Topic-context Dependency on Continuous Semantic Reconstruction of Language from fMRI Signals

Fermin Travi[*,†], **Agustín Delmagro**[*], **Diego Fernández Slezak**[*,†], **Bruno Bianchi**[*,†], **Juan E Kamienkowski**[*,†,‡]

[*]Instituto en Ciencias de la Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires

[†]Departamento de Computación, Facultad de Ciencias Exactas y Naturales, Universidad de Buenos Aires

[‡]Maestría de Explotación de Datos y Descubrimiento del Conocimiento, Universidad de Buenos Aires

ftravi, bbianchi, juank@dc.uba.ar

Abstract

Recent advances in neural language decoding have enabled reconstruction of perceived speech from fMRI signals using large language models and brain encoders. While these systems achieve impressive semantic fidelity, the extent to which training data biases signal encoding and the resulting decoded output remains unclear. Here, we investigate how topic-specific training constrains language decoding by partitioning 84 auditory stories into two semantic clusters and training brain encoders on topic-specific subsets across three participants. We find that encoders trained exclusively on one topic systematically bias decoded outputs toward their training domain: family-related stories are reclassified as such when reconstructed with family-trained encoders, but not otherwise. Moreover, topic-constrained encoders perform significantly above chance only within their training topic and fall below random baseline when decoding out-of-topic stimuli. These findings underscore the need for more diverse and semantically rich training data, while also raising questions about the breadth of semantic variability that current brain encoders can effectively capture.

Keywords: language, fMRI, voxelwise encoding

1. Introduction

The convergence of increasingly powerful large language models (LLMs) with growing fMRI data availability has enabled the development of neural language decoding systems that reconstruct perceived or imagined speech from brain activity (Tang et al., 2023; Antonello and Huth, 2024; Antonello et al., 2023). Broadly, the architecture of these decoders typically involves training language-to-BOLD signal encoders as voxelwise linear regressions on extensive, subject-specific datasets. During decoding, an LLM generates plausible text continuations, which are then encoded to predicted voxel activity; the continuation whose predicted activity best matches the observed brain signal is selected as the decoded output. Remarkably, decoded texts often capture the semantic content of stimuli with high fidelity (Tang et al., 2023). These systems have advanced both language neuroscience (by enabling more detailed accounts of how language is represented in the brain (Antonello et al., 2021; Chen et al., 2024, 2025; Gwilliams et al., 2025)) and natural language processing (by improving and deepening our understanding of how LLMs process language (Toneva and Wehbe, 2019; Zhou et al., 2024)), exemplifying the productive interplay between AI and neuroscience.

Despite this success, an important question remains: to what extent does the semantic content represented in the stimuli systematically bias signal encoding and its downstream decoding? While recent work by Chen et al. (2025) has examined topic-related functional specialization in the brain (demonstrating that different brain regions show preferential activation for distinct topics), this focuses on neural organization rather than on how training constraints shape the encoder's interpretation of brain signals. Here, we directly address how semantic biases in training data propagate through to decoded outputs. We partition stimuli into two distinct topics using a largely unsupervised approach, then train brain encoders on topic-specific subsets and analyze the resulting biases in decoded text. Our findings reveal that topic-constrained training produces decoders that systematically reconstruct stimuli as belonging to their training topic, while performing below chance on stimuli from other semantic domains. These results demonstrate that language encoders can become over-specialized to narrow semantic domains during training, systematically biasing decoded outputs toward those domains. This finding has important implications for both the reliability of brain-computer interface applications and the interpretations of language processing results.

Proceedings of the 15th Workshop on Cognitive Modeling and Computational Linguistics (CMCL) @ LREC 2026, pages 286–292
16 May 2026.

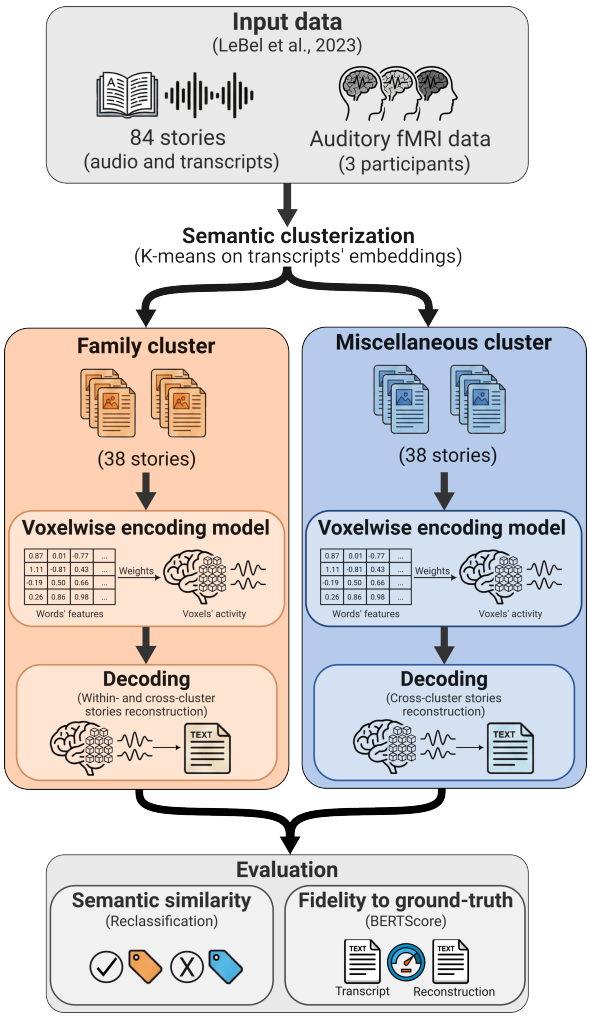

Figure 1: Methodological pipeline for topic-specific language decoding. Narrative stories are partitioned via semantic clusterization into Family (orange) and Miscellaneous (blue) topics. For each cluster, voxelwise encoding models (GPT-2 based) are trained and used to reconstruct stories from brain activity. Reconstructions are evaluated for thematic consistency and linguistic fidelity against ground-truth transcripts using BERTScore.

2. Methods

To evaluate how semantic topics influence auditory fMRI signal encoding and downstream language decoding, we conducted a semantic classification on the 84 stories from the LeBel et al. (2023) dataset, focusing on the data from three participants who listened to all the stories. First, we divided these stories into two semantic clusters (*Family* and *Miscellaneous*). Then, we trained topic-specific voxelwise encoding models to map the text stimuli to the observed brain activity. The decoding process then used these same encoding models to reconstruct the held-out stories: candidate text was generated, mapped to predicted voxel activity via the encoding model, and compared to the observed fMRI signal to select the most likely continuation (Tang et al., 2023). We evaluated the resulting reconstructions based on their semantic similarity to their original cluster and their fidelity to the ground-truth transcript (Fig. 1).

2.1. Dataset

The data analysed pertain to BOLD fMRI responses from the dataset released by LeBel et al. (2023), where participants listened to natural, narrative stories extracted from the podcasts The Moth Radio Hour (77 stories) and Modern Love (7 stories) (LeBel et al., 2023). We focused our analysis on the three participants (S1, S2, and S3) that listened to all 84 stories (over 16 hours). These stories possess up to five different tags that describe their content, with the exception of five stories that contain no tags. These were tagged by providing their transcript and the set of available tags to Gemini 2.5 Pro. The most frequent tag is 'Family' (30 stories), followed by 'Relationships' (21 stories), 'Health/Medicine' (17 stories) and 'Jobs/Employment' (16 stories). We kept only the most frequent tag to describe the main topic of each story (Fig 2A), combining 'Health/Medicine' and 'Jobs/Employment' into 'Health/Jobs' and assigning 'Other' to those stories that did not possess any frequent tag.

2.2. Semantic Clusterization

To classify the stories according to their semantic content in a largely unsupervised manner, we made use of OpenAI's `text-embedding-3-small` embedding model to transform the stories' transcripts into 1536 dimensional text embeddings. These were then used as input to a K-Means algorithm (K=2, scikit-learn v1.7.0).

2.3. Voxelwise encoding models

We made use of the code base released by LeBel et al. (2023) to train the voxelwise encoding models, albeit with some changes. Instead of employing GPT-1 (Radford et al., 2018) as the base LLM of the encoding model, we made use of GPT-2 Small (Radford et al., 2019) (approximately 124 million parameters) fine-tuned to use word level tokenization[1], as it was shown to perform more similarly to humans (Vaidya et al., 2023). The encoding model hyperparameters remained unchanged (layer nine was used for extracting text features, with a context window of five words). For within-cluster evaluation, we performed leave-one-out cross-validation

[1] https://github.com/HuthLab/lm-repeating-text

to maximize the amount of training data in the encoding model. Training time for each fold averaged four hours, while decoding time of each story averaged two hours. This was performed on an Intel Core i7-11700 CPU, NVIDIA GeForce RTX 3060 12GB vRAM GPU, and 64GB DDR4 memory.

2.4. Decoding stories from fMRI signals

Decoding stories from fMRI consists of two stages (Tang et al., 2023). First, a voxelwise encoding model is trained to map a text stimulus to predicted brain activity, as described above. Second, this same encoding model is used to decode the held-out stories: at each time step, the language model generates up to 200 possible text continuations (beginning with a fixed set of start words). Repetitions are filtered out of these extension words, to which we added the filtering of the unknown token to prevent it from derailing text generation. Each candidate continuation is then encoded into predicted voxel activity using the trained encoding model, and the candidate whose predicted activity best matches the observed fMRI signal is selected. This process is repeated iteratively to reconstruct the full story. A random model is defined by replacing the encoder model with a random weighing of the possible continuations.

The resulting reconstructions are then evaluated by recategorizing them into the previously defined semantic clusters and by comparing them against the ground-truth transcription. The latter is done in each time window by means of BERTScore (Zhang et al., 2020) using `microsoft/deberta-xlarge-mnli`, as it tops the benchmarks in human evaluation. BERTScore is an automatic evaluation metric that measures the similarity between generated text and reference text using pre-trained contextual embeddings. Instead of relying on exact string matching, it computes the sum of maximum pairwise cosine similarities between the token embeddings of the candidate and reference sentences. By employing greedy matching, each token is aligned to its most similar counterpart in the other sentence to compute precision, recall, and an F1 measure. An inverse document frequency (idf) file is provided in the dataset for evaluation, containing 1000 sentences for each text. These values are incorporated to apply importance weighting to token matches. However, two of the stories are missing ("From Boyhood to Parenthood" and "Where there's smoke"), so they were left out of the evaluation.

3. Results

The stories' classification into topics was primarily unsupervised. First, text embeddings were computed from the stories' transcripts to generate lower-dimensional semantic descriptions. Subsequently, unsupervised clustering (K-means, K=2; silhouette score 0.37) was applied to these embeddings, dividing the stories into two topics. One semantic cluster, designated *Family,* predominantly contains family-related narratives, while the other, *Miscellaneous,* comprises a mixed set of stories, as indicated by their primary tags (Table 1 and Fig 2A). To ensure equal cluster size, eight randomly selected stories from the Miscellaneous cluster were excluded. Word clouds (Fig 2B) of the main content words for each cluster further support this distinction, clearly showing a high frequency of kinship terms (e.g., "father," "mother," "dad," "daughter") in the Family cluster and mixed terms in the other.

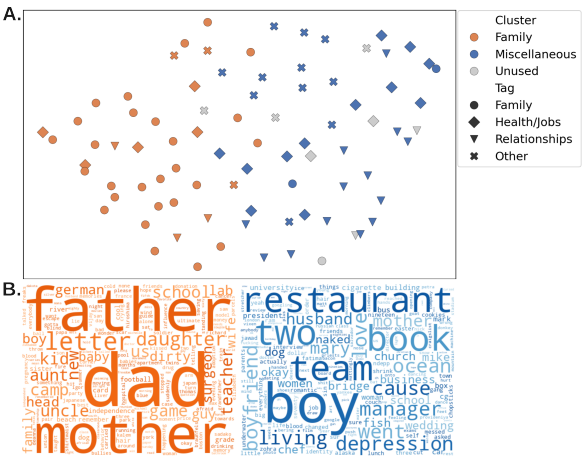

Figure 2: **A.** UMAP components of the stories' embeddings, clusterized with K-means (K=2; 0.37 silhouette score). The most frequent tag was selected for each story, as they usually possess more than one. **B.** Word cloud of the Family stories cluster (left) and Miscellaneous stories cluster (right).

Using this division, we trained voxelwise encoders (Tang et al., 2023) for all three participants on the Family stories using leave-one-out cross-validation (37 stories per training fold). This yielded 114 individual encoding models (38 per participant), plus six models trained on entire clusters (two per participant). These encoders were then used to decode either the held-out story (within-topic decoding) or the full held-out cluster (cross-topic decoding). We evaluated the resulting transcriptions in two ways: their semantic similarity to the original cluster and their fidelity to the ground-truth story transcription (see Section 2.4).

To assess semantic similarity, we computed text embeddings of the decoded transcriptions and clustered them using the previously fitted K-means algorithm. When decoded with the Family-trained encoder, 36 of 38 Family stories remained in the same cluster for at least one participant. In contrast, when using an encoder trained exclusively on

	Family cluster	Miscellaneous cluster
Number of stories	38	38
Total words (average)	68,293 (1,797)	74,531 (1,961)
Total unique words	6,081	6,644
Most common tags	Family (27), Health (10)	Relationships (13), Jobs (11)
Shared unique words	51% (3,108)	47% (3,108)

Table 1: Descriptive statistics of the stories in each semantic cluster.

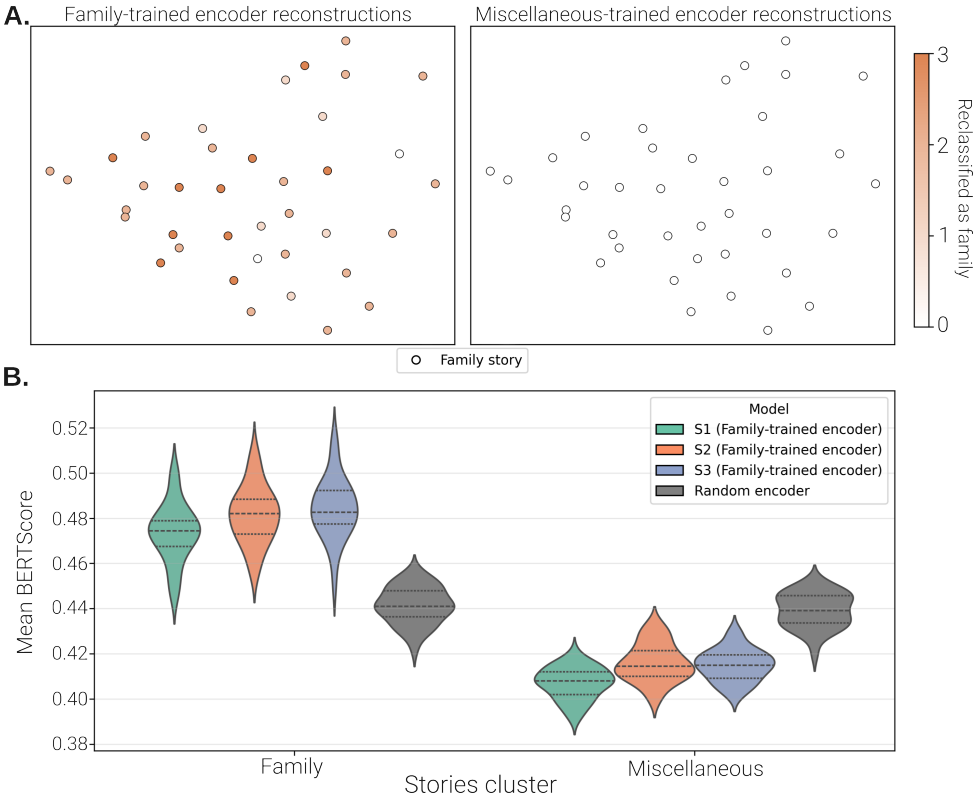

Figure 3: **A.** Clustering of the reconstructed Family stories using an encoder trained only on Family stories (left) and one trained only on Miscellaneous stories (right). Color intensity corresponds to how many reconstructions (of all three participants) have been clusterized as Family. **B.** Reconstruction performance as measured with BERTScore on both the Family and Miscellaneous stories using the Family encoder of all three participants. In grey, the performance of a random encoder is depicted.

Miscellaneous stories, all reconstructions shifted from the Family cluster to the Miscellaneous cluster (Table 2 and Fig 3A). This reveals a strong dependency on training topics: encoders are biased to interpret fMRI signals as referring to their specific semantic training context.

Encoder	Ground-truth stories cluster	Reconstructed Family	Reconstructed Misc.
Family-trained	Family	76	38
	Miscellaneous	0	114
Misc.-trained	Family	0	114

Table 2: Stories categorization after being reconstructed with Family-trained and Miscellaneous-trained encoders.

To assess reconstruction fidelity, we computed BERTScore (Zhang et al., 2020) between ground-truth transcriptions and decoded text in each time window, using a random encoder as baseline (Tang et al., 2023). Figure 3B shows average BERTScore for each story across participants when using the Family-trained encoder to reconstruct both within-cluster (Family) and cross-cluster (Miscellaneous) stories. The Family-trained encoder significantly outperformed the random encoder for Family stories (mean BERTScore: 0.47, 0.48, 0.48 ± 0.13 for S1, S2, S3, respectively, vs. 0.44 ± 0.01 for random; $p < 0.001$, Wilcoxon rank-sum test), but performed significantly worse on Miscellaneous cluster stories (mean: 0.41, 0.42, 0.42 ± 0.01 vs. 0.44 ± 0.01; $p < 0.001$, Wilcoxon rank-sum test). These results demonstrate limited generalization of brain encoders trained on specific topics, and

caution against overinterpreting their predictions when applied to out-of-domain fMRI signals.

Finally, we examined the most prominent brain regions ranked by Pearson correlation between Family-trained encoder predictions and ground-truth fMRI signals from Family stories, averaged across stories for each participant (Fig. 4). To achieve this, we kept the resulting top 10% of voxels and parcelled them into ROIs using Schaefer's atlas (2mm³, 400 ROIs) (Schaefer et al., 2018). A strong dominance from the left hemisphere can be observed in S3 (who yielded the highest correlation), particularly in brain regions associated with the language and speech perception networks (Fedorenko et al., 2024; Damera et al., 2023), encompassed in the atlas by the Somatomor and temporal Default mode networks (Table 3). This is consistent with prior work on language decoding from fMRI signals (Tang et al., 2023), and also provides a more nuanced understanding of the functional aspects involved. The average 90th percentile correlation was 0.16 (std 0.02; 8,113 voxels) for S1, 0.18 (std 0.03; 9,426 voxels) for S2, and 0.20 (std 0.05; 9,556 voxels) for S3.

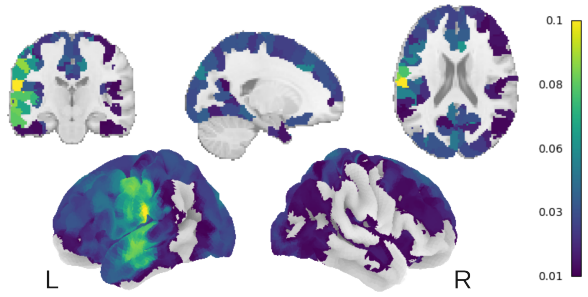

Figure 4: Mapping to Schaefer's atlas (Schaefer et al., 2018) with 400 ROIs of the 10% voxels with the highest correlation between its activity and the estimation of the Family encoder on Family stories in S3.

Network	Hem.	N°	Corr.
Somatomotor	L	11	0.11
Somatomotor	L	16	0.09
Default Mode Temporal	L	8	0.09
Somatomotor	L	15	0.09
Somatomotor	L	13	0.08
Default Mode Temporal	L	4	0.08
Default Mode Temporal	L	2	0.08
Somatomotor	L	2	0.08
Somatomotor	L	12	0.07
Somatomotor	L	1	0.07

Table 3: ROIs with the highest correlation in S3. Hem. refers to hemisphere, and N° refers to the ROI number.

4. Discussion

In this work, we explored how specific semantic domains influence the decoding of language from fMRI signals. Voxelwise encoding models are becoming increasingly more prominent in the field (Dupré la Tour et al., 2025), helping to unveil localized semantic coding in the brain (Chen et al., 2025) and to leverage data-driven models for testing of language neuroscience hypotheses (Antonello et al., 2023). Conversely, alignment of LLMs to predictive models of brain signals is proving to be a productive endeavour, as language processing performance of the former improves in accordance to its alignment with the latter (Toneva and Wehbe, 2019; Zhou et al., 2024). As such, understanding the potential biases, limitations, and mitigation strategies of these tools becomes essential. Our results reveal a strong dependence on the topics included in their training, where narrowing them to specific contexts (e.g., "Family") forces the interpretation of brain activity as consistently relating to those domains, resulting in degraded performance on out-of-domain content. Even though the case studied here is taken to an extreme, it serves to highlight an aspect that has generally been overlooked in the field: the potential biases that may incur if the semantic domains employed are not diverse enough.

It is a well-established principle in machine learning that divergence between training and test data yields degraded performance. In the standard case, a model trained on a narrow distribution generalizes poorly to unseen domains, producing outputs that are simply less accurate or less confident (Quiñonero-Candela et al., 2022). In the results presented here, however, the effect is qualitatively different: topic-constrained encoding models do not merely fail to generalize, but actively reconstruct out-of-domain stimuli in the direction of their training domain, with out-of-topic performance falling below the random baseline. This points to a form of domain bias that is stronger than a simple loss of generalization: the model does not become uninformative in the face of unfamiliar input, but rather imposes a systematic distortion, pulling reconstructions toward a familiar but incorrect semantic space. This distinction matters because it implies that deploying narrowly trained encoding models on ecologically diverse stimuli could introduce structured errors that are harder to detect than mere noise and, potentially, harder to correct for.

What constitutes a diverse enough semantic domain, however, remains an open question. A related consideration raised by this work is how generalizable these tools can effectively be. Given that BOLD signals are inherently slow and spatially coarse, voxelwise encoding models may face in-

herent constraints on the breadth of semantic diversity they can capture. Moving forward, a stronger emphasis should be made both to the variety and lexical richness of stimuli used during data collection and to systematic evaluation of cross-domain generalization.

Limitations

Some of the limitations on this work stem from the usage of a previously published dataset, which was not specifically designed to test the hypothesis inquired here. In particular, topics are not clearly divided, nor balanced. The semantic cluster Miscellaneous comprises stories with diverse topics, instead of a single one, which would have helped to contrast the effects of training in one specific semantic domain versus another. Another limitation pertains to the amount of training data: dividing stories in two independent sets implies halving the number of samples, yielding a substantial loss in performance. Even though we have tried to maximize the amount of training data by performing leave-one-out cross validation, it involved a great cost in compute time (nearly 2,000 hours or 82 days).

Ethics statement

This study utilizes a previously published, de-identified fMRI dataset. As the data was collected under protocols approved by the original host institution's Institutional Review Board (IRB) and is publicly available for research use, this work did not require additional ethics committee approval.

5. Acknowledgements

J.E.K. received research grants from CONICET (PIP 11220220100240CO). F.T., B.B., D.F.S., and J.E.K. were funded by the CONICET and UBA.

6. Bibliographical References

Richard Antonello and Alexander Huth. 2024. Predictive Coding or Just Feature Discovery? An Alternative Account of Why Language Models Fit Brain Data. *Neurobiology of Language (Cambridge, Mass.)*, 5(1):64–79.

Richard Antonello, Javier Turek, Vy Vo, and Alexander Huth. 2021. Low-dimensional structure in the space of language representations is reflected in brain responses. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, pages 8332–8344, Red Hook, NY, USA. Curran Associates Inc.

Richard Antonello, Aditya Vaidya, and Alexander Huth. 2023. Scaling laws for language encoding models in fMRI. *Advances in Neural Information Processing Systems*, 36:21895–21907.

Catherine Chen, Tom Dupré la Tour, Jack L. Gallant, Daniel Klein, and Fatma Deniz. 2024. The cortical representation of language timescales is shared between reading and listening. *Communications Biology*, 7(1):284. Publisher: Nature Publishing Group UK London.

Jiaqi Chen, Richard Antonello, Kaavya Chaparala, Coen Arrow, and Nima Mesgarani. 2025. Quantifying Semantic Functional Specialization in the Brain Using Encoding Models of Natural Language. In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, pages 77–90, Albuquerque, New Mexico, USA. Association for Computational Linguistics.

Srikanth R. Damera, Lillian Chang, Plamen P. Nikolov, James A. Mattei, Suneel Banerjee, Laurie S. Glezer, Patrick H. Cox, Xiong Jiang, Josef P. Rauschecker, and Maximilian Riesenhuber. 2023. Evidence for a Spoken Word Lexicon in the Auditory Ventral Stream. *Neurobiology of Language*, 4(3):420–434.

Tom Dupré la Tour, Matteo Visconti di Oleggio Castello, and Jack L. Gallant. 2025. The Voxelwise Encoding Model framework: A tutorial introduction to fitting encoding models to fMRI data. *Imaging Neuroscience*, 3:imag_a_00575.

Evelina Fedorenko, Anna A. Ivanova, and Tamar I. Regev. 2024. The language network as a natural kind within the broader landscape of the human brain. *Nature Reviews Neuroscience*, 25(5):289–312.

Laura Gwilliams, Alec Marantz, David Poeppel, and Jean-Rémi King. 2025. Hierarchical dynamic coding coordinates speech comprehension in the human brain. *Proceedings of the National Academy of Sciences*, 122(42):e2422097122.

Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G. Huth. 2023. A natural language fMRI dataset for voxelwise encoding models. *Scientific Data*, 10(1):555. Publisher: Nature Publishing Group.

Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, Neil D. Lawrence, Michael I. Jordan, and Thomas Dietterich, editors. 2022.

Dataset Shift in Machine Learning. Neural Information Processing series. MIT Press, Cambridge, MA, USA.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving Language Understanding by Generative Pre-Training.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners.

Alexander Schaefer, Ru Kong, Evan M. Gordon, Timothy O. Laumann, Xi-Nian Zuo, Avram J. Holmes, Simon B. Eickhoff, and B. T. Thomas Yeo. 2018. Local-Global Parcellation of the Human Cerebral Cortex from Intrinsic Functional Connectivity MRI. *Cerebral Cortex (New York, N.Y.: 1991)*, 28(9):3095–3114.

Jerry Tang, Amanda LeBel, Shailee Jain, and Alexander G. Huth. 2023. Semantic reconstruction of continuous language from non-invasive brain recordings. *Nature Neuroscience*, 26(5):858–866. Publisher: Nature Publishing Group.

Mariya Toneva and Leila Wehbe. 2019. Interpreting and improving natural-language processing (in machines) with natural language-processing (in the brain). In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc.

Aditya Vaidya, Javier Turek, and Alexander Huth. 2023. Humans and language models diverge when predicting repeating text. In *Proceedings of the 27th Conference on Computational Natural Language Learning (CoNLL)*, pages 58–69, Singapore. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. ArXiv:1904.09675 [cs].

Yuchen Zhou, Emmy Liu, Graham Neubig, Michael J. Tarr, and Leila Wehbe. 2024. Divergences between Language Models and Human Brains. *Advances in Neural Information Processing Systems*, 37:137999–138031.