Modality-Agnostic Representations are Widespread Across the Cortex

Mitja Nikolaus (mitja.nikolaus@cnrs.fr)
Université de Toulouse, CNRS, CerCo, Toulouse, France

Milad Mozafari
Torus AI, Toulouse, France

Nicholas Asher
Université de Toulouse, IRIT, Toulouse, France

Leila Reddy
Université de Toulouse, CNRS, CerCo, Toulouse, France

Rufin VanRullen
Université de Toulouse, CNRS, CerCo, Toulouse, France

Abstract

Humans are able to perform tasks that require manipulation of inputs regardless of how these signals were perceived by the brain. This can be achieved thanks to representations that are agnostic to the stimulus modality. Previous work that attempted to localize such modality-agnostic representations has not yet led to conclusive results, with different studies proposing varying sets of candidate regions. These analyses have largely relied on relatively small-scale fMRI datasets with predefined sets of stimulus categories. In our work, we leveraged a new large-scale multimodal fMRI dataset of 6 subjects watching both diverse images and short text descriptions of such images to localize modality-agnostic representations. To this end, we performed a searchlight analysis with decoders trained by mapping brain activity patterns to the latent space of pretrained deep neural networks. We identified regions in which it is possible to decode both stimulus modalities in a modality-agnostic way (i.e., with a single decoder applied to brain responses from images or text). We found that large areas of the brain contain modality-agnostic representations, particularly in the left hemisphere. Our study highlights the importance of naturalistic stimuli and large-scale datasets for insightful analyses of representations in the human brain.

Introduction

Modality-agnostic patterns are representations that are abstracted away from particularities of specific modalities such as vision and language. A range of theories have been developed to explain how and where in the human brain such abstract representations are created (Baars, 1993; Damasio, 1989; A. Martin, 2016; Barsalou, 2016; Ralph, Jefferies, Patterson, & Rogers, 2017). Previous studies that aimed to localize modality-agnostic patterns did not always agree on the exact location and extent of such regions (Vandenberghe, Price, Wise, Josephs, & Frackowiak, 1996; Shinkareva, Malave, Mason, Mitchell, & Just, 2011; Devereux, Clarke, Marouchos, & Tyler, 2013; Fairhall & Caramazza, 2013; Jung, Larsen, & Walther, 2018). A major limitation of these studies is their reliance on a predefined set of stimulus categories, which stands in contrast to the complex stimuli we perceive in our everyday life.

Here, we analyzed SemReps-8K, a large-scale multimodal fMRI dataset of 6 subjects each viewing more than 8,000 stimuli which are presented separately in one of two modalities, as images or as descriptive captions (of such images). Based on this dataset with naturalistic stimuli, we designed a searchlight analysis to localize modality-agnostic regions in the cortex. More specifically, we used modality-agnostic decoders that are specifically trained to leverage modality-agnostic patterns by exposing them to brain imaging data from two modalities. We require that above-chance decoding of both modalities is possible using such decoders. Additionally, we trained modality-specific decoders for both modalities and evaluated them in a cross-decoding setup (testing a vision-trained decoder on captions, and vice-versa) to ensure that the patterns transfer between modalities in both directions.

Despite these strict requirements our method allowed us to identify a large network of regions with modality-agnostic patterns in the brain.

Methods

fMRI Experiment and Preprocessing

The experiment involved 6 subjects (2 female, all right-handed and fluent English speakers). Functional data as well as anatomical images was acquired using a 3T Philips ACHIEVA scanner (10 sessions, each with 13-16 runs). During each run a subject was shown images and captions in random alternation (stimulus presentation: 2.5s; inter-stimulus interval: 1s), and was instructed to press a button whenever the stimulus matched the immediately preceding one (one-back matching task). Images and captions were taken from the COCO dataset (Lin et al., 2014). As training set, a random subset of images and another random subset of captions were selected for each subject. As test set, a shared subset of 140 stimuli (70 images and 70 captions) was presented repeatedly to each subject.[1]

Preprocessing was performed using SPM 12 (Ashburner et al., 2014). For each subject, slice time correction and realignment were applied. All functional scans were coregistered to a manually corrected anatomical scan and afterwards transformed to MNI305 space. In order to obtain beta-values for each training and test stimulus, a GLM was fit for each subject. Finally, the data was transformed to surface space. Further details on the fMRI experiment and dataset can be found in the paper accompanying the dataset release (Nikolaus et al., 2025).

Localizing Modality-Agnostic Patterns

We use modality-specific decoders, trained only on brain imaging data of a single modality, and modality-agnostic decoders trained on brain imaging data from multiple modalities, and therefore allowing for decoding of stimuli irrespective of their modality.

Modality-agnostic regions should contain patterns that generalize between stimulus modalities. Therefore, such regions should allow for decoding of stimuli in both modalities using a single modality-agnostic decoder, i.e. the decoding performance for images and captions should be above chance. We additionally added two conditions to control that the representations directly transfer between the modalities, by training two modality-specific decoders and evaluating them in a cross-decoding setup, i.e. we require that their performance in the modality they were not trained on is also above chance.

[1] Note that for each stimulus presented to the subject (e.g. an image), we also have access to the corresponding stimulus in the other modality (the corresponding caption), allowing us to extract multimodal stimulus features based on both image and caption.

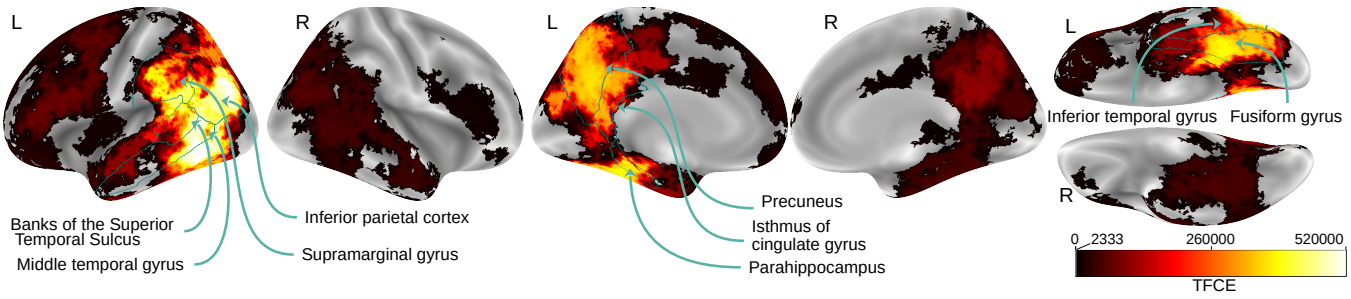

Figure 1: Searchlight results for modality-agnostic regions. Maps thresholded for $p < 10^{-4}$ (corresponding to a TFCE value of 2333). Anatomical regions with highest cluster values are annotated based on the Desikan-Killiany atlas (Desikan et al., 2006).

All decoders were trained by fitting ridge regression models that take fMRI beta-values as input and predict latent representations extracted from a pretrained deep learning model.[2]

Preliminary experiments using the whole brain data for decoding showed that decoders based on multimodal ImageBind (Girdhar et al., 2023) features leads to the highest decoding performance (across a large set of tested models), in the subsequent experiments we therefore used these features.

For each vertex, we defined a searchlight with a fixed size by selecting the 750 closest vertices, corresponding to an average radius of ~ 10mm. We trained and evaluated a modality-agnostic decoder and modality-specific decoders for both modalities for each searchlight location and each subject, providing us with 4 pairwise accuracy scores for each location on the cortex.[3] Then we performed t-tests to identify locations in which the decoding performance is above chance ($acc > 0.5$). We aggregated all 4 comparisons by taking the minimum of the 4 t-values at each location, and performed TFCE (Smith & Nichols, 2009) to identify modality-agnostic clusters. To estimate the statistical significance of the resulting clusters we performed a permutation test ($n = 10K$).

Results and Discussion

The results of the searchlight analysis (Figure 1) reveal that modality-agnostic patterns can be found in a widespread left-lateralized network across the brain.

All areas with high cluster values confirm findings from previous studies: The left precuneus (Shinkareva et al., 2011; Fairhall & Caramazza, 2013; Popham et al., 2021; Handjaras et al., 2016), posterior cingulate / retrosplenial cortex (Fairhall & Caramazza, 2013; Handjaras et al., 2016), supramarginal gyrus (Shinkareva et al., 2011), inferior parietal cortex (Man, Kaplan, Damasio, & Meyer, 2012; Vandenberghe et al., 1996; Shinkareva et al., 2011; Devereux et al., 2013; Popham et al., 2021; Simanova, Hagoort, Oostenveld, & van Gerven, 2014; Handjaras et al., 2016), superior temporal sulcus (Man et al., 2012), middle and inferior temporal gyrus (Vandenberghe et al., 1996; Shinkareva et al., 2011; Fairhall & Caramazza, 2013; Devereux et al., 2013; Simanova et al., 2014; Handjaras et al., 2016), fusiform gyrus (Vandenberghe et al., 1996; Moore & Price, 1999; Bright, Moss, & Tyler, 2004; Shinkareva et al., 2011; Fairhall & Caramazza, 2013; Simanova et al., 2014), and parahippocampus (Vandenberghe et al., 1996). However, previous studies have led to contradicting results regarding the locality of modality-agnostic regions (they identified varying subsets of these regions), probably due to the limited number and artificial nature of stimuli employed. Our method identified *all* of the aforementioned regions as regions with modality-agnostic patterns, highlighting the advantage of our searchlight method and the large multimodal dataset in which subjects are viewing photographs of complex natural scenes and reading full English sentences.

The left-lateralization of our results can be seen as support for theories that link modality-agnostic representations to (lexical-) semantic representations (Simmons & Barsalou, 2003; Binder, Desai, Graves, & Conant, 2009; Meschke & Gallant, 2024). Relatedly, a range of studies that aimed to identify brain regions linked to semantic/conceptual representations found evidence for such regions that overlap to a high degree with the regions identified in our study (Fernandino et al., 2016; C. B. Martin, Douglas, Newsome, Man, & Barense, 2018; Carota, Nili, Pulvermüller, & Kriegeskorte, 2021; Fernandino, Tong, Conant, Humphries, & Binder, 2022; Tong et al., 2022).

In the future, we plan to investigate how modality-agnostic patterns are modulated by attention, by analyzing additional test data from the same subjects in which they were instructed to pay attention to only one of the modalities in specific runs.

This abstract is based on a manuscript currently under review (Nikolaus et al., 2025), a preprint is available at https://doi.org/10.1101/2025.06.08.658221 .

[2] The regularization hyperparameter α was optimized using 5-fold cross validation on the training set.

[3] In the case of cross-modal decoding (e.g. mapping an image stimulus into the latent space of a language model), a trial was counted as correct if the caption corresponding to the image (according to the ground-truth in COCO) was closest.

Acknowledgments

This research was funded by grants from the French Agence Nationale de la Recherche (ANR: AI-REPS grant number ANR-18-CE37-0007-01 and ANITI grant number ANR-19-PI3A-0004) as well as the European Union (ERC Advanced grant GLoW, 101096017). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

We thank the Inserm/UPS UMR1214 Technical Platform for their help in setting up and for the acquisitions of the MRI sequences.

References

Ashburner, J., Barnes, G., Chen, C.-C., Daunizeau, J., Flandin, G., Friston, K., … Moran, R. (2014). SPM12. *Wellcome Trust Centre for Neuroimaging*.

Baars, B. J. (1993). *A cognitive theory of consciousness*. Cambridge University Press.

Barsalou, L. W. (2016). On Staying Grounded and Avoiding Quixotic Dead Ends. *Psychonomic Bulletin & Review*, *23*(4), 1122–1142.

Binder, J. R., Desai, R. H., Graves, W. W., & Conant, L. L. (2009). Where Is the Semantic System? A Critical Review and Meta-Analysis of 120 Functional Neuroimaging Studies. *Cerebral Cortex*, *19*(12), 2767–2796.

Bright, P., Moss, H., & Tyler, L. (2004). Unitary vs multiple semantics: PET studies of word and picture processing. *Brain and Language*, *89*(3), 417–432.

Carota, F., Nili, H., Pulvermüller, F., & Kriegeskorte, N. (2021). Distinct fronto-temporal substrates of distributional and taxonomic similarity among words: evidence from RSA of BOLD signals. *NeuroImage*, *224*, 117408.

Damasio, A. R. (1989). The Brain Binds Entities and Events by Multiregional Activation from Convergence Zones. *Neural Computation*, *1*(1), 123–132.

Desikan, R. S., Ségonne, F., Fischl, B., Quinn, B. T., Dickerson, B. C., Blacker, D., … Killiany, R. J. (2006). An automated labeling system for subdividing the human cerebral cortex on MRI scans into gyral based regions of interest. *NeuroImage*, *31*(3), 968–980.

Devereux, B. J., Clarke, A., Marouchos, A., & Tyler, L. K. (2013). Representational Similarity Analysis Reveals Commonalities and Differences in the Semantic Processing of Words and Objects. *The Journal of Neuroscience*, *33*(48).

Fairhall, S. L., & Caramazza, A. (2013). Brain Regions That Represent Amodal Conceptual Knowledge. *Journal of Neuroscience*, *33*(25), 10552–10558.

Fernandino, L., Binder, J. R., Desai, R. H., Pendl, S. L., Humphries, C. J., Gross, W. L., … Seidenberg, M. S. (2016). Concept Representation Reflects Multimodal Abstraction: A Framework for Embodied Semantics. *Cerebral Cortex*, *26*(5), 2018–2034.

Fernandino, L., Tong, J.-Q., Conant, L. L., Humphries, C. J., & Binder, J. R. (2022). Decoding the information structure underlying the neural representation of concepts. *Proceedings of the National Academy of Sciences*, *119*(6).

Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K. V., Joulin, A., & Misra, I. (2023). ImageBind: One Embedding Space To Bind Them All. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 15180–15190).

Handjaras, G., Ricciardi, E., Leo, A., Lenci, A., Cecchetti, L., Cosottini, M., … Pietrini, P. (2016, July). How concepts are encoded in the human brain: A modality independent, category-based cortical organization of semantic knowledge. *NeuroImage*, *135*, 232–242.

Jung, Y., Larsen, B., & Walther, D. B. (2018, June). Modality-Independent Coding of Scene Categories in Prefrontal Cortex. *Journal of Neuroscience*, *38*(26), 5969–5981.

Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., … Zitnick, C. L. (2014). Microsoft COCO: Common Objects in Context. In *Computer Vision – ECCV 2014* (Vol. 8693, pp. 740–755).

Man, K., Kaplan, J. T., Damasio, A., & Meyer, K. (2012). Sight and Sound Converge to Form Modality-Invariant Representations in Temporoparietal Cortex. *Journal of Neuroscience*, *32*(47), 16629–16636.

Martin, A. (2016). GRAPES—Grounding representations in action, perception, and emotion systems: How object properties and categories are represented in the human brain. *Psychonomic Bulletin & Review*, *23*(4), 979–990.

Martin, C. B., Douglas, D., Newsome, R. N., Man, L. L., & Barense, M. D. (2018). Integrative and distinctive coding of visual and conceptual object features in the ventral visual stream. *eLife*, *7*.

Meschke, E., & Gallant, J. (2024). Mapping Multimodal Conceptual Representations within the Lexical-Semantic Brain System. In *CCN.*

Moore, C. J., & Price, C. J. (1999). Three Distinct Ventral Occipitotemporal Regions for Reading and Object Naming. *NeuroImage*, *10*(2), 181–192.

Nikolaus, M., Mozafari, M., Berry, I., Asher, N., Reddy, L., & VanRullen, R. (2025). *Modality-Agnostic Decoding of Vision and Language from fMRI.* bioRxiv. doi: 10.1101/2025.06.08.658221

Popham, S. F., Huth, A. G., Bilenko, N. Y., Deniz, F., Gao, J. S., Nunez-Elizalde, A. O., & Gallant, J. L. (2021). Visual and linguistic semantic representations are aligned at the border of human visual cortex. *Nature Neuroscience*, *24*(11), 1628–1636.

Ralph, M. A. L., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, *18*(1), 42–55.

Shinkareva, S. V., Malave, V. L., Mason, R. A., Mitchell, T. M., & Just, M. A. (2011). Commonality of neural representations of words and pictures. *NeuroImage*, *54*(3), 2418–2425.

Simanova, I., Hagoort, P., Oostenveld, R., & van Gerven, M. A. J. (2014). Modality-Independent Decoding of Semantic Information from the Human Brain. *Cerebral Cortex*, *24*(2), 426–434.

Simmons, W. K., & Barsalou, L. W. (2003). The similarity-in-topography principle: Reconciling theories of conceptual deficits. *Cognitive Neuropsychology*, *20*(3-6), 451–486.

Smith, S. M., & Nichols, T. E. (2009). Threshold-free cluster enhancement: Addressing problems of smoothing, threshold dependence and localisation in cluster inference. *NeuroImage*, *44*(1), 83–98.

Tong, J., Binder, J. R., Humphries, C., Mazurchuk, S., Conant, L. L., & Fernandino, L. (2022). A Distributed Network for Multimodal Experiential Representation of Concepts. *Journal of Neuroscience*, *42*(37), 7121–7130.

Vandenberghe, R., Price, C., Wise, R., Josephs, O., & Frackowiak, R. S. (1996). Functional anatomy of a common semantic system for words and pictures. *Nature*, *383*(6597).