Language, Computation and Cognition Project Report

Jeremy Jornet (jeremyj@campus.technion.ac.il)
Technion, Israel Institute of Technology

Abstract

This work investigates the neural decoding of linguistic meaning from brain activity using static and contextual semantic embeddings. We first addressed three main scientific questions: (1) Does the choice of static embedding model affect decoding accuracy for single-word and sentence-level stimuli? (2) Do contextual embeddings such as BERT provide a closer match to neural representations of sentence meaning than static embeddings? (3) To what extent can semantic decoders generalize across individuals? Building upon the foundational framework established by Pereira et al. (2018), we present both replication and extensions of their neural semantic decoding methodology. Our results indicate that Word2Vec static embeddings yielded performance comparable to GloVe. Then, contextual BERT embeddings yield significantly superior performance compared to static embedding models (GloVe, Word2Vec) in sentence-level decoding tasks and show stronger alignment with brain activity patterns. However, despite these advances in single-subject decoding, semantic decoders fail to generalize meaningfully across individuals, consistently performing at chance level (90% mean rank). Pairwise analysis reveals that certain participant pairs may share similar representational spaces, suggesting the potential for subgroup-specific rather than universal cross-subject decoding approaches. This work points to plausible next steps for advancing inter-subject semantic decoding: further analyses on a larger scale using robust statistical methods to identify pairwise shared semantic representations, as well as the use of alignment methods to bring fMRI data into a common representational space across individuals.[1]

Introduction

Understanding how language is represented in the brain is a central challenge in cognitive neuroscience and computational linguistics. A landmark study by Pereira et al. proposed a method for decoding linguistic meaning from brain activation patterns using distributional semantic models. In their work, the authors trained subject-specific decoders to predict 300-dimensional GloVe vectors (Pennington et al., 2014) from fMRI responses elicited while participants read words or sentences, or viewed pictures. This approach demonstrated that it is possible to reconstruct semantic representations from neural activity, and further showed that the decoded vectors preserved meaningful semantic structure.

At the time of their study, only static word embedding models were available. These models assign a single vector to each word, regardless of context, thus failing to capture important aspects of language. Since late 2018, the release of contextual word embedding models—such as BERT (Devlin et al., 2019)—has opened new avenues for studying how context-sensitive meaning is encoded in the brain. These models generate distinct representations for the same word depending on its surrounding context, potentially offering a more neurally plausible account of semantic representation.

We began by reproducing Experiment 1 using an alternative static word embedding model, Word2Vec, instead of GloVe. We found that Word2Vec achieved results very similar to GloVe in decoding single-word semantic representations from fMRI data, suggesting that the choice of a different embedding model may not substantially impact performance within the class of static embeddings. We also reproduced Experiments 2 and 3 from Pereira et al., which involved predicting semantic representations of sentences from brain activity using a decoder trained on single-word fMRI data from Experiment 1.

We then derived a new version of Experiment 2. Instead of decoding sentence meaning from individual concept representations, we trained a decoder directly on the sentence-level fMRI data. We compared performance using static GloVe vectors and contextual BERT vectors as targets. This allowed us to address the following research question: *Do contextual embeddings provide a better match to brain activity than static embeddings when decoding sentence-level meaning?* Our results showed that BERT consistently outperformed GloVe in this setting, both in terms of decoding accuracy and representational alignment with brain activity.

In addition to these decoding analyses, we inverted the direction of the original study's main analysis. While Pereira et al. decoded embeddings from brain activity, we trained *brain encoder models* that map both static (GloVe) and contextual (BERT) sentence vectors to fMRI responses. This approach allows us to examine which voxels are significantly associated with different types of embedding representations and how accurately they predict voxel-level activation patterns. Our results revealed that BERT-based embeddings are not only more predictive overall, but also show stronger and more widespread voxelwise associations, suggesting that contextual embeddings are more aligned with the brain's representation of sentence meaning.

Finally, we extended *Experiment 1* from Pereira et al. by exploring the research question: *Can a brain decoder generalize semantic representations across subjects?* This *inter-subject decoder* approach addresses a fundamental limitation of current subject-specific models and raises the question about the possibility of constructing shared neural-semantic representations that are robust across individuals. This is both scientifically and practically important: it tests the generalizability of semantic representations across brains, addressing the fundamental question of whether there exists a shared neural encoding of meaning across individuals. Demonstrating a common representational structure would suggest that semantic processing relies on consistent neural patterns across people. Previous studies have explored the feasibility of decoding semantic information across individuals us-

[1] Code available at `https://github.com/jeremyjrnt/LaCC`

ing text-only stimuli. Anderson et al. (2018) introduced an integrated neural decoder capable of predicting semantic content from fMRI data across participants. They demonstrated successful cross-participant decoding, suggesting the presence of partially shared semantic representations across individuals. Similarly, a follow-up study published in *Frontiers in Human Neuroscience* extended this line of research using sentence-level fMRI data and leave-one-subject-out evaluation. The authors reported that certain semantic dimensions—particularly those grounded in concrete experiential content—were more consistently decodable across brains, while more abstract features showed greater individual variability. While previous work reported above-chance decoding, our brain inter-subject decoders did not generalize meaningfully across participants, performing at chance level. Our analysis outcome suggests that common semantic representations across all subjects may be absent, and that meaningful decoding likely requires subject-specific training. These results also point to the possible need for additional methods such as alignment or shared representational vector space to better align fMRI data across individuals for inter-subject semantic decoding. An additional noteworthy finding was that pairwise decoder transfer exhibited significant deviations from chance-level performance for certain participant pairs, suggesting that semantic representations may be shared within specific subgroups of the population, rather than being uniformly shared across all individuals.

Data

We used the fMRI dataset introduced by Pereira et al., which is publicly available at `https://web.mit.edu/evlab//sites/default/files/documents/index2.html`. It includes recordings from 16 English-speaking participants. Each participant was scanned while reading or seeing linguistic stimuli across three experiments. For consistency in data formatting, we retained 15 out of the 16 subjects: the first subject's data were excluded due to a different preprocessing format.

The dataset includes neural responses to three types of stimuli:

- **Experiment 1:** 180 individual concepts (mostly nouns, adjectives, and verbs) presented in three disambiguating conditions: sentence context, image, and word cloud.
- **Experiment 2:** 384 sentences grouped into 96 passages, spanning 24 broad topics (e.g., musical instruments, disasters). Each topic comprises 4 passages, and each passage contains 4 sentences.
- **Experiment 3:** 243 sentences grouped into 72 passages from 24 new topics not overlapping with Experiment 2. Each topic includes 3 passages, with 3 to 4 sentences per passage.

Each sentence or word was shown multiple times (typically 3), and voxel activity was averaged across repetitions. Brain responses are represented as vectors of voxels per stimulus.

Across experiments, the nature of the stimuli and the decoding setup differ:

- Experiment 1 focuses on isolated word-level concepts with disambiguating context, making it suitable for single-word decoding and encoder modeling.
- Experiment 2 uses topic-coherent factual passages composed of declarative sentences, enabling analyses of semantic composition and topic generalization.
- Experiment 3 includes a mix of factual and narrative-style passages across new topics, introducing greater diversity in linguistic structure and content.

The concepts used in the stimuli are semantically diverse, ranging from highly concrete and imageable (e.g., *bag*, *prison*) to abstract and non-visual (e.g., *relationship*, *kindness*).

For semantic representation, we used three types of embeddings:

- **GloVe** (Pennington et al., 2014): 300-dimensional static word embeddings.
- **Word2Vec**: another static model, also producing 300-dimensional vectors.
- **BERT** (Devlin et al., 2019): contextual embeddings producing 768-dimensional vectors.

Experiments and Results

This section is divided into two main parts. The first part, **Structured Tasks**, follows the analyses required as part of the replication and extension of the original experiments presented in Pereira et al.. These tasks include standard decoding and encoding evaluations based on their experimental design.

The second part, **Open-Ended Task**, reflects our own design choices and research questions. It aims to go beyond the original framework by exploring an additional hypothesis : the inter-subject generalization.

We employ the cross-fold mean-rank evaluation framework used in Pereira et al. throughout this work.

Structured Tasks

Sentence Decoding We began by applying a Word2Vec-based decoder to Experiment 1, replacing the original GloVe embeddings used by Pereira et al.. Like GloVe, the Word2Vec vectors were 300-dimensional. We followed the exact same training setup as in our GloVe-based model from Homework 3, where GloVe achieved a global mean rank of 61.91. Word2Vec achieved very similar generalization performance, with a **global mean rank of 61.08**, substantially better than the chance level of 90. As shown in Figure 1, the average rank was consistently below the chance threshold across all 18 folds, highlighting the robustness of Word2Vec embeddings for modeling brain-based semantic representations.

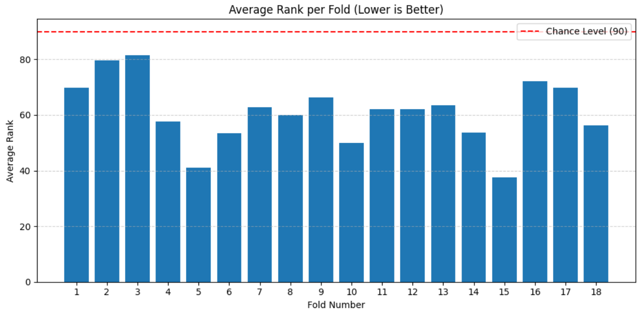


Figure 1: Average rank score per fold for the Word2Vec-based decoder on Experiment 1 (180 concept decoding). Lower values indicate better performance. The red dashed line indicates the chance level (rank = 90). Most folds achieve substantially better-than-chance decoding accuracy, confirming that Word2Vec captures semantic information aligned with fMRI representations.

Before proceeding with the replication and extension of some of the original experiments, we first describe the key similarities and differences between Analyses 1, 2, and 3 in Pereira et al.. Understanding well these analyses helps clarify how our work builds upon and extends the original study. Experiments 1, 2, and 3 in Pereira et al. share a common decoder and semantic space. In all three analyses, the same subject-specific ridge regression decoder trained on isolated concept representations from Experiment 1 is used to map brain activity to 300-dimensional GloVe word embeddings. Decoding performance is evaluated using consistent metrics such as rank accuracy and pairwise classification. Despite these shared elements, the experiments differ in several ways. **Experiment 1** focuses on decoding individual concepts from fMRI activity. A set of 180 words is presented to participants in three formats: as a sentence, an image, and a word cloud. In contrast, **Experiments 2 and 3** tests whether the decoder trained from Experiment 1 can generalize from isolated word representations to structured sentence-level contexts. The semantic structures are centered on *topics*. Each topic is associated with four short passages, each covering a specific element of that topic, and each passage includes four sentences. In total, Experiment 2 includes 384 sentences while Experiment 3 contains 243 sentences. The sentence stimuli used in these two experiments were constructed independently from the materials in Experiment 1 and independently from each other. In **Experiment 2**, all passages are written in a Wikipedia-style encyclopedic tone, whereas **Experiment 3** introduces greater diversity, combining both Wikipedia-style texts and first-/third-person narrative passages. These two sentence-level experiments maintain similar levels of within- and between-topic and semantic similarities. Sentence vectors in both Experiments 2 and 3 are computed by averaging GloVe embeddings over the content words in each sentence. Evaluation also differ between Experiment 1 and Experiments 2 and 3. In Experiment 1, decoding accuracy is assessed by comparing predicted vectors for individual words. In contrast, Experiments 2 and 3 involve sentence-level evaluations use a multi-level pairwise classification task with increasing semantic difficulty: (1) between two sentences from the same passage, (2) between two sentences from the same topic, and (3) between two sentences from different topics. This evaluation introduces more fine-grained difficulty levels than the word-level pairwise tests in Experiment 1. Rank accuracy is also employed in Experiments 2 and 3 to assess overall generalization and is computed at the sentence level. Lastly, the number of participants varies across experiments: Experiment 1 includes 16 subjects, Experiment 2 includes 8, and Experiment 3 includes 6.

We reproduced Experiments 2 and 3 by using the learned GloVe-based decoder model from Experiment 1 to decode sentence-level semantic representations. For each sentence, we computed its vector representation by averaging the GloVe embeddings of the content words. Decoding performance was evaluated using the rank accuracy metric. In Experiment 2, which included 384 sentences, the decoder achieved an average rank of **156.93**, compared to a chance level of 192. In Experiment 3, involving 243 sentences, the decoder achieved a stronger average rank of **100.74**, below the chance level of 121.5. These results confirm that the decoder trained on single-word fMRI responses was able to generalize to sentence-level stimuli, achieving better-than-chance predictions in both cases.

We further analyzed decoding performance at the topic level by averaging the rank scores of sentences within each topic. In Experiment 2, the best-performing topics included *landscape* (93.38), *clothing* (117.94), and *disaster* (119.06), suggesting that the decoder performed well on perceptually grounded, concrete categories. In contrast, it struggled on topics such as *profession* (218.44), *animal* (203.50), and *fish* (204.19), which may contain more abstract or diffuse content. In Experiment 3, the best-decoded topics were *bone fracture* (43.73), *law school* (46.00), and *gambling* (61.20). These topics likely benefited from consistent vocabulary and structured descriptions. On the other hand, decoding performance was lowest for topics such as *beekeeping* (177.75), *pharmacist* (130.00), and *owl* (127.50), which are more technical or specialized topics that may be less familiar to participants and harder to describe using precise or consistent vocabulary.

Sentence Decoding We trained a decoder model on the dataset from Experiment 2 using two types of sentence representations: static GloVe vectors, obtained by averaging the embeddings of the content words in each sentence, and contextual BERT embeddings, computed by averaging the token representations from the final hidden layer of the BERT-base model. Since the default dimensionality of BERT embeddings is 768, we also evaluated a PCA-projected version of BERT reduced to 300 dimensions, in order to match the dimensionality of the GloVe embeddings and ensure a fair comparison across embedding sizes. We evaluated sentence decoding performance on Experiment 2 using three types

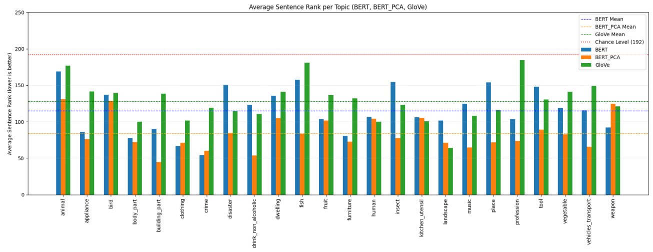


Figure 2: Average sentence rank per topic on Experiment 2 using three types of sentence embeddings: BERT (blue), BERT reduced to 300 dimensions via PCA (orange), and GloVe (green). Lower ranks indicate better decoding performance. The red dotted line indicates the chance level (192), while dashed lines show the global mean for each method. BERT PCA consistently outperformed both original BERT and GloVe across most topics, suggesting that contextual information combined with dimensionality compression yields more brain-aligned semantic representations.

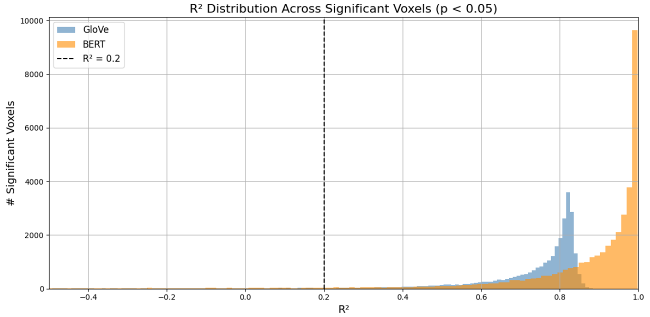


Figure 3: R^2 distribution across significantly predicted voxels ($p < 0.05$) for GloVe (blue) and BERT (orange) encoder models. The vertical dashed line marks the $R^2 = 0.2$ threshold used to indicate strong voxelwise prediction. BERT exhibits a broader and more right-skewed distribution, with a higher density of voxels near $R^2 = 1.0$, suggesting more robust and widespread alignment with brain activity compared to GloVe.

of sentence embeddings: GloVe (300-dimensional), BERT (768-dimensional), and BERT reduced to 300 dimensions via PCA. All three models achieved strong generalization, with average ranks well below the chance level of 192: GloVe obtained a global mean sentence rank of 128.12, BERT 114.70, and BERT PCA 84.13. As shown in Figure 2, contextual embeddings outperformed the static GloVe embeddings in this decoding task. Surprisingly, the PCA-reduced BERT embeddings achieved a substantially lower (better) mean rank than both the original BERT. This suggests that dimensionality reduction via PCA may have enhanced the signal-to-noise ratio by compressing the most informative semantic dimensions while avoiding redundant or noisy features.

Brain Encoder Model We then reversed the previous decoding experiment by training a brain *encoder* model to predict human neural signals directly from semantic embeddings. The goal of this analysis is to evaluate how well semantic information contained in different types of embeddings explains brain activity. Specifically, we compared static GloVe vectors and contextual BERT embeddings using the dataset from Experiment 2, which includes 384 sentences and their corresponding fMRI activation patterns.

For each embedding type, we fit a separate linear regression model for each voxel in the brain, resulting in one model per voxel. This voxelwise encoding approach allows us to examine both (i) how many voxels are significantly associated with the information embedded in the sentence vectors, and (ii) how accurately the activation in those voxels can be predicted from the semantic features.

Several prior studies have adopted heuristic or statistically justified thresholds on the coefficient of determination (R^2) to identify brain voxels that are meaningfully predicted by semantic models. For instance, Güçlü & van Gerven (2017) explicitly apply an R^2 threshold of 0.1 to select voxels for visualization and interpretation in visual encoding tasks.

In the domain of semantic encoding, studies such as Huth et al. (2016) and Pereira et al. rely on *voxel-wise permutation testing* to establish the statistical significance of voxel-wise prediction performance. Following this approach, we evaluate the significance of each voxel's R^2 value by comparing it against a null distribution obtained via stimulus permutation, using 100 random permutations. Voxels with $p < 0.05$ are considered significantly predicted. Among these, I further examine how well the model predicts neural responses by evaluating the distribution of R^2 scores, focusing in particular on the subset of significant voxels with $R^2 > 0.2$—a threshold commonly used in prior work to identify strongly predicted voxels (e.g., Huth et al., 2016; Naselaris et al., 2011; Guntupalli et al., 2016).

For the GloVe-based model, **30,116 out of 185,866 voxels** (16.20%) were significantly associated with the information embedded in the word vectors at $p < 0.05$. Among these, **28,903 voxels** (95.97%) had an R^2 value greater than 0.2, indicating strong semantic encoding.

In comparison, the BERT-based model significantly associated **39,167 voxels** (21.07%) with the semantic embeddings, and among these **37,122 voxels** (94.78%) exceeded the $R^2 > 0.2$ threshold. This indicates that BERT captures semantic information across a broader portion of the brain compared to GloVe. R^2 distributions for both models are provided in Figures 7 and 8 in the appendix.

As shown in Figure 3, significant voxels in the BERT model tend to exhibit higher R^2 values, with many approaching 1.0, whereas the GloVe model shows a narrower distribution concentrated mostly between 0.6 and 0.85. This suggests that BERT not only explains a greater number of voxels but also does so with higher predictive power.

Open Ended Task

The open-ended task was designed to extend the scope of the original experiments by exploring new analyses not covered in Pereira et al..

We investigated whether a semantic decoder trained on the brain activity of one group of participants can generalize to a different participant. This question directly addresses the possibility of shared neural-semantic representations across subjects, and whether a decoder trained on one brain can effectively predict meaning from another.

For this analysis, we used the fMRI data from Experiment 1, which includes brain responses to 180 individual concepts, and the original 300-dimensional GloVe embeddings used throughout the paper as the target semantic representations. One challenge in building the inter-subject decoder was that the fMRI data vectors did not have the same dimensionality across the 15 subjects. To address this issue, we applied transformations to the fMRI vectors of each subject in order to standardize their dimensionality, ensuring that all brain activation vectors had the same length across subjects. This step was necessary to enable shared and transferred decoder model across participants. We applied a PCA projection to each subject's fMRI matrix, reducing it to a 180-dimensional space[2] .

We conducted three experiments under three main paradigms to evaluate the generalizability of inter-subject decoding:

1. **Full-data training:** We trained the decoder on the full fMRI data of 14 participants using a 17/1 cross-validation scheme, and evaluated its performance on the held-out 15th participant.

2. **Averaged-data training:** We first averaged the fMRI data across the 14 participants for each stimulus, then trained the decoder on this averaged data using the same 17/1 scheme, and tested on the held-out participant.

3. **Pairwise transfer:** For each pair of participants (i, j), we trained a decoder on participant i using 17/1 cross-validation and directly tested the learned decoder on participant j's data.

All decoding performances were evaluated using rank accuracy, following the same method as in Experiment 1 of Pereira et al., where a lower rank indicates better alignment between the predicted and target semantic vectors. We recall that chance level in this setting is 90, given 180 concepts. The mean rank scores reported in this section correspond to the global mean of the average rank scores obtained across all 18 folds of the 17/1 cross-validation evaluation.

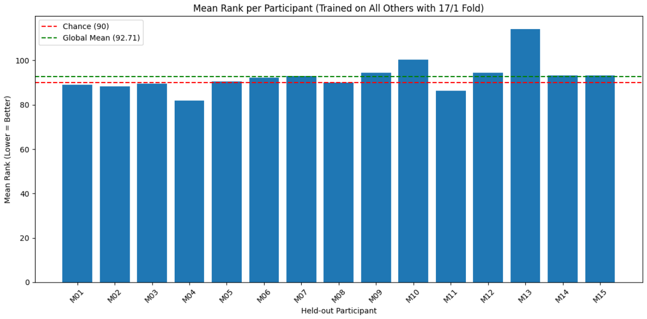


Figure 4: Mean rank per participant for the full-data training paradigm using 17/1 cross-validation. Each bar represents the average decoding rank when a participant was held out as the test subject and the model was trained on the other 14 participants. The red dashed line indicates chance level (90), and the green dashed line shows the global mean rank (92.71). Most scores hover near the chance level, indicating limited generalization of the decoder across subjects.

Full-data training The goal of the full-data training paradigm was to examine whether increasing the quantity of training data—without aggregation—could improve inter-subject decoder performance. Unlike the averaged-data approach, where each concept was represented by a single group-level fMRI vector, this setup preserved the individual variability of each subject and explicitly increased the number of training examples. Specifically, for each of the 180 concepts, the decoder was trained on all 14 individual subject representations, resulting in a much larger training set. The hope was that this expanded dataset would allow the model to better capture generalizable patterns across diverse brains and lead to improved performance when decoding on a held-out participant.

The global mean rank across all participants was **92.71**, slightly above the chance level of 90. As shown in Figure 4, the rank scores for all subjects is around 90, highlighting the fact that the decoder failed to generalize meaningfully across participants and produced near-chance predictions. Despite training on a large amount of data (14 subjects), the decoder could not generalize meaningfully to unseen participants. Simply increasing the amount of training data by pooling multiple participants is not sufficient for better inter-subject generalization in our case.

Averaged-data training The averaged data paradigm was designed to assess whether averaging could smooth out individual noise and reveal shared semantic representations that generalize more effectively than those trained on a single subject. In this approach, we averaged the fMRI responses across 14 participants for each of the 180 concepts in Experiment 1, producing a single group-level brain response per stimulus. Three variants of this aggregation were implemented: the arithmetic mean, the geometric mean, and the harmonic mean. The motivation for using multiple averaging schemes

[2] The PCA solution was chosen in agreement with the course Teaching Assistants, who supported it as a sufficient approach to resolve the dimensionality mismatch across subjects and proceed with the experimental work.

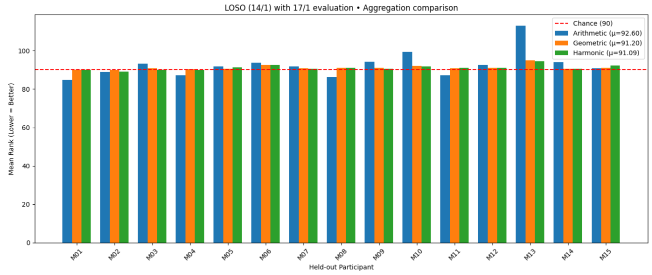


Figure 5: Mean rank of arithmetic, geometric, and harmonic mean aggregation strategies in the Averaged-Data Training Paradigm. Bars show the mean rank (lower = better) for each held-out participant under a 14/1 participant training scheme with 17/1 fold evaluation. The red dashed line indicates chance performance (rank = 90). Global means (μ) per average variant are reported in the legend. Both geometric and harmonic means yielded slightly lower average ranks than the arithmetic mean. However, all aggregation methods performed close to chance level.

was to explore whether different ways of attenuating individual variability would affect the extracted shared semantic signal. The arithmetic mean equally weights all participants responses, providing a straightforward estimate of the central tendency. The geometric mean is less sensitive to large outliers and can emphasize proportional rather than absolute differences across participants. The harmonic mean, being more influenced by smaller values, can down-weight disproportionately high voxel responses, potentially enhancing robustness to extreme activations. A decoder was then trained on each type of aggregated dataset and tested on the held-out participant.

The global mean ranks across all participants were highly similar for the three aggregation methods : arithmetic **92.60**, geometric **91.20**, and harmonic **91.09**. As shown in Figure 5, the rank scores for nearly all participants and all averaging schemes hovered close to the chance level of 90, with only minor, non-significant fluctuations between methods. This pattern indicates that, regardless of whether arithmetic, geometric, or harmonic averaging was used, aggregating fMRI responses across participants did not yield meaningful gains in decoding performance, and the averaged-data paradigm failed to generalize effectively to a held-out participant.

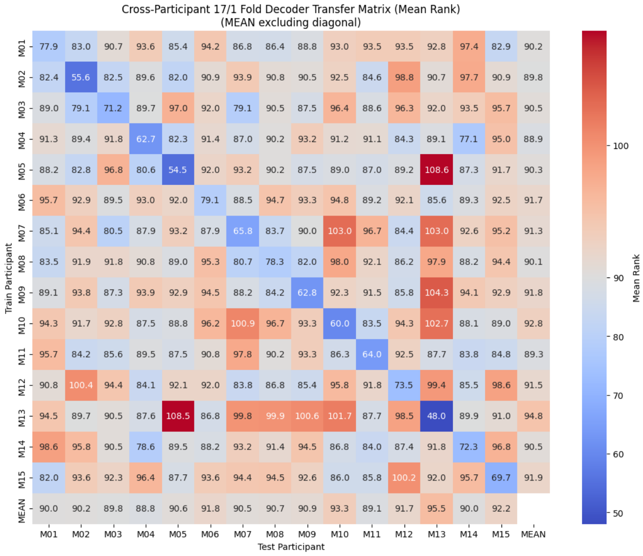
The patterns observed in Figure 5 strongly resembles that of the averaged-data paradigm shown in Figure 4, with both plots exhibiting a consistent distribution of scores near chance level across participants. This similarity further suggests that neither increasing the quantity of training data nor averaging across subjects resolves the fundamental challenge of inter-subject generalization in semantic brain representations.

Pairwise transfer In our previous analyses, the leave-one-participant-out approach for the test paradigms failed to achieve meaningful generalization, suggesting that a universal decoder trained across all other participants may not capture subject-independent semantic representations effectively. This motivated us to explore whether inter-subject generalization might occur in a more localized, pairwise fashion. In other words, our goal was then to evaluate whether a decoder trained on one participant could be successfully transferred to another. Beyond overall performance, this setup allows us to examine more nuanced patterns: for example, whether certain participants consistently receive or provide better-than-chance decoders. It would be particularly interesting to identify pairs where the decoder performs significantly above or below chance, as this could suggest closer representational similarity between certain individuals.

The mean values in the row and column summaries are very close to the chance level (90), as shown in Figure 6, indicating that, overall, the pairwise transfer setup fails to consistently generalize across pairs of participants.

An interesting observation is that some specific participant pairs deviate notably from the chance level. For example, the pair M13 $\rightarrow$ M05 achieves a mean rank of 108.55 (absolute deviation of 18.55), while M14 $\rightarrow$ M04 yields a mean rank of 78.58 (absolute deviation of 11.42). These and other extreme cases are detailed in Table 1 in the Appendix. We note that, for pairs with notable deviations below chance level, the scores remain substantially higher than those observed in within-subject decoding, suggesting that even if they share some aspects of semantic representation, this alignment may only be partial. It remains unclear whether such deviations are simply due to variance or if they reflect genuine cases where certain participant pairs share a particularly similar (or dissimilar) semantic representational space. This question motivated the following analysis.

We considered generating *null decoders* by permuting the concept labels in the training data. Such permutations disrupt the natural alignment between voxel activity patterns and their corresponding semantic vectors, effectively breaking any meaningful mapping between brain responses and concepts. By repeatedly training decoders on these permuted datasets, we obtain a null distribution of pairwise transfer scores that reflects performance expected by chance under identical experimental conditions. The idea is that, if the distribution of scores (and in particular its variance) for the null decoders differs notably from that of the true decoders, it would suggest that the observed extreme deviations in the real data are not some random fluctuations. Instead, such differences would point to specific participant pairs that share unusually similar, or conversely unusually dissimilar, semantic representational spaces. As an initial, exploratory assessment, we visually compared the off-diagonal elements of the true pairwise transfer score matrix with those obtained from a randomly permuted dataset, using the same color scale for both (Figures 9 and 10 in the Appendix). This qualitative inspection reveals that the permuted-concept scores exhibit less pronounced color extremes, suggesting that the null decoders produce fewer strongly deviating values. To complement this



	M01	M02	M03	M04	M05	M06	M07	M08	M09	M10	M11	M12	M13	M14	M15	MEAN
M01	77.9	83.0	90.7	93.6	85.4	94.2	86.8	86.4	88.8	93.0	93.5	93.5	92.8	97.4	82.9	90.2
M02	82.4	55.6	82.5	89.6	82.0	90.9	93.9	90.8	90.5	92.5	84.6	98.8	90.7	97.7	90.9	89.8
M03	89.0	79.1	71.2	89.7	97.0	92.0	79.1	90.5	87.5	96.4	88.6	96.3	92.0	93.5	95.7	90.5
M04	91.3	89.4	91.8	62.7	82.3	91.4	87.0	90.2	93.2	91.2	91.1	84.3	89.1	77.1	95.0	88.9
M05	88.2	82.8	96.8	80.6	54.5	92.0	93.2	90.2	87.5	89.0	87.0	89.2	108.6	87.3	91.7	90.3
M06	95.7	92.9	89.5	93.0	92.0	79.1	88.5	94.7	93.3	94.8	89.2	92.1	85.6	89.3	92.5	91.7
M07	85.1	94.4	80.5	87.9	93.2	87.9	65.8	83.7	90.0	103.0	96.7	84.4	103.0	92.6	95.2	91.3
M08	83.5	91.9	91.8	90.8	89.0	95.3	80.7	78.3	82.0	98.0	92.1	86.2	97.9	88.2	94.4	90.1
M09	89.1	93.8	87.3	93.9	92.9	94.5	88.2	84.2	62.8	92.3	91.5	85.8	104.3	94.1	92.9	91.8
M10	94.3	91.7	92.8	87.5	88.8	96.2	100.9	96.7	93.3	60.0	83.5	94.3	102.7	88.1	89.0	92.8
M11	95.7	84.2	85.6	89.5	87.5	90.8	97.8	90.2	93.3	86.3	64.0	92.5	87.7	83.8	84.8	89.3
M12	90.8	100.4	94.4	84.1	92.1	92.0	83.8	86.8	85.4	95.8	91.8	73.5	99.4	85.5	98.6	91.5
M13	94.5	89.7	90.5	87.6	108.5	86.8	99.8	99.9	100.6	101.7	87.7	98.5	48.0	89.9	91.0	94.8
M14	98.6	95.8	90.5	78.6	89.5	88.2	93.2	91.4	94.5	86.8	84.0	87.4	91.8	72.3	96.8	90.5
M15	82.0	93.6	92.3	96.4	87.7	93.6	94.4	94.5	92.6	86.0	85.8	100.2	92.0	95.7	69.7	91.9
MEAN	90.0	90.2	89.8	88.8	90.6	91.8	90.5	90.7	90.9	93.3	89.1	91.7	95.5	90.0	92.2	

Figure 6: Pairwise decoder transfer matrix across participants using a 17/1 fold training scheme. Each cell (i, j) shows the average rank score when a decoder is trained on participant i and tested on participant j using Experiment 1 data. Lower values (cooler colors) indicate better decoding performance. The diagonal shows within-subject performance; off-diagonal values reflect cross-subject generalization. While some participant pairs show modestly lower-than-chance rank scores, most cross-subject transfers yield near or above-chance results, indicating limited generalizability of decoders across individuals.

observation, Figure 11 in the Appendix presents the distributions of scores from the true and permuted conditions plotted against each other. The distribution derived from the true decoders displays a greater spread, with more values in the tails, indicating the presence of participant pairs whose transfer performance deviates more substantially from the chance level compared to what is expected under the null model. Given that both the true and permuted scores have means close to the chance expectation (90), we sought to determine whether their distributions differ in the tails, with a particular focus on potential differences in variance. To this end, we conducted an F-test for equality of variances, repeated over 50 independent permutations of the concept labels. Permutation testing is widely adopted for assessing statistical significance in neuroimaging, where it is used to construct null models by label shuffling and to compare the resulting statistical properties with those from the original data (Nichols & Holmes, 2002).

Repeating the test across multiple permutations serves to ensure robustness against random fluctuations in any single randomization and provides a more stable estimate of the null distribution's variability. For each permutation, we compared the variance of the true score distribution to that of the corresponding permuted distribution, considering p-values below 0.05 as statistically significant. The results show significant variance differences in **48 out of 50** tests (96.0%), indicating that the variance of the true scores is reliably distinct from that of the null decoders. This suggests that certain participant pairs may indeed exhibit similar or dissimilar semantic representational spaces, leading to more extreme transfer scores than expected under the null model.

Discussion and Conclusions

In this project, we replicated and extended prior work showing that linguistic meanings can be decoded from fMRI brain activation patterns using semantic models. Building upon Pereira et al. (2018), we tested various semantic representations and decoding paradigms to explore neural encoding of language.

Our results showed that Word2Vec embeddings performed comparably to GloVe embeddings in decoding individual word meanings, with no substantial advantage for either model. When decoding sentence-level stimuli, we observed that visually grounded topics were decoded more accurately than abstract and diffuse ones, suggesting that some aspects of concepts play a significant role in decoding success.

An extension of prior work involved training decoders directly on sentence-level fMRI data. We found that contextual embeddings (e.g., BERT), particularly when dimensionally reduced via PCA, improved decoding accuracy relative to static embeddings. This advantage was further supported by voxel-wise brain encoding models, where BERT representations outperformed GloVe in both voxel coverage and predictive power. These results suggest that contextual embeddings like BERT offer a closer alignment with brain representations of sentence meaning compared to static embeddings. The advantage of BERT, implies that sentence-level neural representations are context-dependent.

The inter-subject analysis diverged from previously reported findings in the literature, showing no evidence of generalization under the full-data, averaged-data, or global pairwise transfer paradigms. Nevertheless, the pairwise transfer results revealed significant extreme deviations from chance levels, suggesting the potential existence of universal decoders within specific subgroups of the population. As a consequence, certain groups appear to share common intra-group semantic representations, yielding high decoding accuracy within these clusters, while exhibiting pronounced dissimilarities between groups, which in turn leads to markedly reduced cross-group performance. Future work could extend these findings by leveraging large-scale inter-subject datasets and conducting more detailed statistical analyses to rigorously validate subgroup effects and quantify the reliability of observed cross-group differences.

Our conclusions should be interpreted with caution. The use of PCA alone to reduce and align sentence embeddings across participants may distort individual differences, as it does not ensure a shared representational subspace. Dimension-wise features may be lost due to misalignment between subjects embeddings. Future work should also include alignment techniques to better preserve cross-subject correspondence. Improving vector alignment across participants remains a key direction for pursuing inter-subject decoders and requires deeper engagement with the fMRI literature on cross-subject alignment methods.

Acknowledgments

I thank the teaching assistants Keren Gruteke-Klein and Omer Shubi for their guidance and helpful discussions during this project.

References

Anderson, A. J., Binder, J. R., & Fernandino, L. (2018). An integrated neural decoder of linguistic and experiential meaning. *Journal of Cognitive Neuroscience*, *31*(4), 503–517.

Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Güçlü, U., & van Gerven, M. A. (2017). Modeling the dynamics of human brain activity with recurrent neural networks. *Frontiers in Computational Neuroscience*, *11*, 7. doi: 10.3389/fncom.2017.00007

Guntupalli, J. S., Hanke, M., Halchenko, Y. O., Connolly, A. C., Ramadge, P. J., & Haxby, J. V. (2016). A model of representational spaces in human cortex. *Cerebral Cortex*, *26*(6), 2919–2934. doi: 10.1093/cercor/bhw068

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. doi: 10.1038/nature17637

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

Nichols, T. E., & Holmes, A. P. (2002). Nonparametric permutation tests for functional neuroimaging: A primer with examples. *Human Brain Mapping*, *15*(1), 1–25. doi: 10.1002/hbm.1058

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (emnlp)* (pp. 1532–1543).

Pereira, F., Lou, B., Pritchett, B., Ritter, S., Gershman, S. J., Kanwisher, N., . . . Fedorenko, E. (2018). Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, *9*(1), 1–13.

Appendix

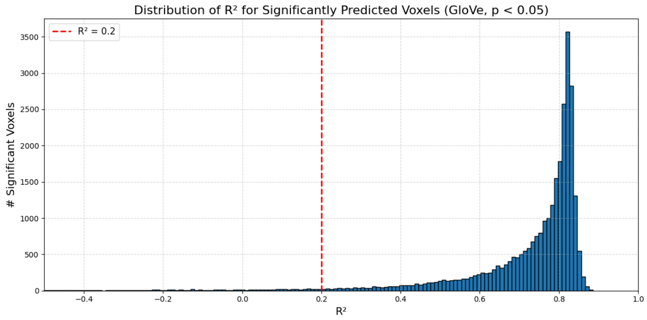


Figure 7: R^2 distribution for the GloVe-based brain encoder model. Voxels shown are significantly predicted at $p < 0.05$ using permutation testing.

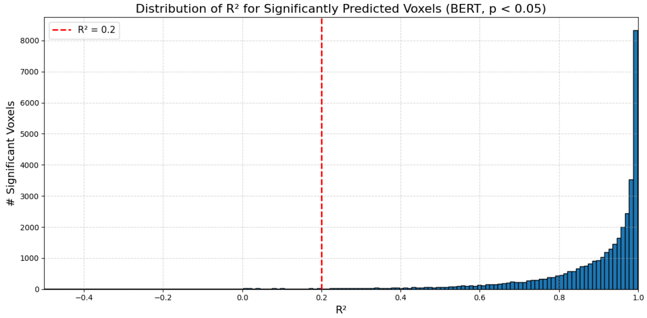


Figure 8: R^2 distribution for the BERT-based brain encoder model. Voxels shown are significantly predicted at $p < 0.05$ using permutation testing.

Table 1: Extreme pairs (|score−90| > 10), retaining the score closest to 90 between symmetric pairs, sorted by absolute deviation from 90.

Train	Test	Score	—Dev—	Direction
M13	M05	108.55	18.55	M13 → M05
M07	M13	103.04	13.04	M07 → M13
M13	M10	101.68	11.68	M13 → M10
M14	M04	78.58	11.42	M14 → M04
M10	M07	100.94	10.94	M10 → M07
M03	M02	79.12	10.88	M03 → M02
M03	M07	79.14	10.86	M03 → M07
M13	M09	100.57	10.57	M13 → M09
M12	M02	100.37	10.37	M12 → M02
M15	M12	100.19	10.19	M15 → M12

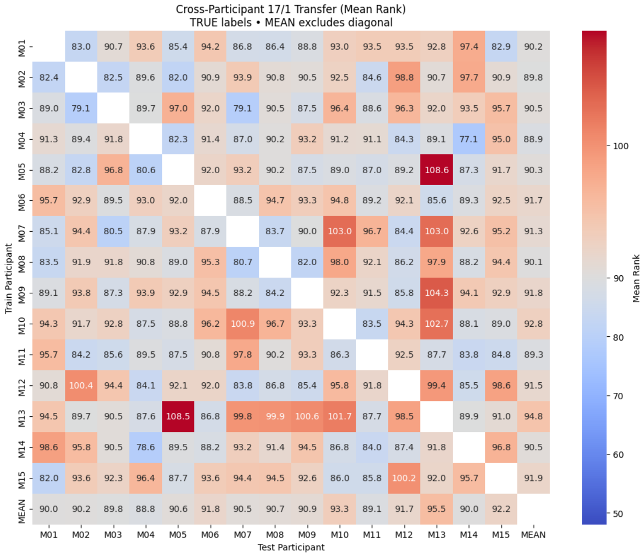


Figure 9: Cross-participant 17/1 transfer decoding results (mean rank accuracy) using true semantic labels. Each cell shows the average rank of the correct sentence when a decoder trained on one participant (rows) is tested on another participant (columns). Lower values indicate better decoding performance. The rightmost column shows the mean across all test participants for each training participant, and the bottom row shows the mean across all training participants for each test participant; means exclude the diagonal (within-participant).

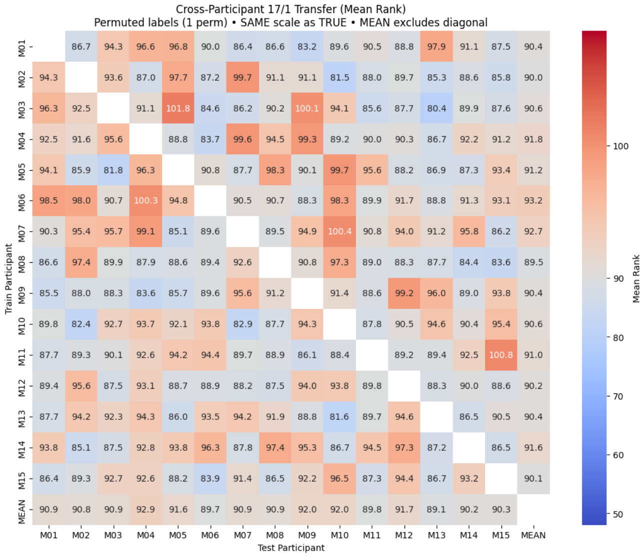


Figure 10: Cross-participant 17/1 transfer decoding results (mean rank accuracy) with permuted semantic labels (single permutation). Each cell shows the average rank of the correct sentence after random shuffling of semantic vectors, using the same color scale as the true-label condition for direct comparison. Lower values indicate better decoding performance. The rightmost column and bottom row show means across test and training participants respectively, excluding within-participant (diagonal) results.

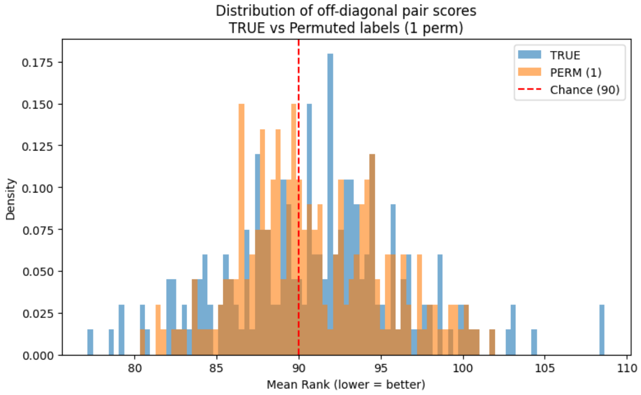


Figure 11: Distribution of mean rank scores for all off-diagonal train–test participant pairs under true semantic labels (blue) and a single permutation of labels (orange). The red dashed line marks the expected chance level (mean rank = 90). While both distributions are centered near chance, the true-label scores show a broader spread, indicating greater variability across participant pairs.