Mapping Brains with Language Models: A Survey

Antonia Karamolegkou[1], **Mostafa Abdou**[2], **Anders Søgaard**[1]
[1]University of Copenhagen, [2] Princeton University
antka@di.ku.dk, ma4231@princeton.edu, soegaard@di.ku.dk

Abstract

Over the years, many researchers have seemingly made the same observation: Brain and language model activations exhibit *some* structural similarities, enabling linear partial mappings between features extracted from neural recordings and computational language models. In an attempt to evaluate how much evidence has been accumulated for this observation, we survey over 30 studies spanning 10 datasets and 8 metrics. How much evidence has been accumulated, and what, if anything, is missing before we can draw conclusions? Our analysis of the evaluation methods used in the literature reveals that some of the metrics are less conservative. We also find that the accumulated evidence, for now, remains ambiguous, but correlations with model size and quality provide grounds for cautious optimism.

1 Introduction

Advances in neuroimaging technologies have made it possible to better approximate the spatiotemporal profile of the computations responsible for language in the brain (Poldrack and Farah, 2015; Avberšek and Repovš, 2022). At the same time, advances in natural language processing have produced language models (LMs) with high performance in many tasks (Min et al., 2021).

This progress has motivated scientists to start using state-of-the-art LMs to study neural activity in the human brain during language processing (Wehbe et al., 2014b; Huth et al., 2016; Schrimpf et al., 2021; Toneva et al., 2022b; Caucheteux and King, 2022). Conversely, it has also prompted NLP researchers to start using neuroimaging data to evaluate and improve their models (Søgaard, 2016; Bingel et al., 2016; Toneva and Wehbe, 2019; Hollenstein et al., 2019; Aw and Toneva, 2023).

At the conceptual core of these studies lies the suggestion that representations extracted from NLP models can (partially) explain the signal found in neural data. These representations can be based on co-occurrence counts (Mitchell et al., 2008; Pereira et al., 2013; Huth et al., 2016) or syntactic and discourse features (Wehbe et al., 2014a,b). Later studies use dense representations such as word embeddings (Anderson et al., 2017; Pereira et al., 2018; Toneva and Wehbe, 2019; Hollenstein et al., 2019) and recurrent neural networks to extract contextual stimuli representations (Qian et al., 2016; Jain and Huth, 2018; Sun et al., 2019). More recently, transformer-based architectures have been shown to align even better with neural activity data (Gauthier and Levy, 2019; Schrimpf et al., 2021; Oota et al., 2022b).

Such work shows that LMs can be trained to induce representations that are seemingly predictive of neural recordings or features thereof. However, pursuing the literature, it quickly becomes clear that these papers all rely on different experimental protocols and different metrics (Minnema and Herbelot, 2019; Hollenstein et al., 2020; Beinborn et al., 2023). So questions are: How much evidence has really been accumulated in support of structural similarities between brains and LMs? And more importantly, what exactly, if anything, drives this alignment, and what are we to understand from it? After gathering all the studies, we examine their evaluation metrics and their interrelationships, providing discussions on the corresponding findings.

Contributions Our study provides four major contributions for the wider NLP audience: (a) a detailed review of the literature on mappings between fMRI/MEG recordings and representations from language models; (b) an overview of the datasets and mapping methods; (c) an analysis of the evaluation setups that have been used to link neural signals with language models and how they relate; (d) a discussion of what drives this representational alignment and what we, as a field, can make of it going forward.

Findings of the Association for Computational Linguistics: ACL 2023, pages 9748–9762
July 9-14, 2023 ©2023 Association for Computational Linguistics

Terminology First, a brief note on terminology: Neural response measurements refer to recordings of the brain activity of subjects reading or listening to language. We focus on (a) functional magnetic resonance imaging (fMRI), which measures neuronal activity via blood oxygenation level-dependent contrast and has a high spatial resolution but poor temporal resolution (3–6s) and (b) magnetoencephalography (MEG), which involves the measurement of the magnetic field generated by the electrical activity of neurons in the cortex, providing a more accurate resolution of the timing of neuronal activity.

Voxels refer to the smallest unit of data in a neuroimage, being the three-dimensional equivalent of a pixel in two-dimensional images (Gerber and Peterson, 2008). Finally, we use brain decoding to refer to predicting stimuli from brain responses (i.e. *reading the brain*). Brain encoding will then refer to predicting brain responses from stimuli. Whereas decoding models serve as a test for the presence of information in neural responses, encoding models can be interpreted as process models constraining brain-computational theories (Kriegeskorte and Douglas, 2019).

2 Datasets

To infer a mapping between language models and brains, researchers rely on datasets in which brain activity is recorded in response to linguistic stimuli. In some studies, the stimuli are single words (Mitchell et al., 2008; Anderson et al., 2017) or sentences displayed on a screen (Pereira et al., 2018). In others, participants read longer stories (Wehbe et al., 2014a; Bhattasali et al., 2020; Nastase et al., 2021) or listened to speech or podcasts (Huth et al., 2016; Antonello et al., 2021). Table 1 lists publicly available datasets that have been used in the context of mapping language models to and from recordings of brain response. Differences between the datasets –the number of participants, the equipment, the experimental setup, pre-processing steps, and probabilistic corrections – should lead us to expect some variation in what researchers have concluded (Hollenstein et al., 2020).

3 How to predict brain activity?

In this section, we survey work in which neural responses are predicted from linguistic representations. Such work typically aims to shed light on how language functions in the brain. One of the earliest studies exploring the mapping between brain and language representations is by Mitchell et al. (2008), who trained a linear regression model on a set of word representations extracted from 60 nouns using 115 semantic features based on co-occurrence statistics, to predict the corresponding fMRI representations of the same nouns. They use pair-wise matching accuracy evaluation, extracting two words w and w' for evaluation, and showed that the predicted fMRI for a word w was closer to the real fMRI image for w than to the real fMRI image for w', at above-chance levels. Mitchell et al. (2008) also report percentile rank results, ranking predicted fMRI images by similarity with the real image of w. We discuss how the metrics relate in §6.

The dataset of Mitchell et al. (2008)is also used by Murphy et al. (2012), who extract linguistic features from part-of-speech taggers, stemmers, and dependency parsers, showing that dependency parsers are the most successful in predicting brain activity. They also use leave-2-out pair-matching as their performance metric.

Later on, Wehbe et al. (2014a) moved on to predicting brain activation patterns for entire sentences rather than for isolated words. They recorded fMRI neural response measurements while participants read a chapter from *Harry Potter and the Sorcerer's Stone*, then extracted a set of 195 features for each word (ranging from semantic, syntactic properties to visual and discourse-level features) to train a comprehensive generative model that would then predict the time series of the fMRI activity observed when the participants read that passage. Leave-2-out pair-matching accuracy is used for evaluation.

Huth et al. (2016), in contrast, use fMRI recordings of participants listening to spoken narrative stories, representing each word in the corpus as a 985-dimensional vector encoding semantic information driven by co-occurrence statistics. They train per-voxel linear regression models and evaluate their predicted per-word fMRI images by their per-voxel Pearson correlation with the real fMRI images, showing that 3-4 dimensions explained a significant amount of variance in the FMRI data.

Wehbe et al. (2014b) are among the first to use neural language models, using *recurrent* models to compute contextualized embeddings, hidden state vectors of previous words, and word probabilities. They run their experiments of MEG recordings

	Data	Authors	Method	N subjects
1	60 Nouns	Mitchell et al. (2008)	fMRI	9
2	Harry Potter Dataset	Wehbe et al. (2014a)	fMRI	8
3	Harry Potter Dataset	Wehbe et al. (2014b)	MEG	8
4	The Moth Radio Hour Dataset	Huth et al. (2016)	fMRI	7
5	Pereira Dataset	Pereira et al. (2018)	fMRI	16
6	Mother of all Unification Studies	Schoffelen et al. (2019)	fMRI, MEG	204
7	Natural Stories Audio Dataset	Zhang et al. (2020)	fMRI	19
8	Narratives	Nastase et al. (2021)	fMRI	345
9	Podcast dataset	Antonello et al. (2021)	fMRI	5
10	The Little Prince Datasets	Li et al. (2022)	fMRI	112

Table 1: The opensource datasets that were used in the studies surveyed. Numbering used in Table 2.

of participants reading Harry Potter, obtained in a follow-up study to Wehbe et al. (2014a). From the three sets of representations, they then train linear regression models to predict the MEG vectors corresponding to each word, and the regression models are then evaluated by computing pair-matching accuracy.

Similarly, Søgaard (2016) evaluates static word embeddings on the data from Wehbe et al. (2014a), learning linear transformation from word embeddings into an fMRI vector space. The predictions are evaluated through mean squared error (MSE).

Jain and Huth (2018) evaluate recurrent language models against the fMRI dataset from Huth et al. (2016). Their findings show that contextual language model representations align significantly better (to brain response) compared to static word embedding models. Their evaluation metric is the total sum of explained variance[1]

Following this, Schwartz et al. (2019) use attention-based transformer language models for brain mapping. They finetune BERT (Devlin et al., 2019)to predict neural response measurements from the Harry Potter dataset, showing that the fine-tuned models have representations that encode more brain-activity-relevant language information than the non-finetuned models. They rely on pair-matching accuracy as their performance metric.

As in Søgaard (2016), Zhang et al. (2020) map static word embeddings into the vector space of the neural response measurements (fMRI). They introduce a new dataset of such measurements from subjects listening to natural stories. They rely on explained variance as their performance metric.

Toneva and Wehbe (2019) evaluate word and sequence embeddings from 4 recurrent and attention-based transformer language models, using the Harry Potter fMRI dataset. They evaluate models across layers, context lengths, and attention types, using pairwise matching accuracy as their performance metric. In a later study, Toneva et al. (2022a) induce compositional semantic representations of "supra-word meaning" which they then use to predict neural responses across regions of interest, evaluating their models using Pearson correlation.

Also using the Harry Potter data, Abnar et al. (2019) evaluate five models, one static and four contextualized, relying on a variant of representational similarity analysis (Kriegeskorte et al., 2008). The results suggest that models provide representations of local contexts that are well-aligned to neural measurements. However, as information from further away context is integrated by the models, representations become less aligned to neural measurements.

In a large-scale study, Schrimpf et al. (2021) examine the relationships between 43 diverse state-of-the-art neural network models (including embedding models, recurrent models, and transformers) across three datasets (two fMRI, one electrocardiography). They rely on a metric they term Brain Score which involves normalising the Pearson correlation by a noise ceiling. Their results show that transformer-based models perform better than recurrent or static models, and larger models perform better than smaller ones.

Similarly, in Caucheteux and King (2022), the Schoffelen et al. (2019) fMRI and MEG datasets are used to compare a variety of transformer architectures. They study how architectural details, training settings, and the linguistic performance of these models independently account for the generation of brain correspondent representations. The results suggest that the better language models are at predicting words from context, the better their

[1] The squared Pearson correlation coefficient. We will not distinguish between studies using Pearson correlation and studies using explained variance. See Appendix A.2.

activations linearly map onto those of the brain.

Antonello et al. (2021) evaluate three static and five attention-based transformer models, in combination with four fine-tuning tasks and two machine translation models. They train linear regression models to evaluate their word-level representations against a new fMRI dataset from participants listening to podcast stories. They find a low-dimensional structure in language representations that can predict brain responses. In a similar setting, Antonello and Huth (2022) examine why some features fit the brain data better arguing that the reason is that they capture various linguistic phenomena.

Reddy and Wehbe (2021) evaluate syntactic features in conjunction with BERT representations, finding that syntax explains additional variance in brain activity in various parts of the language system, even while controlling for complexity metrics that capture processing load.

In a series of studies Caucheteux et al. (2021, 2022b,a) investigate GPT2's activations in predicting brain signals using the Nastase et al. (2021) dataset. Their evaluation metric is Brain Score (Schrimpf et al., 2018). To determine which factors affect the brain encoding Pasquiou et al. (2022) examine the impact of test loss, training corpus, model architecture, and fine-tuning in various models using the Li et al. (2022) dataset. They evaluate model performance using Pearson Correlation.

Oota et al. (2022a) study the impact of context size in language models on how they align with neural response measurements. They use the Nastase et al. (2021) dataset and evaluate recurrent and attention-based transformer architectures. In a later study, Oota et al. (2022b) use the Pereira et al. (2018) dataset and evaluate BERT-base models (fine-tuned for various NLP tasks). They showed that neural response predictions from ridge regression with BERT-base models fine-tuned for coreference resolution, NER, and shallow syntactic parsing explained more variance for Pereira et al. (2018) response measurements. On the other hand, tasks such as paraphrase generation, summarization, and natural language inference led to better encoding performance for the Nastase et al. (2021) data (audio). Using the same dataset, in Oota et al. (2022c) it is shown that the presence of surface, syntactic, and semantic linguistic information is crucial for the alignment across all layers of the language model. They use pairwise matching accuracy and/or Pearson correlation as their performance metrics in these studies.

Aw and Toneva (2023) extract feature representations from four attention-based transformer models. They evaluate the impact of fine-tuning on the BookSum dataset (Kryscinski et al., 2021). All models are used to predict brain activity on the Harry Potter data. Pairwise matching accuracy and Pearson correlation are their performance metrics. Merlin and Toneva (2022) focus more narrowly on variants of GPT-2, showing that improvements in alignment with brain recordings are probably not because of the next-word prediction task or word-level semantics, but due to multi-word semantics. Their reported metric is Pearson correlation.

Intermediate summary The above studies differ in many respects. Several metrics are used: pairwise-matching accuracy,[2] Pearson correlation (or Brain Score), mean squared error, and representational similarity analysis. Even studies that report the same performance metrics are not directly comparable because they often report on results on different datasets and use slightly different protocols, e.g., Murphy et al. (2012) and Wehbe et al. (2014b). Beinborn et al. (2023) compare various encoding experiments and receive very diverse results for different evaluation metrics. The diversity of metrics and data renders a direct comparison difficult. To remedy this, we consider how the metrics compare in §6.

4 How to predict linguistic stimuli?

Decoding models work in the other direction and aim to predict linguistic features of the stimuli from recordings of brain response. Pereira et al. (2018) introduce a decoder that predicts stimuli representation of semantic features given fMRI data.They introduce a novel dataset of neural responses aligned with annotation of concrete and abstract semantic categories (such as pleasure, ignorance, cooking etc.). They evaluate static word embeddings by applying ridge regression to predict per-word fMRI vectors. A separate regression model is trained per dimension, allowing for dimension-wise regularization. The model is evaluated in terms of pairwise matching accuracy, but also in terms of percentile rank, adapted to the decoding scenario.

[2] Some papers (Wehbe et al., 2014b; Toneva and Wehbe, 2019; Aw and Toneva, 2023) use a variant of pairwise-matching accuracy, in which the model has to discriminate between two averages of 20 random predicted neural response measurements. We do not distinguish between the two variants.

Gauthier and Levy (2019) also train linear regression models which map from the response measurements in Pereira et al. (2018), but to representations of the same sentences produced by the BERT language model finetuned on different natural language understanding tasks. The regression models are evaluated using two metrics: mean squared error and average percentile rank. Their results show that fine-tuning with different NLU objectives leads to worse alignment and that, somewhat surprisingly, the only objective which does lead to better alignment is a scrambled language modeling task where the model is trained to predict scrambled sentences.

Minnema and Herbelot (2019) re-examine the work of Pereira et al. (2018) using various metrics (pairwise matching accuracy, percentile rank, cosine distance, R^2, RSA), comparing decoder models (ridge regression, perceptron, and convolutional neural networks).[3] They show that positive results *are only obtained using pairwise matching accuracy*.

Abdou et al. (2021) investigate whether aligning language models with brain recordings can be improved by biasing their attention with annotations from syntactic or semantic formalisms. They fine-tune the BERT models using several syntacto-semantic formalisms and evaluate their alignment with brain activity measurements from the Wehbe et al. (2014a) and Pereira et al. (2018) datasets. Their results – obtained using Pearson correlation as performance metric – are positive for two in three formalisms.

Zou et al. (2022) propose a new evaluation method for decoding, a so-called *cross-modal cloze task*. They generate the data for the task from the neural response measures in Mitchell et al. (2008) and Wehbe et al. (2014a). The task itself amounts to a cloze task in which the context is prefixed by the fMRI image of the masked word. They evaluate models using precision@k. Note how this task is considerably easier than linearly mapping from language model representations into fMRI images, and precision@k results therefore cannot be compared to those obtained in other settings. Their best precision@1 scores are around 0.3, but only marginally (0.03) better than a unimodal LM.

Finally, Pascual et al. (2022) try a more realistic setup by predicting language from fMRI scans of subjects not included in the training. They use the (Pereira et al., 2018) dataset and evaluate the regression models based on pairwise accuracy and precision@k (or top-k accuracy). They propose evaluating with direct classification as a more demanding setup to evaluate and understand current brain decoding models.

Intermediate summary Decoding studies also differ in many respects. Several metrics are used: pairwise-matching accuracy, Pearson correlation, percentile rank, cosine distance, precision@k, and representational similarity analysis; and several datasets are used. Gauthier and Ivanova (2018) criticize the evaluation techniques of decoding studies and suggest adopting task and mechanism explicit models. It is of particular interest to our study that both Minnema and Herbelot (2019) only report positive results for pairwise matching accuracy compared to other metrics. This suggests pairwise matching accuracy is a less conservative metric (and maybe less reliable).

5 Performance Metrics

We present the evaluation metrics used in the above studies and discuss how they relate. See Table 2 for a summary of metrics and corresponding studies.

Mitchell et al. (2008) introduce **pairwise matching accuracy**. Because of their small sample size, they use a leave-2-out cross-validation, which later work also adopted. The metric is a binary classification accuracy metric on a balanced dataset, so a random baseline converges toward 0.5. Many studies have relied on this metric, both in encoding and decoding (see Table 2).[4]

Pearson correlation Pearson correlation is another widely used metric in the studies surveyed above, measuring the linear relationship between variables, and providing insight into the strength and direction of their association. Huth et al. (2016), compute Pearson correlation between predicted and actual brain responses using Gaussian random vectors to test statistical significance. Resulting p-values are corrected for multiple comparisons within each subject using false discovery rate (FDR) (Benjamini and Hochberg, 1995). Others have used Bonferroni correction (Huth et al., 2016) or block-wise permutation test (Adolf et al., 2014) to evaluate the statistical significance of the

[3] Only the former two are linear and relevant for this meta-study.

[4] The method is often referred to as 2v2 Accuracy. The variant that averages across 20 images, is then referred to as 20v20 Accuracy.

Authors	Data	E/D	Acc	P/B	Rank	MSE	RSA	Cos Sim	P@k
Mitchell et al. (2008)	1	E	✓		✓			✓	
Murphy et al. (2012)	1	E	✓						
Wehbe et al. (2014a,b)	2,3	E	✓	✓					
Huth et al. (2016)	4	E		✓					
Søgaard (2016)	2	E				✓			
Pereira et al. (2018)	5	D	✓		✓				
Jain and Huth (2018)	4	E		✓					
Toneva and Wehbe (2019)	2	E	✓						
Gauthier and Levy (2019)	5	D			✓	✓	✓		
Minnema and Herbelot (2019)	5	D	✓		✓		✓	✓	
Schwartz et al. (2019)	3	E	✓						
Abnar et al. (2019)	2	E					✓		
Zhang et al. (2020)	7	E	✓	✓					
Schrimpf et al. (2021)	5	E	✓	✓					
Abdou et al. (2021)	2,5	D		✓					
Reddy and Wehbe (2021)	2	E		✓					
Antonello et al. (2021)	9	E		✓					
Antonello and Huth (2022)	9	E		✓					
Caucheteux et al. (2021, 2022b,a)	8	E		✓					
Caucheteux and King (2022)	6	E		✓					
Zou et al. (2022)	1,2	D							✓
Oota et al. (2022a,b,c)	8	E	✓	✓					
Pasquiou et al. (2022)	10	E	✓						
Toneva et al. (2022a)	2	E		✓					
Merlin and Toneva (2022)	2	E		✓					
Pascual et al. (2022)	5	D	✓						✓
Aw and Toneva (2023)	2	E	✓	✓					

Table 2: Overview of what studies rely on what data and what performance metrics. See Table 1 for dataset numbering. **E/D:** Encoding or decoding model **Acc:** Pairwise matching accuracy. **P/B:** Pearson correlation or BrainScore. **Rank:** percentile rank. **MSE:** mean squared error. **RSA:** representational similarity analysis. **CosSim:** cosine similarity. **P@k:** precision@k. Schrimpf et al. (2021) used a non-public fMRI dataset, too.

correlation (Zhang et al., 2020). Some report R^2 (explained variance) instead of or in addition to correlation coefficients (Minnema and Herbelot, 2019; Reddy and Wehbe, 2021). Others have adopted a more elaborate extension of Pearson correlation, namely BrainScore (Schrimpf et al., 2018). BrainScore is estimated on held-out test data, calculating Pearson's correlation between model predictions and neural recordings divided by the estimated ceiling and averaged across voxels and participants.

Percentile rank was first used for encoding (Mitchell et al., 2008), but can also be used for decoding (Pereira et al., 2018; Gauthier and Levy, 2019; Minnema and Herbelot, 2019). In encoding, the predicted brain image for w is ranked along the predicted images for a set of candidate words w' by their similarity to the real (ground truth) image for w. The average rank is then reported. For decoding, they rank word vectors rather than neural response images. Note the similarity metric is unspecified, but typically cosine distance is used.

Mean squared error, the average of the squared differences between word vectors and neural responses, was first used for encoding in Søgaard (2016) on a held-out test split. It was also used by Gauthier and Levy (2019).

Representational similarity analysis (RSA) was introduced in Kriegeskorte et al. (2008) as a non-parametric way to characterize structural alignment between the geometries of representations derived from disparate modalities. RSA abstracts away from activity patterns themselves and instead computes representational similarity matrices (RSMs), which characterize the information carried by a given representation method through global similarity structure. A rank correlation coefficient is computed between RSMs derived from the two spaces, providing a summary statistic indicative of the overall representational alignment between them. Being non-parametric, RSA circumvents many of the various methodological weaknesses (such as over fitting, etc.). Gauthier and Levy (2019), Minnema and Herbelot (2019), and Abnar et al. (2019) apply (variations of) RSA to investigate the relations between different model components, and then to study the alignment of these components with brain response.

Cosine similarity was used in Mitchell et al.

(2008) to select between the candidate images in pairwise matching accuracy, as well as in percentile rank and RSA, but the raw cosine similarities between predicted and real images or embeddings can also be used as a metric. Minnema and Herbelot (2019) use this metric to quantify how close the predicted word vectors are to the target. Finally, Zou et al. (2022) use **precision@k**, a standard metric in other mapping problems, e.g., cross-lingual word embeddings (Søgaard et al., 2019).

Comparisons Most metrics are used to evaluate both encoding and decoding models (pairwise matching accuracy, Pearson correlation, percentile rank, MSE, RSA, cosine distance). Results for two of the most widely used metrics – pairwise matching accuracy[5] and percentile rank – tend to be around 0.7–0.8 with generally better results for more recent architectures and larger LMs. To draw conclusions across studies relying on different metrics, we need to investigate which metrics are more conservative, and how different metrics relate.

Pairwise matching accuracy vs. Pearson correlation It seems that pairwise matching accuracy tends to increase monotonically with Pearson correlation. Consider three sets of distances over corresponding point sets, A, B, and C. If A and B are more strongly linearly correlated than A and C, under an optimal linear mapping Ω (minimizing point-wise squared error distance), $\mathbb{E}[(a - b\Omega)^2] > \mathbb{E}[(a - c\Omega)^2]$. Even in this conservative setting in our synthetic experiments in Appendix A.1, the correlation between matching accuracy and percentile rank was very high, ~0.9.

Pairwise matching accuracy vs. percentile rank Both metrics have random baseline scores of 0.5, and they will converge in the limit. If a has a percentile rank of p in a list $\mathcal{A}$, it will be higher than a random member of $\mathcal{A}$ p percent of the time. In our experiments in Appendix A.1, the correlation converges toward 1.0, with values consistently higher than 0.8 for $N = 100$.

Pairwise matching accuracy vs. precision@k are also positively correlated. Perfect score in one entails perfect score in the other, but precision@k can of course be very small for very high values of pairwise matching accuracy (especially if the set of candidate words is big). Conversely, we can have saturation for high values of k, because matching accuracies higher than $\frac{n-k}{n}$ will mean near-perfect precision@k scores. In practice, precision@k (for low values of k) will be much more conservative, however. The correlation coefficient for $N = 100$ (see Appendix A.1) tends to lie around 0.7.

Relative strength Pairwise Matching Accuracy is a relatively permissible performance metric. To see this, consider the scenario in which all target words can be divided into two equal-sized buckets based on word length (number of characters). Say the neural responses capture nothing but this binary distinction between long and short words, but do so perfectly. Moreover, our mapping method, e.g., linear regression, learns this from training data. Now, from this alone, the pairwise matching accuracy will converge toward $\mu = 0.75$, since our model will do perfectly (1.0) on half of the data, and exhibit random performance (0.5) on the other half. If the neural responses tracked word length (and not just the distinction between short and long words), performance would be even better. In other words, Pairwise Matching Accuracy scores around 0.7-0.8 (observed in the studies above) may only reflect very shallow processing characteristics. The fact that Minnema and Herbelot (2019) only observed good results with this metric, led them to adopt a rather critical stance, for good reasons.

Other metrics are clearly more conservative. For a set of n candidate words, a random mapping will induce a precision@1-score of $\frac{1}{n}$. While hubs may inflate scores for larger values, the metric is extremely conservative for small values of k. However, only Zou et al. (2022) use this metric, and they modify the experimental protocol substantially, making the task much easier by providing additional input to a non-linear model. The small improvement from adding neural response input is interesting, but could potentially be explained by shallow processing characteristics.

They argue that analogy testing would provide a better evaluation protocol:

> one would ideally use standard metrics such as semantic relatedness judgment tasks, analogy tasks, etc. [but] this is not possible due to the limited vocabulary sizes of the available brain datasets

Such evaluation *is* possible on small scale, though, and increasingly larger fMRI datasets are becoming available (see above). Zhang et al. (2020)

[5] When discriminating averages over 20 images (Wehbe et al., 2014b), scores are naturally lower.

have identified analogical reasoning in fMRI brain activation spaces. The analogies are computed using vector offset and probe the systematicity of how semantic relations are encoded. If a model encodes the capital-of relation systematically, we can retrieve the capital of Germany by subtracting the fMRI vector for 'Paris' from the sum of our the fMRI vectors for Germany and France. This is the same kind of analogical reasoning found in language models (Mikolov et al., 2013). Garneau et al. (2021) show that the more language models satisfy analogies, the more isomorphic they are.

So far, it seems that, with the possible exception of Zhang et al. (2020), there is little evidence for structural similarities, beyond what could be induced by shallow processing characteristics, but what about all the studies that report strong Pearson correlations? Per-voxel correlation coefficients are low on average, but across the above studies, typically only around 4-40% of the voxels exhibit significant correlations (Huth et al., 2016; Caucheteux and King, 2022). Since these correlations have been replicated across different datasets, they are generally not disputed, but could still reflect rather shallow processing characteristics.

On a more positive note, several studies show that larger (and better) language models align better with neural response measurements (Schrimpf et al., 2021; Caucheteux and King, 2022). This suggests that language models in the future may align even better with such measurements, possibly reflecting properties of deep processing. Such correlations with model quality and size are positive, making the results reported above more credible.

Generally, the conclusions we can draw from the above studies are somewhat vague. There are two reasons for this: (i) Past studies have relied on permissible (pairwise matching accuracy) and ambiguous (Pearson correlation) performance metrics; and (ii) past studies have relied on small-sized datasets. We believe that *this calls for a meta-analysis of the above studies*. To provide grounds for such a meta-analysis, we have in this section taken steps to compare the metrics used in these studies. We leave it for future work to explore various ways effect sizes can be computed across these studies.

6 Discussion

Many studies, summarized above, aim to compare language model representations with neural response measurements using linear mapping models. Our main reason to focus on linear mapping models is that they quantify the degree of structural similarity (isomorphism). Overall, results suggest that structural similarities between language models and neural responses exist. Furthermore, there is good evidence that alignment has correlated positively with model quality and model size, suggesting a certain level of convergence as language models improve.

What drives alignment? Is alignment driven by deep processing characteristics or by shallow textual characteristics? Classical candidates for shallow ones would be word length, frequency, regularity, and part of speech. Mitchell et al. (2008), for example, only controlled for part of speech. Some authors have presented results to suggest that alignments are driven by syntactic or semantic factors (Abdou et al., 2021; Reddy and Wehbe, 2021; Caucheteux et al., 2021; Zhang et al., 2020), whereas others have claimed some similarities reflect semantic phenomena (Huth et al., 2016; Caucheteux et al., 2021). Others suggest that alignments reflect deeper similarities between model objectives and predictive processing in human brains (Schrimpf et al., 2018; Caucheteux et al., 2022a; Goldstein et al., 2021), but see Antonello and Huth (2022) for a critical discussion of such work.

Linguistically-transparent models that allow for a principled decomposition of a model's components into smaller linguistically meaningful units and models that move towards possible neurobiological implementations of neural computation are likely to be key for answering this question (Hale et al., 2022; Ten Oever et al., 2022). Given the plethora of interpretability methods recently developed, however, we believe that even models which are not intrinsically interpretable can be useful toward this goal.

Do some models align better? Most studies observe that better and larger, contextual models align better with neural responses (Jain and Huth, 2018; Caucheteux and King, 2022). Other improvements include fine-tuning on specific tasks (Oota et al., 2022b; Aw and Toneva, 2023). Pasquiou et al. (2022) outline the impact of model training choices.

What metrics? The inconsistent use of performance metrics makes it hard to compare and interpret the results reported in the literature (Beinborn et al., 2023). We have shown that some metrics are perhaps *too* permissible to detect structural similarities between language models and neural responses. We have argued that precision@k is more conservative than most other metrics. Minnema and Herbelot (2019) have proposed using analogy scores. In the limit (given sufficient analogies), perfect analogical accuracy implies isomorphism (Garneau et al., 2021). So do perfect precision@1 and perfect RSA scores. We, therefore, propose giving priority to these performance metrics, not to conflate shallow processing characteristics with deeper, more semantic properties.

Meta-analysis? Proper meta-analysis is currently hindered by the use of different metrics, but we have taken steps to relate these.

7 Conclusions

We surveyed work on linear mappings between neural response measurements and language model representations, with a focus on metrics. In particular, we surveyed a broad range of 30 studies spanning across 10 datasets and 8 metrics. By examining the metrics, and relating them to one another, we attempt to critically assess the accumulated evidence for structural similarity between neural responses and language model representations. We find that similarities with existing models are limited to moderate, and there is a possibility they might be explained by shallow processing characteristics since there is no standardised methodology for employing controls, but also that positive correlations with model quality and size suggest that language models may exhibit deeper similarities with neural responses in years to come.

Limitations

This work focuses on a specific view of the whole neuro-computational modeling field. We exclude specific angles of research such as non-linear models (Ruan et al., 2016; Qian et al., 2016; Bingel et al., 2016; Anderson et al., 2017; Oota et al., 2018) since we want to evaluate the accumulated evidence for structural similarity (isomorphism) between neural responses and language models. (Ivanova et al., 2022) mention several advantages of using linear mapping models, they are more interpretable and more biologically plausible. They also provide an insightful discussion on mapping model choice, emphasizing the importance of estimating models' complexity over categorizing them as purely linear or nonlinear.

Another limitation is that we do not include speech models (Vaidya et al., 2022; Défossez et al., 2022; Millet et al., 2022) that have been used to map brain representations mostly due to coherency and page-limit restrictions. The survey is also limited to fMRI and MEG data instead of other modalities for two many reasons: (i) fMRI and MEG are used as a combination in many studies (Caucheteux and King, 2022; Schrimpf et al., 2021; Toneva et al., 2022a), and (ii) they offer high spatial resolution and signal reliability (fMRI) and better temporal and spatial resolution (MEG), making them suitable for NLP (Hollenstein et al., 2020). For a survey in encoding and decoding models in cognitive electrophysiology, see Holdgraf et al. (2017).

Ethics Statement

The use of publicly available data in this survey ensures compliance with ethical guidelines while acknowledging the initial consent provided by the participants for data capture and sharing. Participants' consent is a crucial ethical consideration in the collection and sharing of fMRI and MEG data, and the preservation of legal and ethical rights should always be prioritized. By upholding ethical principles, researchers can responsibly contribute to the field of brain encoding and decoding, advancing our understanding of neural processes without compromising individual rights and privacy. Researchers should ensure secure storage, anonymization, and limited access to sensitive neuroimaging data, adhering to data protection regulations and guidelines.

Furthermore, it is essential to prioritize the dissemination of research findings in a responsible manner, with clear and accurate communication that respects the limits and uncertainties of scientific knowledge. Openness and transparency in reporting methods, results, and interpretations contribute to the overall integrity of the research field. Additionally, fostering a culture of collaboration, respect, and acknowledgment of the contributions of participants, colleagues, and the wider scientific community promotes ethical conduct and responsible research practices in brain encoding and decoding. By adhering to these ethical principles, researchers not only advance scientific knowledge but also build public trust, enhance the societal impact of their work, and ensure the long-term sustainability and progress of the field.

Acknowledgements

This work is supported by the Novo Nordisk Foundation. Antonia Karamolegkou was supported by the Onassis Foundation - Scholarship ID: F ZP 017-2/2022-2023'.

References

Mostafa Abdou, Ana Valeria González, Mariya Toneva, Daniel Hershcovich, and Anders Søgaard. 2021. Does injecting linguistic structure into language models lead to better alignment with brain recordings? *ArXiv*, abs/2101.12608.

Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. 2019. 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*, pages 191–203, Florence, Italy. Association for Computational Linguistics.

Daniela Adolf, Snezhana Weston, Sebastian Baecke, Michael Luchtmann, Johannes Bernarding, and Siegfried Kropf. 2014. Increasing the reliability of data analysis of functional magnetic resonance imaging by applying a new blockwise permutation method. *Front. Neuroinform.*, 8:72.

Andrew J. Anderson, Douwe Kiela, Stephen Clark, and Massimo Poesio. 2017. Visually grounded and textual semantic models differentially decode brain activity associated with concrete and abstract nouns. *Transactions of the Association for Computational Linguistics*, 5:17–30.

Richard Antonello and Alexander Huth. 2022. Predictive Coding or Just Feature Discovery? An Alternative Account of Why Language Models Fit Brain Data. *Neurobiology of Language*, pages 1–16.

Richard Antonello, Javier S. Turek, Vy Ai Vo, and Alexander Huth. 2021. Low-dimensional structure in the space of language representations is reflected in brain responses. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 8332–8344.

Lev Kiar Avberšek and Grega Repovš. 2022. Deep learning in neuroimaging data analysis: Applications, challenges, and solutions.

Khai Loong Aw and Mariya Toneva. 2023. Training language models to summarize narratives improves brain alignment. In *The Eleventh International Conference on Learning Representations*.

Lisa Beinborn, Samira Abnar, and Rochelle Choenni. 2023. Robust evaluation of language-brain encoding experiments. In *Gelbukh, A. (eds) Computational Linguistics and Intelligent Text Processing. CICLing 2019.*, volume 13451 of *Lecture Notes in Computer Science*. Springer, Cham.

Yoav Benjamini and Yosef Hochberg. 1995. Controlling the false discovery rate - a practical and powerful approach to multiple testing. *J. Royal Statist. Soc., Series B*, 57:289 – 300.

Shohini Bhattasali, Jonathan R. Brennan, Wen-Ming Luh, Berta Franzluebbers, and John T. Hale. 2020. "the alice dataset: fmri dataset to study natural language comprehension in the brain".

Joachim Bingel, Maria Barrett, and Anders Søgaard. 2016. Extracting token-level signals of syntactic processing from fMRI - with an application to PoS induction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 747–755, Berlin, Germany. Association for Computational Linguistics.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Remi King. 2021. Disentangling syntax and semantics in the brain with deep networks. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1336–1348. PMLR.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. 2022a. Long-range and hierarchical language predictions in brains and algorithms. *Nature Human Behaviour*, abs/2111.14232.

Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. 2022b. Deep language algorithms predict semantic comprehension from brain activity. *Nature Scientific Reports*, 12.

Charlotte Caucheteux and Jean-Rémi King. 2022. Brains and algorithms partially converge in natural language processing. *Communications Biology*, 5:134.

Alexandre Défossez, Charlotte Caucheteux, Jeremy Rapin, Ori Kabeli, and Jean-Rémi King. 2022. Decoding speech from non-invasive brain recordings. Working paper or preprint.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Nicolas Garneau, Mareike Hartmann, Anders Sandholm, Sebastian Ruder, Ivan Vulic, and Anders Søgaard. 2021. Analogy training multilingual encoders. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(14):12884–12892.

Jon Gauthier and Anna A. Ivanova. 2018. Does the brain represent words? an evaluation of brain decoding studies of language understanding. *ArXiv*, abs/1806.00591.

Jon Gauthier and Roger Levy. 2019. Linking artificial and human neural representations of language. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 529–539, Hong Kong, China. Association for Computational Linguistics.

Andrew J Gerber and Bradley S Peterson. 2008. What is an image? *Journal of the American Academy of Child and Adolescent Psychiatry*, 47(3):245–248.

Ariel Goldstein, Zaid Zada, Eliav Buchnik, Mariano Schain, Amy Price, Bobbi Aubrey, Samuel A. Nastase, Amir Feder, Dotan Emanuel, Alon Cohen, Aren Jansen, Harshvardhan Gazula, Gina Choe, Aditi Rao, Se Catherine Kim, Colton Casto, Lora Fanda, Werner Doyle, Daniel Friedman, Patricia Dugan, Lucia Melloni, Roi Reichart, Sasha Devore, Adeen Flinker, Liat Hasenfratz, Omer Levy, Avinatan Hassidim, Michael Brenner, Yossi Matias, Kenneth A. Norman, Orrin Devinsky, and Uri Hasson. 2021. Thinking ahead: spontaneous prediction in context as a keystone of language in humans and machines. *bioRxiv*.

John T. Hale, Luca Campanelli, Jixing Li, Shohini Bhattasali, Christophe Pallier, and Jonathan R. Brennan. 2022. Neurocomputational models of language processing. *Annual Review of Linguistics*, 8(1):427–446.

Christopher R Holdgraf, Jochem W Rieger, Cristiano Micheli, Stephanie Martin, Robert T Knight, and Frederic E Theunissen. 2017. Encoding and decoding models in cognitive electrophysiology. *Frontiers in Systems Neuroscience*, 11:61.

Nora Hollenstein, Maria Barrett, and Lisa Beinborn. 2020. Towards best practices for leveraging human language processing signals for natural language processing. In *Proceedings of the Second Workshop on Linguistic and Neurocognitive Resources*, pages 15–27, Marseille, France. European Language Resources Association.

Nora Hollenstein, Antonio de la Torre, Nicolas Langer, and Ce Zhang. 2019. CogniVal: A framework for cognitive word embedding evaluation. In *Proceedings of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 538–549, Hong Kong, China. Association for Computational Linguistics.

Alexander G Huth, Wendy A de Heer, Thomas L Griffiths, Frédéric E Theunissen, and Jack L Gallant. 2016. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, 532(7600):453–458.

Anna A Ivanova, Martin Schrimpf, Stefano Anzellotti, Noga Zaslavsky, Evelina Fedorenko, and Leyla Isik. 2022. Beyond linear regression: mapping models in cognitive neuroscience should align with research goals. *Neurons, Behavior, Data analysis, and Theory*, 1.

Shailee Jain and Alexander Huth. 2018. Incorporating context into language encoding models for fmri. In *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc.

Nikolaus Kriegeskorte and Pamela Douglas. 2019. Interpreting encoding and decoding models. *Current opinion in neurobiology, 55*, page 67–179.

Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. 2008. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2:4.

Wojciech Kryscinski, Nazneen Rajani, Divyansh Agarwal, Caiming Xiong, and Dragomir R. Radev. 2021. Booksum: A collection of datasets for long-form narrative summarization. *ArXiv*, abs/2105.08209.

Jixing Li, Shohini Bhattasali, Shulin Zhang, Berta Franzluebbers, Wen-Ming Luh, R Nathan Spreng, Jonathan R Brennan, Yiming Yang, Christophe Pallier, and John Hale. 2022. Le petit prince multilingual naturalistic fMRI corpus. *Scientific Data*, 9(1):530.

G. Merlin and M. Toneva. 2022. Language models and brain alignment: beyond word-level semantics and prediction. *arXiv*.

Tomas Mikolov, Scott Wen-tau Yih, and Geoffrey Zweig. 2013. Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT-2013)*. Association for Computational Linguistics.

Juliette Millet, Charlotte Caucheteux, Pierre Orhan, Yves Boubenec, Alexandre Gramfort, Ewan Dunbar, Christophe Pallier, and Jean-Remi King. 2022. Toward a realistic model of speech processing in the brain with self-supervised learning. In *Advances in Neural Information Processing Systems*.

Bonan Min, Hayley Ross, Elior Sulem, Amir Pouran Ben Veyseh, Thien Huu Nguyen, Oscar Sainz, Eneko Agirre, Ilana Heinz, and Dan Roth. 2021. Recent advances in natural language processing via large pre-trained language models: A survey. Workingpaper.

Gosse Minnema and Aurélie Herbelot. 2019. From brain space to distributional space: The perilous journeys of fMRI decoding. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 155–161, Florence, Italy. Association for Computational Linguistics.

Tom M. Mitchell, Svetlana V. Shinkareva, Andrew Carlson, Kai-Min Chang, Vicente L. Malave, Robert A. Mason, and Marcel Adam Just. 2008. Predicting human brain activity associated with the meanings of nouns. *Science*, 320(5880):1191–1195.

Brian Murphy, Partha Talukdar, and Tom Mitchell. 2012. Selecting corpus-semantic models for neurolinguistic decoding. In **SEM 2012: The First Joint Conference on Lexical and Computational Semantics – Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 114–123, Montréal, Canada. Association for Computational Linguistics.

Samuel Nastase, Yun-Fei Liu, Hanna Hillman, Asieh Zadbood, Liat Hasenfratz, Neggin Keshavarzian, Janice Chen, Christopher Honey, Yaara Yeshurun, Mor Regev, Mai Nguyen, Claire H. C. Chang, Christopher Baldassano, Olga Lositsky, Erez Simony, Michael Chow, Yuan Leong, Paula Brooks, Emily Micciche, and Uri Hasson. 2021. The "narratives" fmri dataset for evaluating models of naturalistic language comprehension. *Scientific Data*, 8:250.

Subba Reddy Oota, Frederic Alexandre, and Xavier Hinaut. 2022a. Long-term plausibility of language models and neural dynamics during narrative listening. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, volume 44.

Subba Reddy Oota, Jashn Arora, Veeral Agarwal, Mounika Marreddy, Manish Gupta, and Bapi Surampudi. 2022b. Neural language taskonomy: Which NLP tasks are the most predictive of fMRI brain activity? In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3220–3237, Seattle, United States. Association for Computational Linguistics.

Subba Reddy Oota, Manish Gupta, and Mariya Toneva. 2022c. Joint processing of linguistic properties in brains and language models. *arXiv*.

Subba Reddy Oota, Naresh Manwani, and Raju S. Bapi. 2018. fmri semantic category decoding using linguistic encoding of word embeddings. In *Neural Information Processing*, pages 3–15, Cham. Springer International Publishing.

Damian Pascual, Béni Egressy, Nicolas Affolter, Yiming Cai, Oliver Richter, and Roger Wattenhofer. 2022. Improving brain decoding methods and evaluation. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1476–1480.

Alexandre Pasquiou, Yair Lakretz, John Hale, Bertrand Thirion, and Christophe Pallier. 2022. Neural language models are not born equal to fit brain data, but training helps. In *ICML 2022 - 39th International Conference on Machine Learning*, page 18, Baltimore, United States.

Francisco Pereira, Matthew M. Botvinick, and Greg Detre. 2013. Using wikipedia to learn semantic feature representations of concrete concepts in neuroimaging experiments. *Artificial intelligence*, 194:240–252.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature Communications*, 9.

Russell A. Poldrack and Martha J. Farah. 2015. Progress and challenges in probing the human brain. *Nature*, 526(7573):371–379.

Peng Qian, Xipeng Qiu, and Xuanjing Huang. 2016. Bridging lstm architecture and the neural dynamics during reading. *In Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence, p. 1953–1959*, abs/1604.06635.

Aniketh Janardhan Reddy and Leila Wehbe. 2021. Can fmri reveal the representation of syntactic structure in the brain? In *Advances in Neural Information Processing Systems*, volume 34, pages 9843–9856. Curran Associates, Inc.

Yu-Ping Ruan, Zhen-Hua Ling, and Yu Hu. 2016. Exploring semantic representation in brain activity using word embeddings. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 669–679, Austin, Texas. Association for Computational Linguistics.

J.M. (Jan Mathijs) Schoffelen, Robert Oostenveld, Nietzsche Lam, Julia Udden, Annika Hultén, and P. (Peter) Hagoort. 2019. Mother of unification studies, a 204-subject multimodal neuroimaging dataset to study language processing.

Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua B. Tenenbaum, and Evelina Fedorenko. 2021. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences*, 118(45):e2105646118.

Martin Schrimpf, Jonas Kubilius, Ha Hong, Najib Majaj, Rishi Rajalingham, Elias B. Issa, Kohitij Kar, Pouya Bashivan, Jonathan Prescott-Roy, Kailyn Schmidt, Daniel L. K. Yamins, and James J. DiCarlo. 2018. Brain-score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*.

Dan Schwartz, Mariya Toneva, and Leila Wehbe. 2019. *Inducing Brain-Relevant Bias in Natural Language Processing Models*. Proceedings of the 33rd International Conference on Neural Information Processing Systems, Red Hook, NY, USA.

Anders Søgaard. 2016. Evaluating word embeddings with fMRI and eye-tracking. In *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pages 116–121, Berlin, Germany. Association for Computational Linguistics.

Anders Søgaard, Ivan Vulić, Sebastian Ruder, and Manaal Faruqui. 2019. *Cross-Lingual Word Embeddings*, 2 edition. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, United States.

Jingyuan Sun, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2019. Towards sentence-level brain decoding with distributed representations. In *Proceedings of the Thirty-Third AAAI Conference on Artificial Intelligence and Thirty-First Innovative Applications of Artificial Intelligence Conference and Ninth AAAI Symposium on Educational Advances in Artificial Intelligence*, AAAI'19/IAAI'19/EAAI'19. AAAI Press.

Sanne Ten Oever, Karthikeya Kaushik, and Andrea E Martin. 2022. Inferring the nature of linguistic computations in the brain. *PLoS computational biology*, 18(7):e1010269.

Mariya Toneva, Tom M. Mitchell, and Leila Wehbe. 2022a. Combining computational controls with natural text reveals new aspects of meaning composition. *Nature Computational Science volume 2*.

Mariya Toneva and Leila Wehbe. 2019. *Interpreting and Improving Natural-Language Processing (in Machines) with Natural Language-Processing (in the Brain)*. Curran Associates Inc., Red Hook, NY, USA.

Mariya Toneva, Jennifer Williams, Anand Bollu, Christoph Dann, and Leila Wehbe. 2022b. Same cause; different effects in the brain.

Aditya R. Vaidya, Shailee Jain, and Alexander G. Huth. 2022. Self-supervised models of audio effectively explain human cortical responses to speech. In *International Conference on Machine Learning*.

Leila Wehbe, Brian Murphy, Partha Talukdar, Alona Fyshe, Aaditya Ramdas, and Tom Mitchell. 2014a. Simultaneously uncovering the patterns of brain regions involved in different story reading subprocesses. *PLoS One*, 9(11):e112575.

Leila Wehbe, Ashish Vaswani, Kevin Knight, and Tom Mitchell. 2014b. Aligning context-based statistical models of language with brain activity during reading. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 233–243, Doha, Qatar. Association for Computational Linguistics.

Yizhen Zhang, Kuan Han, Robert Worth, and Zhongming Liu. 2020. Connecting concepts in the brain by mapping cortical representations of semantic relations. *bioRxiv*.

Shuxian Zou, Shaonan Wang, Jiajun Zhang, and Chengqing Zong. 2022. Cross-modal cloze task: A new task to brain-to-word decoding. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 648–657, Dublin, Ireland. Association for Computational Linguistics.

A Appendix

A.1 Metric Correlations

We used the following synthetic experiment to estimate the correlations between some of the most widely used performance metrics:

(i) Generate n random numbers and sort them to produce the list $\mathcal{A}$.

(ii) Sample $\frac{n}{10}$ items $\mathcal{B}$ of $\mathcal{A}$ at random.

(iii) For $\epsilon \in \{\frac{1}{100}, \ldots, \frac{100}{100}\}$, evaluate $\mu_{b \in \mathcal{B}}$ for $\langle b, \epsilon \cdot b \rangle$ for all metrics.

In other words, for a noise level ϵ, we evaluate predicted images or word vectors $\epsilon \cdot b$ against true images or word vectors b relative to a set of target images/vectors of 99 candidate words. This experiment is easily repeated to estimate reliable coefficients.

A.2 Correlation - Explained Variance

In this study, we do not distinguish between studies using Pearson correlation and studies using explained variance.

Pearson Correlation can be defined as:

$$r = \frac{\sum_{i=1}^{n}(x_i - \overline{x})(y_i - \overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i - \overline{y})^2}}$$

where:

- r is the Pearson correlation coefficient between variables X and Y
- x_i and y_i are individual data points for variables X and Y
- $\overline{x}$ and $\overline{y}$ are the means of variables X and Y
- n is the sample size.

The proportion of variance explained by the correlation is represented by r^2. The correlation coefficient (r) measures the strength and direction of the linear relationship, while the coefficient of determination ($R^2 = r^2$) represents the proportion of the variance explained by the independent variable(s) in the dependent variable.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
8

☑ A2. Did you discuss any potential risks of your work?
Left blank.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Left blank.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

□ B1. Did you cite the creators of artifacts you used?
No response.

□ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

□ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

□ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

□ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

□ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☒ Did you run computational experiments?

Left blank.

□ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

□ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

□ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

□ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

□ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

□ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

□ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.