Christopher J. Diak
PSY2554R: Complex Thought and Cooperation
Professor Joshua D. Greene
Harvard University
12/23/2020

Infinite Use of Finite Means:

The Promise and Limits of the Cognitive Neuroscience of Language

What do we learn from fMRI studies of linguistic comprehension and production that behavioral, computational, and theoretical linguistics cannot teach us? Uttal (2001) and Coltheart (2006) provided substantive critiques of the interpretation of functional neuroimaging data for high-level cognitive processes, centering on the concepts of *localization* and *implementation*, respectively. In light of these critiques, here we examine a recent functional neuroimaging study of verb- patient binding in human left-middle temporal gyrus (Frankland and Greene, 2019). Building on previous studies of abstract semantic variable binding in left mid-superior temporal cortex (Frankland and Greene, 2015) the authors conducted two fMRI experiments aiming to understand the neural representation of sentence structure and meaning. While the authors acknowledge that there is no necessary isomorphism between the "abstract, symbolic descriptions useful for linguistic analysis and the neural representations over which computations are performed," they proceed with a method designed to tease out the patterns of activation that would be elicited by two distinct symbolic levels of representation. These two levels respectively encode 1) variable-binding of nouns to role (i.e. agent or patient) and 2) whether the arguments of semantic variables are internal or external to the verb-phrase. Based on the activation patterns observed in the left middle temporal gyrus and the known association of this area with sentence processing (Dronkers et al. 2004, Hickok and Poeppel, 2007), the authors take positions simultaneously on a theoretical issue in linguistics and the neural representation of meaning in the brain. Our task in this paper is to examine the methodological assumptions that would justify such a move and ask whether it is warranted in light of the critiques listed above from Uttal and Coltheart.

The Problem of Localization

In the opening chapter of *The New Phrenology*, Uttal claims—wrongly I believe—that the very posing of the mind-body problem arises from a particular set of philosophical assumptions (i.e. monism, naturalism, materialism) that must be accepted—explicitly or implicitly—by cognitive neuroscientists:

> *Simply put, every study of the localization issue and every theory about it is premised on the idea that variations in the psychological domain are in some very direct way related to variations in the neurological domain. Make no mistake, such directness is tantamount to identity. To believe otherwise is to deny a tight enough correlation between brain activity and measures of mind to permit drawing any conclusions about the ways in which one influences the other.*[1]

By positioning cognitive neuroscience in this way, Uttal is making two interrelated intellectual moves. First, he is making a *descriptive* claim about the nature of cognitive neuroscience as a field dedicated to localizing cognitive processes in the brain. Secondly, he is making a *prescriptive* claim about the theoretical vocabularies with which cognitive neuroscientists ought to be concerned. Namely, those of monism, naturalism, and materialism. While we differ on the prescriptive claim, Uttal and I do agree on the description of the field of cognitive neuroscience. The question that lies at the heart of the field is whether "psychological processes can be adequately defined and isolated in a way that permits them to be associated with particular brain locales."[2] If we can't define "language" in a way that lends itself to scientific explanation *and* association with particular locales in the brain, then we concede there can be no cognitive neuroscience of language. This is why Uttal believes "the preeminent problem in achieving a general solution to the localization issue lies in defining the psychological processes and mechanisms for which loci are being sought."[3] It's a conceptual issue, not an empirical problem that fuels Uttal's critique. Fortunately for cognitive neuroscientists, "language" may be precisely the kind of cognitive process that is amenable to formal exposition in terms of mathematical laws on the one hand, and neurobiological localization on the other.

---

[1] Uttal, W.R. (2001). *The New Phrenology: The Limits of Localizing Processes in the Brain*, 4. MIT Press.
[2] Ibid, 1.
[3] Ibid, 16.

It is precisely our difference on the question of which theoretical vocabularies *ought* to concern cognitive neuroscientists which leads to my fundamental disagreement with Uttal about the prospects and limits of the field as described. After Turing (1950), Fodor and Pylyshyn (1988) and Chomsky (1955, 1982, 1992), I'm committed to conceptualizing the brain in terms of a computational architecture that functions like a classical computer at the cognitive level for certain cognitive processes, such as language and other forms of generative or compositional thought. In other words, I'm committed to a using a theoretical vocabulary that asserts the existence of a symbolic level of mental representation for language and complex thought. Following Fodor and Pylyshyn (1988) and Frankland and Greene (2019) I reiterate that there is no (known) necessary isomorphism between the "abstract, symbolic descriptions useful for linguistic analysis and the neural representations over which computations are performed." And this fact lies at the heart of Uttal's critique of the ubiquity of the localization thesis in cognitive neuroscience. Higher-order cognitive processes like language and the computational thereof may be instantiated in the brain in a manner that more closely resembles a connectionist network, even if they compute over discrete, functional units called symbols. In response to a study that attempts to locate a mechanism for a particular cognitive function in the prefrontal cortex, Uttal suggests

> *a totally different conceptualization of the localization problem, one that offers, in place of a specific function being precisely localized (i.e., instantiated, represented, or encoded) in a particular place, the idea of one center contributing to the operation of a complex system of nodes and loci that are collectively responsible for the behavior.*[4]

While even Fodor and Pylyshyn (1988) admit it is possible that a classical cognitive architecture could be implemented in a connectionist network, Uttal's formulation presents us with another serious problem for the cognitive neuroscience of language. If the brain *does* use a connectionist body to implement a classical mind, can facts about the physical constitution of the brain be used to draw inferences about psychological or linguistic theory?

[4] Ibid, 18.

The Problem of Implementation

Coltheart (2006) suggests that for the time being the answer is "no" or "not yet." Rather than ask whether psychological processes can be defined in such a way as to allow localization in the brain, Coltheart asks whether there are facts about the brain which are relevant to evaluating psychological theories.[5] Psychological and linguistic theories tend *not* to posit empirical hypotheses about the structure or organization of the brain. Standard practice in cognitive neuroscience moves in precisely the opposite direction, beginning with a theoretical commitment about some cognitive process, proceeding with an attempt to localize that process, predicated on a non-neurobiological theory, in accordance with the strictures set by the established cognitive neuroscience literature. In the case of Frankland and Greene (2019), the region of interest (left MTG) was selected on the basis of prior literature (Pallier et al., 2011; Frankland and Greene, 2015; Dehaene et al., 2015; Fedorenko et al. 2017; Nelson et al., 2017; Ding et al. 2017) suggesting cortical regions around the left sylvian fissure are involved in compositional thought. The primary finding of their study is described as a representational asymmetry in middle temporal gyrus that catalogs 1) the differences between nouns in the role of agent or patient and 2) whether the argument of a semantic variable is internal or external to the verb-phrase. Coltheart would point out that the theoretical models upon which the interpretation of their results rest (Williams, 1981; Marantz, 1984; Grimshaw, 1990; Kratzer, 1996) do *not* make empirical predictions about the left middle temporal gyrus. If we represent two linguistic theories with T_a and T_b and allow T_a to be a theory that marks a distinction between *agent* and *patient* semantic variable types concerning whether they tend to be external or internal to the verb phrase, and T_b does *not* make this distinction, do Frankland and Greene's (2019) results warrant rejecting T_b? Or, put less forcefully, do the facts about the brain presented in Frankland and Greene (2019) have any relevance to constructing or evaluating psychological or linguistic theories? Following Block (1990) and Morton (1984), Coltheart voices skepticism on the subject.

[5] Coltheart, M. (2006). "Perhaps Functional Neuroimaging Has Not Told Us Anything About the Mind (So Far). *Cortex*, 42, 422.

This skepticism toward cognitive neuroscience has a structural parallel in the arguments that Fodor and Pylyshyn (1988) advanced to deflate connectionism as a theory of cognitive architecture. Just as we are asking whether facts about the organization of matter in the brain have anything to say about psychological theory, Fodor and Pylyshyn questioned whether the existence of a physically connectionist architecture in the brain had any bearing on the question of the language of thought hypothesis. "It is, in particular, perfectly possible that nonrepresentational neurological states are interconnected in the ways described by Connectionist models *but that the representational states themselves are not*," they wrote.[6] While analogies can always be misleading, we should ask whether we are in an analogous situation with respect to current study. Are the patterns of activation identified in the MTG by Frankland and Greene (2019) in principle consistent with the implementation of semantic theories that do not involve a representational asymmetry between agent and patient roles? To decide this question let's turn to a critical analysis of the experimental methodology.

In the first experiment, 25 human subjects underwent high-resolution fMRI while reading simple sentences of the form AGENT-VERB-PATIENT, constructed from a stock of 5 verbs and 4 nouns. The sentences were generated such that two distinct nouns always occupied the positions of agent and patient, yielding 60 total propositions. A machine learning algorithm called a linear discriminant classifier was trained by sorting patterns of activation based on *agent* and *patient* noun roles for a particular verb-role conjunction. In other words, the training data consisted of sentences with only one verb and the four nouns (dog, cat, man, girl) in each role relative to the verb. The classifier was then tested on the patterns of activation elicited by stimuli containing the remaining verbs. Separate classifier algorithms were trained on each verb. The maps generated by each classifier were then averaged into a single map of activation patterns. The resulting statistical analysis therefore was designed to identify patterns of activity elicited only by verb-agent or verb-patient conjunctions, regardless of the particular verb or noun. Only the results for verb-patient conjunctions were statistically significant. By designing the experiment in this

[6] Fodor, J. and Pylyshyn, Z. (1988). Connectionist Cognitive Architectures: A Critical Analysis. *Cognition*, 28 1-2, 6.

way, the authors claimed any statistically significant patterns of activation would encode information about the verb-role conjunctions in question. In the second experiment, the authors attempted to "predict the BOLD signal in MTG as participants read novel propositions that are composed of familiar components… including models that predicted BOLD signal as a *combination* of multiple learned representations."[7] The authors used models composing 1) verb-patient and general agent parameters and 2) verb-agent and general patient parameters to predict BOLD signals on novel sentences and found that only the first model elicited statistically significant patterns of activation in MTG. This asymmetry was taken to encode a representational difference between the relationship of the verb-to-agent and verb-to-patient in this region of the temporal cortex.

Cognition and Computation

The question of how information is encoded in neural activity — and the mechanisms involved in encoding at different physical scales — is a major unsolved problem in neuroscience. If it is the case that the way information about verb-patient and verb-agent conjunction is encoded in the brain is accessible via fMRI at this scale, then the patterns of activity may very well reflect the neurological structures which implement the computations underlying that encoding process. And if we conceive of the brain from a computational-representational perspective, meaning, that we take an abstracted view of the kinds of processes that must be at work to encode and process the information that goes into the cognitive processes we care about, then we can attempt to unify the vocabulary of the neurosciences with that of theoretical linguistics or psychology. Coltheart's critique asks us to build theories and hypotheses with specifically neurological consequences that can be falsified. While it is unclear whether the authors have succeeded in doing so here, the kind of work in which they are engaged could be fruitful for future theorists and experimenters who wish to construct a unified discipline of computational neurolinguistics. Cognitive neuroscientific theories of language developed in terms of computational architecture would generate empirical predictions about both the linguistic structures we might predict *a priori* and the kinds of neural mechanisms that could support and implement those computations.

[7] Frankland and Greene (2019). 3.

References

Coltheart, M. (2006). "Perhaps Functional Neuroimaging Has Not Told Us Anything About the Mind (So Far). *Cortex*, 42.

Fodor, J.A., Pylyshyn, Z.W. (1988). *Connectionism and cognitive architecture: a critical analysis.* Cognition 28 (1-2), 3-71.

Frankland, S.M., Greene, J.D. (2019). *A representational asymmetry for composition in the human left-middle temporal gyrus.* 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

Uttal, W.R. (2001). *The New Phrenology: The Limits of Localizing Processes in the Brain*. MIT Press.