Language Models That Accurately Represent Syntactic Structure Exhibit Higher Representational Similarity To Brain Activity

Abraham Jacob Fresen[1] (bram.fresen@student.uva.nl), Rochelle Choenni[1], Micha Heilbron[2], Willem Zuidema[1], Marianne de Heer Kloots[1] (m.l.s.deheerkloots@uva.nl)
[1]Institute for Logic, Language and Computation [2]Amsterdam Brain & Cognition
University of Amsterdam, 1012 WP Amsterdam, The Netherlands

Abstract

We investigate whether more accurate representation of syntactic information in Transformer-based language models is associated with better alignment to brain activity. We use fMRI recordings from a large dataset (MOUS) of a Dutch sentence reading task, and perform Representational Similarity Analysis to measure alignment with 2 mono- and 3 multilingual language models. We focus on activity in a region known for syntactic processing (the Left posterior Medial Temporal Gyrus). We correlate model-brain similarity scores with the accuracy of dependency structures extracted from model internal states using a labelled structural probe. We report three key findings: 1) Accuracy of syntactic dependency representations correlates with brain similarity, 2) The link between brain similarity and dependency accuracy persists regardless of sentence complexity, although 3) Sentence complexity decreases dependency accuracy while increasing brain similarity. These results highlight how interpretable, linguistic features such as syntactic dependencies can mediate the similarity between language models and brains

Keywords: Artificial Intelligence; Cognitive Neuroscience; Linguistics; Natural Language Processing; fMRI

Introduction

In recent years, the Transformer (Vaswani et al., 2017) has become the ubiquitous architecture for state-of-the-art language models. While the linguistic performance of these language models (LMs) is unrivalled (Qiu et al., 2020), it is still debated to which extent the representations that drive this performance are analogous to those employed by the human brain (e.g., Blank, 2023). Recent studies have shown that the internal states of Transformer models show high similarity to human brain activity during the reading of isolated sentences (Caucheteux & King, 2022; Schrimpf et al., 2021), natural listening (Caucheteux, Gramfort, & King, 2021; Schrimpf et al., 2021), and natural conversation (Cai, Hadjinicolaou, Paulk, Williams, & Cash, 2023). However, there is still an ongoing debate on which linguistic features are represented by LMs, and how this relates to their ability to predict—or correlate with—the human brain.

Early experiments, using *static* embeddings, found improvements in neural predictivity when embeddings were enriched with dependency information. Murphy, Talukdar, and Mitchell (2012) and Abnar, Ahmed, Mijnheer, and Zuidema (2017) compared many linear voxelwise encoding models based on linguistic features derived from different types of word co-occurrence. For example, a part-of-speech model was based on co-occurrences of words that shared both surface form and part-of-speech tag, and a dependency model included the full dependency parse in the co-occurrence computation. Embeddings incorporating the dependency relations between words were found to be best predict brain activation.

More recently, the importance of representing syntactic information for neural predictivity has also been investigated with modern LMs based on the Transformer architecture. Abdou, Gonzalez, Toneva, Hershcovich, and Søgaard (2021) explore whether tuning the models' attention mechanism according to several syntactic and semantic formalisms (including dependency parses) improves brain alignment, but find mixed results across two fMRI datasets. On another fMRI dataset, Oota, Gupta, and Toneva (2023) have shown that the removal of syntactic (phrase-structure) features from language model representations leads to a larger drop in brain alignment than the removal of surface-level or semantic features. In contrast, findings by Kauf, Tuckute, Levy, Andreas, and Fedorenko (2023) have indicated that not syntactic, but semantic content is the most important driver of brain similarity in Transformers. Moreover, an earlier study by Gauthier and Levy (2019) found that finetuning models for scrambled language modelling enhanced brain decoding performance, while presumably abolishing syntactic information.

In short, current findings remain inconclusive on how alignment between language models and brain activity is affected by the accurate representation of syntactic information generally, and of dependency relations specifically.

The Present Study

The present study explores the relationship between the ability of transformers to accurately represent dependency structures and their representational similarity (RS) to the brain.

Previous work on alignment between language models and brains has primarily used models trained on English texts, comparing them to a relatively limited pool of datasets with brain activity recorded during English language comprehension. Here we instead focus on Dutch, allowing us to test how common findings generalize to a different language as well as a different large-scale dataset of multimodal neural signals recorded during sentence processing (Schoffelen et al., 2019). As the brain similarity of off-the-shelf pretrained Transformer to this dataset is relatively unexplored, we decided to compare a range of different models. We expected that accurate Dutch sentence representations could be obtained both from *monolingual* models trained on Dutch texts exclusively as well as *multilingual* models which create a shared rep-

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

resentation space (Chang, Tu, & Bergen, 2022) from being trained on a mixture of texts in different languages. Hence, we compare five pretrained Transformer models, including both Dutch monolingual (*BERTje*, *RobBERT*) and multilingual variants (*mBERT*, *XLM-RoBERTa*, *XLM-V*).

Next to their brain alignment, we examine the *dependency information* encoded by these Transformer models using DepProbe (Müller-Eberstein, van der Goot, & Plank, 2022). This tool allows us to extract dependency parses from the models' internal states, and assess their accuracy through the Labelled Attachment Score (LAS) metric.

We investigate the RS of Transformer models by comparing their internal activations to fMRI data collected from a subset of participants performing a sentence reading task in Dutch (Schoffelen et al., 2019). Given our interest in the representation of syntactic dependencies specifically, we chose to focus our analyses on alignment to activity in the Left posterior Middle Temporal Gyrus (LpMTG), a region of interest functionally associated with syntactic processing (Hagoort & Indefrey, 2014; Tyler, Cheung, Devereux, & Clarke, 2013; Uddén et al., 2022). We hypothesized that alignment to brain activity in the LpMTG should be correlated with accuracy at representing dependency structures.

In line with our hypothesis, we find that Transformer layers with higher LAS show higher RS with fMRI data. This relationship disappears in a control condition where individual word meaning is preserved, while word order is perturbed: layer activations resulting from scrambled inputs show lower RS scores in general, and the obtained RS scores do not correlate to the LAS of extracted dependency structures. Finally, we explore how syntactic complexity, as measured by left-branching complexity (LBC), affects RS, LAS, and their relationship. We find that syntactic complexity differentially impacts these measurements: for more complex sentences, LAS is lower, RS is higher, and their correlation persists.

Materials and Methods

Brain Data and Stimuli

We selected fMRI data from the Mother Of all Unification Studies (MOUS), a dataset collected by researchers from Radboud University in the Netherlands (Schoffelen et al., 2019). The MOUS dataset includes brain signals of 204 participants recorded using fMRI and MEG while they read or listened to Dutch sentences or word lists. For our present study, we only analyzed fMRI data recorded during sentence reading, i.e., the processing of textual stimuli. To limit the computational resources needed for fMRI preprocessing and our experiments, we further restricted our analyses to a subset of participants who were presented with a particular subset of 60 sentences (specifically, those labelled "scenario 5"). We chose this subset because no data collection problems were reported in Schoffelen et al. for this subset, though inspection of the event logs revealed that one participant was not presented with all stimuli. We used only data from the remaining participants (N = 16) for our further analyses.

fMRI Preprocessing and Region Of Interest Extraction

All preprocessing was performed with the fMRIprep pipeline using default parameters (Esteban et al., 2019). Within this pipeline, pre-processing of the structural MRI data was conducted using Freesurfer's recon-all pipeline (Dale, Fischl, & Sereno, 1999). Specifically, the resulting BOLD timeseries were detrended and deconfounded from 18 variables, which included the six estimated head-motion parameters (transx, y, z, rotx, y, z), the first six noise components calculated using anatomical CompCorr, and six DCT-basis regressors using nilearn's cleanimg pipeline (Abraham et al., 2014). After preprocessing, ROI extraction of the LpMTG was performed in FSL (Jenkinson, Beckmann, Behrens, Woolrich, & Smith, 2012) by applying a 10mm spherical mask using the MNI coordinates reported in Uddén et al. (2022) (−50, −36, −2). Correct placement of the mask was visually verified using the Harvard-Oxford cortical and subcortical structural atlases (Frazier et al., 2005). BOLD signals were selected using a delay of 4 seconds after stimulus onset to account for the hemodynamic response time and averaged over TRs within each sentence (4-7 TRs depending on the sentence length).

Syntactic Complexity

The sentences presented to the participants in the MOUS dataset were either syntactically complex, containing a relative clause, or syntactically simpler, containing a main clause and a simple subordinate clause. Uddén et al. (2022) quantified the syntactic complexity of the MOUS sentences by their *left-branching complexity* (LBC). This value is equal to the maximum number of dependents that are at the point of reading not yet assigned to a verb, while parsing the sentence from left to right (see Figure 2). The manually annotated sentences with their corresponding LBC values as used in our study were provided by the authors of Uddén et al. (2022).

Transformer sentence embeddings

As noted above, we compared sentence embeddings from five different Transformer models. Two of these models were monolingual: BERTje (de Vries et al., 2019), and RobBERT (Delobelle, Winters, & Berendt, 2020), and three were multilingual: mBERT (Devlin, Chang, Lee, & Toutanova, 2019), XLM-R (Conneau et al., 2019), and XLM-V (Liang et al., 2023). All embeddings were generated using the pre-trained versions. While all models rely on the same architecture, they vary in training regimes and model size, see Table 1. In particular, RobBERT and XLM-R are robustly optimized versions of BERTje and mBERT, that omit the Next Sentence Prediction task, and XLM-V improves over XLM-R by vastly increasing its vocabulary size.

All models generate internal activations for each token in the text input. To construct sentence embeddings for our further analyses, we averaged over all token activations within sentences. Because earlier research has shown that brain-like representations are mostly found in the middle layers of Transformer models (Schrimpf et al., 2021; Kumar et al.,

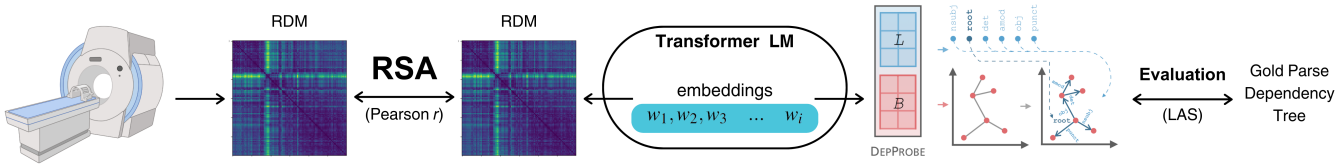


Figure 1: Schematic overview of the procedure. Participants read sentences while their brain signals were recorded with fMRI. The same sentences were fed through the five transformers, whose embeddings were then used with DepProbe to infer the dependency information represented across layers. The dependency information is then compared with silver parses to arrive at the LAS for each layer. Next, we performed the RSA. The embeddings from each layer of the transformers and the brain signals were represented in representational dissimilarity matrices based on the cosine distances between sentence representations. Finally, the RDMs were compared with one another using the Pearson correlation coefficient. We could then compare the Pearson r with the accuracy of the represented dependency information in terms of LAS across layers. Adapted with permission from Müller-Eberstein et al. (2022).

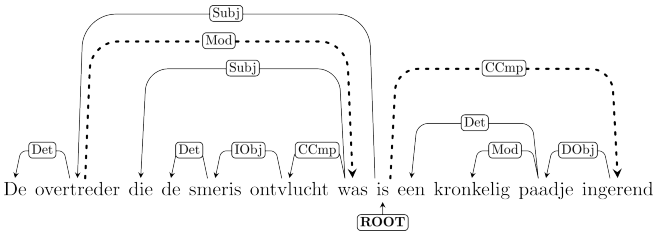


Figure 2: Example of a parsed complex sentence from the MOUS dataset, "De overtreder die de smeris ontvlucht was is een kronkelig paadje ingerend" (The offender who had escaped from the cop ran into a winding path). The maximum number of open verbal dependencies in this sentence is 4, during the retrieval of the word "ontvlucht". Reproduced with permission from Harbusch et al. (2015).

Model	#lgs	V	P	Data	Dutch size
BERTje	1	30k	109M	Various sources	12 GB
mBERT	104	110k	178M	Wikipedia	Unknown
RobBERT	1	40k	117M	OSCAR	39 GB
XLM-r	100	250K	278M	CC-100	29.3 GB
XLM-v	100	1M	778M	CC-100	29.3 GB

Table 1: Details on Transformer models used in this study. Number of languages on which the model is trained (#lgs), Vocabulary size (V), Number of Parameters (P). Both BERTje and mBERT, and RobBERT and XLM-r form minimal pairs that follow the same training procedure but differ in the number of languages they were pretrained on.

2022; Goldstein et al., 2023), we restrict our analyses to embeddings extracted from layers 4-11.

Transformer dependency representations

To assess the representation of dependency information in our selected Transformer models, we made use of *DepProbe*[1] — a recent diagnostic tool from the natural language processing field, which reconstructs fully directed and labelled dependency parses from the vector representations generated by neural LMs (Müller-Eberstein et al., 2022). DepProbe builds on the 'structural probe' introduced by Hewitt and Manning (2019), which extracts undirected dependency graphs from LM representations by applying the minimum spanning tree algorithm to the distances between all possible pairs of words within sentences (Jarník, 1930; Prim, 1957). On top of extracting this undirected dependency graph, DepProbe also predicts the labels governing the dependencies. Each dependency relation r_i, connected to word w_i, is described by a single label l_k (examples of labels are `Det`, `Subj`, `Mod`; see Figure 2). DepProbe hence adds a linear transformation to extract the probability that relation r_i is of label l_k given word w_i, i.e. $p(r_i = l_k|w_i)$. The direction of dependency relations is then determined by the location of the root; the head of the relation between words w_i and w_j is simply the word with the lowest number of connections to the root. This complete process, combining the extracted dependency structures with their relational information, enables DepProbe to extract fully labelled and directed dependency graphs.

The accuracy of these graphs can subsequently be computed by comparison to a set of silver (correctly parsed by a dependency parser) dependency structures. As our accuracy metric, we use the *Labelled Attachment Score* (LAS), which is the percentage of correctly placed and labelled directed edges in the extracted dependency parse.

Both the structural and relational components of DepProbe require supervised training on a parsed corpus. Müller-Eberstein et al. (2022) achieve good results across 13 languages by training and evaluating DepProbe on treebanks from the Universal Dependencies (UD) project (Nivre et al., 2020). For our purposes, we trained DepProbe on the Dutch dependency data in the Alpino treebank (van der Beek, Bouma, Malouf, & van Noord, 2001) from the UD database, which contains over 13000 annotated Dutch sentences. Following Müller-Eberstein et al. (2022), we also experimented with extracting structural and relational information from different layers (one 'structural' and one 'relational' layer), specifically the layers later used in our brain-similarity analyses (4-11). Thus, for each transformer model, DepProbe was trained using the embeddings from every possible combina-

[1] `https://github.com/personads/depprobe`

tion of layers, while keeping the other training hyperparameters identical to the settings used by Müller-Eberstein et al..

To confirm DepProbe's functioning, we evaluated its performance against the test split of the Alpino treebank, using the LAS metric. The results followed our expectations, with the LAS peaking in the middle to late layers. Furthermore, the results indicated that changing the relational layer had relatively little effect on the resulting LAS compared to the structural layer; thus, hereafter all DepProbe results are averaged over the relational layers.

Analysis procedure

A schematic overview of our analysis procedure can be found in Figure 1. We first used our trained DepProbe models to extract labelled and directed dependency graphs for the MOUS sentences, from our five Transformer models of interest. We used a neural parser (the biaffine parser; Dozat & Manning, 2016) to obtain silver parses for the same sentences as a topline comparison. We then computed LAS scores to evaluate the accuracy of the extracted DepProbe graphs for each Transformer model against the obtained silver parses.

To assess the alignment between the Transformer sentence embeddings and the fMRI signals, we used Representational Similarity Analysis (RSA; Kriegeskorte, Mur, & Bandettini, 2008). RSA enables comparison between very different systems by transforming their representations to a dissimilarity space over a common set of stimuli, which can then be directly correlated between systems. This technique has been applied before to assess the alignment of Transformer embeddings to fMRI signals recorded during narrative reading (e.g., Abnar, Beinborn, Choenni, & Zuidema, 2019). For our present study, representations were compared on the sentence level. The extracted sentence embeddings from all five Transformer models were permuted to match the stimulus presentation order for each participant. We constructed representational dissimilarity matrices (RDMs) based on cosine distances between sentence representations, i.e. between model embeddings on the Transformer side, and between voxel activation values on the fMRI side. Separate RDMs were created for each model layer (4-11) and each participant. For our main RS analyses, we computed the Pearson's correlation coefficient (r) between model and fMRI RDMs as constructed over all sentence stimuli. For our analyses on sentence complexity, we grouped sentence stimuli by LBC value (into 4 groups, ranging from LBC 1 to 4) and computed RS separately for each group. Finally, we averaged RS scores over participants to arrive at our layerwise measure of model-brain similarity.

A correlation between a Transformer model's brain-similarity and LAS measures could in principle arise through spurious features that correlate with both LAS and fMRI activity, but do not relate to the model's representation of syntax. To confirm that our results can be ascribed to the encoding of syntactic information in model layers, we compared our results to a control condition similar to analyses performed by Kauf et al. (2023). In this condition, the models were fed perturbed versions of the sentences that did not contain meaningful segments of three or more consecutive words. In this way, most of the syntactic information available to our models was abolished, and the remaining representational similarity can be ascribed to lexical meaning rather than syntactic structure.

Results

Model Layers That Accurately Represent Syntactic Dependencies Are More Brain-like

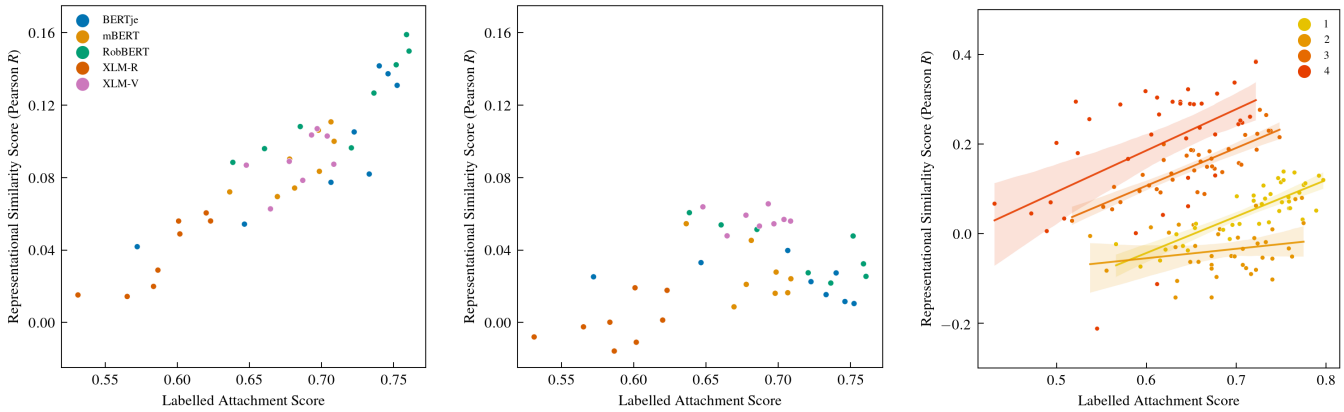
Figure 3a plots the brain alignment against grammatical information for all 5 language models included in our study. Evident from this plot is that embeddings that yield higher LAS scores also yield higher RS scores (Figure 3a). A mixed effect model analysis with language model as random intercept reveals a significant positive relationship between LSA and RS score (b = 0.55, t = 11.76, SE = 0.05, p $<$.001, 95% CI = [0.46 - 0.64]). This effect dissipates in the control condition (b = -0.09, t = -1.91, SE = 0.05, p = .06), indicating that the observed correspondence between RS and LAS measures cannot be ascribed to word-level features (Figure 3b).

A similar pattern can be observed by comparing the layerwise patterns in Figure 4. Across model layers, representational similarity follows largely the same pattern as LAS. The RS scores from the control condition do not follow the same pattern (Figure 4c). A paired-samples t-test comparing RS scores for each model layer between conditions showed a significant difference between the unperturbed condition (M = 0.086, SD = 0.036) and the control condition (M = 0.029, SD = 0.022), t(39) = 10.44, p $<$.001. The magnitude of the differences in the means (mean difference = 0.057, 95% CI: [0.046 - 0.068]) was very large (Cohen's d = 1.84). Thus, perturbing the word order of the input sentences significantly decreased RS, implying that word meaning alone cannot account for the brain alignment of the transformer models.

Examining Figures 3 and 4, Dutch monolingual models (BERTje and RobBERT) appear to surpass multilingual models (XLM-R, XLM-V, mBERT) in terms of both LAS and RS scores. However, upon closer inspection, it seems that XLM-R may be primarily responsible for this difference. Thus, though brain-similarity varies across models generally, it does not seem to specifically differ between models trained in mono- vs. multilingual settings.

Syntactic Complexity Influences Both Brain Similarity And Dependency Accuracy But Not Their Relationship

The MOUS dataset includes sentences with different levels of syntactic complexity, indexed by their Left-Branching Complexity (LBC) score. We investigated whether syntactic complexity affected the RS score, the LAS, and the relationship between the LAS and RS scores (see Figure 3c). For each of the four LBC values, we built a linear mixed effect model with LM as a random intercept and LAS as a fixed effect predictor of RS. The results of this analysis revealed a positive



(a) LAS and RS scores in the unperturbed condition.

(b) LAS and RS scores in the control condition.

(c) LAS and RS scores when the sentences are grouped by syntactic complexity.

Figure 3: The LAS and RS scores for each layer from 4-11 of each model. In the unperturbed condition 3a, there exists a positive relationship between LAS and RS scores while this pattern is absent in the control condition, 3b. This shows that model layers that accurately represent dependency information show more brain similarity. This pattern persists if we group the sentences based on their syntactic complexity, 3c.

relationship between LAS and RS for all LBC values, see Table 2.

LBC	b	t	SE	p	95% CI
LBC 1	0.80	7.51	0.11	$< .001$	[0.59 - 1.00]
LBC 2	0.30	2.19	0.13	$< .05$	[0.02 - 0.56]
LBC 3	0.73	6.98	0.10	$< .001$	[0.52 - 0.99]
LBC 4	1.20	3.92	0.31	$< .001$	[0.43 - 1.81]

Table 2: Results of the linear mixed models for each LBC value with LM as a random intercept and LAS as a fixed effect predictor of RS score.

To investigate whether syntactic complexity influences RS scores, we performed a Kruskal-Wallis H test to evaluate differences in RS scores across LBC groups. Results showed a significant difference in RS scores across groups, $H(3) = 94.07$, $p < .001$. Post-hoc comparisons using a Dunn test with a Bonferroni correction for multiple comparisons indicated that RS scores were significantly larger for LBC 1 than for LBC 2 sentences ($Z = 3.62$, $p < .001$), but RS scores between LBC 3 and LBC 4 sentences did not differ significantly ($Z = -0.91$, $p = 1$). In all other pairwise comparisons, larger LBC values corresponded to higher RS scores ($Z > 4$, $p < .001$).

To investigate whether syntactic complexity influences LAS, we again employed a Kruskal-Wallis H test to evaluate differences in LAS across LBC groups. Results indicated a significant difference in LAS across groups, $H(3) = 45.45$, $p < .001$. These results indicate that syntactic complexity does influence LAS overall. Post-hoc comparisons using a Dunn test with a Bonferroni correction for multiple comparisons revealed that the LAS of adjacent LBC values did not differ significantly. However, for the other pairwise comparisons it holds that higher LBC values lead to lower LAS ($Z < -4$, $p < .001$). Thus, we can conclude that syntactic complexity does not alter the observed relationship between accurate dependency representation and brain similarity. Yet, we found a surprising pattern showing that syntactic complexity lowers dependency accuracy while increasing brain similarity.

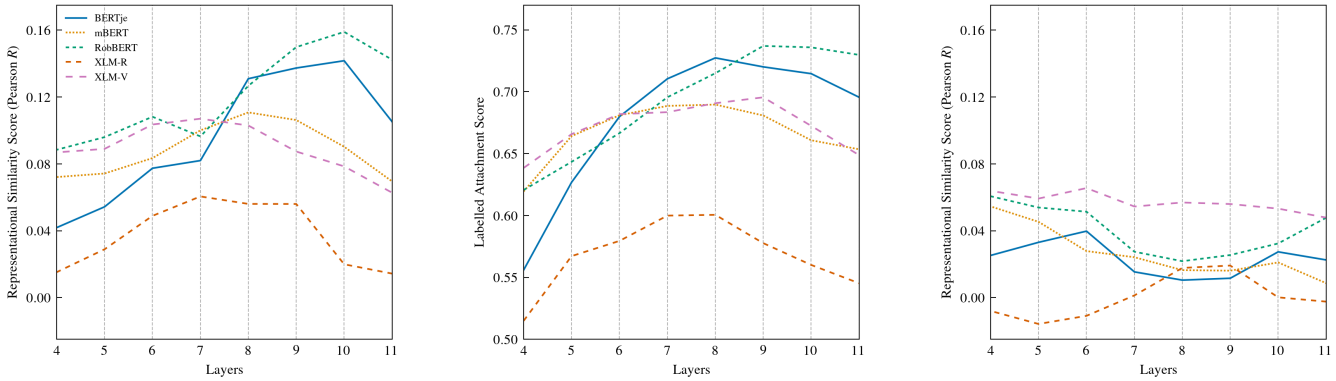
Discussion

Our study revealed three key findings. First, there is a positive correlation between the accuracy with which transformers represent dependency structures and their brain similarity. Second, the relationship between the accuracy of dependency representations and brain similarity persists across different syntactic complexities. Third, we observed that increasing syntactic complexity strengthens the correlation between model representations and brain activity, but reduces the accuracy of dependency representations. This suggests a complex interaction between model performance, syntactic complexity, and brain-like representation accuracy. Additionally, comparing performance between models, it seems that distinctions in brain similarity and accuracy of dependency representations may be model-specific rather than dependent on the linguistic diversity of the training data.

Accuracy of Dependency Representations and Brain Score

Our results show a strong positive relationship between the accuracy with which Transformer model layers represent syntactic dependencies and their similarity to the brain. We specifically found a positive relationship between the Labelled Attachment Score (LAS) of dependency structures extracted from model layers and the Representational Similarity (RS) between those layers and the LpMTG, a brain area associated with syntactic processing (Hagoort & Indefrey, 2014).

In line with earlier research investigating the importance of syntactic information in model-brain alignment (Murphy et



(a) The RS scores for each model across layers in the unperturbed condition

(b) The LAS for each model plotted across layers.

(c) The RS scores for each model across layers in the control condition

Figure 4: Comparison between the RS scores across all model layers in the unperturbed (4a) and control condition (4c), and the LAS across all model layers (4b). The LAS and RS scores in the unperturbed condition follow largely the same pattern, while this pattern is absent in the control condition.

al., 2012; Abnar et al., 2017; Oota et al., 2023), our findings suggest that for the purpose of modelling human language processing, it is beneficial to choose models that accurately represent syntactic dependencies. This contrasts with recent findings by Kauf et al. (2023) which suggested that semantic rather than syntactic information drives brain similarity. However, our current findings are confined to the LpMTG, in which activity has been observed to increase as a function of syntactic complexity (Uddén et al., 2022). Since the MOUS sentences are designed for their syntactic complexity, this may emphasize the importance of dependency representations for brain similarity. Many comparative studies between language models and brain data are focused on the language network as a whole which may obscure more fine-grained divisions of labour within the network. Future research could attempt to map individual LM components to more specific ROIs, enabling a more detailed understanding of how various functional representations within LMs relate to the brain. Furthermore, our results are based on RSA while Kauf et al. (2023) used trained regressions to map between models and brains. Few published studies directly compare these methods and how they influence the results of model-brain similarity measurements; a systematic and up-to-date comparison across popular models, datasets and mapping techniques used in the field is needed to further investigate these diverging results (Beinborn, Abnar, & Choenni, 2023). Unlike most previous studies, the present study utilized a Dutch dataset, which may limit direct comparisons. However, it extends the findings Caucheteux and King (2022) on the same dataset. Future studies should investigate whether these results are consistent across various datasets and languages. While our results demonstrate a correlation between accurate representation of dependencies and brain similarity, we cannot assert that these more accurate dependency representations *cause* higher brain similarity. Future research could employ a multifactor analysis to explore whether there are alternative explanations for the correlation we observed.

Comparing individual model performances, it is surprising that XLM-R scored relatively poorly in both dependency representation and brain similarity, as this model has been found to outperform mBERT across a range of different languages and tasks (Conneau et al., 2019). Our findings are inconclusive about any systematic difference between mono- and multilingual models in either measure; future research could investigate this with a more diverse set of models, brain areas, and linguistic features.

Syntactic Complexity

Across models, our results suggest an interesting dissociation in how syntactic complexity affects LAS (negatively) versus how it affects RS (positively). Yet, when grouping sentences by their Left-Branching Complexity values, the positive correspondence between LAS and RS scores remains. The negative effect of sentence complexity on LAS could arise from syntactically more complex sentences being more difficult to parse. Indeed, some MOUS sentences are complex enough that even native speakers could struggle to identify dependencies on their first reading. An explanation for the effect of syntactic complexity on RS may lie in the function of the LpMTG. Uddén et al. (2022) found that the LpMTG becomes more active when a sentence is syntactically more complex. Syntactically more intricate sentences could therefore lead to a stronger signal which would, in turn, result in a higher signal-to-noise ratio (Welvaert & Rosseel, 2013). Therefore, the higher RS scores associated with higher syntactic complexity may be due to an increase in detectable signal.

Overall, we successfully demonstrate how interpretable stimulus features as well as model interpretability analyses can reveal interesting dynamics underlying the similarity between language models and brain activity. We hope this approach inspires further work into a deeper understanding of language processing in Transformers as well as human brains.

Acknowledgements

We would like to thank Julia Uddén for providing the linguistic annotations of the MOUS and for her insightful comments that significantly enhanced this research. We also extend our thanks to all co-authors of the MOUS paper for their efforts in developing the dataset that was crucial for this study.

References

Abdou, M., Gonzalez, A. V., Toneva, M., Hershcovich, D., & Søgaard, A. (2021, January). *Does injecting linguistic structure into language models lead to better alignment with brain recordings?* arXiv. Retrieved from http://arxiv.org/abs/2101.12608 (arXiv:2101.12608 [cs]) doi: 10.48550/arXiv.2101.12608

Abnar, S., Ahmed, R., Mijnheer, M., & Zuidema, W. H. (2017). Experiential, distributional and dependency-based word embeddings have complementary roles in decoding brain activity. In *Workshop on cognitive modeling and computational linguistics.*

Abnar, S., Beinborn, L., Choenni, R., & Zuidema, W. (2019, August). Blackbox Meets Blackbox: Representational Similarity & Stability Analysis of Neural Language Models and Brains. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP* (pp. 191–203). Florence, Italy: Association for Computational Linguistics. Retrieved 2020-01-23, from https://www.aclweb.org/anthology/W19-4820 doi: 10.18653/v1/W19-4820

Abraham, A., Pedregosa, F., Eickenberg, M., Gervais, P., Mueller, A., Kossaifi, J., ... Varoquaux, G. (2014). Machine learning for neuroimaging with scikit-learn. *Frontiers in Neuroinformatics*, *8*. Retrieved from https://www.frontiersin.org/articles/10.3389/fninf.2014.00014/full doi: 10.3389/fninf.2014.00014

Beinborn, L., Abnar, S., & Choenni, R. (2023). Robust evaluation of language–brain encoding experiments. In A. Gelbukh (Ed.), *Computational linguistics and intelligent text processing* (pp. 44–61). Cham: Springer Nature Switzerland.

Blank, I. A. (2023, November). What are large language models supposed to model? *Trends in Cognitive Sciences*, *27*(11), 987–989. Retrieved 2024-01-31, from https://www.sciencedirect.com/science/article/pii/S1364661323002024 doi: 10.1016/j.tics.2023.08.006

Cai, J., Hadjinicolaou, A. E., Paulk, A. C., Williams, Z. M., & Cash, S. S. (2023). Natural language processing models reveal neural dynamics of human conversation. *bioRxiv*. Retrieved from https://www.biorxiv.org/content/early/2023/03/11/2023.03.10.531095 doi: 10.1101/2023.03.10.531095

Caucheteux, C., Gramfort, A., & King, J.-R. (2021, November). Model-based analysis of brain activity reveals the hierarchy of language in 305 subjects. In *Findings of the association for computational linguistics: Emnlp 2021* (pp. 3635–3644). Punta Cana, Dominican Republic: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2021.findings-emnlp.308 doi: 10.18653/v1/2021.findings-emnlp.308

Caucheteux, C., & King, J.-R. (2022, 02). Brains and algorithms partially converge in natural language processing. *Communications Biology*, *5*, 134. doi: 10.1038/s42003-022-03036-1

Chang, T. A., Tu, Z., & Bergen, B. K. (2022). The geometry of multilingual language model representations. *ArXiv*, *abs/2205.10964*. Retrieved from https://api.semanticscholar.org/CorpusID:248987203

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., ... Stoyanov, V. (2019). Unsupervised cross-lingual representation learning at scale. *CoRR*, *abs/1911.02116*. Retrieved from http://arxiv.org/abs/1911.02116

Dale, A. M., Fischl, B., & Sereno, M. I. (1999). Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage*, *9*(2), 179-194. Retrieved from http://www.sciencedirect.com/science/article/pii/S1053811998903950 doi: 10.1006/nimg.1998.0395

Delobelle, P., Winters, T., & Berendt, B. (2020, November). RobBERT: a Dutch RoBERTa-based Language Model. In *Findings of the association for computational linguistics: Emnlp 2020* (pp. 3255–3265). Online: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2020.findings-emnlp.292 doi: 10.18653/v1/2020.findings-emnlp.292

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4171–4186). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N19-1423 doi: 10.18653/v1/N19-1423

de Vries, W., van Cranenburgh, A., Bisazza, A., Caselli, T., van Noord, G., & Nissim, M. (2019, December). *BERTje: A Dutch BERT Model.* arXiv. Retrieved from http://arxiv.org/abs/1912.09582 (arXiv:1912.09582 [cs]) doi: 10.48550/arXiv.1912.09582

Dozat, T., & Manning, C. D. (2016). Deep biaffine attention for neural dependency parsing. *CoRR*, *abs/1611.01734*. Retrieved from http://arxiv.org/abs/1611.01734

Esteban, O., Markiewicz, C., Blair, R. W., Moodie, C., Isik, A. I., Erramuzpe Aliaga, A., ... Gorgolewski, K. J. (2019). fMRIPrep: a robust preprocessing pipeline for functional MRI. *Nature Methods*, *16*, 111–116. doi: 10.1038/s41592-018-0235-4

Frazier, J. A., Chiu, S., Breeze, J. L., Makris, N., Lange,

N., Kennedy, D. N., ... Biederman, J. (2005). Structural brain magnetic resonance imaging of limbic and thalamic volumes in pediatric bipolar disorder. *American Journal of Psychiatry*, *162*(7), 1256-1265. Retrieved from https://doi.org/10.1176/appi.ajp.162.7.1256 (PMID: 15994707) doi: 10.1176/appi.ajp.162.7.1256

Gauthier, J., & Levy, R. (2019, November). Linking artificial and human neural representations of language. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 529–539). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://aclanthology.org/D19-1050 doi: 10.18653/v1/D19-1050

Goldstein, A., Ham, E., Nastase, S. A., Zada, Z., Grinstein-Dabus, A., Aubrey, B., ... Hasson, U. (2023). Correspondence between the layered structure of deep language models and temporal structure of natural language processing in the human brain. *bioRxiv*. Retrieved from https://www.biorxiv.org/content/early/2023/02/21/2022.07.11.499562 doi: 10.1101/2022.07.11.499562

Hagoort, P., & Indefrey, P. (2014). The neurobiology of language beyond single words. *Annual review of neuroscience*, *37*, 347-62.

Harbusch, K., van den Bosch, A., & Kempen, G. (2015). *Hand corrected mous dependency trees.* (Retrieved through personal communication)

Hewitt, J., & Manning, C. D. (2019, June). A structural probe for finding syntax in word representations. In *Proceedings of the 2019 conference of the north American chapter of the association for computational linguistics: Human language technologies, volume 1 (long and short papers)* (pp. 4129–4138). Minneapolis, Minnesota: Association for Computational Linguistics. Retrieved from https://aclanthology.org/N19-1419 doi: 10.18653/v1/N19-1419

Jarník, V. (1930). *O jistém problému minimálním: (z dopisu panu o. bor°uskovi)*. Mor. přírodovědecká společnost. Retrieved from https://books.google.nl/books?id=bOYpnwEACAAJ

Jenkinson, M., Beckmann, C. F., Behrens, T. E., Woolrich, M. W., & Smith, S. M. (2012). Fsl. *Neuroimage*, *62*(2), 782–790.

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. *bioRxiv*. Retrieved from https://www.biorxiv.org/content/early/2023/05/06/2023.05.05.539646 doi: 10.1101/2023.05.05.539646

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*. Retrieved from https://www.frontiersin.org/article/10.3389/neuro.06.004.2008

Kumar, S., Sumers, T. R., Yamakoshi, T., Goldstein, A., Hasson, U., Norman, K. A., ... Nastase, S. A. (2022). Reconstructing the cascade of language processing in the brain using the internal computations of a transformer-based language model. *bioRxiv*. Retrieved from https://www.biorxiv.org/content/early/2022/06/09/2022.06.08.495348 doi: 10.1101/2022.06.08.495348

Liang, D., Gonen, H., Mao, Y., Hou, R., Goyal, N., Ghazvininejad, M., ... Khabsa, M. (2023, 01). Xlm-v: Overcoming the vocabulary bottleneck in multilingual masked language models.

Müller-Eberstein, M., van der Goot, R., & Plank, B. (2022, May). Probing for labeled dependency trees. In *Proceedings of the 60th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 7711–7726). Dublin, Ireland: Association for Computational Linguistics. Retrieved from https://aclanthology.org/2022.acl-long.532 doi: 10.18653/v1/2022.acl-long.532

Murphy, B., Talukdar, P. P., & Mitchell, T. M. (2012). Selecting corpus-semantic models for neurolinguistic decoding. In *International workshop on semantic evaluation.*

Nivre, J., de Marneffe, M.-C., Ginter, F., Hajič, J., Manning, C. D., Pyysalo, S., ... Zeman, D. (2020, May). Universal Dependencies v2: An evergrowing multilingual treebank collection. In *Proceedings of the twelfth language resources and evaluation conference* (pp. 4034–4043). Marseille, France: European Language Resources Association. Retrieved from https://aclanthology.org/2020.lrec-1.497

Oota, S. R., Gupta, M., & Toneva, M. (2023). *Joint processing of linguistic properties in brains and language models.*

Prim, R. C. (1957). Shortest connection networks and some generalizations. *The Bell System Technical Journal*, *36*(6), 1389-1401. doi: 10.1002/j.1538-7305.1957.tb01515.x

Qiu, X., Sun, T., Xu, Y., Shao, Y., Dai, N., & Huang, X. (2020). Pre-trained models for natural language processing: A survey. *Science China Technological Sciences*, *63*, 1872 - 1897.

Schoffelen, J. J. M., Oostenveld, R., Lam, N., Udden, J., Hultén, A., & Hagoort, P. P. (2019). Mother of unification studies, a 204-subject multimodal neuroimaging dataset to study language processing. Retrieved from https://doi.org/10.34973/37n0-yc51 doi: 10.34973/37n0-yc51

Schrimpf, M., Blank, I., Tuckute, G., Kauf, C., Hosseini, E., Kanwisher, N., ... Fedorenko, E. (2021, 11). The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, *118*, e2105646118. doi: 10.1073/pnas.2105646118

Tyler, L., Cheung, T., Devereux, B., & Clarke, A. (2013).

Syntactic computations in the language network: Characterizing dynamic network properties using representational similarity analysis. *Frontiers in Psychology*, *4*. Retrieved from https://www.frontiersin.org/articles/10.3389/fpsyg.2013.00271 doi: 10.3389/fpsyg.2013.00271

Uddén, J., Hultén, A., Schoffelen, J.-M., Lam, N., Harbusch, K., van den Bosch, A., ... Hagoort, P. (2022, 09). Supramodal Sentence Processing in the Human Brain: fMRI Evidence for the Influence of Syntactic Complexity in More Than 200 Participants. *Neurobiology of Language*, *3*(4), 575-598. Retrieved from https://doi.org/10.1162/nol_a_00076 doi: 10.1162/nol_a_00076

van der Beek, L., Bouma, G., Malouf, R., & van Noord, G. (2001). The alpino dependency treebank. In *The clinician.*

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., ... Polosukhin, I. (2017). *Attention is all you need.*

Welvaert, M., & Rosseel, Y. (2013). On the definition of signal-to-noise ratio and contrast-to-noise ratio for fmri data. *PLoS ONE*, *8*. Retrieved from https://api.semanticscholar.org/CorpusID:3965703