Between-Subjects Neural Decoding of Phrase Representations Outperforms Word2Vec Decoding

Elizabeth A. Shay[1], Benjamin D. Zinszer[2], Rajeev D. S. Raizada[1]
[1]University of Rochester, [2]University of Texas at Austin

Introduction

Semantic decoding

- High levels of consistency in representations for individual words between subjects (e.g., Haxby et al., 2011; Raizada & Connolly, 2012)
- Non-neural representational models of individual words can decode neural representations of words *and sentences* (Anderson et al., 2016)
- Little work has been done looking at consistency of high-level representations across participants

Semantic composition

- Understanding composition has implications for variety of fields outside cognitive neuroscience
- Previous work debates best composition function (mathematical combinations) for words (treated as vectors)
- Variety of successful options (e.g., addition in Anderson et al., 2016; multiplication in Chang et al., 2009)

This study

- We apply Raizada and Connolly's (2012) similarity-space approach to explore between-subjects consistency of adjective-noun phrases
- We compare between-subjects decoding to model-based decoding using element-wise addition (e.g., Anderson et al., 2016) to model composition in phrases, using Word2Vec (Mikolov et al., 2013)

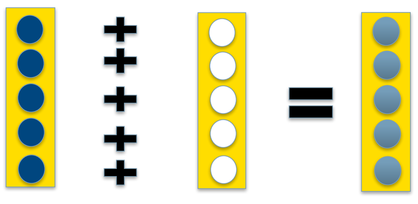
Methods

Participants & Stimuli

- fMRI scans, 15 participants reading 60 phrases
  - 6 adjectives: 4 colors, 2 sizes
    big, small, black, green, red, white
    (taken from Baroni & Zamparelli, 2010)
  - 10 nouns: 5 body parts, 5 buildings
    arm, eye, foot, hand, leg, apartment, barn, church, house, igloo
    (taken from Mitchell et al., 2008)

Analyses

- *n*-fold (leave-one-subject-out) and *k*-fold cross validations for neural decoding
- Pairwise decoding for each pair of phrases
- Model-based similarity matrix created using Word2Vec (Mikolov et al., 2013) vectors and element-wise addition (e.g., Anderson et al., 2016) to create phrases



Element-wise addition process

- Same analyses performed on each region in the Harvard-Oxford Atlas (96 regions, Desikan et al., 2006)

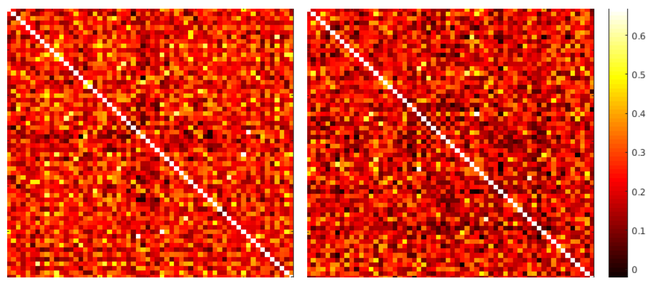
This work was conducted while EAS was partially funded by the Center for Language Sciences, University of Rochester

Contact: EAS – eshay@bcs.rochester.edu

Results

Between-Subjects Decoding

- Average LOOCV accuracy:
  95% (chance = 50%, $p < 0.001$)
- Average *k*-folds cross-validation (n=8 training set, n=7 test set, 6435 folds):
  99% (chance = 50%, $p < 0.001$)

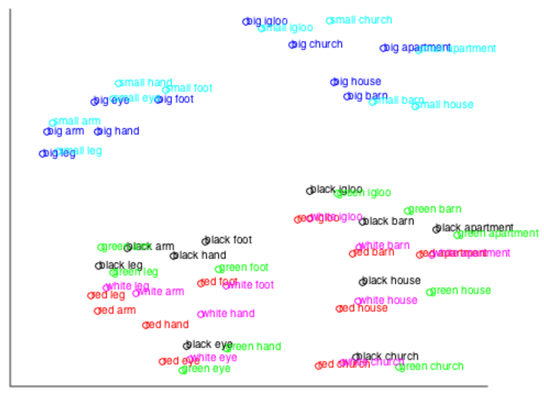


Heat map of brain phrase similarity matrix for an example training set (left) and test set (right) for *k*-fold cross-validation

Word2Vec (Semantic Model)-Based Decoding

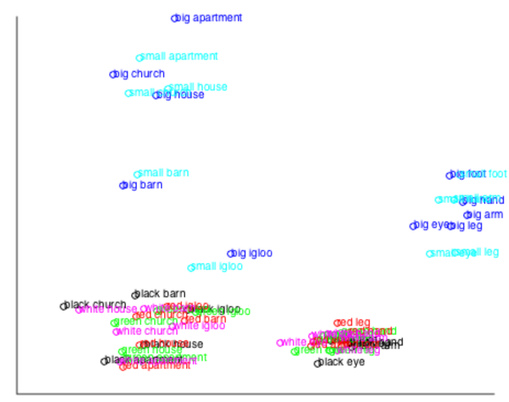
- Mean pairwise decoding for each subject:
  37% (chance = 50%, $p < 0.05$)
- Pairwise decoding of average across subjects:
  37% ($p < 0.05$)

MDS Plot of Word2Vec Phrases
(averaged adjective and noun vectors)



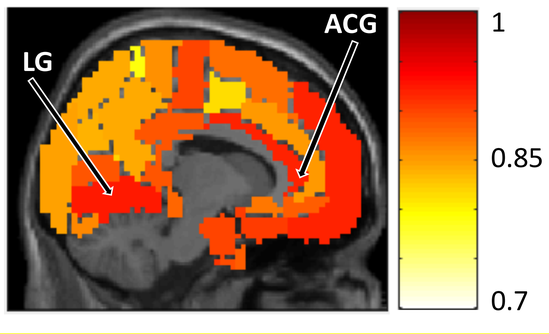
Adjective Coloring: Big Small Black Green Red White

MDS Plot of Neural Phrase Representations



Harvard-Oxford Atlas Region Analysis

- All regions produced above-chance decoding between subjects
- High accuracy regions include bilateral lingual gyrus (LG) and bilateral anterior cingulate gyrus (ACG)



Slice x=12

Left LG = 92%
Right LG = 93%
Left ACG = 90%
Right ACG = 92%

Discussion

Explaining the High Decoding Rates

- Decoding accuracy at the phrase level can be partly explained by successful discrimination of only the noun or adjective in a phrase.
  Ex. BIG HOUSE vs. RED HAND could be discriminated on the basis of color vs. size or building vs. bodypart, even if the individual words are not successfully decoded
- Knowing the category of the adjective/noun is sufficient for 75% of phrase pairs
- Adjective-only (word-level) between-subjects decoding was around 85% ($p < 0.05$)
- Noun-only (word-level) between-subjects decoding around 41% ($p > 0.05$)
- Remaining decoding accuracy comes from the interaction between adjective and noun in the phrase

Composition

- Element-wise addition of Word2Vec doesn't translate to neural phrase representations
  - Likely composition function choice
  - Average of Word2Vec representations brings pairwise decoding to 57% ($p < 0.05$)
- Between-subjects decoding accuracy far exceeds accuracy achieved using Word2Vec composition representations
  - Brain-to-brain regularities are far greater than accounted for by Word2Vec composed representations
- LG and ACG previously implicated in conceptual combination and more active in response to conceptually complex of confusable items (Baron & Osherson, 2011)
  - High level of overlap in words and categories in phrases necessitated a fine-grained conceptual discrimination to accurately classify the phrases
  - Subtle conceptual combinations may underlie the successful decoding between-subjects

Conclusions

- Extends Raizada & Connolly (2012): other participants remain the best model of decoding another individual's brain
- Supports important assumption in cognitive neuroscience: multivariate patterns of neural activity describe real representational regularities across brains of individuals
- Element-wise addition (at least of Word2Vec) is not a good candidate for composition of adjective and noun representations in the brain

REFERENCES

Anderson et al. (2016). *Cereb Cortex,* 1-17.
Baron & Osherson (2011). *NeuroImage, 55*(4), 1847-1852.
Baroni & Zamparelli (2010). *EMNLP-ACL 2010* (pp. 1183-1193).
Chang et al. (2009). *IJCNLP f the AFNLP* (pp. 638-646).
Desikan et al. (2006). *NeuroImage, 31*, 968-980.
Haxby et al. (2011). *Neuron, 72*, 404-416.
Mikolov et al. (2013. *Adv. neu. info. proc.* (pp. 3111-3119).
Mitchell et al. (2008). *Science, 320*, 1191-1195.
Raizada & Connolly (2012). *Cognitive Neurosci, 24*(4), 868-877.