Which Sense Dominates Multisensory Semantic Understanding? A Brain Decoding Study

Dandan Huang[1], **Lu Cao**[2,3], **Zhenting Li**[1,*], **Yue Zhang**[2,3]
[1]Zhejiang Lab, [2]School of Engineering, Westlake University
[3]Institute of Advanced Technology, Westlake Institute for Advanced Study
{huangdandan, zhenting.li}@zhejianglab.com, {caolu, zhangyue}@westlake.edu.cn

Abstract

Decoding semantic meanings from brain activity has attracted increasing attention. Neurolinguists have found that semantic perception is open to multisensory stimulation, as word meanings can be delivered by both auditory and visual inputs. Prior work which decodes semantic meanings from neuroimaging data largely exploits brain activation patterns triggered by stimulation in cross-modality (i.e. text-audio pairs, text-picture pairs). Their goal is to develop a more sophisticated computational model to probing what information from the act of language understanding is represented in human brain. While how the brain receiving such information influences decoding performance is underestimated. This study dissociates multisensory integration of word understanding into written text, spoken text and image perception respectively, exploring the decoding efficiency and reliability of unisensory information in the brain representation. The findings suggest that, in terms of unisensory, decoding is most successful when semantics is represented in pictures, but the effect disappears in the case of congeneric words which share a related meaning. These results reveal the modality dependence and multisensory enhancement in the brain decoding methodology.

Keywords: semantic understanding, brain decoding, sensory modality, fNIRS

1. Introduction

Brain decoding is a complex task that involves both neuroscience and computational linguistics. Pereira et al. (2001) presented the first neural network to distinguish brain activation patterns in reading tasks. Since then, in-depth explorations have been conducted to demonstrate that semantic clues are encoded in neural patterns and can be decoded by extrinsic representational models (Mitchell et al., 2004, 2008; Murphy et al., 2009; Anderson et al., 2013, 2017; Wang et al., 2020; Srikant et al., 2022; Murphy et al., 2022). The primary approach is to establish a predictive relationship between the neural activation recorded by neuroimaging equipment and the word distributional representation produced by embedding models (Pennington et al., 2014; Peters et al., 2018; Devlin et al., 2019).

As a prerequisite for brain decoding, neural activation needs to be recorded with highly-controlled stimuli. Part of studies adopted plain texts as stimuli (Pereira et al., 2018; Murphy et al., 2022). While others exploited text-picture pairs (Mitchell et al., 2008) or pictures with auditory words (Zinszer et al., 2017) as stimuli. The intuition behind these studies is that semantic perception is open to both auditory and visual inputs, as word meanings can be conveyed through both modalities. The neural patterns collected in these studies are either monomodal or induced by cross-modal integration of semantic perception as a whole. However, the impact of how semantic information conveyed in each modality affects brain decoding is less well understood.

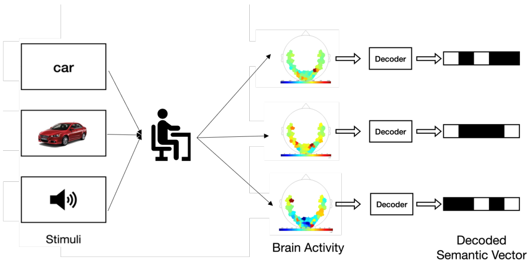

Figure 1: Brain decoding methodology. We first collect human brain activation in response to unimodal stimuli, then train decoders which take neural patterns as input and output corresponding semantic vectors, thereby predicting the presented stimuli.

In this study, we ask whether neural activation triggered by same semantic meaning in different sensory modalities contains equivalent information which is reliable and efficient for neural decoding. The focus is on processing semantically clear stimuli through reading a particular word, viewing a drawing of object, or hearing a spoken word respectively, which activate different unisensory modality separately (Figure 1). Following Zinszer et al. (2017) and Cao et al. (2021), we collect data by fNIRS and train linear regression models to map neural activation into representations of words produced by GloVe (Pennington et al., 2014). Zinszer et al. (2017) and Cao et al. (2021) have revealed that semantic representations triggered by multisensory integration are encoded in fNIRS neural data. We further extend and precise their results to the

---

*Zhenting Li is the corresponding author.

LREC-COLING 2024, pages 17557–17563

brain signals of three separate modality, decomposing the neural bases of language understanding.

Empirically, we have two salient findings. First, decoding efficiency diverges significantly as semantic meaning is presented in different modality. Image perception can encode feasible information with words in different classes, but written or spoken text fails, which shows the **modality dependence** in brain decoding. Second, multisensory information shows decodibility in both between- and within-category conditions, but unisensory in the form of picture perception shows decoding efficiency only in between-category condition, which is easier than within-category condition. This highlights the role of **multisensory enhancement** in decoding semantic clues. We publicly release our gathered fNIRS neural pattern datasets for future research[1].

2. Methods

The basic idea is to learn a mapping from brain activation patterns to particular semantic dimensions. Figure 1 describes the experimental design.

Brain activity We exploit fNIRS (functional near-infrared spectroscopy) for neuroimaging. Measuring brain activity through non-invasive methods (e.g. fMRI, EEG, MEG and fNIRS) has no skull transgression and can be setup outside clinical environments with low risk and high flexibility, thus has garnered significant attention from both neuroscience researchers and natural language processing experts. fMRI, EEG and MEG have been studied extensively and have a wealth of datasets available (Bhattasali et al., 2020; Oseki and Asahara, 2020; Zou et al., 2022), but datasets for fNIRS are relatively scarce. Our research aims to contribute to the community by collecting brain signals through fNIRS and making the dataset publicly available. This work holds the potential to broaden the scope of non-invasive brain activity research and enhance our understanding of the relationship between brain activity and natural language processing.

For the purpose of functional neuroimaging, fNIRS uses near-infrared spectrum, which is emitted by sources, propagating through the scalp, and then received by detectors, to estimate blood oxygenation changes in the cortical surface. Hemoglobin is a significant absorber of near-infrared light, thus changes in light absorption can be used to measure changes in oxygenated-hemoglobin (*oxy-Hb*) and deoxygenated-hemoglobin (*deoxy-Hb*) concentration, which is response to neural activity [2] (Watanabe et al., 2017). To date, the decoding ability of fNIRS data on semantic information has also been proved in brain mapping studies (Emberson et al., 2016; Zinszer et al., 2017; Mercure et al., 2020; Cao et al., 2021). These studies have demonstrated the feasibility of decoding *oxy-Hb* density from the brain activation patterns to semantic vectors produced by neural networks.

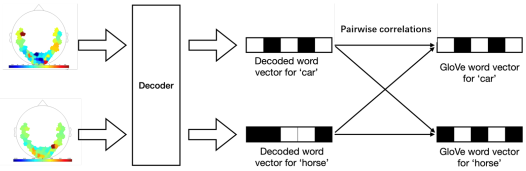

Figure 2: Each trained model is tested by first predicting word vectors for the two held-out fNIRS images and then matching them to the corresponding GloVe vectors.

Semantic vectors The stimuli are single words without sentential contex, so we select the word level embedding model GloVe (Pennington et al., 2014) to estimate semantic representations of each stimulus. GloVe has successfully served as a semantic representation in prior work which decodes linguistic meaning from neural brain data (Pereira et al., 2018; Gauthier and Ivanova, 2018; Abnar et al., 2019; Gauthier and Levy, 2019; Zou et al., 2022). Neural network representations capturing language information distributed across a high-dimensional space have been put forward, but their improvements in brain decoding are marginal at best. For example, Gauthier and Levy (2019) takes BERT (Devlin et al., 2019) and its fine-tuned variants as embedding models. Results show that none of them yield significant increases in brain decoding performance, while syntax-light representations do. Cao et al. (2021) finds the limits of decoding fine-grained semantic clues encoded in high-dimensional embeddings from fNIRS neuroimaging, suggesting that a relatively lower dimensional GloVe word embeddings (i.e. 50) can achieve better decoding performance for fNIRS patterns. Recently, Zou et al. (2022) also find that BERT embedding does not capture semantics well compared to GloVe in a fMRI-based brain-to-word decoding task. Based on these findings, we exploit GloVe representation and let $e(w_i)$ be the 50-dimensional GloVe embedding for stimulus word w_i. The word embeddings are obtained through a two-step process: global matrix factorization and local context windows. As GloVe provides pre-trained word embeddings for various languages and corpus sizes, we can directly run it. For an input token w_i, the output of the GloVe model is a corresponding 50-dimensional contextualized representation $e(w_i)$.

---

[1] https://github.com/hddbang/fNIRS

[2] Hemodynamic response to neural activation consists of an increase in *oxy-Hb* and an antiphase decrease in *deoxy-Hb*.

Decoding methodology Following early studies (Mitchell et al., 2008; Pereira et al., 2018), we use ridge regression to train a linear decoder $\delta : H_i \rightarrow e(w_i)$ for each subject, predicting the 50-dimensional GloVe word vectors given neural data by minimizing the cost function:

$$J = ||\delta H_i - e(w_i)||_2^2 + \alpha||\delta||^2, \tag{1}$$

where H_i is *oxy-Hb* concentration transferred from near-infrared light wavelength in response to the i^{th} stimulus, and α is a regularization hyperparameter.

Each linear regression model is trained and evaluated by the *leave-two-out* pairwise classification (Mitchell et al., 2008). Given N stimuli and its corresponding brain imaging, each time we use $N-2$ samples for training and the remaining two for validation. As Figure 2 shows, for stimuli pair (w_1, w_2), we predict vectors $(p(w_1), p(w_2))$ from brain patterns and match them to corresponding GloVe word vectors $(e(w_1), e(w_2))$. The *cosine similarity* is used for comparing whether each predicted vector has more similarity with its respective GloVe vector or the left out vector:

$$\begin{aligned} match\,[p(w_1) = e(w_1), p(w_2) = e(w_2)] = \\ cosine(p(w_1), e(w_1)) + \\ cosine(p(w_2), e(w_2)). \end{aligned} \tag{2}$$

If the decoded vector is more similar to its respective GloVe vector than the alternative one, we deem the classification correct. The training and testing processes repeat for C_N^2 times. The correct classification percentage represents model accuracy.

Baseline Chance level accuracy for matching the left-out neural data to words is 0.50. Following prior work which adopt a ramdon baseline (Cao et al., 2021; Zou et al., 2022), we take random scrambled pairs as a baseline to enhance the reliability of results. In this setting, the brain activities and word vectors are randomly shuffled.

3. Experimental Setting

Participants Nine right-handed native speakers (four males, mean age 21) are enrolled for the study. None of them has motor or neurological disorders.

Category	Exemplar
animal	cat, dog, horse, cow, panda
vehicle	car, train, aircraft, truck, bicycle

Table 1: Exemplars used in the experiment. The selection criteria is word familiarity in daily life to avoid ambiguity and difficulty in understanding.

	Text	Image	Audio	Zinszer et al.
Acc	0.48	0.62	0.50	0.66
RSP	0.52	0.48	0.52	\

Table 2: Decoding performance across subjects. Acc denotes the average accuracy of models. RSP denotes the accuracy of random scrambled pairs.

Procedure We present subjects with ten stimuli drawn from two broad categories (Table 1). Each subject is presented successively with stimuli in the format of text, picture and audio. During each condition, there is a break at least 60 minutes to avoid semantic priming effects[3] (Sperber et al., 1979). The task for participants is to passively view in the first two rounds and listen in the last round, trying to perceive the meaning as stimuli presented. Each textual and pictorial stimulus presentation lasts for 3 seconds and is followed by a 10-second rest period. Each audio stimulus is naturally stopped and followed by a 10-second rest period. Subjects are instructed to fixate on an X on the screen center during rest period. The stimuli are permutated randomly and repeat 7 times in each session.

fNIRS measurement and preprocessing We use NIRx NIRScout fNIRS system[4] to measure subjects' blood oxygenation changes throughout the experiment. As shown in Figure 3, the multichannel fNIRS system comprises eight sources and seven detectors, resulting in 22 measurement channels arranged in the left hemisphere. Cerebral hemodynamic responses from fNIRS do not vary significantly across recording regions in either left or right hemisphere (Cao et al., 2021). While in the left hemisphere, there are cortical areas known to be involved in language processing: temporo-parietal cortex and inferior frontal cortex thought to subserve phonological decoding, and occipito-temporal cortex and visual word form area thought to subserve orthographic processing. Thus the left hemisphere would be our regions of interest.

Following Cao et al. (2021), we set the sampling rate as 7.8Hz and perform data preprocessing with nirsLab (Xu et al., 2014). Data preprocessing includes artifacts removal, 0.01∼0.1Hz bandpass filtering, *oxy-Hb* and *deoxy-Hb* concentration computation according to the modified Beer-Lambert law (Kocsis et al., 2006) .

4. Results

We train separate decoders for each participants. Results are validated by permutation test, with statistics being created by permutation test 1000 times. The significance level is 0.05.

Overall decoding As Table 2 shows, the average cross-validated accuracy is 0.48, 0.62, 0.50 for text,

---

[3] Semantic priming refers to a facilitation of responding that occurs as a result of the preceding presentation of a semantically related prime.

[4] https://nirx.net/nirscout

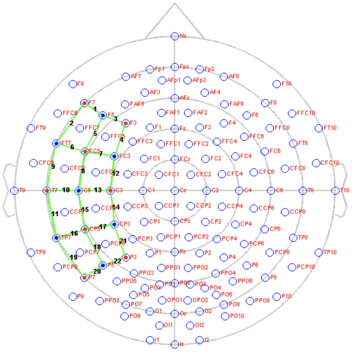

Figure 3: fNIRS probe arrangement. The eight red circles and seven blue circles represent laser sources and detectors respectively. The green lines represent the path between sources and detectors (i.e. channels). There are 22 channels in total.

image and audio stimuli respectively. Image perception exhibits decoding performance significantly above the RSP baseline ($p < 0.003$) and chance levels, while written or spoken text is inferior to the baseline ($p > 0.05$), failing to show equivalent decoding effect. Zinszer et al. (2017) reports an accuracy of 0.66 with multi-sensory inputs in the format of picture plus auditory word, which is slightly better than unisensory image perception in the current study. Alone with previous studies (Palatucci et al., 2009; Pereira et al., 2018; Murphy et al., 2012) which use other neuroimage equipment to demonstrate the feasibility of using distributed semantic representations to probe meaning representations with multisensory integration of word understanding in the brain, this result shows *modality dependence* in brain decoding methodology. We concluded that decoding is most successful when triggered by multisensory stimuli, then unisensory pictorial stimuli. Textual and spoken word alone cannot stimulate enough neural features for decoding.

Between vs within-category decoding Our stimuli are organized into two broad categories, five for animals and five for vehicles. We hypothesized that category-based differences may contribute to decoding accuracy when choosing between items in different categories. To test whether decoding accuracy of unisensory word understanding relies on category differences, we divide the pairwise decoding trials into between-category and within-category conditions, thereupon examine the accuracy of each set of trials. In the between-category case, the two held-out test words come from different groups (e.g. *cat* versus *car*), while in the within-category case, the two held-out test words come from the same category (e.g. *cat* versus *dog*). Under cross-modality setting, Zinszer et al. (2017) reports that average within- and between-category accuracies do not significantly differ. Cao et al. (2021) also reports a robust differentiation power both in within-category and between-category conditions. We compare the performance of our models trained on unisensory information when predicting words in the same or divergent semantic categories. As shown in Table 3, textual and audio stimuli fail to demonstrate decoding feasibility in both between-category and within-category settings, consistent with the overall decoding performance (as shown in Table 2). For the picture stimuli, the decoding accuracy is 0.75 for between-category, significantly higher than chance level ($p < 0.001$). However, it drops dramatically to 0.47 in the within-category condition, even worse than chance level. The decoding effect vanishes for unisensory modality in the harder within-category case in our study, which sees evidence to suggest that there is *multisensory enhancement* in the brain decoding.

	Text	Image	Audio
between-category	0.49	0.75	0.49
within-category	0.46	0.47	0.49

Table 3: Decoding performance across semantic categories.

Activation pattern For reasons behind the decoding advantages of visual stimuli over other modalities, we trace back to the brain activation patterns triggered by stimuli in different unisensory modality. As Figure 4 shows, the *oxy-Hb* concentration fluctuates when the stimuli onset and reaches the extremum in 2-4 seconds before coming back to original level. The image stimuli induces the highest *oxy-Hb* intensity at 5.46μm, followed closely by text stimuli at 4.81μm, while sound stimuli responds the most quickly and causes the weakest activation levels at 3.27μm. The faster reaction of sound accords with previous findings that word learners are faster at acquiring phonology than orthography because they are better at store phonological representation (Dehaene, 2009). But the disparity in *oxy-Hb* concentration may reflect the less specificity in phonological representations and the greater distinctiveness for visual stimuli. As for visual stimuli, image perception induces stronger *oxy-Hb* fluctuation than text, we assume that one important factor is the modality asymmetry: that a pictorial representation for a word conception is more likely to contain cross-modality information.

5. Conclusion

We present an empirical study to understand modality influence for semantic understanding and brain decoding. We collect brain activity data that dissociates multisensory integration of word meaning. Results suggest that 1) different perception modalities induce different decoding effects, highlighting the importance of considering modality in brain decoding research; 2) in terms of unisensory, image-induced patterns can be reliably decoded,

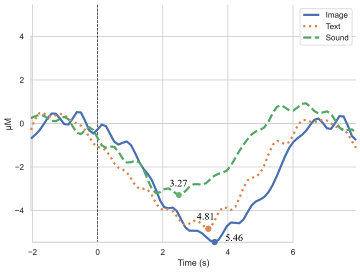

Figure 4: Visual inspection of the average brain activation patterns induced by text, image and sound stimuli. The stimuli onset at time 0, before which the subject is at a resting state and the *oxy-Hb* concentration during this period is the benchmark for comparison.

whereas textual and auditory stimulation fail; 3) unisensory modality cannot show robust decoding effects for words within same semantic category, indicating the importance of multisensory enhancement in brain decoding. These confirm our hypotheses concerning the modality dependence and multisensory enhancement in semantic decoding.

6. Acknowledgements

We appreciate all anonymous reviewers for their constructive comments. This work is supported by the Key R&D Program of Zhejiang (No.2022C04006).

Limitations

This work is a pilot study to evaluate the feasibility of brain decoding by dissociating multisensory integration of language understanding. Due to expenses and difficulty in managing human experiments, the study is limited in word understanding with a small dataset and has not extended to sentence-level research. The amount of data we adopted is relatively small but comparable to previous literature. For example, Zinszer et al. (2017) used eight stimuli, and Cao et al. (2021) also used eight stimuli in his pilot study. In the future work, we will enlarge the dataset and expand to sentence-level research.

Ethics Statement

We honor the ACL Ethics Policy. The study was approved by the local ethics committee. All subjects participated for payment, and gave informed consent in accordance with the procedure approved by the institutional Review Board. No part of the study procedures and analyses were pre-registered prior to the research being conducted. No private data or non-public information was used in this work.

7. Bibliographical References

Samira Abnar, Lisa Beinborn, Rochelle Choenni, and Willem Zuidema. 2019. Blackbox meets blackbox: Representational similarity and stability analysis of neural language models and brains. *arXiv preprint arXiv:1906.01539*.

Andrew J. Anderson, Elia Bruni, Ulisse Bordignon, Massimo Poesio, and Marco Baroni. 2013. Of words, eyes and brains: Correlating image-based distributional semantic models with neural representations of concepts. In *Proc. of EMNLP*, pages 1960–1970. ACL.

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 ACL*, 5:17–30.

Shohini Bhattasali, Jonathan Brennan, Wen-Ming Luh, Berta Franzluebbers, and John Hale. 2020. The alice datasets: fmri & eeg observations of natural language comprehension. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 120–125.

Lu Cao, Dandan Huang, Yue Zhang, Xiaowei Jiang, and Yanan Chen. 2021. Brain decoding using fnirs. In *Proc. of the AAAI*, volume 33, pages 7047–7054.

Stanislas Dehaene. 2009. Reading in the brain. *New York*.

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.

Lauren L Emberson, Benjamin D Zinszer, Rajeev DS Raizada, and Richard N Aslin. 2016. Decoding the infant mind: multichannel pattern analysis (mcpa) using fnirs. *bioRxiv*, page 061234.

Jon Gauthier and Anna Ivanova. 2018. Does the brain represent words? an evaluation of brain decoding studies of language understanding. *arXiv preprint arXiv: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.

Laszlo Kocsis, Peter Herman, and Andras Eke. 2006. The modified beer–lambert law revisited. *Physics in Medicine & Biology*, 51(5):N91.

Evelyne Mercure, Samuel Evans, Laura Pirazzoli, Laura Goldberg, Harriet Bowden-Howl, Kimberley Coulson-Thaker, Indie Beedie, Sarah Lloyd-Fox, Mark H Johnson, and Mairéad MacSweeney. 2020. Language experience impacts brain activation for spoken and signed language in infancy: insights from unimodal and bimodal bilinguals. *Neurobiology of Language*, 1(1):9–32.

Tom M Mitchell, Rebecca Hutchinson, Radu S Niculescu, Francisco Pereira, Xuerui Wang, Marcel Just, and Sharlene Newman. 2004. Learning to decode cognitive states from brain images. *Machine learning*, 57(1-2):145–175.

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.

Alex Murphy, Bernd Bohnet, Ryan McDonald, and Uta Noppeney. 2022. Decoding part-of-speech from human eeg signals. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2201–2210.

Brian Murphy, Marco Baroni, and Massimo Poesio. 2009. EEG responds to conceptual stimuli and corpus semantics. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 619–627, Singapore. Association for Computational Linguistics.

Brian Murphy, Partha Talukdar, and Tom Mitchell. 2012. Selecting corpus-semantic models for neurolinguistic decoding. In **SEM 2012*, pages 114–123. ACL.

Yohei Oseki and Masayuki Asahara. 2020. Design of bccwj-eeg: Balanced corpus with human electroencephalography. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 189–194.

Mark Palatucci, Dean Pomerleau, Geoffrey E Hinton, and Tom M Mitchell. 2009. Zero-shot learning with semantic output codes. In *Advances in neural information processing systems*, pages 1410–1418.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *EMNLP (EMNLP)*, pages 1532–1543.

Francisco Pereira, Marcel Just, and Tom Mitchell. 2001. Distinguishing natural language processes on the basis of fmri-measured brain activation. In *European Conference on Principles of Data Mining and Knowledge Discovery*, pages 374–385. Springer.

Francisco Pereira, Bin Lou, Brianna Pritchett, Samuel Ritter, Samuel J Gershman, Nancy Kanwisher, Matthew Botvinick, and Evelina Fedorenko. 2018. Toward a universal decoder of linguistic meaning from brain activation. *Nature communications*, 9(1):1–13.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Richard D Sperber, Charley McCauley, Ronnie D Ragain, and Carolyne M Weil. 1979. Semantic priming effects on picture and word processing. *Memory & Cognition*, 7(5):339–345.

Shashank Srikant, Ben Lipkin, Anna Ivanova, Evelina Fedorenko, and Una-May O'Reilly. 2022. Convergent representations of computer programs in human and artificial neural networks. *Advances in Neural Information Processing Systems*, 35:18834–18849.

Shaonan Wang, Jiajun Zhang, Nan Lin, and Chengqing Zong. 2020. Probing brain activation patterns by dissociating semantics and syntax in sentences. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34:9201–9208.

Hama Watanabe, Yoshihiko Shitara, Yoshinori Aoki, Takanobu Inoue, Shinya Tsuchida, Naoto Takahashi, and Gentaro Taga. 2017. Hemoglobin phase of oxygenation and deoxygenation in early brain development measured using fnirs. *Proceedings of the National Academy of Sciences*, 114(9):E1737–E1744.

Yong Xu, Harry L Graber, and Randall L Barbour. 2014. nirslab: a computing environment for fnirs neuroimaging data analysis. In *Biomedical optics*, pages BM3A–1. Optical Society of America.

Benjamin D Zinszer, Laurie Bayet, Lauren L Emberson, Rajeev DS Raizada, and Richard N Aslin. 2017. Decoding semantic representations from functional near-infrared spectroscopy signals. *Neurophotonics*, 5(1):011003.

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.