. https://arxiv.org/abs/2207.03380

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z.,

Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830. https://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Pennington, J., Socher, R., & Manning, C. (2014). Glove: Global Vectors for Word Representation. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 1532–1543. https://doi.org/10.3115/v1/D14-1162

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., & Fedorenko, E. (2018a). Toward a universal decoder of linguistic meaning from brain activation [Number: 1 Publisher: Nature Publishing Group]. *Nature Communications*, *9*(1), 963. https://doi.org/10.1038/s41467-018-03068-4

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., Botvinick, M., & Fedorenko, E. (2018b). Toward a universal decoder of linguistic meaning from brain activation [Bandiera_abtest: a Cc_license_type: cc_by Cg_type: Nature Research Journals Number: 1 Primary_atype: Research Publisher: Nature Publishing Group Subject_term: Computational science;Neural decoding Subject_term_id: computational-science;neural-decoding]. *Nature Communications*, *9*(1), 963. https://doi.org/10.1038/s41467-018-03068-4

Pulvermüller, F. (2013). Semantic embodiment, disembodiment or misembodiment? In search of meaning in modules and neuron circuits. *Brain and Language*, *127*(1), 86–103. https://doi.org/10.1016/j.bandl.2013.05.015

Radford, A., Wu, J., Child, R., Luan, D., Amodei, D., Sutskever, I., et al. (2019). Language models are unsupervised multitask learners. *OpenAI blog*, *1*(8), 9. https://d4mucfpksywv.cloudfront.net/better-language-models/language_models_are_unsupervised_multitask_learners.pdf

Raichle, M. E. (2015). The brain's default mode network [Publisher: Annual Reviews]. *Annual review of neuroscience*, *38*, 433–447. https://pubmed.ncbi.nlm.nih.gov/25938726/

Regev, M., Honey, C. J., Simony, E., & Hasson, U. (2013). Selective and Invariant Neural Responses to Spoken and Written Narratives. *Journal of Neuroscience*, *33*(40), 15978–15988. https://doi.org/10.1523/JNEUROSCI.1580-13.2013

Russin, J., Jo, J., O'Reilly, R. C., & Bengio, Y. (2019). *Compositional generalization in a deep seq2seq model by separating syntax and semantics* [ARXIV.1904.09708]. https://doi.org/10.48550/ARXIV.1904.09708

Santi, A., & Grodzinsky, Y. (2010). Fmri adaptation dissociates syntactic complexity dimensions. *NeuroImage*, *51*(4), 1285–1293. https://doi.org/https://doi.org/10.1016/j.neuroimage.2010.03.034

Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E. A., Kanwisher, N., Tenenbaum, J., & Fedorenko, E. (2020). *Artificial Neural Networks Accurately Predict Language Processing in the Brain* (tech. rep.) [Company: Cold Spring Harbor Laboratory Distributor: Cold Spring Harbor Laboratory Label: Cold Spring Harbor Laboratory Section: New Results Type: article]. MIT. https://doi.org/10.1101/2020.06.26.174482

Shetreet, E., & Friedmann, N. (2014). The processing of different syntactic structures: Fmri investigation of the linguistic distinction between wh-movement and verb

movement. *Journal of Neurolinguistics*, *27*(1), 1–17. https://www.sciencedirect.com/science/article/abs/pii/S0911604413000468

Siegelman, M., Blank, I. A., Mineroff, Z., & Fedorenko, E. (2019). An attempt to conceptually replicate the dissociation between syntax and semantics during sentence comprehension [Publisher: Elsevier]. *Neuroscience*, *413*, 219–229. https://www.sciencedirect.com/science/article/pii/S0306452219304026

Simony, E., Honey, C. J., Chen, J., Lositsky, O., Yeshurun, Y., Wiesel, A., & Hasson, U. (2016). Dynamic reconfiguration of the default mode network during narrative comprehension [Number: 1 Publisher: Nature Publishing Group]. *Nature Communications*, *7*(1), 12141. https://doi.org/10.1038/ncomms12141

Sinha, K., Jia, R., Hupkes, D., Pineau, J., Williams, A., & Kiela, D. (2021). Masked language modeling and the distributional hypothesis: Order word matters pre-training for little. *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2888–2913. https://doi.org/10.18653/v1/2021.emnlp-main.230

Sperry, R. W. (1961). Cerebral Organization and Behavior: The split brain behaves in many respects like two separate brains, providing new research possibilities. [Publisher: American Association for the Advancement of Science]. *Science*, *133*(3466), 1749–1757. https://pubmed.ncbi.nlm.nih.gov/17829720/

Stromswold, K., Caplan, D., Alpert, N., & Rauch, S. (1996). Localization of syntactic comprehension by positron emission tomography. *Brain and Language*, *52*(3), 452–473. https://doi.org/https://doi.org/10.1006/brln.1996.0024

Toneva, M., & Wehbe, L. (2019). Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In H. Wallach, H. Larochelle, A. Beygelzimer, F. d. Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in Neural Information Processing Systems 32* (pp. 14954–14964). Curran Associates, Inc. Retrieved October 12, 2020, from

http://papers.nips.cc/paper/9633-interpreting-and-improving-natural-language-processing-in-machines-with-natural-language-processing-in-the-brain.pdf

Ullman, M. T. (2004). Contributions of memory circuits to language: The declarative/procedural model. *Cognition*, *92*(1), 231–270. https://www.sciencedirect.com/science/article/pii/S0010027703002324

Vigliocco, G. (2000). Language processing: The anatomy of meaning and syntax. *Current Biology*, *10*(2), R78–R80. https://doi.org/https://doi.org/10.1016/S0960-9822(00)00282-7

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., ... 1.0 Contributors, S. (2020). Scipy 1.0: Fundamental algorithms for scientific computing in python. *Nature Methods*, *17*(3), 261–272. https://doi.org/10.1038/s41592-019-0686-2

Wehbe, L., Murphy, B., Talukdar, P., Fyshe, A., Ramdas, A., & Mitchell, T. (2014). Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses [Publisher: Public Library of Science]. *PloS one*, *9*(11). https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0112575

Wolf, T. (2020). *Huggingface* [To appear] [http://huggingface.co]. https://huggingface.co/

Xu, J., Kemeny, S., Park, G., Frattali, C., & Braun, A. (2005). Language in context: Emergent features of word, sentence, and narrative comprehension. *NeuroImage*, *25*(3), 1002–1015. https://doi.org/10.1016/j.neuroimage.2004.12.013

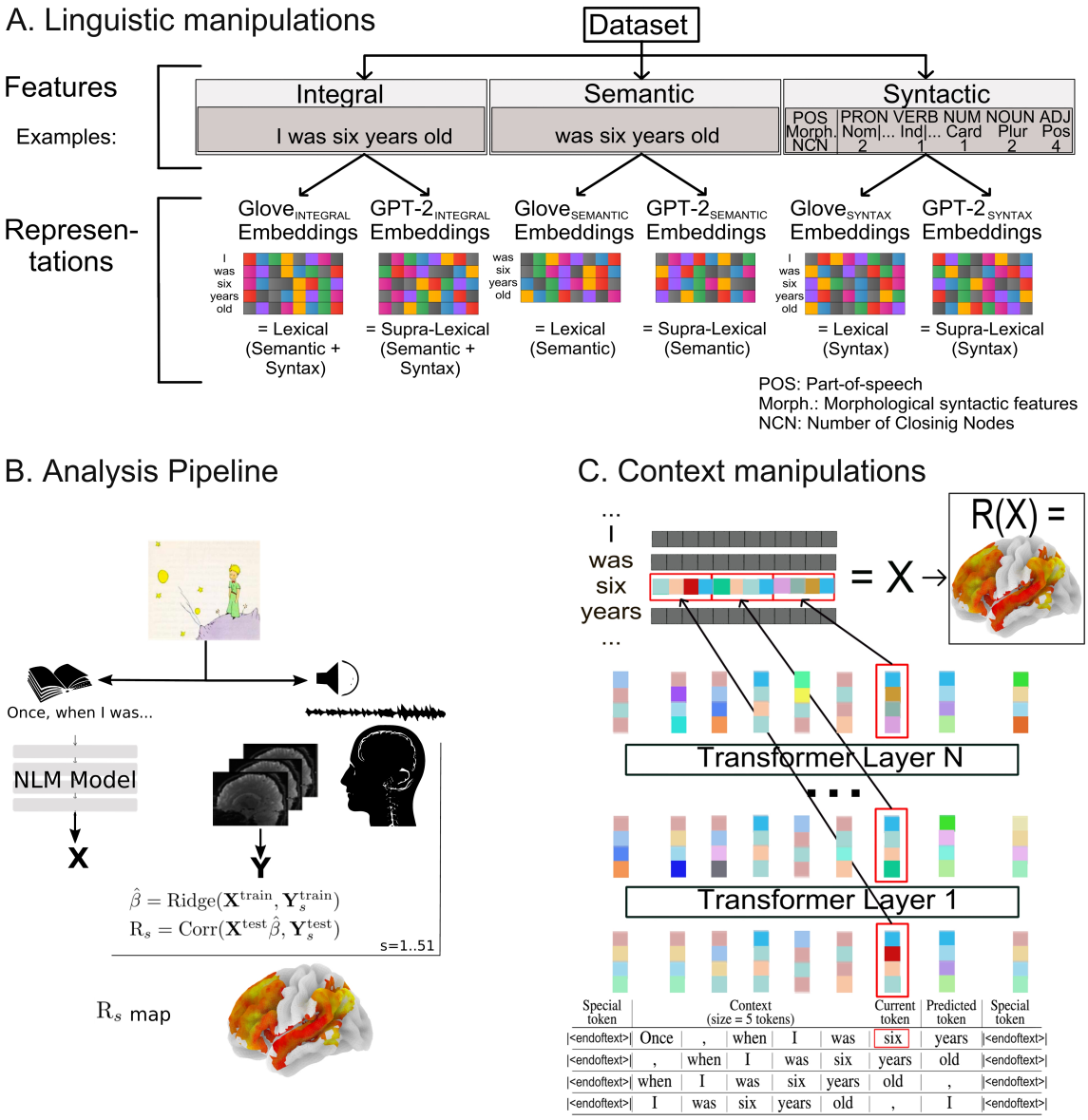


Figure 1

Experimental setup A) *A corpus of novels was used to create a dataset from which we extracted three different sets of features: (i) Integral features, comprising all tokens (words+punctuation); (ii) Semantic features, comprising only the content words; (iii) Syntactic features, comprising syntactic characteristics (Part-of-speech, Morphological syntactic characteristics, Number of Closing Nodes) of all tokens. GloVe and GPT-2 models were trained on each feature space.* ***B)*** *fMRI scans of human participants listening to an audio-book were obtained. The associated text transcription was input to Neural models, yielding embeddings that were convolved with an haemodyamic kernel and fitted to brain activity using a Ridge-regression. Brain maps of cross-validated correlation between encoding models' predictions and fMRI time-series were computed.* ***C)*** *To study sensitivity to context, a GPT-2 model was trained and tested on input sequences of bounded context length (5, 15 and 45). The resulting representations were then used to predict fMRI activity.*

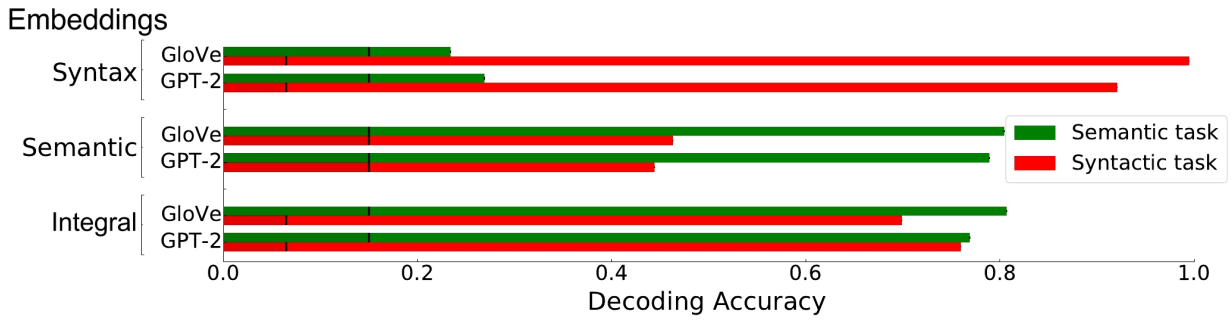


Figure 2

Decoding syntactic and semantic information from words embeddings. *For each dataset and model type (Glove and GPT-2), logistic classifiers were set up to decode either the syntactic or the semantic categories of the words from the text of* The Little Prince. *Chance-level was assessed using dummy classifiers and is indicated by black vertical lines.*

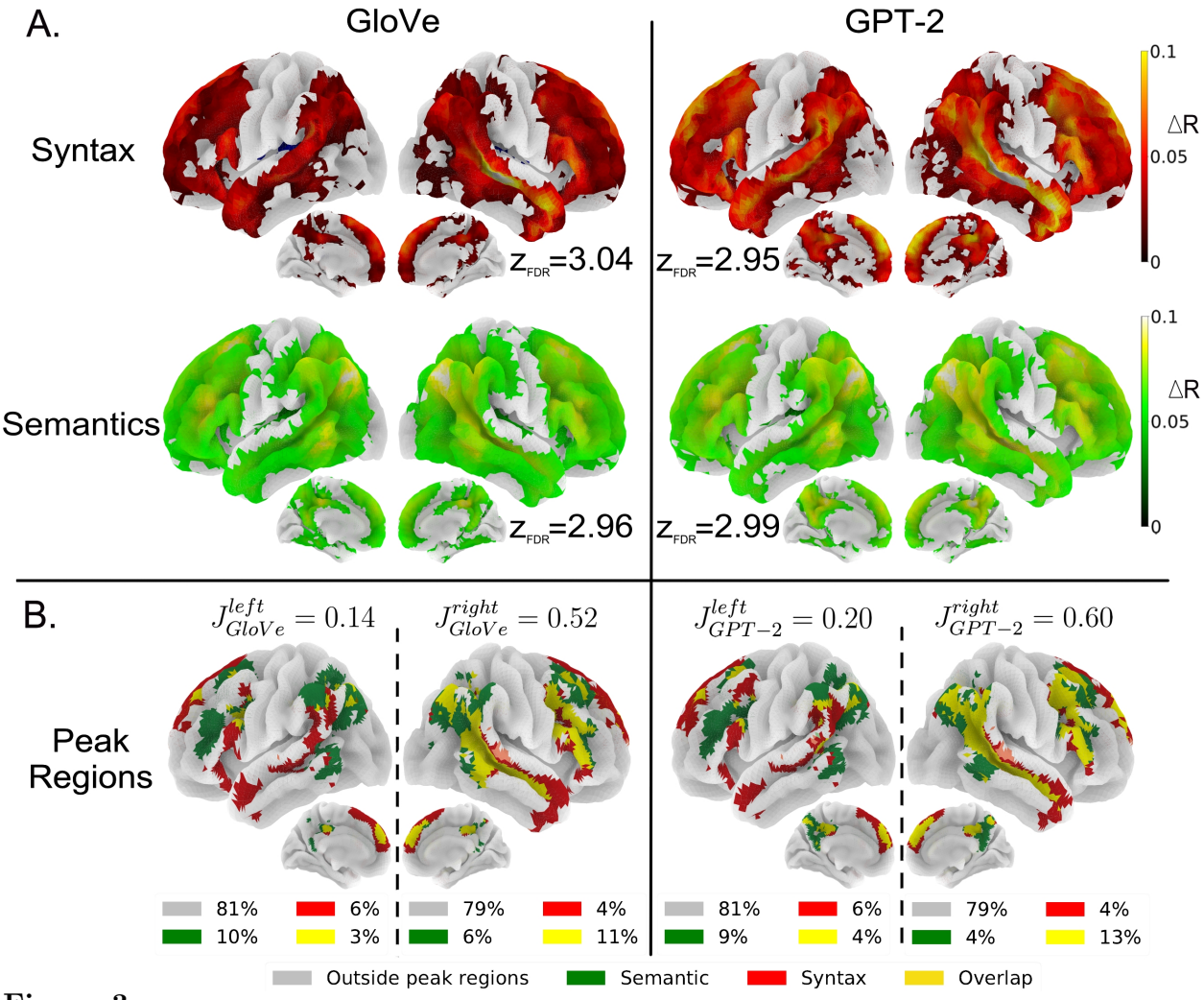


Figure 3

Comparison of the ability of GloVe and GPT-2 to fit brain data when trained on either the semantic or the syntactic features. ***A)*** *Significant increase in R scores relative to the baseline model for GloVe (a non contextual model) and GPT-2 (a contextual model), trained either on the Syntactic features or on the Semantic features (voxel-wise thresholded group analyses; N=51 subjects; corrected for multiple comparisons with a FDR approach $p < 0.005$; for each figure z_{FDR} indicates the significance threshold on the Z-scores).* ***B)*** *Bilateral spatial organisation of syntax and semantics highest R scores. Voxels whose R score belong in the 10% highest R scores (in green for models trained on the semantic features, and in red for models trained on the syntactic features) are projected onto brain surface maps for GloVe and GPT-2 (overlap in yellow and other voxels in grey). Jaccard score for each hemisphere are computed, i.e. the ratio between the size of the intersection and the size of the union of semantics and syntax peak regions; the proportion of voxels of each category are displayed for each hemisphere and model.*

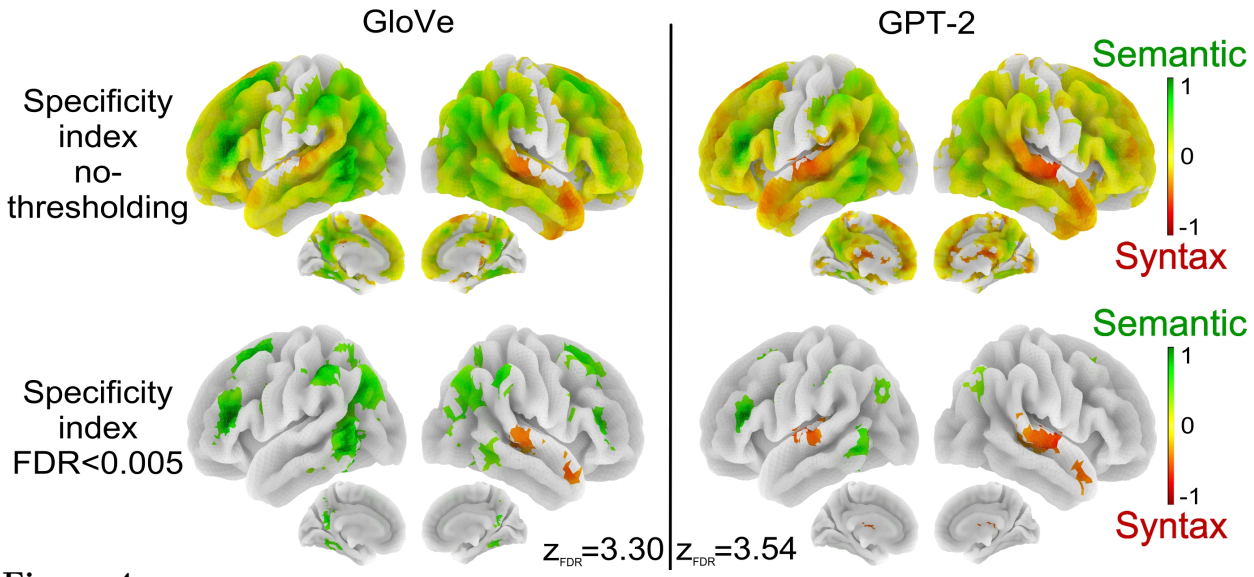


Figure 4

Voxels' sensitivity to syntactic and semantic embeddings. *Voxels' specificity indexes are projected onto brain surface maps reflecting how much semantic information helps to better fit the time-courses of a voxel compared to syntactic information; the greener the more the voxel is categorized as a semantic voxel, the redder the more the voxel is categorized as a syntactic voxel. Yellow regions are brain areas where semantic and syntactic information lead to similar R score increases. The top row displays specificity indexes in voxels where there was a significant effect for semantic or syntactic embeddings in Fig.3A. The bottom row is the voxel-wise thresholded group analyses; N=51 subjects; corrected for multiple comparisons with* $FDR < 0.005$ *(for each figure* z_{FDR} *indicates the significance threshold on the Z-scores).*

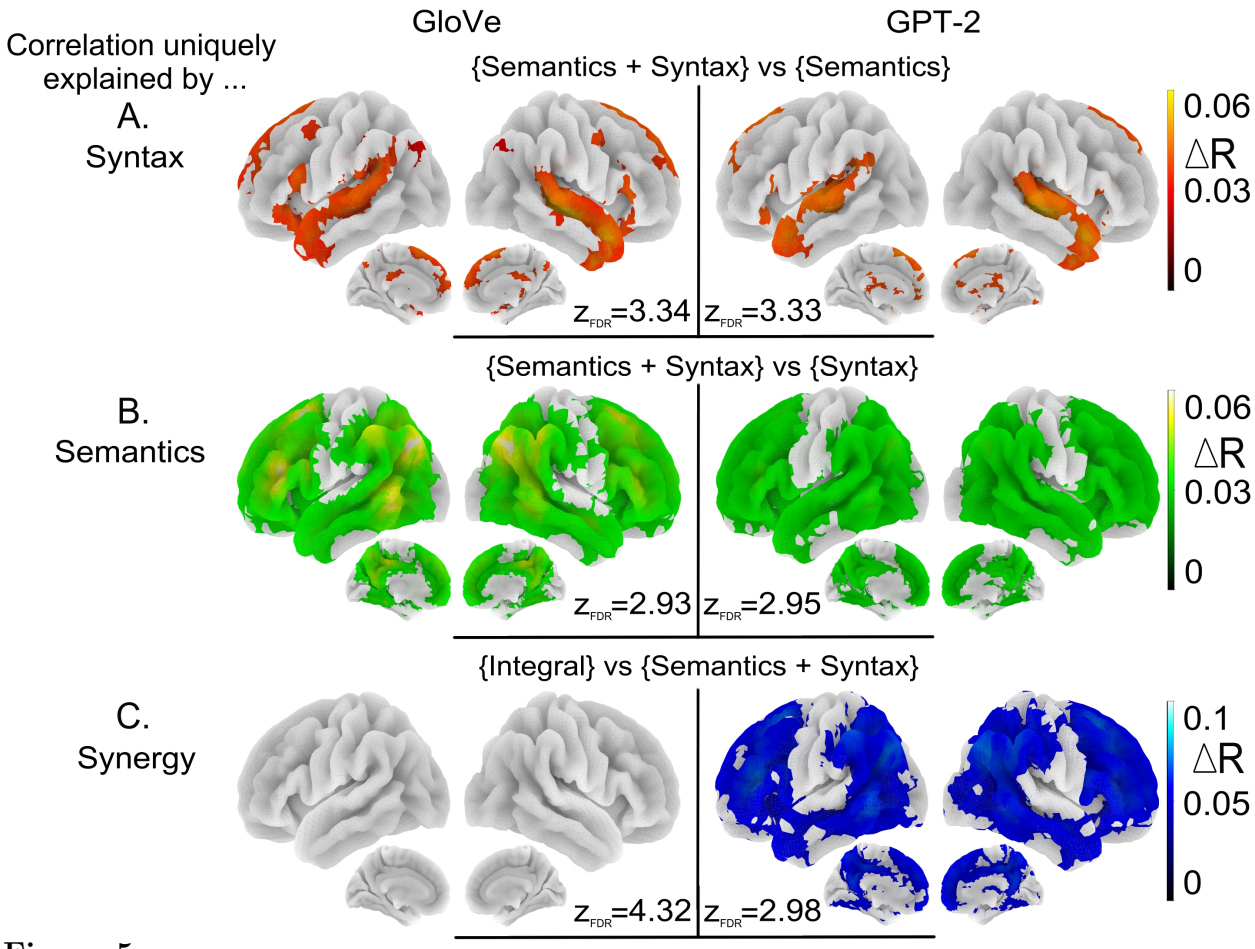


Figure 5

Correlation uniquely explained by each embeddings A) *Increase in R scores relative to the semantic embeddings when concatenating semantic and syntactic embeddings in the encoding model.* ***B)*** *Increase in R scores relative to the syntactic embeddings when concatenating semantic and syntactic embeddings in the encoding model.* ***C)*** *Increase in R scores relative to the concatenated semantic and syntactic embeddings for the integral embeddings. These maps are voxel-wise thresholded group analyses; N=51 subjects; corrected for multiple comparisons with a FDR approach* $p < 0.005$*; for each figure* z_{FDR} *indicates the significance threshold on the Z-scores.*

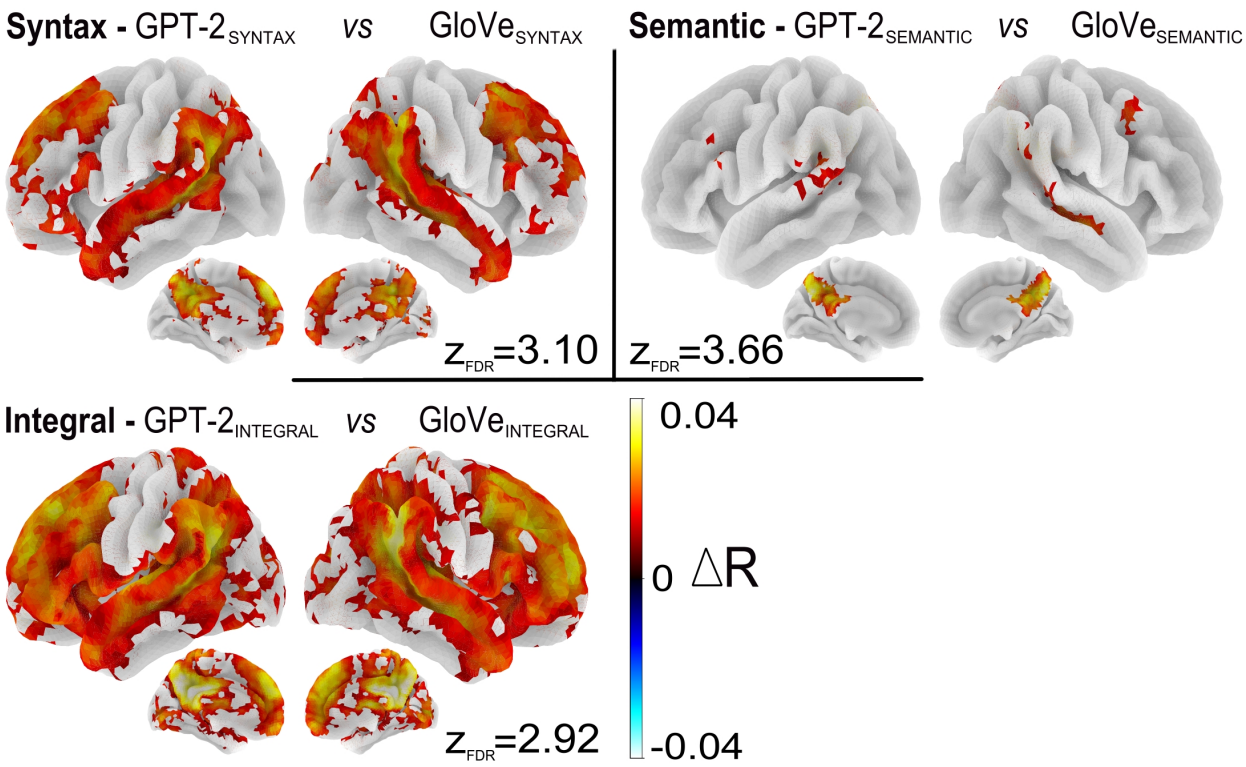


Figure 6

Comparison of lexical and supra-lexical processing levels. *Brain regions that are significantly better predicted by GPT-2 (in red) compared to Glove, when trained on syntactic features (top left), semantic features (top right) and integral features (bottom left). Maps are voxel-wise thresholded group analyses; N=51 subjects; corrected for multiple comparisons with a FDR approach* $p < 0.005$*; for each figure* z_{FDR} *indicates the significance threshold on the Z-scores.*

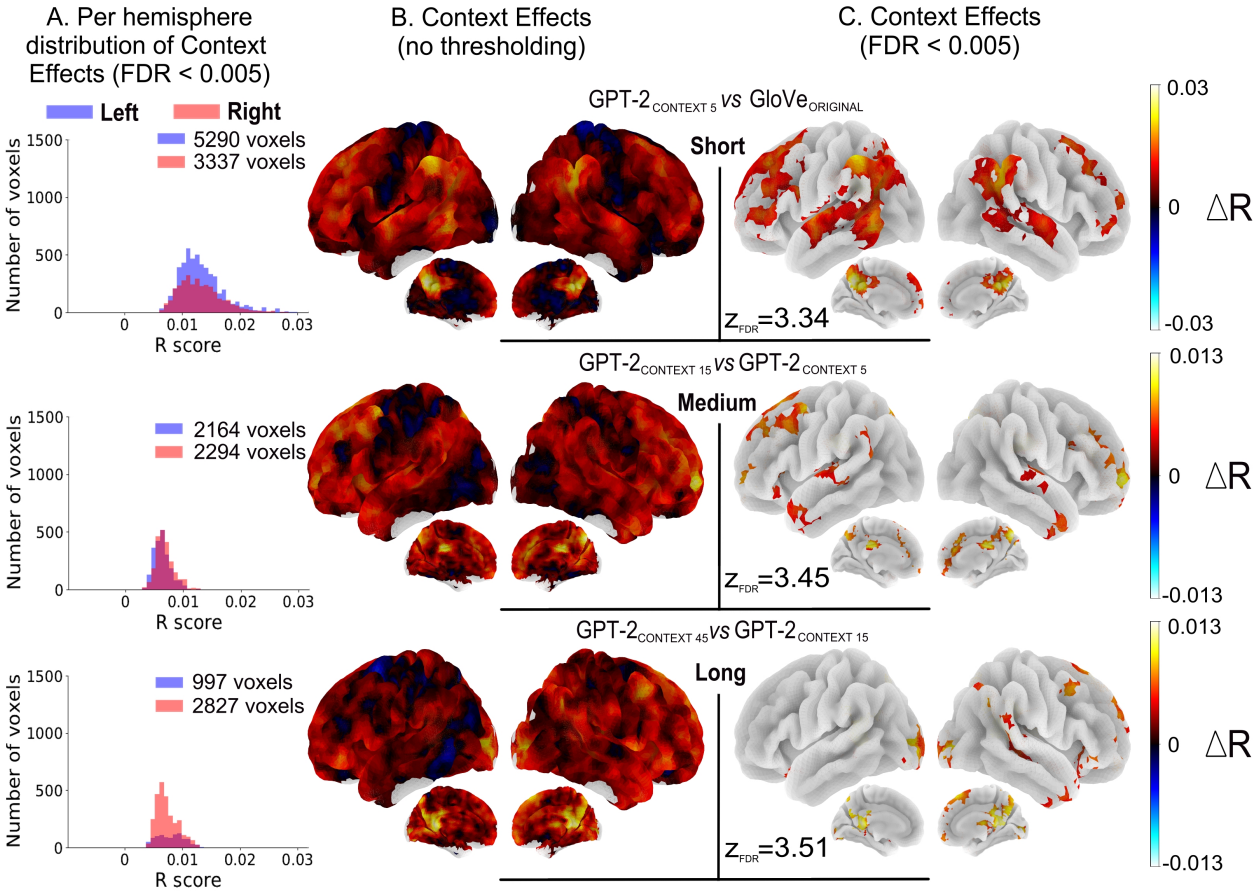


Figure 7

Integration of context at different levels of language processing. ***A)*** *Per hemisphere histograms of significant context effects after group analyses (N=51 subjects); thresholded at p<0.005 voxel-wise, corrected for multiple comparisons with the FDR approach.* ***B)*** *Uncorrected group averaged surface brain maps representing R scores increases when fitting brain data with models leveraging increasing sizes of contextual information.* ***C)*** *Corrected group averaged surface brain maps representing R scores increases when fitting brain data with models leveraging increasing sizes of contextual information; thresholded at p<0.005 voxel-wise, corrected for multiple comparisons with the FDR approach (for each figure z_{FDR} indicates the significance threshold on the Z-scores).* ***(top row)*** *Comparison of the model trained with 5 tokens of context (GPT-$2_{Context-5}$) with the non-contextualized GloVe.* ***(middle row)*** *Comparison of the models respectively trained with 15 (GPT-$2_{Context-15}$) and 5 (GPT-$2_{Context-5}$) tokens of context.* ***(bottom row)*** *Comparison of the models respectively trained with 45 (GPT-$2_{Context-45}$) and 15 (GPT-$2_{Context-15}$) tokens of context.*

Appendix A

Models training

We trained GloVe and GPT-2 on syntactic or semantic features by adapting both vocabulary size and the associated tokenizer. Table A1 recapitulates information about the training of the models used. Table A2 provides examples of the features extracted from a short passage. After feature extraction, a vocabulary listing all possible feature instances is created for each feature type. A unique id is then associated to each element of the vocabulary. The tokenizer converts each feature to its unique id. Finally, the model is fed sequences of ids and learns to perform its task.

Table A1

Models hyperparameters

Models Models	Number of tokens	Number of unique words	Context window	Vector size	Number of epochs	Number of layers
GloVe Syntax	980 M	1190	15	768	20	NaN
GPT-2 Syntax	980 M	1190	512	768	5	4
GloVe Semantics	370 M	91880	15	768	20	NaN
GPT-2 Semantics	370 M	91880	512	768	5	4
GloVe Integral	980 M	92945	15	768	20	NaN
GPT-2 Integral	980 M	92945	512	768	5	4

Table A2

Examples of input sequences given to the neural language models when trained on the different feature spaces.

		Input sequence						
Integral Features		The	sixth	planet	was	ten	times	larger
Syntactic Features	Part-of-Speech	DET	ADJ	NOUN	VERB	NOUN	NOUN	ADJ
	Morphology	Definite=Def|PronType=Art	Degree=Pos	Number=Sing	Ind|Sing|Past|Person=3|Fin	Number=Card	Number=Plur	Degree=Cmp
	Number of Closing Nodes	1	1	2	1	1	2	2
Semantic Features	Content words	–	sixth	planet	–	ten	times	larger

The Morphology field contains a list of morphological features, with vertical bar (|) as list separator and with underscore to represent the empty list. All features represent attribute-value pairs, with an equals sign (=) separating the attribute from the value. In addition, features are selected from the universal feature inventory (https://universaldependencies.org/u/feat/index.html) and are sorted alphabetically by attribute names. It is possible that a feature has two or more values for a given word: Case=Acc,Dat. In this case, the values are sorted alphabetically.

Note: for display purposes, the morphology attribute values were removed for 'was', it was originally equal to 'Mood=Ind|Number=Sing|Person=3|Tense=Past|VerbForm=Fin'.

Appendix B

Context-limited models

Using the same original collection of English novels from Project Gutenberg, we trained three GPT-2 models to probe context integration. More precisely, we restricted the preceding context (size $k = 5, 15$ or 45 tokens) given to the GPT-2 models during training on the "*Integral dataset*".

When training GPT-2 with a limited amount of contextual information, each input sequence contained $k + 5$ tokens: a special token at the beginning, k context tokens, the current token for which we retrieve the activations in order to fit fMRI brain data, the token that is predicted by the current token and the 2 special tokens at the end (the last special end-of-sentence token is always preceded by a token encoding a blank space, we omitted it in the following table).

Table B1

Examples of context-limited input sequences given to GPT-2 for the analyses on context-integration. Here the context size k is equal to 5.

Special token	Context (size = 5 tokens)					Current token	Predicted token	Special token
|<endoftext>|	Once	,	when	I	was	six	years	|<endoftext>|
|<endoftext>|	,	when	I	was	six	years	old	|<endoftext>|
|<endoftext>|	when	I	was	six	years	old	,	|<endoftext>|
|<endoftext>|	I	was	six	years	old	,	I	|<endoftext>|

Appendix C

Removing absolute position information in GPT-2 trained on semantic features

For the GTP-2 model trained on the semantic features, small modifications had to be made to the model architecture in order to remove all residual syntax. By default, GPT-2 encodes the absolute positions of tokens in sentences. As word ordering might contain syntactic information, we had to make sure that it could not be leveraged by GPT-2 by means of its positional embeddings, yet keeping information about word proximity as it influences semantics. We achieved it by slightly modifying the architecture of GPT-2: we first removed the default positional embeddings, and added to the attention scores embeddings encoding relative positions between input tokens. Indeed, just removing positional embeddings would have led to a bag-of-words model. By adding these embeddings encoding relative position to the attention scores a token will weight the attention granted to another token depending on their distance. By doing so, information about absolute and relative positions is removed from tokens' embeddings as it is not directly added to the tokens' hidden states. The following explains how this operation was performed. Let $\mathbf{c}_W = (c_{w_1}, \ldots, c_{w_m})$ be a sequence of m tokenized content words. $\mathbf{c}_W$ is then fed to a n_{layers} transformer with n_{heads} of dimension d_{heads} that first build an embedding representation $\mathbf{E}_i, i = 1..m$ (of size $d = d_{heads} * n_{heads}$) to which it appends (by default) a position embedding $\mathbf{p}_i, i = 1..m$ (of size d) for each token. To remove all syntactic content, the first step is to discard the previously mentioned positional embeddings $\mathbf{p}_i, i = 1..m$. However stopping here would only lead to a bag-of-word model where a given token might be influenced similarly by an adjacent token or one far away. As a consequence, we had to weight the attention score granted to a token depending on its relative distance.

The attention operation can be described as mapping a query (Q) and a set of key-value (K, V) pairs to an output, where the query, keys, values, and output are all vectors (generally packed into matrices). The output is computed as a weighted sum of the values, where the weight assigned to each value is computed by a compatibility function of

the query with the corresponding key. We thus modify the classical attention operation:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}((\mathbf{Q}\mathbf{K}^T)/\sqrt{d_k})\mathbf{V}$$

by adding the previously described relative positional embedding $\mathbf{W}$ in the attention mechanisms:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}((\mathbf{Q}\mathbf{K}^T + \mathbf{W})/\sqrt{d_k})\mathbf{V}$$

To build $\mathbf{W}$, we first defined the matrix $\mathbf{D} = (n - 1 + j - i)_{i,j=1..m} \in \mathbb{R}^{m \times m}$ (encoding the number of tokens separating two tokens in the input sequence shifted by $n - 1$) for each input sequence $\mathbf{c}_W$, where n is the maximal input size. $\mathbf{D}$ is then embedded using a lookup table that stores an embedding of size (d_{head}) for each possible value of $\mathbf{D}$, giving $\mathbf{U}$ ($\in \mathbb{R}^{m \times m \times d_{head}}$).

Finally, the weights assigned to the value vectors are adjusted using the embedded relative distances between tokens $\mathbf{W}$ ($\in \mathbb{R}^{n_{heads} \times m \times m}$), defined as:

$$W_{i,j,k} = \sum_{d=1}^{d_{head}} K_{i,j,d} U_{j,k,d}$$

By doing so, we were able to weight words interactions depending on their relative distance in the input sequence, while removing all absolute positional information from tokens hidden-states.

Appendix D

Convergence of the language models during training

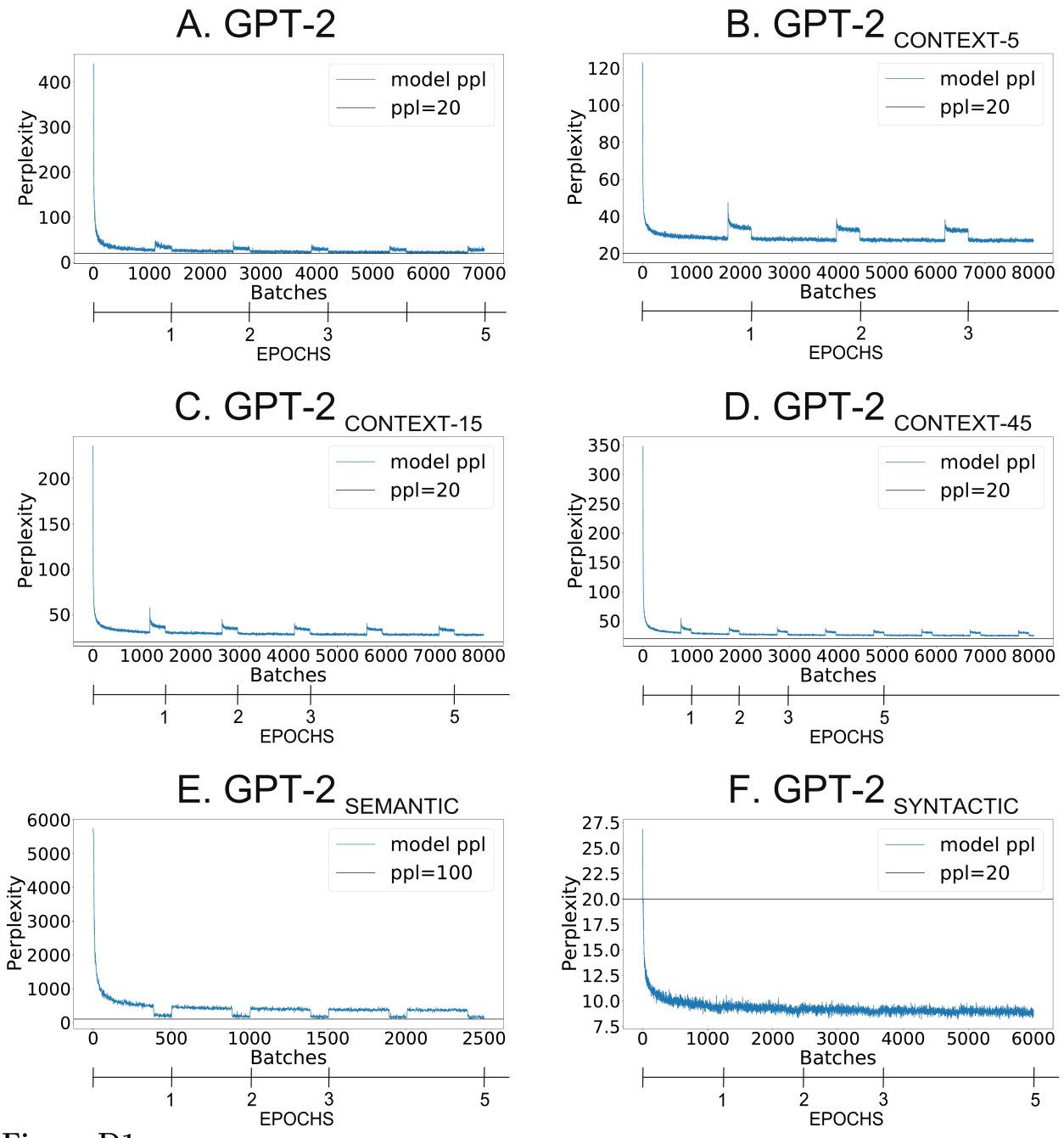


Figure D1

GPT-2 convergence during training. *The models represented in panels A to D were trained on the integral features. Models in panels E and F were respectively trained on the semantic and syntactic features. Models were trained until no further improvement could be observed on the validation set.*

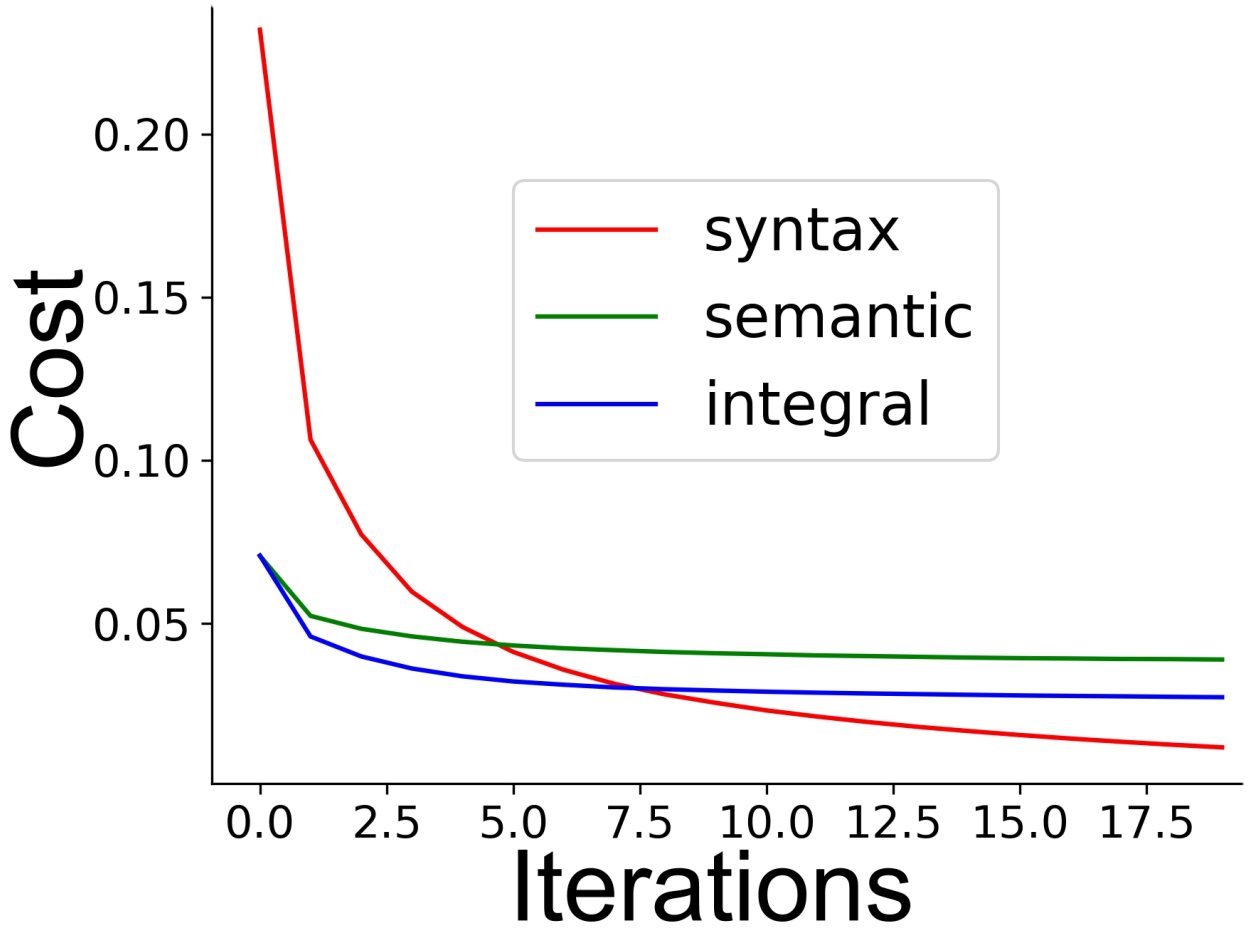


Figure D2

GloVe convergence during training. *The models represented were trained on the integral features (blue), semantic features (green) or syntactic features (red). Models were trained until no further improvement could be observed on the validation set.*

Appendix E

Decoding individual syntactic features

Appendix 1-Fig.E1 shows the decoding accuracies of GloVe and GPT-2 models when trained on one of the three datasets. The syntactic labels varies from encompassing all categories (Morph+POS+NCN), to only one of them.

It is important to highlight that the decoding performance of the semantic models on the syntactic decoding task primarily relies on *Morph*. In contrast, the decoding of the Part-of-speech or the Number of Closing Nodes (NCN) are at chance level. This suggests that information related to the gender or plurals might be encoded by both syntactic and semantic embeddings. In addition, there are indeed more semantic labels compared to the syntactic ones. Consequently, the space occupied by syntactic embeddings is relatively smaller than that of semantic embeddings. As a result, it is relatively easier to project the larger semantic space onto the syntactic embedding space.

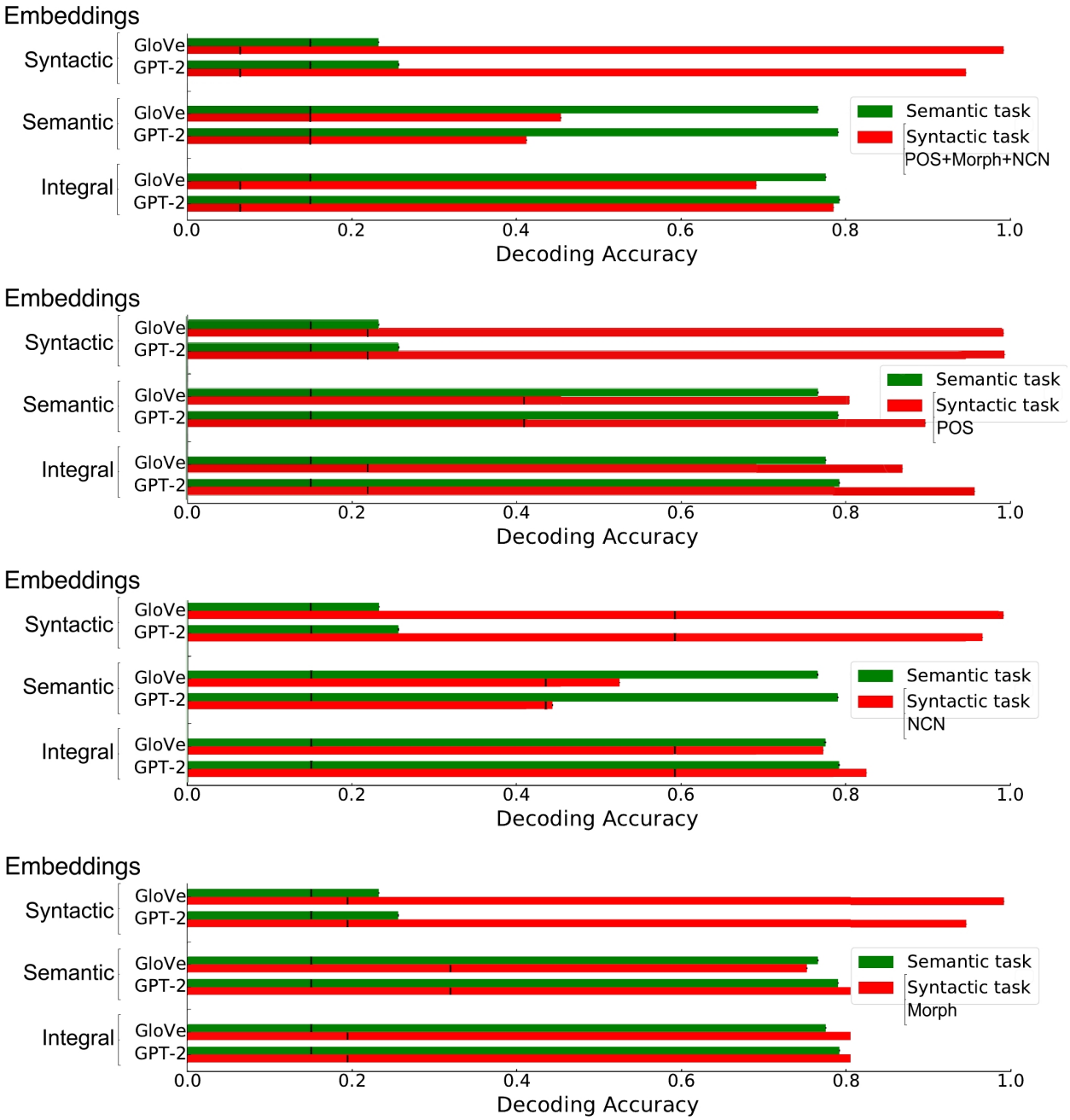


Figure E1

Decoding syntactic information from words embeddings. *For each dataset and model type (Glove and GPT-2), logistic classifiers were set up to decode either the syntactic or the semantic categories of the words from the text of The Little Prince. Chance-level was assessed using dummy classifiers and is indicated by black vertical lines. From top to bottom: The syntactic label is i) the triplet (POS, Morph, NCN), ii) the POS, iii) the NCN, iv) the Morph.*

Appendix F

Mapping NLM activations to brain data

Given two non-linear transformations φ_1 (the neural language model that takes as input the sentence and from which we extract latent representations) and φ_2 (the brain that takes as input the sentence and from which we extract voxels' activations) and an input sequence $\mathbf{w} = (\mathrm{w}_1, \ldots, \mathrm{w}_M)$, we define $\mathbf{Y}_s = \varphi_2(\mathbf{w}) \in \mathbb{R}^{N \times V}$ and $\mathbf{X} = \varphi_1(\mathbf{w}) \in \mathbb{R}^{M \times d}$, and we aimed at finding a linear transformation from $\mathbf{X}$ to $\mathbf{Y}_s$, where d is the dimension of the model, V is the number of brain voxels, and N the number of fMRI scans acquired. One issue is that $\mathbf{X}$ and $\mathbf{Y}_s$ don't have the same sampling frequency: $\mathbf{X}$ being defined at word-level while $\mathbf{Y}_s$ has been re-sampled at the fMRI acquisition frequency, every 2 seconds. To map $\mathbf{X}$ to $\mathbf{Y}_s$ we first need to temporally align them, taking the dynamic of the fMRI BOLD signal into account, and then determine a linear spatial mapping between the convolved and re-sampled $\mathbf{X}$ and $\mathbf{Y}_s$. Using the standard model-based encoding approach to modelling fMRI signals (Huth et al., 2016; Naselaris et al., 2011; Pasquiou et al., 2022), we first convolve each column of $\mathbf{X}$ with the *SPM* haemodynamic kernel (K), which corresponds to the profile of the fMRI BOLD response following a Dirac stimulation, and then sub-sampled the signal to match the sampling frequency of $\mathbf{Y}_s$, giving $\tilde{\mathbf{X}} = S_{ub}(K \circ X)$, with S_{ub} the sub-sampling operator. Finally, we learn the linear spatial mapping between $\tilde{\mathbf{X}}$ and $\mathbf{Y}_s$ using a nested cross-validated L2-regularized (aka Ridge) univariate linear encoding model. More precisely, for each voxel $\mathbf{y}_s^v$, we learn a linear projection $\hat{\boldsymbol{\beta}}_s^v$ from $\tilde{\mathbf{X}}$ to $\mathbf{y}_s^v$ using a nested cross-validated L2-regularized univariate linear encoding model whose general solution is given by:

$$\hat{\boldsymbol{\beta}}_{\boldsymbol{s}}^{\boldsymbol{v}} = arg\min_{\boldsymbol{\beta_s}} \|\mathbf{y}_s^v - \boldsymbol{\beta}_{\boldsymbol{s}}^T \mathbf{X}\|^2 + \lambda \|\boldsymbol{\beta}_{\boldsymbol{s}}\|_2^2 \text{ i.e. } \hat{\boldsymbol{\beta}}_s = \mathrm{Ridge}(\mathbf{X}, \boldsymbol{Y_s})$$

The latter stage resulted for each model and each run into a design matrix $\mathbf{X}$ of size $N \times d$. Given a neural language model, we gave the associated nine design-matrices to a nested cross-validated L2-regularized univariate linear encoding model to fit the fMRI brain data (of size $N \times V$). To evaluate model performance and the optimal regularization parameter

λ^*, we used a nested cross-validation procedure: we split each participant's dataset into training, validation and test sets, such that the training set included 7 out of the 9 experiment runs, and the validation and test sets contained one of the two remaining sessions. We evaluated model performance using Pearson correlation coefficient R, which is a measure of the linear correlation between encoding models' predicted time-courses and the actual time-courses. For each subject and each voxel, we first determined λ^* by comparing R_{valid} for 10 different values of λ, linearly spaced in log-scale between 10^{-3} and 10^4. We then calculated R_{test} for λ^*. Finally, we repeated this procedure 9 times, using cross-validation. This resulted in 9 R_{test} values that we then averaged to produce a single R_{test} map for the participant. We evaluated the quality of the mapping for subject s in voxel v using Pearson correlation:

$$R(X)_s^v = \mathrm{Corr}(\mathbf{Y_s^v}, \hat{\boldsymbol{\beta}}_\mathbf{s}^\mathbf{v}\mathbf{X})$$

Appendix G

The Basic Features baseline model

To assess the specific impact of NLMs' embeddings, the maps shown in Fig.3 report *increases in R values* relative to a *baseline model* which comprised three variables of non-interest:

- acoustic energy (root mean squared of the audio signal sampled every 10ms)
- word offsets (one event at each word offset)
- log of the lexical frequency of each word (modulator of the words events).

More generally, as we looked at increases in R scores between models, the baseline model was appended to all other models studied in order to cancel out the effects of the 3 features of non-interest. Appendix 1-Fig.G1 below displays the cross-validated correlations obtained from this baseline model.

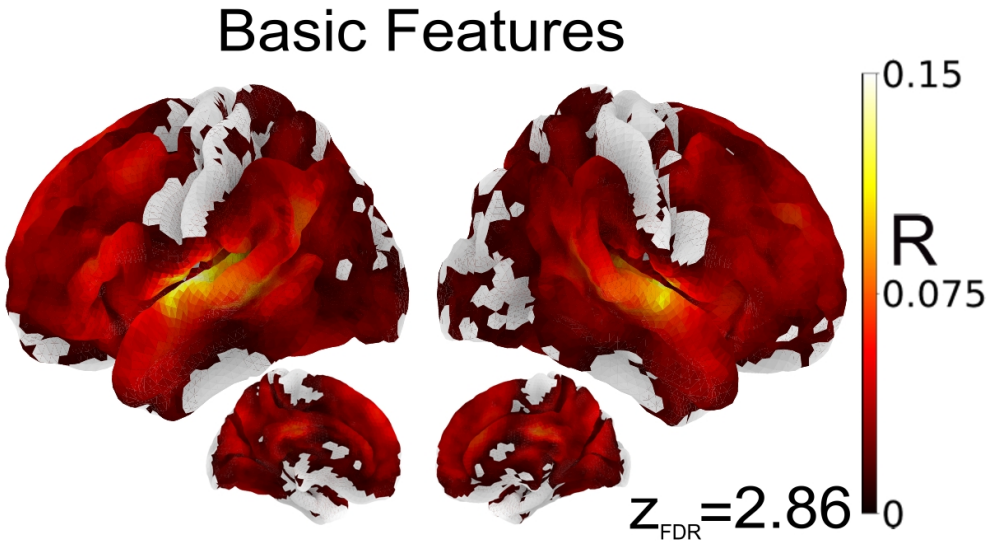


Figure G1

Brain regions showing significant activations for the Basic Features baseline model. *Using the Basic Features (BF) baseline model to fit fMRI brain data, we displayed voxels where there was a significant correlation (voxel-wise thresholded group analyses; N=51 subjects; corrected for multiple comparisons with a FDR approach $p < 0.005$; z_{FDR} is the FDR threshold on the z-scores). The effects from the Basic Features baseline model were discarded from all the analyses in the paper.*

Appendix H

Brain fit of GloVe and GPT-2 when trained on the Integral Features

Appendix 1-Fig.H1 shows the increase in R, relative to the baseline model, provided by the GloVe and GPT-2 models trained on the Integral Features, that is, the intact text.

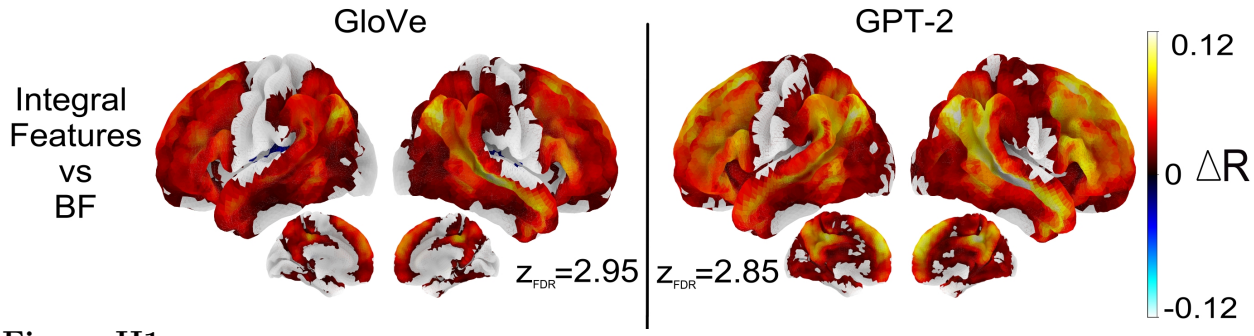


Figure H1

Brain regions showing significant *R* ***score increases compared to the Baseline Model for GloVe and GPT-2 when trained on the Integral Features.*** *Increases in R scores relative to the baseline model for GloVe (a non contextual model) and GPT-2 (a contextual model), trained on the Integral features (voxel-wise thresholded group analyses; N=51 subjects; corrected for multiple comparisons with a FDR approach* $p < 0.005$*;* z_{FDR} *is the FDR threshold on the z-scores).*

Appendix I

R Scores Distribution for GloVe and GPT-2 Trained on Semantic or Syntactic Features

Appendix 1-Fig.I1 below shows the averaged (across participants) voxels distribution, of the increase in R scores obtained from GloVe and GPT-2 models on semantic or syntactic features, relative to the baseline model.

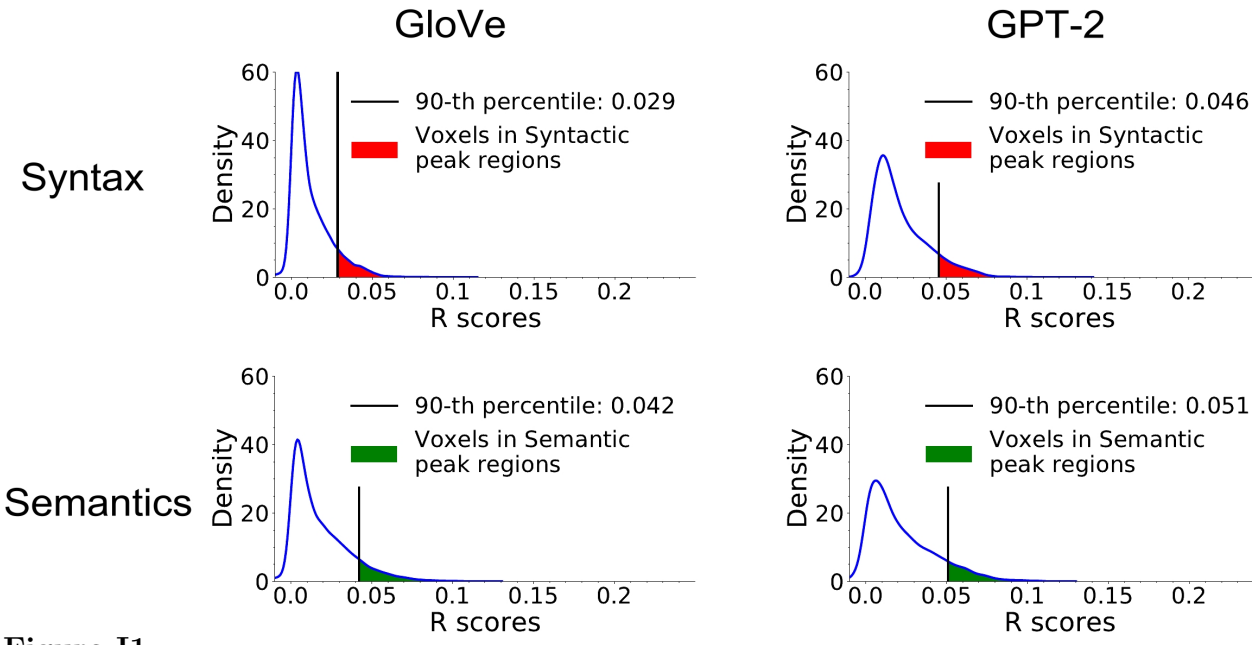


Figure I1

Distribution of *R* ***scores derived from GloVe and GPT-2 semantic and syntactic embeddings.*** *The 90th-percentile of the R scores distribution is highlighted with a vertical black line and used to select voxels for the peak regions analyses.*

Appendix J

Models trained on Semantic features vs models trained on Syntactic features

Appendix 1-Fig.J1 shows the differences in R scores between the semantic and syntactic models, for Glove and GPT-2. Correcting for multiple comparisons (N=51; $p < 0.005$ after FDR correction), we observed significant differences in favor of the syntactic embeddings in the STG, and significant differences in favor of the semantic embeddings in the pMTG, the AG and the IFS and SFS.

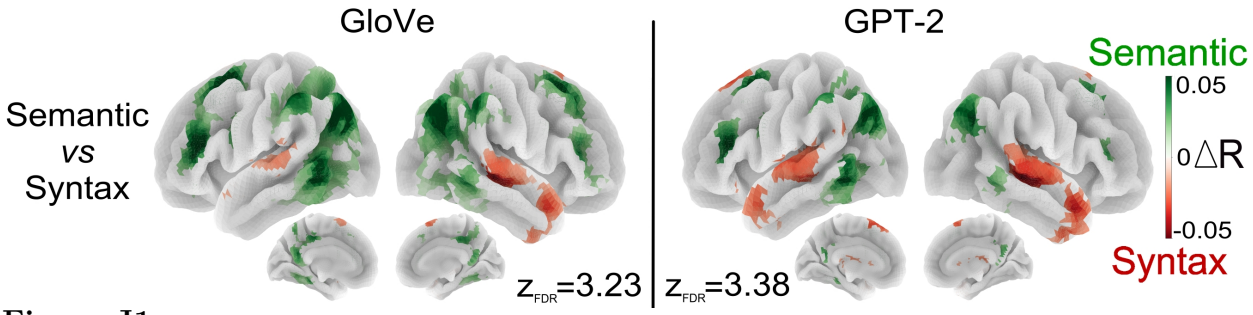


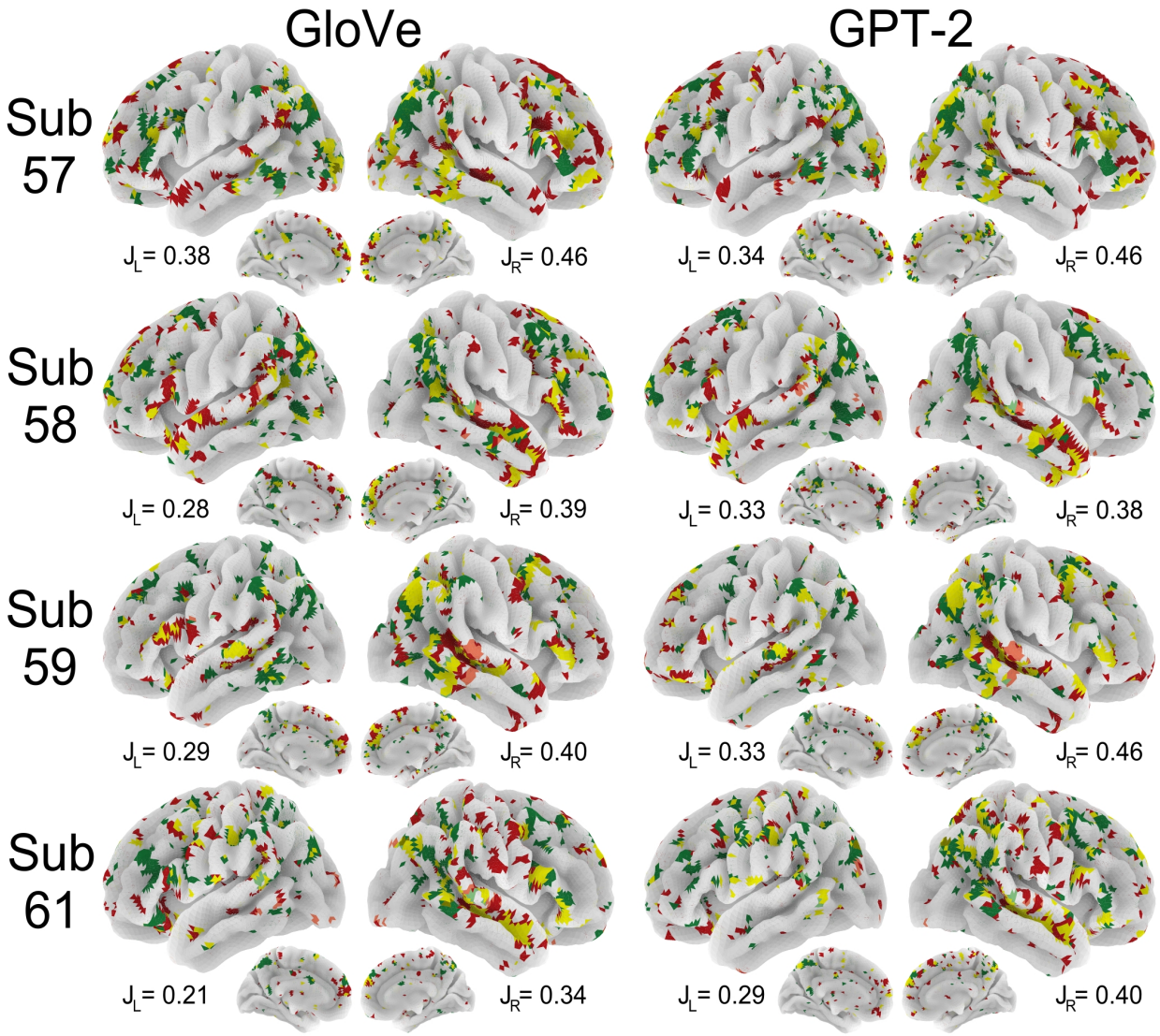
Figure J1

Comparison of the models trained on Semantic features with the models trained on Syntactic features. *Significant R score differences between the models trained on Semantic features and the models trained on Syntactic features. The brain regions that are better fitted by the former model appear in green, while the regions better fitted by the latter model appear in red. (All these maps represent voxel-wise thresholded group analyses; N=51 subjects; corrected for multiple comparisons with a FDR approach* $p < 0.005$*).*

Appendix K

Semantic and syntactic peak regions at the subject-level

Appendix 1-Fig.K1 shows the peak regions analysis for GloVe and GPT-2 for the first 10 subjects. The figure shows syntactic peak regions around temporal regions and the dmPFC and semantic peak regions around the pMTG, AG, IFS and Precuneus. Appendix 1-Fig.K4 shows the distribution of Jaccard scores across subjects, separating the left hemisphere (in red) from the right (in blue).



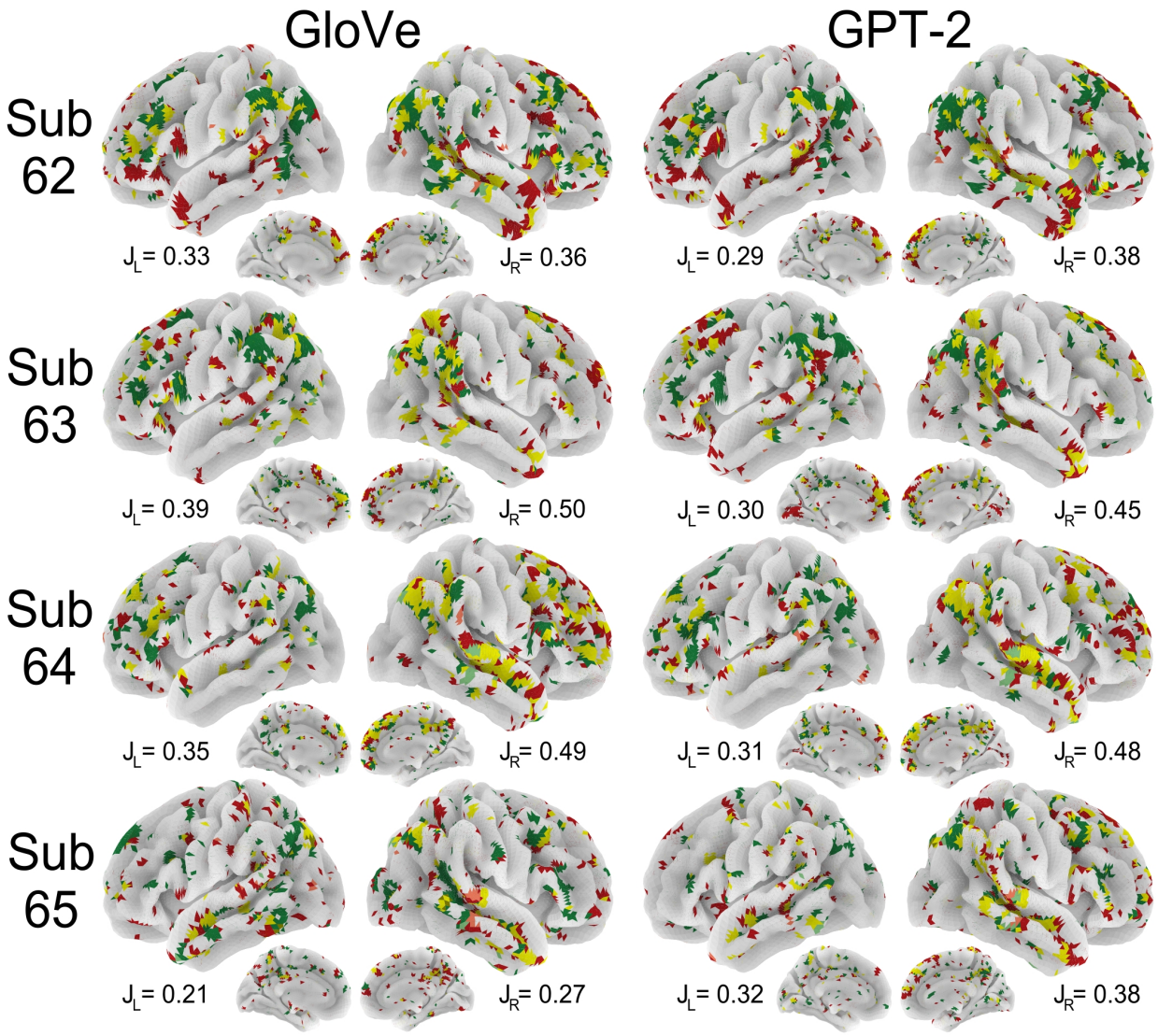


Figure K1

Peak regions of syntax and semantics across subjects. *Bilateral spatial organisation of syntax and semantics highest R scores for the first 10 English subjects of The Little Prince fMRI corpus. Voxels whose R score belong in the 10% highest R scores (in green for models trained on the semantic features, and in red for models trained on the syntactic features) are projected onto brain surface maps for GloVe and GPT-2 (overlap in yellow and other voxels in grey). Jaccard score for each hemisphere are computed, i.e. the ratio between the size of the intersection and the size of the union of semantics and syntax peak regions.*

Appendix 1-Fig.K2 and Appendix 1-Fig.K3 show that the subject-level and group-level analyses are coherent with syntactic peak regions around the Temporal regions, the IFG and dmPFC, and semantic peak regions around the TPJ and the

Precuneous/posterior Cingulate gyri.

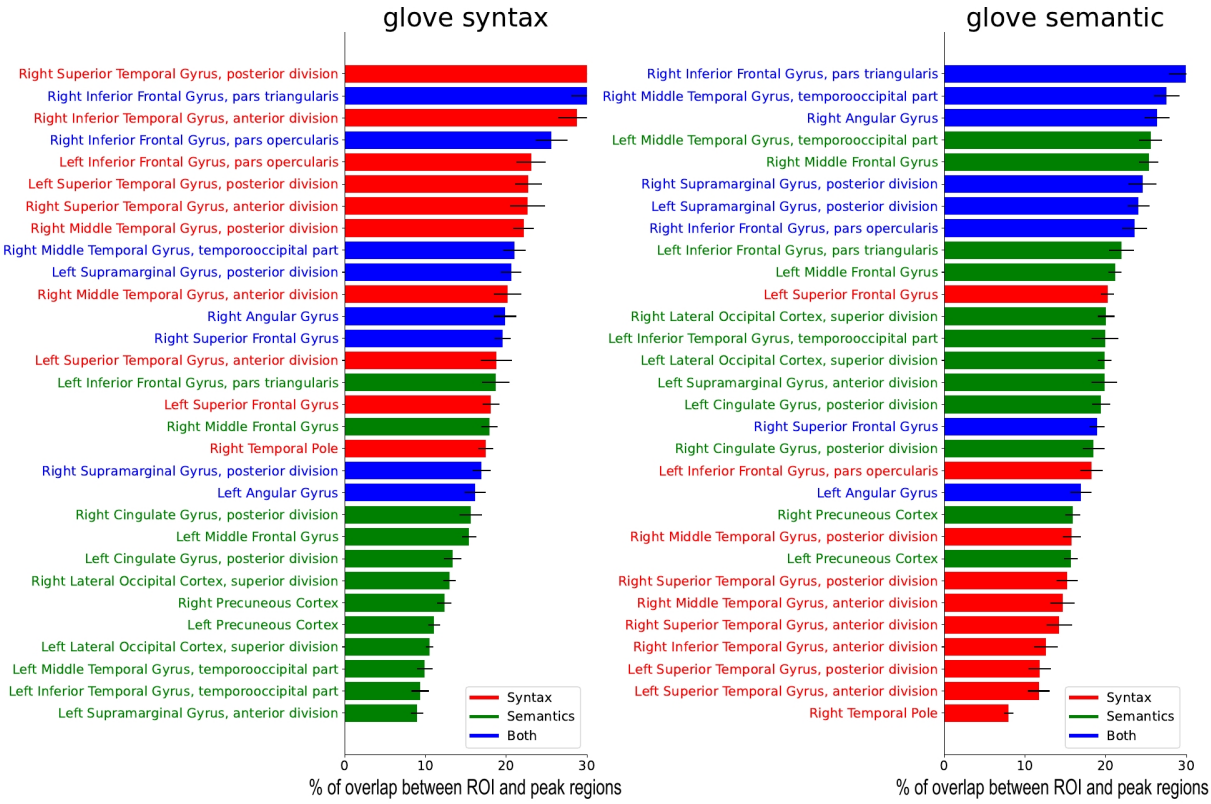


Figure K2

Overlap between Harvard-Oxford ROIs and syntactic/semantic peak regions, averaged across subjects (for GloVe). *Percentage of voxels of the Harvard-Oxford ROIs that belong to the syntactic peak regions (left) and semantic peak regions (right), averaged across the 51 English subjects. The error bars display the standard error to the mean. Regions in red were identified as syntactic peak regions in the group-level analysis, while regions in green were identified as semantic peak regions. Regions in blue belong to both.*

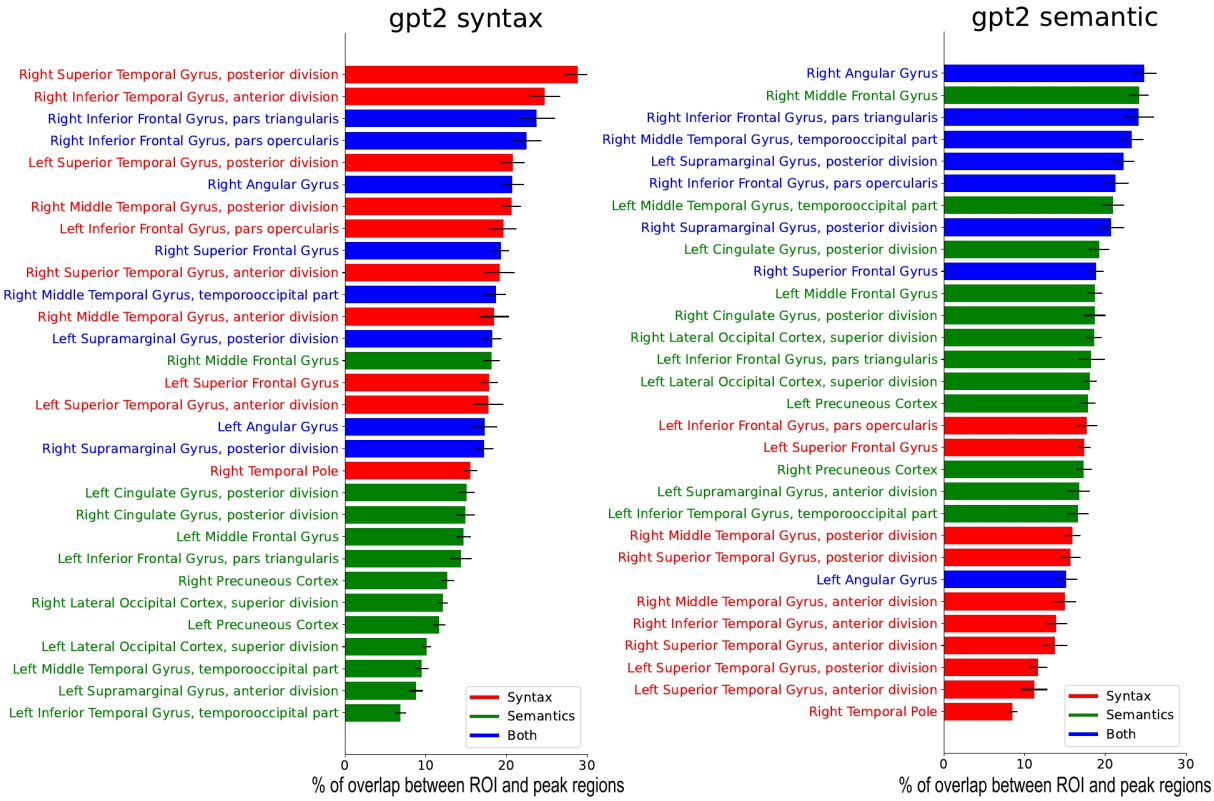


Figure K3

Idem but for GPT-2

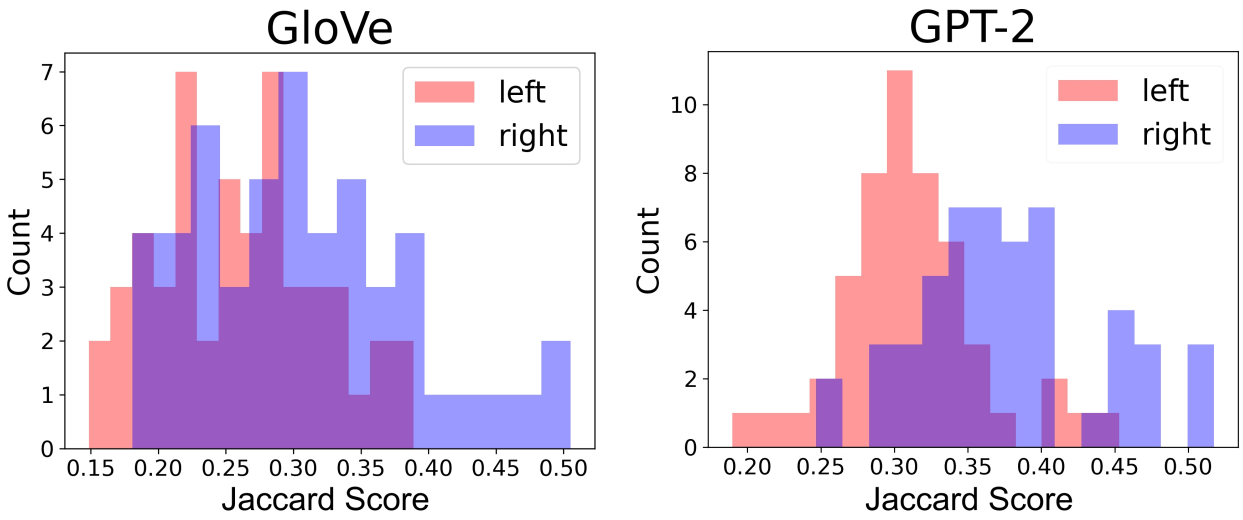


Figure K4

Jaccard scores distribution across subjects. *Distribution of the Jaccard scores across the 51 English participants, in red for the left hemisphere and blue for the right.*

Appendix L

Layer-wise analysis

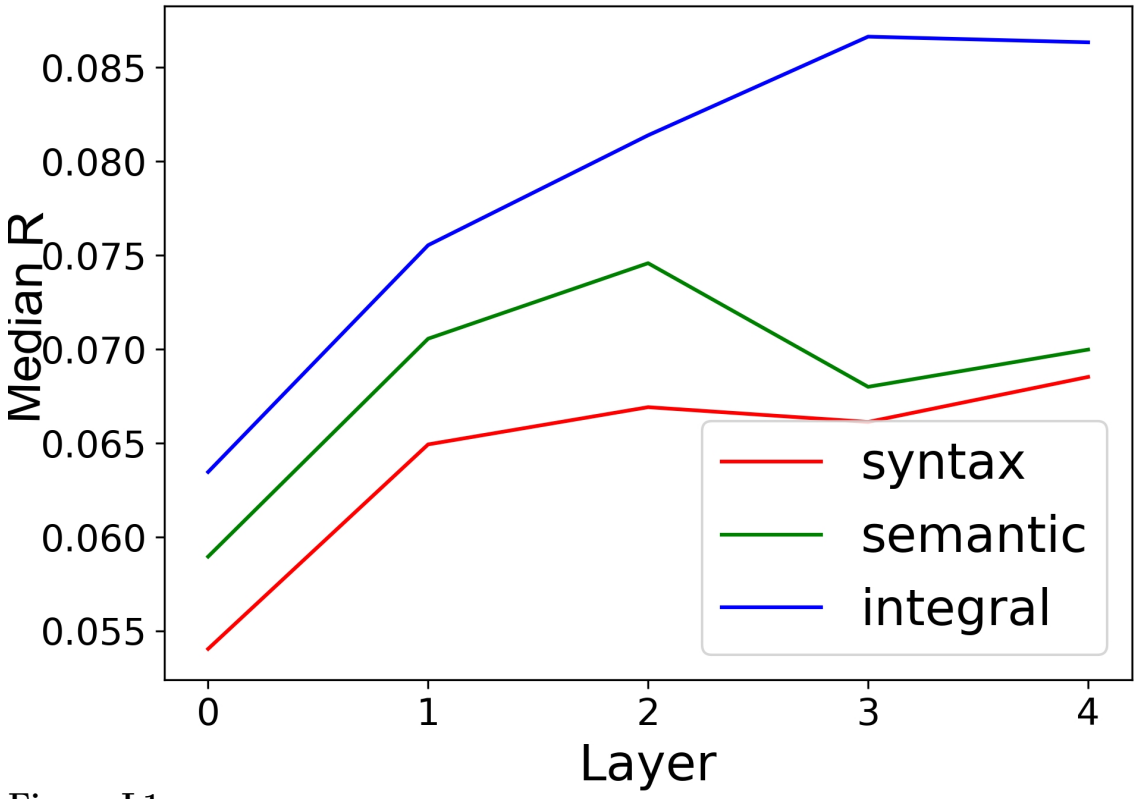


Figure L1

Layer-wise analysis of the models trained on integral/semantic/syntactic features. *Impact of layer depth on the predictive power of GPT-2 when trained on the integral features (blue), the syntactic features (red) and the semantic features (green).*

We further demonstrate the relevance of using the late middle layers in the transformer models' architecture. We display the impact of layer depth on the, per-region, predictive power of BERT models[1] having different total number of layers.

[1] made available by GOOGLE at https://github.com/google-research/bert

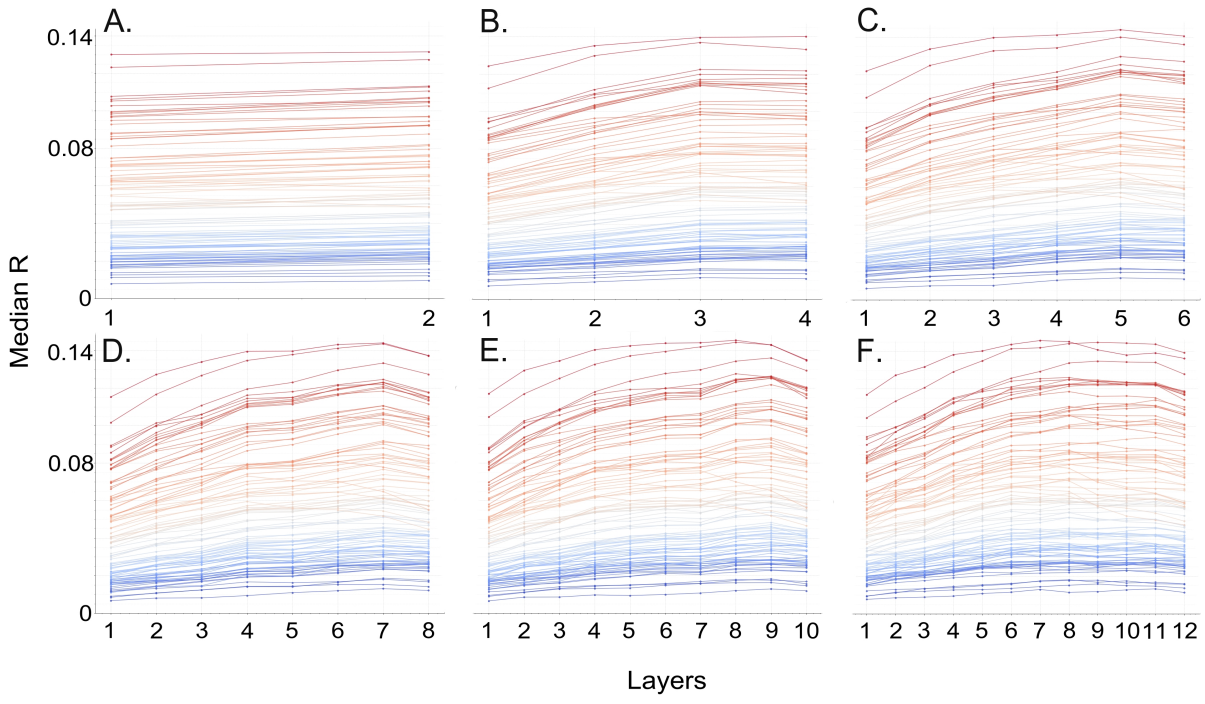


Figure L2

Impact of layer depth on the, per-region, predictive power of BERT models having different total number of layers. *Impact of layer depth on the, per-region, predictive power of BERT models. A) 2-layer BERT, B) 4-layer BERT, C) 6-layer BERT, D) 8-layer BERT, E) 10-layer BERT, F) 12-layer BERT. Brain scores (median R values) were computed across voxels inside brain regions defined by the Harvard-Oxford atlas; each line corresponds to a region.*

Appendix M

Meta-Analysis based on Neurosynth

We used the *Neurosynth* database (https://github.com/neurosynth/neurosynth) to perform a meta-analysis of brain regions that appeared in fMRI articles containing the words 'syntactic' or 'semantic' in their abstract. Using a frequency threshold of 0.05, the keyword *semantic* yielded 626 articles, while *syntactic* yielded 128 articles.

The *meta.MetaAnalysis* function from the neurosynth package was then used to create association test maps for syntax and semantics. These maps display voxels that are reported more often in articles that mention the keyword than articles that do not. Such association test maps indicate whether or not there's a non-zero association between activation of the voxel in question and the use of a particular term in a study. We fused the maps associated to *syntactic* and *semantic*, thresholded with a False Discovery Rate set to 0.01, to produce Fig.M1.

In Fig.M1, we present the outcome of a meta analysis of the literature based on the search for the keywords 'syntactic' and 'semantic' in the Neurosynth database. This analysis, albeit somewhat simplistic, reveals the brain regions most often associated with syntax and semantics.

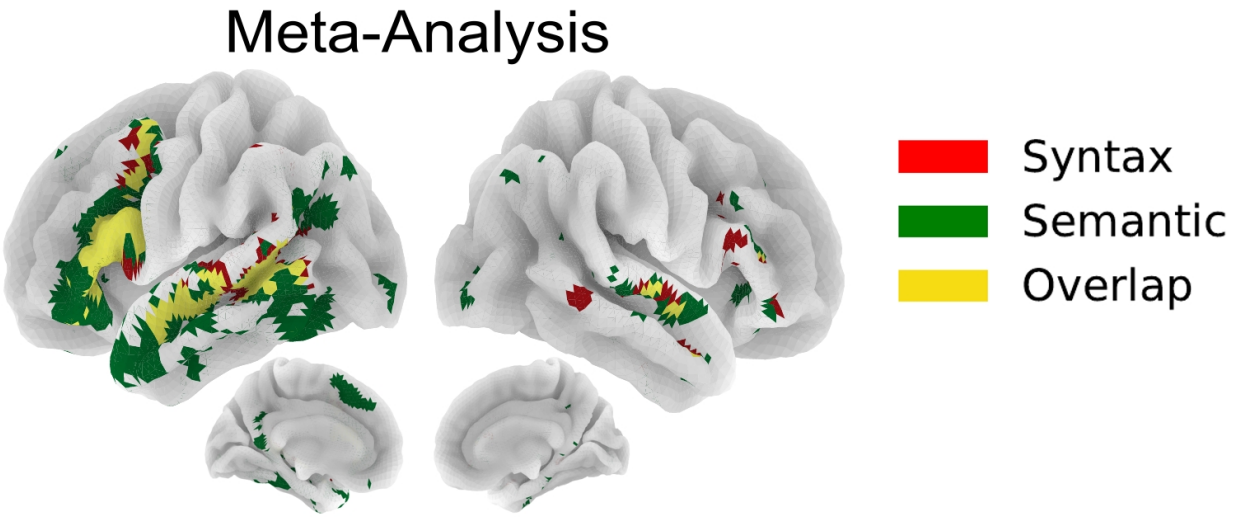


Figure M1

Association maps for the terms "semantic" and "syntactic" in a meta-analysis using Neurosynth *(http://neurosynth.org) The association test map for syntactic (resp. semantic) displays voxels that are reported more often in articles that include the term syntactic (resp. semantic) in their abstracts than articles that do not (FDR correction of 0.01).*

Appendix N

*

Brain Regions abbreviations

- STG: superior Temporal Gyrus
- STS: superior Temporal Sulcus
- TP: Temporal Pole
- IFG: inferior Frontal Gyrus
- IFS: inferior Frontal Sulcus
- DMPC: Dorso-Medial Prefrontal Cortex
- pMTG: posterior Middel Temporal Gyrus
- TPJ: temporo-parietal junction
- pCC: posterior Cingulate Cortex
- AG: Angular Gyrus
- SMA: Supplementary Motor Area