Multiscale mapping of linguistic information in the human brain

Shinji NISHIMOTO[1,2,3]

1. Graduate School of Frontier Biosciences, The University of Osaka, Osaka, Japan.
2. Graduate School of Medicine, The University of Osaka, Osaka, Japan
3. Center for Information and Neural Networks (CiNet). National Institute of Information and Communications Technology, Osaka, Japan

E-mail: *Nishimoto.shinji.fbs@osaka-u.ac.jp*

Abstract: Language operates across diverse contexts, supporting situational understanding, long-term story comprehension, and real-time dialogue. Here, by analyzing human brain activity evoked during film viewing and spontaneous conversation, we show that multi-level LLM embeddings reveal distinct semantic networks across contexts as well as distinctive representations for language production and comprehension.

Keywords: Neural Representations, Neuroimaging, Natural Language Processing

Introduction

Human experience is a seamless tapestry of meanings drawn from what we see, hear, and say—and from the ever-shifting social and narrative contexts that frame those sensations. Every glance at a crowded street, every overheard remark, and every exchange of dialogue with a friend arrives already embedded in layers of semantic structure that stretch across multiple sensory channels and time-scales. Our brains must therefore integrate auditory, visual, linguistic, and contextual cues on the fly, distilling them into coherent representations that allow us to understand both the world and one another.

Over the past decade, the convergence of large-scale neuroimaging and large language models (LLMs) has begun to illuminate this integration process. By coupling fMRI, iEEG, and MEG recordings with embeddings extracted from state-of-the-art LLMs, researchers can now predict brain activity while participants listen to narrated stories, watch richly annotated films, or even navigate virtual environments. These voxel-wise encoding and decoding techniques have yielded striking cortical maps of meaning, emotion, and higher-order cognition, demonstrating that language-derived semantic vectors reliably track activity in wide cortical areas and also be used to even decode semantic contents from brain activity [1].

Yet most existing studies still examine a single stream of information in isolation: speech without images, comprehension without production, or narrative text stripped of its audiovisual context. As a result, we still know surprisingly little about how the brain simultaneously represents diverse semantic signals—or whether auditory and visual meanings converge on shared neural codes or remain partially segregated across modality-specific pathways.

In today's keynote, I will present two new lines of work that address this gap by probing brain activity under naturalistic, multimodal conditions. First, I will show how we disentangle auditory, visual, and narrative (contextual) contributions while participants view hours of movies and TV dramas, revealing both shared and distinct topographies for each semantic tier [2]. Second, I will turn to real-time conversation, contrasting the neural signatures of speaking versus listening to uncover how production and comprehension networks encode partly overlapping but functionally specialized meaning spaces [3]. Together, these studies suggest that the brain constructs a multi-layered semantic landscape: some dimensions serve as common currency across senses and communicative roles, while others preserve modality- or speaker-specific nuance.

Methods

Participants and data acquisition: We collected functional magnetic-resonance-imaging (fMRI) data from two independent cohorts of healthy adults. All participants provided written informed consent in accordance with the institutional review board of National Institute of Information and Communications Technology. Scans were acquired on a 3 T MRI with a multiband echo-planar imaging sequence (TR = 1.0 s, voxel size = 2 mm isotropic) while subjects either (i) watched 8.3 h of audiovisual material or (ii) engaged in 3 h of unscripted dialogue with an experimenter. A T1-weighted anatomical image (MPRAGE, 1 mm isotropic) was also obtained for each participant.

Movie-viewing experiment: Stimuli comprised excerpts from feature films and television dramas spanning multiple genres (drama, comedy, sci-fi, animation, action). For every shot we produced a rich, hierarchical annotation set that included (1) the exact spoken transcript, (2) an objective visual scene description, (3) higher-level narrative context (background story, character goals), and (4) spatiotemporal metadata such as setting, location, and time of day. Annotation was performed by trained human raters and aligned to TR resolution.

Conversational experiment: Each participant held spontaneous, face-to-face conversations with an experimenter on 27 pre-defined topics covering daily life, personal preferences, science, and current events. No script or cue cards were provided. Audio was recorded with noise-canceling microphones, transcribed with Microsoft Azure Speech-to-Text, and manually corrected. The final corpus consists of time-stamped turn-by-turn transcripts separated by speaker (self vs. partner).

Computational modeling: We extracted latent semantic representations from a battery of large language models (e.g., GPT, Llama) and multimodal vision–language models (e.g., CLIP, Llava). Each embedding was temporally aligned to the

fMRI time series (down-sampled to TR resolution), and z-scored within runs. We then fit voxel-wise ridge-regularized linear encoding models to predict BOLD activity from the concatenated embeddings. Model performance was quantified by Pearson correlation (r) between predicted and observed time-courses in left-out runs using a cross-validation scheme.

Results

In movie viewing data, embeddings extracted from large language and vision–language models predicted fMRI responses more accurately than conventional lexical or acoustic features. Speech-based, vision-based, and story-level linguistic embeddings each corresponded to partially distinct cortical zones, and gains were greatest for long-timescale narrative content in movie scenes. Combining visual with textual embeddings further improved performance, indicating a shared multimodal semantic code.

In conversational data, contextual GPT embeddings uncovered a common representation shared by speaking and listening at short timescales (words and single sentences). Outside this core, production depended on shorter integration windows, whereas comprehension benefited from longer ones, pointing to distinct temporal demands for real-time utterance planning versus discourse integration.

Discussion

Our findings indicate that the brain hosts a multi-layered semantic architecture: speech, vision, and narrative each give rise to partially overlapping yet distinguishable neural codes, while higher-order representations knit these channels together into a unified experience. In other words, meaning is neither wholly shared across modalities nor strictly siloed; instead, it is distributed across tiers that can be selectively accessed or integrated as task demands shift.

Because the approach hinges on flexible, pre-trained language and vision–language models rather than hand-crafted features, it can be readily extended to a broad range of perceptual and cognitive domains. Applying the same framework to tasks that tap memory, mental imagery, logic, prediction, or self-reflection could yield quantitative models of functions that have so far resisted formal description. Moreover, by tracking how encoding weights evolve over time or differ across individuals, we can begin to chart learning trajectories, developmental trends, and clinically relevant variability within a common analytical space.

Finally, to foster such work we have released the full fMRI datasets—movie viewing and spontaneous dialogue—along with synchronized transcripts, annotations, and code for reproducing all analyses [4][5]. We invite the community to explore these resources, refine our models, and test new hypotheses about how the brain represents the rich tapestry of everyday experience.

Conclusions

In sum, by coupling fMRI recordings obtained under everyday conditions—extended movie viewing and spontaneous dialogue—with latent representations from large language and multimodal vision–language models, we demonstrated that semantic content is encoded in the brain through a partially shared yet functionally specialized architecture that integrates information across modalities and timescales. These results highlight a multimodal semantic code that spans speech, vision, and narrative context while preserving modality-specific and temporally distinct computations. Because our framework relies on general-purpose models rather than hand-crafted features, it can be readily applied to a wider spectrum of perceptual and cognitive tasks, paving the way for more comprehensive models of memory, introspection, and other higher functions, and ultimately informing translational and clinical research.

Acknowledgements

This work was supported by MEXT/JSPS KAKENHI grants as well as JST CREST, ERATO, and AIP Acceleration Research grants.

References

[1] J. Tang et al., “Semantic reconstruction of continuous language from non-invasive brain recordings,” *Nature Neuroscience*, vol. 26, May 2023, doi: 10.1038/s41593-023-01304-9.

[2] Y. Nakagi et al., “Unveiling Multi-level and Multi-modal Semantic Representations in the Human Brain using Large Language Models,” *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Nov. 2024, doi: 10.18653/v1/2024.emnlp-main.1133.

[3] M. Yamashita et al., “Conversational content is organized across multiple timescales in the brain,” *Nature Human Behaviour*, vol. 9, Jun. 2025, doi: 10.1038/s41562-025-02231-4.

[4] H. Q. Yamaguchi, et al., Sept. 26, 2024. “Narrative Movie fMRI Dataset,” distributed by OpenNeuro, doi:10.18112/openneuro.ds005531.v1.0.0.

[5] M. Yamashita, R. Kubo, and S. Nishimoto, June 16, 2025. “Natural Dialogue fMRI Dataset,” distributed by OpenNeuro, doi: doi:10.18112/openneuro.ds004669.v2.0.2.