Challenges for using Representational Similarity Analysis to Infer Cognitive Processes: A Demonstration from Interactive Activation Models of Word Reading

Xuanyi Chen (Xuanyi.Chen@rice.edu)
Department of Cognitive Sciences, Rice University
TX 77005, USA

Randi C. Martin (rmartin@rice.edu)
Department of Psychological Sciences, Rice University
TX 77005, USA

Simon Fischer-Baum (simon.j.fischer-baum@rice.edu)
Department of Psychological Sciences, Rice University
TX 77005, USA

Abstract

Representational Similarity Analysis (RSA) is a powerful tool for linking brain activity patterns to cognitive processes via similarity, allowing researchers to identify the neural substrates of different cognitive levels of representation. However, the ability to map between levels of representation and brain activity using similarity depends on underlying assumptions about the dynamics of cognitive processing. To demonstrate this point, we present three toy models that make different assumptions about the interactivity within the reading system, (1) discrete, feedforward, (2) cascading, feedforward and (3) fully interactive. With the temporal resolution of fMRI, only the discrete, feedforward model provides a straightforward mapping between activation similarity and level of representation. These simulations indicate the need for a cautious interpretation of RSA results, especially with processes that are highly interactive and with neuroimaging methods that have low temporal resolution. The study further suggests a role for fully-fleshed out computational models in RSA analyses.

Keywords: representational similarity analysis; reading, computational models; interactive activation; cascading activation

Introduction

Representational Similarity Analysis (RSA) was first proposed by Kriegeskorte, Mur, and Bandettini in 2008 to bridge the major branches of systems neuroscience by calculating the second-order isomorphism of neural activation patterns, behavioral measures, and/or conceptual and computational models. The method is based on the assumption that items similar to each other at one level of representation should elicit similar neural activity patterns in the brain region responsible for processing that level of representation. One advantage of this technique has been its ability to understand the different kinds of representations the brain uses to process the same stimuli in the same task, by looking at changes in the representational dissimilarity matrix (RDM), or a matrix composed of the distance of neural responses for each stimulus pair, across different brain regions or at different points in time. RSA has been applied to a variety of neuroimaging methods including fMRI, EEG, MEG and ECoG (Kriegeskorte et al., 2008; Cichy & Pantazis, 2017; Chen et al., 2016), and has been widely used to study different cognitive capacities like vision, audition, language, memory, and emotion (see Kriegeskorte & Kievit, 2013 for review).

An appeal of RSA is that these differences in neural similarity structure can be related to different levels of processing in a cognitive model. As a result, RSA may be able to provide a powerful tool for linking brain activity to cognitive operation. Consider, for example, the relatively simple task of reading aloud a single word. This task requires different cognitive levels of representation associated with the stimulus, including an orthographic representation of its spelling, a semantic representation of its meaning, and a phonological representation of its associated pronunciation. At these different levels of representation, different word pairs are represented similarly to each other. The word DOUGH is similar to the word TOUGH at an orthographic level, the word BREAD at a semantic level, and the word SEW at a phonological level. If an fMRI experiment finds a cortical region whose RDM includes low distance – or high similarity – in the pattern of brain activity in response to DOUGH and in response to BREAD, we may be inclined to take this result as evidence that the region is engaged in semantic processing (Fischer-Baum et al., 2017).

However, what we will show in the current paper is that interpreting the link between neural similarity and cognitive level of processing using RSA critically depends on assumptions about the dynamics of cognitive processing. The logic described above holds for discrete, feedforward theories of cognitive processing, in which processing occurs at a single stage, until a selection event occurs and information moves forward to the next level of processing. However, contemporary theories of cognition rarely hold this stage-wise view of processing (Rogers & McClelland, 2014). Instead, theories either assume cascading activation, in which representations at each level activate associated

representations at the next level prior to a selection event (e.g., Goldrick & Blumstein, 2006), or fully interactive activation, in which representations at each level can be activated by both bottom-up connections from earlier levels of representation and top-down connections from later levels of representation (e.g., McClelland & Rumelhart, 1981; Coltheart et al., 2001). As we will show in our modelling work below, with these more complex forms of cognitive dynamics, it becomes more challenging to use similarity to link cognitive levels and neural systems.

This challenge is particularly great for neuroimaging methods that have coarse temporal resolution, like fMRI. Consider again the case of reading aloud a single word, which can be completed in approximately 600ms. Because of both the relatively slow acquisition of fMRI data and the temporal sluggishness of the BOLD signal, whatever measures of similarity we are collecting using this imaging modality are aggregating across the entire process. Interactions between levels of representation that may have different time signatures, for example a level that first receives bottom-up input from the lower-level features of the stimuli and later receives top-down input from later levels of representation, can be washed out in the aggregation process. Here, we focus on the domain of word reading and demonstrate how the interactive nature of cognitive processing and the use of temporally low-resolution neuroimaging methods can result in misleading conclusions based on RSA.

Specifically, we examine recent results using RSA to understand the role of the visual word form area (VWFA) in word reading (Fischer-Baum et al., 2017; Taylor, Davis & Rastle, 2019). The VWFA, located in left ventral occipitotemporal gyrus (lvOT), has been demonstrated to respond to sequences of letters in written words and pseudowords in an abundance of univariate neuroimaging studies (e.g. Dehaene & Cohen, 2011), though there continues to be disagreement about the precise cognitive function that region subserves (e.g. Price & Devlin, 2011). In order to investigate the level of processing carried out by lvOT, Fischer-Baum et al. (2017) used RSA on fMRI data of word reading and found that the neural RDM in lvOT correlated with measures of both orthographic and semantic similarity (see also Wang et al., 2018 and Taylor, Davis & Rastle, 2019). From these results, shown in Figure 1, one may be tempted to conclude that lvOT instantiates both an orthographic and a semantic level of representation. However, Fischer-Baum and colleagues (2017) hypothesized that the correlation between activation patterns in lvOT and semantics could be explained by an interactive processing account in which the lvOT encodes only the cognitive level of orthographic lexical representations. Under this hypothesis, the semantic information feeds back to the orthographic level and activates orthographic lexical representation of semantically related words, over time changing the similarity structure of the space to appear semantic in nature. Fischer-Baum and colleagues further note that this kind of interactive processing is assumed by many computational models of word reading, dating back to the seminal work by McClelland and Rumelhart (1981).

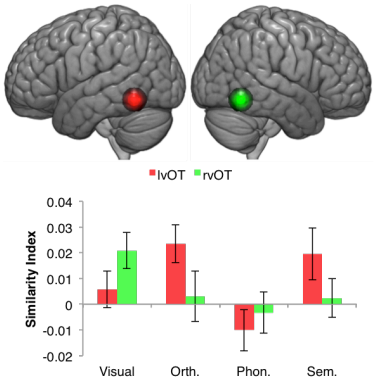

Figure 1: Results from Fischer-Baum et al., 2017. Note that the lvOT shows significant correlation with both orthographic and semantic measures of similarity.

Here, we test this account computationally. We built three simplified models of the reading system, designed to maximally distinguish orthographic, phonological, and semantic representations. These models were identical in their underlying structure, but differ in their dynamics. We then applied RSA to each level of the model, both over time and in an aggregate measure that we take to reflect what is being collected with fMRI. For our specific question of interest, we investigated whether an interactive model with feedback connections from the semantic layer to the orthographic layer would result in significant correlations between the activations in the orthographic layer and the underlying semantic similarity, as predicted by Fischer-Baum et al. (2017). We more generally examined how different kinds of cognitive dynamics can complicate the use of similarity to identify the cognitive level of representation. These results have broad implications for how cognitive neuroscientists can use RSA as a tool for linking cognitive and neural theories – as correlations observed in RSA do not necessarily imply the level of processing carried out in the region analyzed, but could also be due to activation spreading from the downstream or upstream level representation which is localized in a different neural substrate.

Simulation

Model Design

Three models were constructed with the same structure and mappings. The models included five layers: orthographic input units, orthographic lexicon, semantic units, phonological lexicon, and phonological output units (phonemes), as shown in Figure 2a, with connections between each of the subsequent layers as well as direct connections between the orthographic and phonological lexicon. It was roughly structured following the lexical route of the DRC model (Coltheart et al., 2001) though the details

of the model were extensively simplified. Figure 2b demonstrates how English words CALL and CELL would be represented in the current model structure, although the models only included artificial lexical items. All of the models included localist coding of 8 lexical items at the levels of the orthographic and phonological lexicon. There were 12 units at the orthographic input, semantic, and phonological output levels. Lexical items were connected to three units at each of these levels with a weight of one, and had no connections to the other nine units. Every unit at the orthographic input, semantic, and phonological output levels was connected to exactly two lexical items. The mappings were constructed in a way that maximally distinguished orthographic and phonological representations. This construction does not instantiate the orthographic-phonological regularities of a semi-transparent language like English, but it allowed us to better account for the contribution of each level of representation in our analysis. The feature representations were constructed so that each lexical item had three neighbors that shared one of the three units at only one of the three theoretical levels, three neighbors that shared one unit each at two of the three levels, as well as one lexical item that did not have overlapping features with the target lexicon in all three levels. The theoretical RDMs of the orthography, phonology, and semantic representations are shown in Figure 2c, with a Spearman's rank correlation of -0.167 between each two levels of representations.

With this general structure in place, three models were constructed with different dynamics of activation. In the fully interactive model, the connections between layers were set to be bi-directional, so that an active unit in the orthographic lexicon could spread activation to its three associated semantic features, while at the same time each active semantic feature spread activation to its two associated items in the orthographic lexicon. The feedback connections from phonological output units to phonological lexicon level were set to 4 to account for the relatively low level of activation in the phonological units layer. In the feedforward, cascading activation model, there were only unidirectional, feedforward connections between layers, such that active units in the orthographic lexicon could spread their activation to the semantic features, but not vice-versa. In the feedforward, discrete model, there were only unidirectional, feedforward connections, but activation did not spread from one layer to the next until a selection event occurred. In addition, there was a level of orthographic input that had, in all models, feedforward only connections to the level of the orthographic units, with 12 nodes total and a one-to-one mapping between the nodes in orthographic input and the nodes at the level of the orthographic units. This level is set to resemble stimulus presentation.

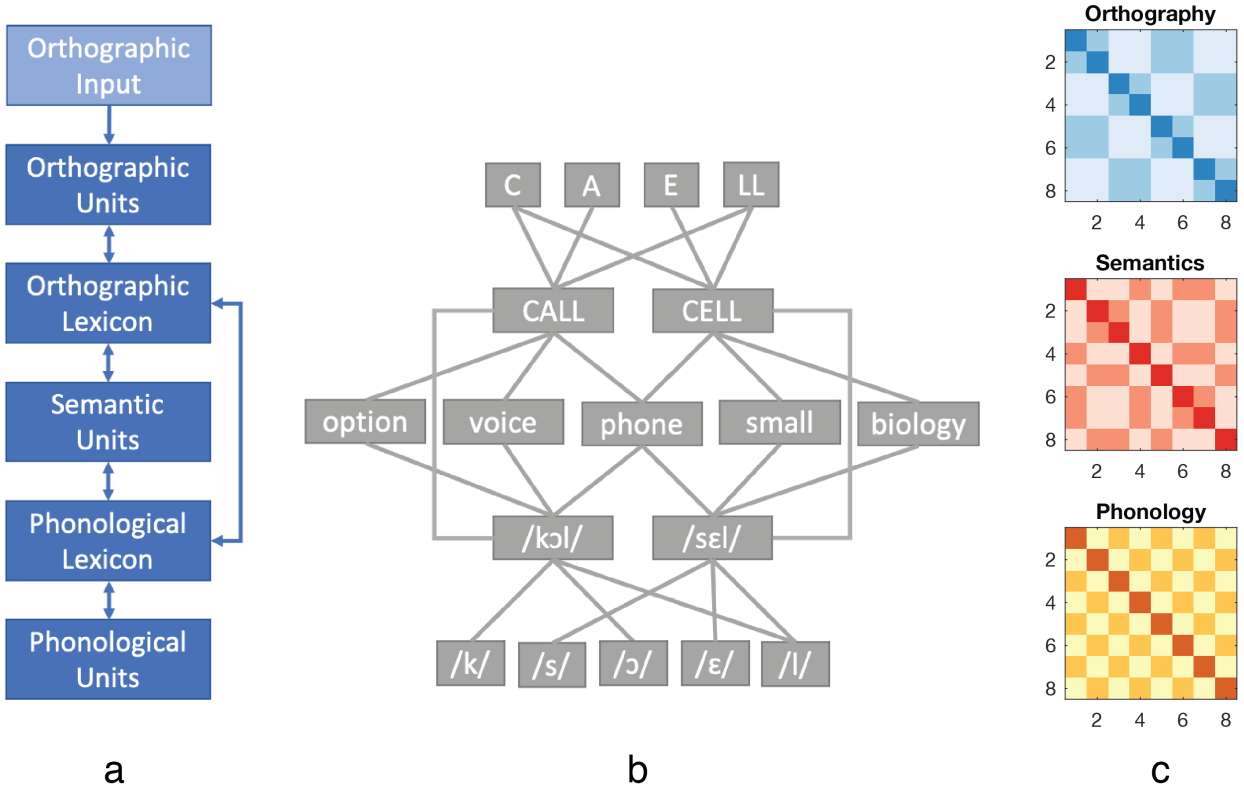

Figure 2: (a) Model structure (b) Example of English words CALL and CELL under the current model structure. Note that these words are not actually included in the models. (c) Similarity between each of the 8 lexical items at each level, with brighter colors reflecting greater similarity between items.

Procedure and Analysis

For the fully interactive and the feedforward cascading models, the process was run for 30 cycles. For the first cycle, the orthographic input was clamped on to a value of 1, after which those values were set to 0. The activation equations followed those of simple spreading activation models (e.g. Dell, 1986):

$$A(j,t_i) = [(1+v)A(j,t_{i-1}) + \sum_{k=1}^{n} A(c_k,t_{i-1})]\,(1-q)$$

where $A(j,t_i)$ is the activation level of node j at time t_i. c_k denotes the nodes that are directly connected to node j.The model further included a normally distributed noise v with a standard deviation of σ, and a decay rate of q.

For the feedforward, discrete model, information at each level accumulates for 5 time-cycles without spreading activation to other layers. For the phonological lexicon layer, which receives input from both the orthographic lexicon level and the semantic level, activation only begins to spread to this level after the selection event at both prior levels is completed. After the five cycles, the top item in the lexical layers or the top three items in the feature layers of orthographic input, semantic, and phonological output levels were selected and their activations set to 15 (for lexical layers) or 7.5, 5, and 2.5 (for feature layers) while other items were all turned to zero. The information of the selected item is then sent to the next layer for the next five cycles.

Simulation of all three models were performed with 1000 trials per word with the σ (the standard deviation of the noise) set at 0.12 and q (the decay rate) set at 0.3. The parameters were selected so that we can simulate a low but observable error rate to be comparable to reading in healthy adults. After simulation, the activation RDMs at each level and cycle were calculated using Spearman's rank correlation. Each of the five activation RDMs at each time point were then correlated with the three orthographic, semantic, and phonological RDMs based on the models' intrinsic structure, again using Spearman's rank correlation.

Subsequently, we calculated an aggregate activation measure across the entire trial time of 30 cycles to provide a measure that is more comparable to the temporally insensitive fMRI results. To account for the difference in the magnitude of raw activations across time and layers, we divided the raw activations at each time point and layer by the mean activation in that layer, and calculated the average of the scaled activation across time. We then constructed an aggregate activation RDM for each of the five levels of representation in the model and correlated each with the three theoretical RDMs.

Results

Temporal Dynamics Figure 3 reports a table of plots showing the temporal dynamics of the Spearman's correlation between each layer and theoretical RDM for the three models.

For the two spreading activation models, the activation patterns of all layers highly correlate with only the orthographic RDM at the initial stage, confirming that the correct orthography of the input stimuli was successfully passed on to the upper layers. For the fully interactive model, different theoretical RDMs showed high correlation with the observed patterns of activity in different layers across time. The orthographic and semantic layers became correlated with the semantic RDM around 10-20 cycles, and subsequently evolved to be correlated with the phonological RDM. Among the three layers, high semantic correlation was first in the semantic layer and was preserved for the longest time, followed by the orthographic lexicon layer. The orthographic units layer began to show semantic correlation at a later time point and the semantic correlation was present for only a brief period. The phonology layers became correlated with the phonological RDM at around 10 cycles, and the high correlations were preserved for the rest of the trial. Additionally, the phonology layers showed a small increase in semantic RDM at the first two cycles to a positive value, but subsequently remained relatively constant.

In comparison, the cascading, feedforward model showed no qualitative temporal change in Spearman's correlation after the first two cycles of processing. All layers were highly correlated with the orthographic RDM and the phonology layers were additionally correlated with the semantic RDM.

Finally, the temporal dynamics of the feedforward, discrete model showed a distinct pattern from the previous two models. In the three feature layers, the activation RDMs consistently correlated with the theoretical RDMs of the level of representation coded in the model, during processing and after selection. In contrast, for the two lexical layers, the activation RDM correlated with the theoretical RDM of the representation of the previous layer during the processing cycles, meaning that the orthographic lexical level was correlated with the orthographic RDM and the phonological lexical level was correlated with the semantic RDM. However, after lexical selection the activation RDMs became correlated with the orthographic RDM due to the few selection errors.

Aggregate activation measure. In order to engage a more direct comparison between the high temporal resolution modeling results and the fMRI result, correlations between the theoretical RDMs and the aggregate activation RDM of each level were computed for the three models (Figure 4). For the fully interactive model, all layers except for the phonological output layer showed initial correlation with the orthographic RDM. The orthographic and semantic layers showed an additional correlation with the semantic RDM, whereas the phonological lexicon layer showed a correlation with the phonological RDM. The phonological units layer showed a primary correlation with the phonological RDM and a secondary correlation with the orthographic RDM. Of note is that, as predicted by Fischer-Baum and colleagues (2017), the level of the orthographic lexicon shows correlation with both orthographic and semantic similarity

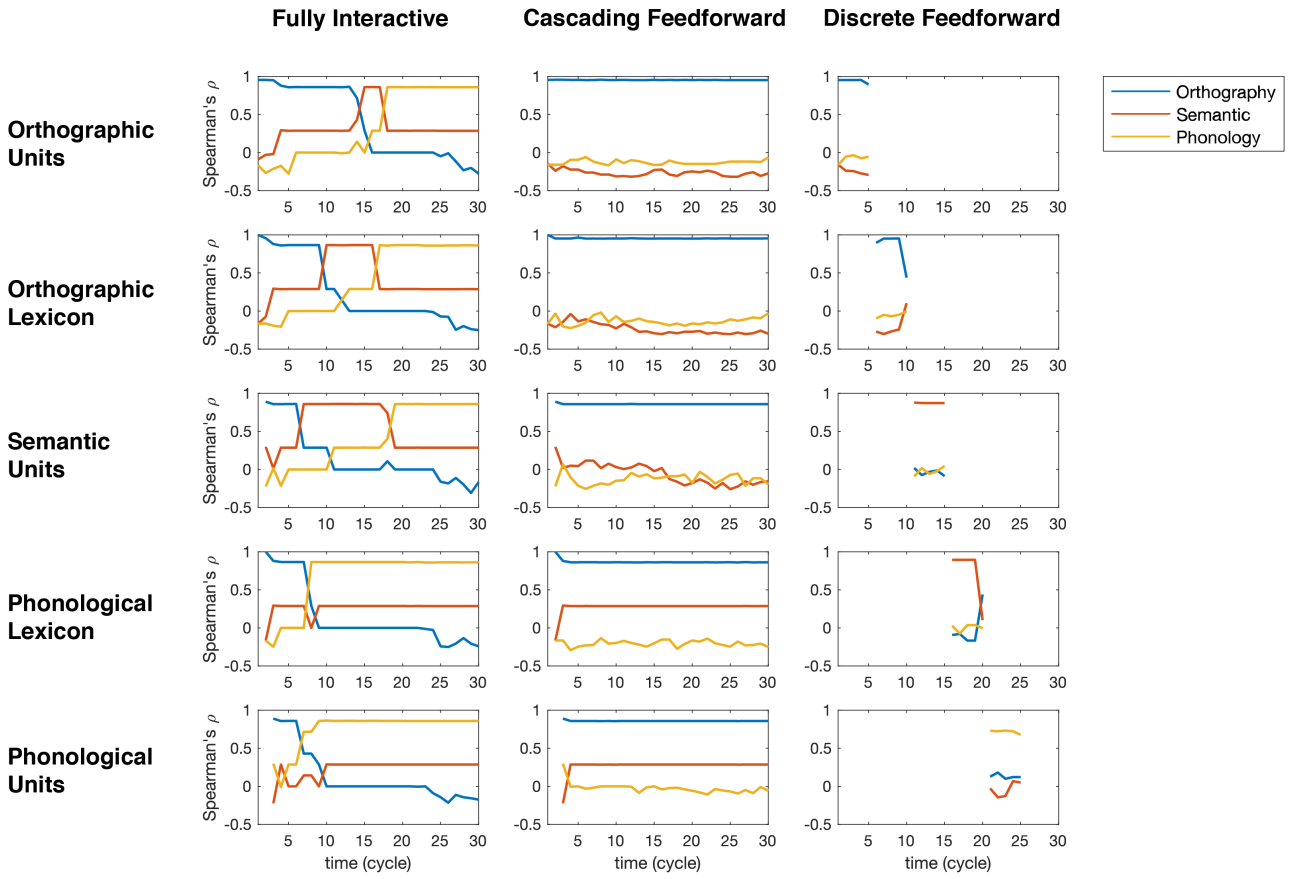

Figure 3. Temporal dynamics of RSA in the fully interactive, cascading feedforward, and discrete feedforward models.

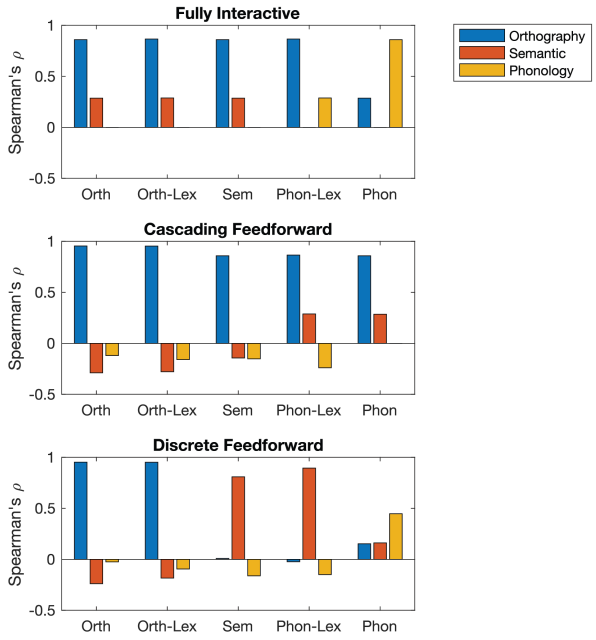

Figure 4. RSA result with an aggregate activation measure.

with the dynamics of interactive activation, though other levels show this pattern as well.

In comparison, for the feedforward, cascading model, all layers showed activation patterns highly correlated with the orthographic RDM. The phonological layers additionally showed correlation with the semantic RDM. Because of the cascading nature, the structure of earlier level of processing is imposed on later levels.

In contrast to the previous two spreading activation models, the discrete, feedforward model showed RSA results more in line with the assumption in neuroimaging literature where similarity in neural activation corresponds to the conceptual similarity of the representation in the target region. All feature layers showed activation RDMs with the expected high correlation to the theoretical RDM of their underlying representations, while the lexical layers showed high correlation with the RDM of the representation in the previous layer. The semantic and phonological units layers showed additional low, above-zero correlations to the orthographic and semantic RDM due to the occasional mis-selection of lexical item in earlier layers.

It is worth noting that the Spearman's correlation measure in the current RSA result should not be understood at its face value. Because the theoretical RDMs are maximally independent and only included two values (similar or

dissimilar), the resulting rank correlation between activation and theoretical RDMs are constrained. This limitation contributes to the abrupt changes in the RSA dynamics, and the lack of phonological effect in aggregate measure despite present at the later stage in the temporal dynamic. Instead, the relative dynamics of the Spearman's correlation is more informative. In addition, the result of the aggregate measure is dependent on the chosen time window. In particular, the phonological layers in the fully interactive model show correlation with semantics, rather than phonology, when focusing only on the earlier cycles.

Discussion and Conclusion

A promise of representational similarity analysis is its ability to link distributed patterns of neural activity to different levels of cognitive processing. The simulation work reported here demonstrates just how much the ability to make this link depends on underlying assumptions about the dynamics of interaction between different levels of processing. A set of simplified models of reading was designed with three different kinds of dynamics: fully interactive, cascading feedforward and discrete feedforward. With the discrete, feedforward model, similarity could clearly isolate different processing levels. But the relationship between similarity and level of representation is more complex in the more interactive models. In the cascading, feedforward model, orthographic similarity dominates all levels of representation, down to the level of phonological output. In the fully interactive model, the higher-level semantic and phonological patterns were observed in layers responsible for orthographic representation. Correlations with orthographic, semantic, and phonological RDMs at different time points were present in the activation dynamics of the lower layers responsible for orthographic and semantic processing, with semantic effects preceding phonological effects. Other models with within-layer inhibition and varying connection weights are beyond the scope of the current study, but are of interest for further pursuit.

The results of the fully interactive model suggest that, as a result of interactivity, the same cognitive level can show a different similarity structure at different time point, evolving from an orthographic, to a semantic, to a phonological composition. However, with imaging techniques that have a low temporal resolution, like fMRI, this fine-grained temporal structure can be washed out. Therefore, if the underlying cognitive system is fully interactive, and if there is a single region that corresponds to a single level of a cognitive architecture, we would expect to see that region show significant correlation with multiple kinds of theoretical measures. This pattern is precisely what was observed in Fischer-Baum et al. (2017), with the lvOT showing significant correlation with both an orthographic and semantic measure of similarity between words in a reading task. While one may want to conclude from this pattern that the cortical region of interest contains neural representations that subserves both an orthographic level or processing and a semantic level of processing (e.g., Wang et al., 2018), our work shows that an alternative explanation is that this region is responsible solely for orthographic processing, in the context of a cognitive theory of reading with feedback interactivity from the semantics to the orthographic levels. Which of these interpretations of the RSA results from the lvOT during reading tasks is correct critically depends on underlying assumptions about the cognitive dynamics of the reading system.

In the face of this problem, we have two suggestions to guide future RSA research. The first is to incorporate more work using neuroimaging modalities with high temporal resolution, like MEG or ECoG. As our simulations show, fully interactive models predict evolving similarity structure over time, which could potentially be observed with these alternative techniques. The second is that research using similarity to bridge cognitive and neural theories should be based on fully implemented computational models of the task. One of the benefits about using RSA in cognitive neuroscience research is that it forces researchers to be explicit about their theoretical assumptions on the underlying cognitive representations. The current work makes it clear that researchers must also be explicit about theoretical assumptions on the underlying dynamics of cognition. The models presented here were highly simplified, with only eight lexical items and a representational structure that maximally distinguished orthographic and phonology. Future work should rely on richer computational models that better capture the domain being studied, as other variables like word frequency, neighborhood size, or regularity of the spelling to sound correspondences may also have unintended consequences on similarity-based analyses.

Finally, the fact that different underlying cognitive dynamics yield different results in a similarity analysis suggests that techniques like RSA could prove useful for testing questions of what the underlying cognitive dynamics are. Dubarry and colleagues (2017) used single trial ECoG data to suggest that cognitive processing is more serial and less parallel than most cognitive theories assume. Previous RSA research has shown selectivity in how the similarity structure in different cortical regions relate to measures based on cognitive levels of representation. Based on our simulations, such a pattern would be unlikely under a cascading, feedforward architecture. By pairing RSA with fully developed cognitive architectures, the technique may be useful for testing questions about the underlying dynamics of cognitive processing, along with questions about the underlying nature of the cognitive representations.

The relationship between similarity structure and cognitive level of representation is complex and depends on questions of cognitive dynamics. Cognitive neuroscience researchers who use RSA to ask questions about the neural basis of different cognitive functions need to exercise caution in how they interpret the results of their study. However, with fully implemented computational models of the task being investigated and imaging techniques with higher temporal resolution, RSA can continue to be a powerful tool for bridging cognitive theories and neural systems.

Acknowledgments

This work was supported by the T.L.L. Temple Foundation Neuroplasticity Laboratory award to Rice University, and the National Science Foundation under Grant no. NSF CAREER SBE-1752751 to S.F-B.

References

Chen, Y., Shimotake, A., Matsumoto, R., Kunieda, T., Kikuchi, T., Miyamoto, S., Fukuyama, H., Takahashi, R., Ikeda, A., & Ralph, M. L. (2016). The 'when'and 'where'of semantic coding in the anterior temporal lobe: Temporal representational similarity analysis of electrocorticogram data. *Cortex*, 79, 1-13.

Cichy, R. M., & Pantazis, D. (2017). Multivariate pattern analysis of MEG and EEG: A comparison of representational structure in time and space. *NeuroImage*, 158, 441-454.

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: a dual route cascaded model of visual word recognition and reading aloud. *Psychological Review*, 108(1), 204.

Dehaene, S., & Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in Cognitive Sciences*, 15, 254-262.

Dell, G. S. (1986). A spreading-activation theory of retrieval in sentence production. *Psychological Review*, 93(3), 283.

Dubarry, A. S., Llorens, A., Trébuchon, A., Carron, R., Liégeois-Chauvel, C., Bénar, C. G., & Alario, F. X. (2017). Estimating parallel processing in a language task using single-trial intracerebral electroencephalography. *Psychological Science*, 28(4), 414-426.

Fischer-Baum, S., Bruggemann, D., Gallego, I. F., Li, D. S.P., & Tamez, E. R. (2017). Decoding levels of representation in reading: A representational similarity approach. *Cortex*, 90, 88-102.

Goldrick, M., & Blumstein, S. E. (2006). Cascading activation from phonological planning to articulatory processes: Evidence from tongue twisters. *Language and Cognitive Processes*, 21(6), 649-683.

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, 17(8), 401-412.

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis – connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, 2, 4.

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: I. An account of basic findings. *Psychological Review*, 88(5), 375.

Price, C. J., & Devlin, J. T. (2011). The interactive account of ventral occipitotemporal contributions to reading. *Trends in Cognitive Sciences*, 15, 246-253.

Rogers, T. T., & McClelland, J. L. (2014). Parallel distributed processing at 25: Further explorations in the microstructure of cognition. *Cognitive Science*, 38(6), 1024-1077.

Taylor, J. S. H., Davis, M. H., & Rastle, K. (2019). Mapping visual symbols onto spoken language along the ventral visual stream. *Proceedings of the National Academy of Sciences*, 116(36), 17723-17728.

Wang, X., Xu, Y., Wang, Y., Zeng, Y., Zhang, J., Ling, Z., & Bi, Y. (2018). Representational similarity analysis reveals task-dependent semantic influence of the visual word form area. *Scientific Reports*, 8(1), 1-10.