Representations of Semantic Relations in the Human Brain During Active Relation Processing

Catherine Chen (cathychen@berkeley.edu)[1], Lily Gong[1], Fatma Deniz[2], Daniel Klein[1], Jack Gallant[1]
[1]UC Berkeley, Berkeley, USA [2]TU Berlin, Berlin, Germany

Abstract

Relations between concepts are a crucial component of human semantic knowledge. Behavioral studies have shown that relations between concepts can be classified into different types, but it is unclear how distinctions between semantic relations are reflected in the human brain. Therefore we conducted a study to characterize brain representations of semantic relations. Six participants each answered over 1000 questions involving six semantic relations while functional magnetic resonance imaging (fMRI) was used to record BOLD responses. Then voxelwise encoding models were used to characterize the selectivity for each semantic relation in each voxel and participant separately. We find that the type of semantic relation accurately predicts brain responses throughout temporal, parietal, and prefrontal cortices. These results suggest that hypothesized distinctions between semantic relations are reflected in brain representations.

Keywords: semantic relations; language; fMRI; BOLD

Introduction

Human conceptual knowledge of objects involves relations between concepts. For example, knowledge of the object "bicycle" involves relations to other concepts such as "wheel", "vehicle", and "transportation". Understanding relations between concepts is thought to be a crucial component for human intelligence, allowing humans to draw inferences, make generalizations, and perform analogical reasoning (Bejar et al., 1991; Chaffin, 1988; Unger & Fisher, 2021). Instances of relations between concepts can be classified into different types of semantic relations (Bejar et al., 1991). For example, instances of the part-whole relation connect objects to their constituent components (e.g., a bicycle has wheels, a guitar has strings), while instances of the hyponym-hypernym relation connect objects to their taxonomic categories (e.g., a bicycle is a type of vehicle, a guitar is a type of musical instrument). Behavioral studies have shown that distinctions between semantic relations are reflected in human similarity judgements (Chaffin & Herrmann, 1984) and can explain human performance on analogical reasoning tasks (Bejar et al., 1991). Semantic relations have been incorporated in classical models of human semantic memory (e.g., Norman & Rumelhart, 1975; Miller & Fellbaum, 1991), and recent work has shown that these semantic relations are also stored in the weights of artificial language models (Bouraoui et al., 2019; Chen et al., 2021; Hernandez et al., 2024). However, it is unclear how semantic relations are processed in the brain.

One possibility is that representations in the brain directly reflect behaviorally derived distinctions between semantic relations. In this case brain representations will generalize across different instances of the same semantic relation, even if each instance involves different objects. Alternatively, brain representations could be organized according to other factors, such as the objects involved in each relation.

A few functional neuroimaging studies have compared brain responses to different semantic relations (Chiang et al., 2021; Wang et al., 2021). However, in those studies different objects were used for different semantic relations. Thus, it is unclear whether observed differences in brain responses reflect distinctions between semantic relations, or distinctions between the objects involved in each semantic relation.

Here we designed a study to investigate brain representations of different semantic relations. Six participants each performed over 1000 trials of a relation-verification task while brain responses were recorded with functional magnetic resonance imaging (fMRI). In each trial participants answered a question about one of six semantic relations. Crucially, the same set of 60 objects was used across all six semantic relations. Voxelwise encoding model weights were used to describe selectivity for each semantic relation at the highest spatial resolution available in the data (Figure 1). We find that the estimated model weights accurately predict brain responses to held-out trials throughout much of the temporal, parietal, and prefrontal cortices. Our results suggest that throughout much of the cerebral cortex brain representations generalize across different instances of the same semantic relation.

Methods

fMRI was used to record blood-oxygen-level dependent (BOLD) activity while six participants each performed over 1000 trials of a relation-verification task. In each trial, a triple of words was shown one at a time via rapid serial visual presentation (RSVP; (Forster, 1970)). Each triple consisted of a semantic relation (e.g., "hypernym"), an object (e.g, "bicycle"), and a potentially related term (e.g, "vehicle"). In each trial the participants pressed a button to indicate whether the triple formed a valid instance of a semantic relation. Half of the trials were valid instances of a semantic relation. Trials included instances of six semantic relations: "hypernym" (e.g., bicycle-vehicle), "location" (e.g., bicycle-garage), "part" (e.g., bicycle-wheel), "symbol" (e.g., bicycle-freedom), "purpose" (e.g., bicycle-transportation), and "material" (e.g., bicycle-aluminum). Trials also included instances of two non-semantic relations: alphabetical ordering (e.g., bicycle-shirt) and wordform match (e.g., bicycle-bicycle).

Voxelwise modeling (VM) was used to model BOLD responses (Wu et al., 2006; Naselaris et al., 2011). First, for

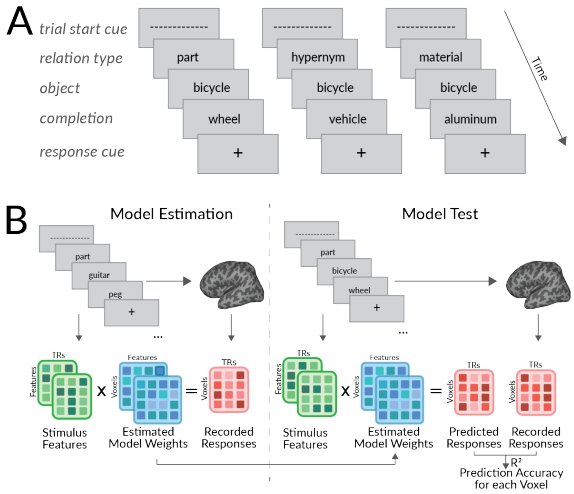

Figure 1: **A**. Experimental paradigm. Participants performed an event-related relation processing experiment while fMRI was used to record BOLD responses. Three example trials are shown. In each trial three words were displayed: a semantic relation (e.g., "part"), an object (e.g., "bicycle"), and a potential completion term (e.g., "wheel"). The participant was instructed to press a button after each triple to indicate whether the triple forms a valid relation. **B**. Modeling framework. Binary encoded stimulus features were constructed to describe the semantic relation of each trial. VM was used to estimate model weights that describe how the type of semantic relation modulates BOLD responses in each voxel. A held-out test set was used to evaluate prediction accuracy.

each semantic relation we constructed a binary feature space that reflects when participants performed trials for that semantic relation. Then we used banded ridge regression to estimate model weights that map from the semantic relation feature spaces to BOLD responses (Nunez-Elizalde et al., 2019; Dupré la Tour et al., 2022). Nuisance feature spaces were included to account for the effect of non-semantic relations, the lexical semantics of each presented word, response times, and visual and motor features. Model weights were estimated separately for each voxel and participant. To evaluate model generalization, estimated model weights were used to predict BOLD responses to a held-out test set that contained triples not used for model estimation. Prediction accuracy was computed as the coefficient of determination (R^2) between predicted and true BOLD responses to the held-out test set. A permutation test with 1000 iterations was used to compute the statistical significance of prediction accuracy. In each iteration the timecourse of true BOLD responses in the held-out test dataset was shuffled in blocks of 10 TRs, and then prediction accuracy was computed between the predicted and shuffled BOLD responses. The permuted prediction accuracies were used as a null distribution to obtain the p-value of prediction accuracy for each voxel separately. The product measure was used to decompose the prediction accuracy of each voxel into

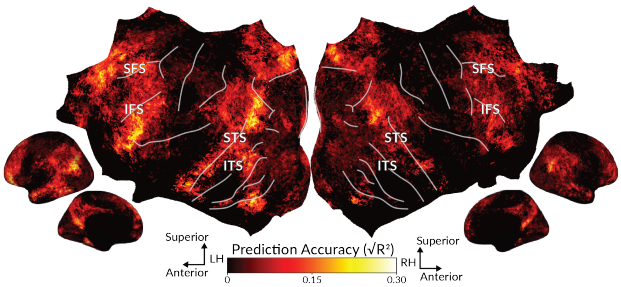

Figure 2: Prediction accuracy of the semantic relation model weights. Group-level results are shown on the flattened surface of the fsAverage brain. Vertex color indicates prediction accuracy. Vertices that were significantly predicted in less than one third of the participants are shown in black. Prediction accuracy is high throughout temporal, parietal, and prefrontal cortices. Thus, brain representations generalize across different instances of each semantic relation.

the contribution of each feature space (Hoffman, 1960; Pratt, 1987; Dupré la Tour et al., 2022; St-Yves & Naselaris, 2018; Chen et al., 2024). To summarize prediction accuracy across the group, results for each participant were projected to a standard template space (fsAverage; (Fischl et al., 1999)) and then the mean over participants was computed for each vertex. To ensure generalization to new participants, data for two participants were not analyzed until the entire experiment and model estimation pipeline was finalized.

Results

To determine whether brain representations generalize across different instances of each same semantic relation, we test how well the semantic relation feature spaces can predict brain responses to held-out instances of each semantic relation. Figure 2 shows the group-level prediction accuracy of the semantic relation feature spaces. Vertices throughout bilateral temporal, parietal, and prefrontal cortices are significantly well predicted ($p < .05$ after false discovery rate correction; (Benjamini & Hochberg, 1995)). Semantic relation feature spaces accurately predict brain responses to held-out trials. Thus, distinctions between the six semantic relations are reflected in brain representations.

Conclusion

Here we compared brain responses to different semantic relations. The type of semantic relation accurately predicted brain responses throughout much of the semantic system. Thus, brain representations generalize across different instances of the same semantic relation. This result suggests that semantic processing in the brain is organized by the type of semantic relation, rather than merely the specific objects involved in each relation. We suggest that separating representations of semantic relations from the specific concepts involved in each instance of a relation enables flexible extension of semantic relation processing to new objects.

Acknowledgments

CC was supported by the National Science Foundation (Nat-1912373 and DGE 1752814) and an IBM PhD Fellowship. FD was supported by the Federal Ministry of Education and Research (BMBF 01GQ1906) and the Berliner ChancengleichheitsProgramm (BCP).

References

Bejar, I. I., Chaffin, R., & Embretson, S. (1991). *Cognitive and psychometric analysis of analogical problem solving*. Springer Science & Business Media.

Benjamini, Y., & Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal statistical society: series B (Methodological)*, *57*(1).

Bouraoui, Z., Camacho-Collados, J., & Schockaert, S. (2019). Inducing relational knowledge from bert. In *Aaai conference on artificial intelligence.*

Chaffin, R. (1988). The nature of semantic relations: a comparison of two approaches. In *Relational models of the lexicon.*

Chaffin, R., & Herrmann, D. J. (1984). The similarity and diversity of semantic relations. *Memory & Cognition*, *12*.

Chen, C., Dupré la Tour, T., Gallant, J. L., Klein, D., & Deniz, F. (2024). The cortical representation of language timescales is shared between reading and listening. *Communications Biology*, *7*(1).

Chen, C., Lin, K., & Klein, D. (2021). Constructing taxonomies from pretrained language models. In *North american chapter of the association for computational linguistics.*

Chiang, J. N., Peng, Y., Lu, H., Holyoak, K. J., & Monti, M. M. (2021). Distributed code for semantic relations predicts neural similarity during analogical reasoning. *Journal of Cognitive Neuroscience*, *33*(3).

Dupré la Tour, T., Eickenberg, M., Nunez-Elizalde, A. O., & Gallant, J. L. (2022). Feature-space selection with banded ridge regression. *NeuroImage*, *264*. doi: 10.1016/j.neuroimage.2022.119728

Fischl, B., Sereno, M. I., Tootell, R. B., & Dale, A. M. (1999). High-resolution intersubject averaging and a coordinate system for the cortical surface. *Human brain mapping*, *8*(4).

Forster, K. I. (1970). Visual perception of rapidly presented word sequences of varying complexity. *Perception & psychophysics*, *8*(4).

Hernandez, E., Sharma, A. S., Haklay, T., Meng, K., Wattenberg, M., Andreas, J., ... Bau, D. (2024). Linearity of relation decoding in transformer language models. *International Conference on Learning Representations*.

Hoffman, P. J. (1960). The paramorphic representation of clinical judgment. *Psychological bulletin*, *57*(2).

Miller, G. A., & Fellbaum, C. (1991). Semantic networks of english. *Cognition*, *41*(1-3).

Naselaris, T., Kay, K. N., Nishimoto, S., & Gallant, J. L. (2011). Encoding and decoding in fMRI. *Neuroimage*, *56*(2).

Norman, D. A., & Rumelhart, D. E. (1975). Explorations in cognition.

Nunez-Elizalde, A. O., Huth, A. G., & Gallant, J. L. (2019). Voxelwise encoding models with non-spherical multivariate normal priors. *Neuroimage*, *197*.

Pratt, J. W. (1987). Dividing the indivisible: Using simple symmetry to partition variance explained. In *Proceedings of the second international tampere conference in statistics, 1987.*

St-Yves, G., & Naselaris, T. (2018). The feature-weighted receptive field: an interpretable encoding model for complex feature spaces. *NeuroImage*, *180*.

Unger, L., & Fisher, A. V. (2021). The emergence of richly organized semantic knowledge from simple statistics: A synthetic review. *Developmental Review*, *60*.

Wang, W.-C., Hsieh, L.-T., Swamy, G., & Bunge, S. A. (2021). Transient neural activation of abstract relations on an incidental analogy task. *Journal of Cognitive Neuroscience*, *33*(1).

Wu, M. C.-K., David, S. V., & Gallant, J. L. (2006). Complete functional characterization of sensory neurons by system identification. *Annu. Rev. Neurosci.*, *29*.