Neurobiology of Language

an open access journal

RESEARCH ARTICLE

Neural Components of Reading Revealed by Distributed and Symbolic Computational Models

Ryan Staples and William W. Graves

Department of Psychology, Rutgers University, Newark, NJ

Keywords: language, reading, fMRI, computational modelling, cognitive neuroscience, orthography

ABSTRACT

Determining how the cognitive components of reading—orthographic, phonological, and semantic representations—are instantiated in the brain has been a long-standing goal of psychology and human cognitive neuroscience. The two most prominent computational models of reading instantiate different cognitive processes, implying different neural processes. Artificial neural network (ANN) models of reading posit nonsymbolic, distributed representations. The dual-route cascaded (DRC) model instead suggests two routes of processing, one representing symbolic rules of spelling–to–sound correspondence, the other representing orthographic and phonological lexicons. These models are not adjudicated by behavioral data and have never before been directly compared in terms of neural plausibility. We used representational similarity analysis to compare the predictions of these models to neural data from participants reading aloud. Both the ANN and DRC model representations corresponded to neural activity. However, the ANN model representations correlated to more reading-relevant areas of cortex. When contributions from the DRC model were statistically controlled, partial correlations revealed that the ANN model accounted for significant variance in the neural data. The opposite analysis, examining the variance explained by the DRC model with contributions from the ANN model factored out, revealed no correspondence to neural activity. Our results suggest that ANNs trained using distributed representations provide a better correspondence between cognitive and neural coding. Additionally, this framework provides a principled approach for comparing computational models of cognitive function to gain insight into neural representations.

INTRODUCTION

To better understand how the brain carries out a cognitive process, we must have robust approaches to both cognitive models and neural functions. Cognitive models provide a mechanistic explanation of cognitive function, and computational implementations of these models provide explicit and testable predictions of how these processes interact (Forstmann et al., 2011). However, these models alone cannot reveal the neural bases or implementation of the modelled processes. Neuroimaging has separately contributed to localizing where in the brain certain aspects of cognition are processed, but localization alone does not explain cognition. Furthermore, in the absence of model-based constraints, the interpretation of fMRI data is radically underconstrained (Haxby et al., 2014). The union of computational cognitive models and neuroimaging, implemented by recent advances in methodology, allows for the quantitative comparison of specific, model-generated predictions to test their biological plausibility. Here we used this approach to better determine the neural and cognitive basis of reading.

Citation: Staples, R., & Graves, W. W. (2020). Neural components of reading revealed by distributed and symbolic computational models. *Neurobiology of Language*, *1*(4), 381–401. https://doi.org/10.1162/nol_a_00018

DOI:
https://doi.org/10.1162/nol_a_00018

Supporting Information:
https://doi.org/10.1162/nol_a_00018

Received: 15 January 2020
Accepted: 29 June 2020

Competing Interests: The authors have declared that no competing interests exist.

Corresponding Author:
Ryan Staples
ryan.staples@rutgers.edu

Handling Editor:
Steven Small

The MIT Press

Reading is a particularly promising domain to investigate, given the existence of well-established computational cognitive models and the rich availability of cognitive neuroscience data. Rather than assume the primacy of a particular model, we took a model comparison approach (Popov et al., 2018) to testing for correspondence between two different computational cognitive models optimized to account for reading performance.

One approach to modelling reading uses artificial neural networks (ANNs). ANN models of reading are distributed, nonsymbolic, and based on connectionist principles. They consist of weighted connections, learned via the backpropagation algorithm, between layers of interconnected, neurally inspired units (Plaut et al., 1996; Seidenberg & McClelland, 1989). ANNs map inputs to outputs using some number of "hidden" layers, in which the statistical regularities of the target outputs are related to those of the inputs (Hinton, 1989).

Another prominent model of reading is the symbolic dual-route cascaded (DRC) model (Coltheart et al., 2001). The DRC model also uses weighted connections between nodes for producing phonological output from orthographic input. Unlike ANN models, the DRC model consists of hand-tuned connections among symbolic representations of explicit rules and lexicons.

Both ANN and DRC models have been designed and tested to replicate aspects of human reading, but the results of these simulations and independent evidence for them have so far not adjudicated between the models. Additionally, neither model matches human behavioral data perfectly. For example, an ANN model accurately reproduces the frequency-regularity interaction found in single word reading (Plaut et al., 1996; Seidenberg & McClelland, 1989), as well as spelling–sound consistency effects and natural variation in nonword pronunciation (Zevin & Seidenberg, 2006). Lesioning of semantic contributions in an ANN model also replicates the symptoms of surface dyslexia, as seen in participants with semantic dementia (Woollams et al., 2007). ANN models, however, have been shown to perform at a below-human level on low-frequency, inconsistent words (Jared, 2002). In contrast, the DRC model succeeds in accounting for the serial, left-to-right nature of reading English (Coltheart et al., 2001). The DRC model is silent, however, on effects of semantic factors such as imageability, the degree to which a word is judged to elicit a sensory impression (Plaut & Shallice, 1993; Strain & Herdman, 1999; Strain et al., 1995), although such effects are beyond the scope of the current investigation.

Both ANN models (Ueno et al., 2011) and the DRC model (Perry et al., 2007, 2010) have been expanded in their capacity to simulate reading behavior since they were originally introduced. Here we elected to use relatively simple versions of the models that are maximally comparable.

Some previous studies have used computational models of reading behavior to inform computational models of neural processing. A series of studies has used ANN models of reading to examine the N400 event-related potential component, thought to index attempted semantic access. Broadly, these results suggest that ANN models with neurobiologically inspired architecture can produce realistic N400 waveforms, account for waveform-modulating variables such as frequency or semantic richness, and perform lexical decision tasks (Cheyette & Plaut, 2017; Laszlo & Armstrong, 2014; Laszlo & Plaut, 2012; Rabovsky et al., 2018).

The present study differs from this earlier work in several ways. First, previous attempts to directly connect models of reading with neuroimaging data have been limited to electrophysiological data. We aimed to extend this literature to fMRI-based word representations. Second, previous work has compared model outputs with neural data, rather than model internal representations. Our approach focused on what the model is computing by examining the intermediate

stage (hidden layer) associated with orthography-to-phonology transforms, as opposed to dealing solely with the output representations. Finally, and most importantly, previous studies have not directly compared representations from both the DRC and ANN models in terms of their fit to neural data. Instead, they have demonstrated that computational mechanisms incompatible with the DRC model can produce a qualitative fit to neural data. We aimed to determine not only whether it is possible to produce a qualitative fit to neural data, but also whether either model produces a better quantitative fit.

Interest in using computational models of reading to interpret neural data on reading is long-standing. Even early functional neuroimaging studies of reading interpreted their results in terms of the ANN and DRC models (Binder et al., 2005; Fiez et al., 1999). Furthermore, in the decades since these models were developed, a considerable amount has been learned about the neural basis of reading. Successful reading involves the integration of visual, orthographic, phonological, and semantic information. Both univariate and multivariate pattern analyses (MVPA) have contributed to localizing and dissociating where in the brain these types of information are processed (Binder et al., 2009; Fischer-Baum et al., 2017; Graves et al., 2010; Liuzzi et al., 2017; Price, 2012; Taylor et al., 2013).

Perhaps the most consistent finding in imaging studies of reading concerns the involvement of the left mid-fusiform gyrus (FG), also referred to as the occipitotemporal cortex (Price, 2012), in orthographic processing. The FG is thought to be an initial part of the pathway from orthography to phonology when reading (Graves et al., 2014; Jobard et al., 2003). Activity in the FG has been shown to follow a spatial gradient, with greater activation to lower-level visual and smaller orthographic features (e.g., single letters or bigrams) in the posterior aspect and tuning to progressively larger orthographic structure (e.g., quadrigrams or whole word forms) in more anterior aspects (Vinckier et al., 2007). The FG shows increased activation as a result of grapheme–phoneme correspondence (GPC) training in children (Brem et al., 2010; Shaywitz et al., 2004) and adults learning novel writing systems (Hashimoto & Sakai, 2004; Taylor et al., 2017). Furthermore, this increased activation to letter strings and words fails to develop in children with dyslexia (Shaywitz et al., 2007; van der Mark et al., 2009), and damage to the mid-FG results in pure alexia (Binder & Mohr, 1992; Damasio & Damasio, 1983; Leff et al., 2001). The mid-FG is sensitive to letter combination probabilities, even when the stimuli are nonpronounceable letter strings, and engages in orthographic processing even when participants are attending to nonlinguistic visual features of the stimuli (Binder et al., 2006). Despite this, the exact computational role of the FG in reading is unclear. The visual word form area hypothesis posits that the FG is specialized for orthographic processing (Cohen et al., 2000; Dehaene & Cohen, 2011; McCandliss et al., 2003). The Interactive Account, on the other hand, suggests that the role of the FG is to integrate phonological and possibly semantic information with bottom-up visual and combinatorial orthographic information (Mano et al., 2013; Price & Devlin, 2011; Twomey et al., 2011). A model-based approach with separate representations for orthography, phonology, and the mapping between them should help clarify the computational role of the FG.

Semantic information is theorized to assist in producing phonology during reading aloud, particularly when reading inconsistent or irregular words (Strain et al., 1995). We briefly summarize relevant information on this critical component of reading. However, strong conclusions regarding semantics are beyond the scope of this study due to the lack of specific semantic representations in the models being tested. Investigations of activation related to semantics in reading reveal a specific, yet widespread network, prominently including the angular gyrus (AG), along with middle and anterior parts of the lateral and ventral temporal cortices (Binder et al., 2009; Taylor et al., 2013). The AG has been frequently implicated in semantic processing

(Binder et al., 2009; but see Humphreys et al., 2015). While its exact role is debated, the AG has been associated with semantic integration (Humphries et al., 2007), processing of semantic features (Binder & Desai, 2011; Wang et al., 2017), more general effects of task difficulty (Humphreys et al., 2015), and overlapping effects of semantics and task difficulty (Mattheiss et al., 2018). The lateral and ventral temporal cortex are involved in the storage and processing of category-related semantic information (Lambon Ralph et al., 2017). Broad swathes of temporal cortex show activation in categorization tasks (Chao et al., 1999; Kable et al., 2005), and lesions result in category-specific processing impairment (Damasio et al., 2004; Hillis & Caramazza, 1991; Lambon Ralph et al., 2007). The left anterior temporal lobe (ATL), though it has been suggested to be part of the default mode network (Buckner et al., 2008; Raichle et al., 2001), demonstrates increased activation compared with rest across stimulus type and modality in semantic tasks (Humphreys et al., 2015). The left ATL has been theorized to map multimodal semantic representations onto phonology (Hoffman et al., 2015).

Phonological information is processed in a cortical network largely distinct from those for orthography and semantics, involving the left posterior superior temporal gyrus (pSTG), the supramarginal gyrus (SMG), and the pars opercularis of the inferior frontal gyrus (IFG). This network is broadly activated when mapping orthography directly onto phonology (Graves et al., 2010; Jobard et al., 2003; Sandak et al., 2004). Supporting this, the pSTG, SMG, and IFG pars opercularis also show decreased activation during reading in subjects with dyslexia (Richlan et al., 2009). Lesions of the SMG have also been shown to produce conduction aphasia (Damasio & Damasio, 1980) and phonological agraphia (Alexander et al., 1992). Deficits in phonological retrieval, as distinct from semantics or comprehension, have also been associated with left-sided lesions focused on the pSTG and SMG (Pillay et al., 2014). Furthermore, an fMRI meta-analysis examining encoding and recall of phonological stimuli found an area of maximum overlap with lesions producing conduction aphasia in the planum temporale, just inferior of the SMG (Buchsbaum et al., 2011). For reading in particular, decreased bigram frequency, likely reflecting difficulty of mapping orthography to phonology, has been related to increased activation in bilateral superior temporal sulcus and posterior middle temporal gyrus (MTG) (Graves et al., 2010).

In addition to the largely localized perspectives on the neural basis of orthographic, semantic, and phonological processing given above, MVPA studies have begun to unravel the structure of neural representations related to these processes. Rothlein and Rapp (2014) demonstrated the presence of separate neural substrates for modality-specific and abstract representation of letters. Recent studies have also found category-specific tuning in the ATL (Malone et al., 2016), semantic representations in the FG (Fischer-Baum et al., 2017; Wang et al., 2018), and orthographic representations in the AG (Fischer-Baum et al., 2017). However, despite these great strides in mapping the neural basis of some of the major cognitive components of reading, there is little consensus on the neural computations these representations may reflect.

The Current Study

The present study aimed to provide the first direct comparison of how well the ANN and DRC models of reading fit fMRI data. Historically, the difficulty of bringing model-based representations and neural representations into the same space prevented direct comparisons. Representational similarity analysis (RSA) provides an elegant solution to this problem (Kriegeskorte & Kievit, 2013; Kriegeskorte et al., 2008). RSA enables the comparison of representations from disparate modalities by transforming those representations into a common space based on stimulus similarity. Using RSA, we related neural data from human participants reading aloud to the internal

representations generated by the ANN and DRC models performing the same orthography-to-phonology task. We predicted that both models of reading would correspond to neural activity in a left-lateralized reading network spanning frontal, temporal, and fusiform gyri. These cortical regions have been repeatedly implicated in studies of reading (Binder et al., 2009; Cattinelli et al., 2013; Fiez & Petersen, 1998; Murphy et al., 2019; Price, 2012; Taylor et al., 2013; Turkeltaub et al., 2002). Furthermore, due to the neurally inspired nature of its architecture and function, including its use of distributed as opposed to localist symbolic representations, we predicted that the ANN model distributed representations would provide a better fit for investigating how the brain computes orthography-to-phonology transforms. In comparing the neural instantiation of the ANN and DRC models, we not only tested their feasibility as mechanistic explanations of reading, but also demonstrated a principled methodology for comparing computational models of cognition.

MATERIALS AND METHODS

Participants

The participants were 18 (13 female) healthy, right-handed adults who spoke English as a first language. The mean age of the participants was 23.2 (*SD*: 3.4). Subjects provided written informed consent following procedures approved by the Medical College of Wisconsin Institutional Review Board, as described in Graves et al. (2010).

Stimuli

464 monosyllabic English words were used as stimuli. These words were selected such that letter length, word frequency, spelling–sound consistency, imageability, bigram frequency, and biphone frequency were uncorrelated. Further details regarding the stimuli can be found in Graves et al. (2010). One word, "hale," was discarded from the original 465 word stimulus list because it was not in the Harm and Seidenberg (2004) stimulus set. This was done to ensure compatibility with planned studies investigating the contribution of semantics to models of reading.

Task

The fMRI task used a fast event-related design with continuous acquisition. Participants viewed a series of randomly presented words. Each word was displayed for 1,000 ms before being replaced by a fixation cross. Participants were instructed to "read each word aloud as quickly and accurately as possible" into an fMRI compatible microphone.

The MRI data were acquired using a 3T GE Excite system (GE Healthcare, Waukesha, WI) using an 8-channel array head radio frequency receive coil. A 134 contiguous axial slice (0.938 × 0.938 × 1.000 mm) T1-weighted anatomical image was acquired using a spoiled-gradient-echo sequence. Functional scans were acquired using a gradient-echo echoplanar imaging (EPI) sequence (echo time = 25 ms, repetition time = 2,000 ms, field of view = 192 mm, matrix = 64 × 64 pixels, voxel dimensions = 3 × 3 × 2.5 mm, gap = 0.5 mm) in an interleaved fashion, resulting in thirty-two axial slices per volume. 240 volumes were collected in each of five runs.

Orthographic and Phonological Representations

The orthographic representations, used as inputs to the ANN model, were the inputs used in Plaut et al. (1996, Table 2). They consisted of 105-unit binary vectors. The presence or absence of a grapheme was indicated with a 1 or 0, and the graphemes were grouped according to whether

they could appear in the onset (first 30 units), vowel (next 27 units), and coda (final 48 units). Multi-letter graphemes such as "ph" were coded such that "p," "h," and "ph" were all set to 1. An orthographic representational dissimilarity matrix (RDM) was computed as the pairwise correlation distance between all orthographic vectors for later use in RSA (see supplementary Figure 1 in the online supporting information located at https://www.mitpressjournals.org/doi/suppl/10.1162/nol_a_00018). The phonological representations, used as the target outputs for training and testing the ANN models, were also binary vectors as used in Plaut et al. (1996, Table 2). They consisted of 61 phoneme slots, also divided into onset, vowel, and coda, where 1 indicated the presence and 0 the absence of a phoneme. The phonemes /ps/, /ks/, and /ts/ were also coded such that both their constituent parts and the combination unit were active. A phonological dissimilarity matrix was computed as the pairwise correlation distance between all phonological vectors. (See supplementary Figure 1 in the online supporting information.)

Computational Models

Two computational models, the feed-forward ANN from Plaut et al. (1996) and the rule-based, symbolic DRC model (Coltheart et al., 2001), were used to simulate human reading. The ANN model was identical to the one in Plaut et al. (1996). The models were trained with inputs, described above, representing 2,998 monosyllabic words. The model had 105 input units. These input units were fully connected to 100 hidden units, which in turn were fully connected to 61 phonological output units (Figure 1). These phonological units were given additional, external input that ramped up as training proceeded. This external input was calculated as a function of the log-compressed frequency for a given word (Plaut et al., 1996, Equation 16), and approximates the growing semantic contribution to phonology that occurs over the development of reading. It further reflects the assumption that higher frequency concepts have stronger semantic representations (Plaut et al., 1996; Woollams et al., 2007). Finally, using frequency as a simplified approximation for semantics also ensures that the ANN model is maximally comparable to the DRC model, which also does not try to account for detailed semantics.

The model was trained for 400 epochs (i.e., 400 presentations of the entire 2,998-word stimulus set) using standard backpropagation of error and tested for generalization on a separate set of nonwords. The number of epochs was chosen using Plaut et al. (1996) as a starting point and then fine-tuned to identify the point at which accuracy on a nonword test set began to decrease. The test set contained 166 monosyllabic nonwords, compiled from Glushko (1979) and McCann and Besner (1987). Accuracy on the training set was assessed by testing the model using the 464 words that the human participants read in the scanner. These words were a subset of the full 2,998-word training corpus. Outputs from the network were rounded to the nearest whole

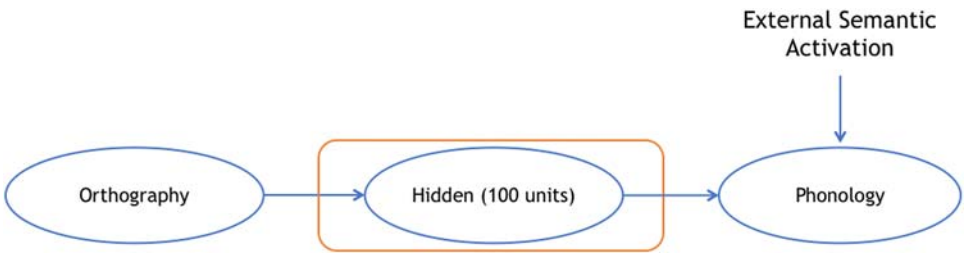

Figure 1. Schematic structure of the artificial neural network model. Word representations for the model were derived from the hidden layer, circled in orange.

number (either 0 or 1) and compared with the appropriate target vector. As real words have well-defined pronunciations, any network output vector that deviated from the appropriate target vector was marked incorrect.

Nonwords, however, do not have a single correct pronunciation. To account for this, a list of plausible alternative pronunciations was generated. For scoring the test set, outputs from the network were rounded to the nearest whole number (either 0 or 1) and compared with the full list of target vectors using the Jaccard similarity index (Levandowsky & Winter, 1971). If the output vector produced by the network was closer to the appropriate target vector than to any other vector in the nonword set, the output was marked correct. If the output vector was closer to the vector of any nonword other than the target, the output was marked incorrect. This reflects the intuition that nonwords do not have well-defined pronunciations. However, it allows for the model to have a reasonably random output, so long as that output is close enough to one of the acceptable pronunciations. Thus, we also tested a stricter scoring method. Rounded model outputs for a given nonword were compared with the list of acceptable pronunciations for that nonword. If the output exactly matched one of these vectors, it was marked correct. The model was run 20 times, with weights re-initialized each time to random uniform values between −0.1 and +0.1. We chose 20 instantiations of the model to roughly match the number of participants in the fMRI data set. This number was expected to be ample, as it is twice that used in similar previous modelling experiments (Welbourne & Lambon Ralph, 2005; Woollams et al., 2007). The stimulus-stimulus correlation matrices generated from the hidden unit layers of these model instantiations were averaged to create a mean ANN RDM.

The second model was the symbolic, rule-based DRC model (Coltheart et al., 2001). Inputs to the DRC model activated visual feature units, which then fed forward to activate letter units. The activation from this letter unit layer then activated two routes in parallel: a lexical-nonsemantic route and a GPC route (Figure 2). The lexical-nonsemantic route contained an orthographic lexicon, in which a local word representation was activated by the parallel activation of earlier letter features. This orthographic lexicon activation then activated the corresponding phonological lexicon representation of the input word. Each of these lexicons had 2,998 units, one for each word in the current corpus. The output of the lexical-nonsemantic route consisted of the model activation value in the unit corresponding to the input word and a 0 in every other position, for each lexicon. The GPC route processed each input word in a serial, left-to-right fashion. When the first letter of an input is processed, the 2,033 preset rules are searched until an appropriate letter-phoneme conversion is located. The next letter then becomes available to the GPC route, and the two-letter string undergoes the same search through rule space. At each stage, rules are searched from largest to smallest matching grapheme. This process is repeated until the input word is named or until the last letter is processed. The output of the GPC route consisted of the model activation for each rule activated by the input word, with a 0 in every other position. Word

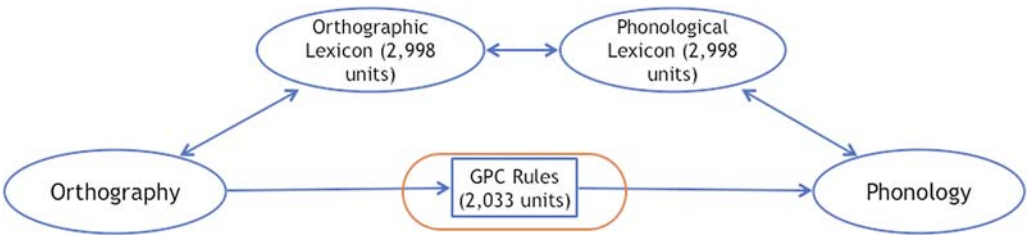

Figure 2. Schematic structure of the dual-route cascaded model. Word representations for the model were derived from the grapheme-phoneme correspondence (GPC) rules, circled in orange.

representations for the DRC model were taken as the vector of activation for the GPC rules. Deriving word representations from these internal components of the model was considered analogous, for purposes of comparison, to representations derived from the hidden unit layer of the ANN. This internal DRC representation consisted of a 2,033-unit sparse vector. Further details of the DRC model can be found in Coltheart et al. (2001).

An important feature of the DRC model is that most words are read by both the lexical and GPC routes, with the GPC route being primary only for unknown words, nonwords, or low frequency regular words (Coltheart et al., 2001). The lexical route has no direct analogue in the ANN model. However, it is possible that representational similarities between the brain and the model might be driven by the lexical route. As such, we also considered 8,029 unit DRC representations (2,033 GPC + 2,998 orthographic lexicon + 2,998 phonological lexicon units), wherein the lexicon representations consisted of a one-hot, localist vector for each word. As the results with this expanded representation (see Supplementary Figure 3 and Supplementary Table 1 in the online supporting information) were nearly identical to the results that did not include the lexical route, we chose the parsimonious approach of focusing on the GPC route, as it is most analogous to the processes in the ANN model. Therefore, we do not consider the lexical route representations further.

fMRI Data Analysis

MRI data were preprocessed using AFNI (http://afni.nimh.nih.gov/afni; Cox, 1996). Data were skull-stripped, slice-timing and motion corrected, and the first six images were discarded to allow for initial saturation. Data were then spatially coregistered.

Voxelwise single-trial effects were estimated using least-squares-sum multiple regression (Mumford et al., 2012), as implemented in the AFNI program 3dLSS. A noise signal calculated from the signal in the lateral ventricles was included along with six motion parameters as covariates of no interest. Resulting coefficient maps were not smoothed during first-level analysis, so as to preserve spatial patterns for RSA analysis. Individual subject data were then transformed into Talairach space (Lancaster et al., 2000).

RSA (Sensitivity Analysis)

To test for correspondence between model-based and brain-based similarity patterns among the stimuli, RSA was performed. This was implemented using PyMVPA software (Hanke et al., 2009). For each model, stimulus features were *z*-scored. An RDM was then generated based on the pairwise correlation distance (1 − Pearson's *r*) between all stimuli. To test for correspondence between this RDM and neural patterns, a searchlight analysis was performed. Each voxel in a cortical gray matter mask iteratively served as the center of a 3-voxel radius sphere (sphere volume: 123 voxels of 3 × 3 × 3 mm each, including the center voxel). A neural RDM was constructed by calculating the pairwise correlation between the beta weights evoked from each stimulus in this sphere. This neural RDM was then compared with the model RDM using Spearman correlation. Use of a rank-based second-order correlation avoids potential differences in correlation means across different types of representations, as recommended by Kriegeskorte et al. (2008). The resulting correlation coefficient was assigned to the center voxel of each sphere. This process was repeated until every voxel in the gray matter mask served as a center voxel once. The individual subject's resulting correlation coefficient maps were spatially smoothed using a 6 mm full-width-half-max (FWHM) kernel before being entered into a one-sample *t* test against a null hypothesis of no correlation without further resampling. The resulting group maps were

Fischer transformed and submitted to cluster correction (voxelwise $p < 0.001$, clusterwise $p = 0.05$, cluster extent = 234.9 mm^3).

Partial Correlations

Separate partial correlations were calculated using the CoSMoMVPA package for MATLAB (Oosterhof et al., 2016). The correlation coefficient maps were then smoothed with a 6 mm FWHM kernel, and *z*-scored. A *t* test was run on the *z*-score maps, and the group *t*-value map was subjected to cluster correction (voxelwise $p < 0.001$, clusterwise $p = 0.05$, cluster extent = 234.9 mm^3).

RESULTS

Human Performance

The mean reaction time was 588 ms (*SD*: 123). Errors in reading aloud were very rare, only 2.6% overall. Further details can be found in Graves et al. (2010).

Model Performance

To determine whether the 20 instantiations of the ANN model achieved an accuracy comparable to human performance, they were tested, as described in Materials and Methods, for the 464 words that humans read in the scanner. Those were a subset of the 2,998 words on which the model was trained. The model obtained 97.2% accuracy on this subset. To test the ability of the models to generalize beyond the training set, they were presented with novel but pronounceable letter strings (hereafter, nonwords). Using Jaccard similarity as a scoring metric, the ANN model attained 76.9% accuracy on the Glushko (1979) nonwords and 91.9% accuracy on the McCann and Besner (1987) nonwords, reaching 84.9% mean accuracy for nonwords overall. Using the stricter exact-match criterion, the ANN was 75.5% accurate on the Glushko (1979) nonwords and 79.5% accurate on the McCann and Besner (1987) nonwords, achieving 77.4% mean overall accuracy. The DRC model made no errors on either nonword set. The dissimilarity matrices derived from the intermediate representations of the ANN and DRC models were significantly correlated ($r = 0.52$, $p < 0.0001$; Table 1). (See Supplementary Figure 2 in the online supporting information for plots of example and full ANN and DRC dissimilarity matrices.)

Imaging Results

Here we focus on comparing model-based word representations with neural representations derived from functional neuroimaging. Univariate analyses of neural activation for these words are reported in Graves et al. (2010) and will not be discussed further.

Table 1. Correlation matrix of the dissimilarity matrices used for representational similarity analysis.

	ANN	DRC	Orthography	Phonology
ANN	1			
DRC	0.52	1		
Orthography	0.9	0.55	1	
Phonology	0.68	0.6	0.63	1

Note. ANN = artificial neural network, DRC = dual-route cascaded model.

Orthography

To test for neural correspondence with the orthographic representations used as inputs to the ANN model, a searchlight RSA was performed. The orthographic input RDM was correlated with neural representations in a left-lateralized network. Specifically, orthographic features correlated with neural activity in the left ATL (lateral and ventral aspects), the orbital IFG, the middle frontal gyrus, the precentral sulcus, the collateral sulcus, the calcarine fissure, and the cuneus. The orthographic RDM was also correlated with neural activity to a lesser extent in the cerebellar tonsil (Figure 3 and Supplementary Table 2 in the online supporting information).

Phonology

Neural correspondence with the phonological ANN output feature representations was also tested using searchlight RSA. Phonological feature representations were correlated with neural representations in the left middle temporal and anterior superior temporal gyri, the middle occipital gyrus, the insula, and the pre- and postcentral gyri. Phonological representations were also related to bilateral neural activity in the anterior inferior temporal and parahippocampal gyri (Figure 4 and Supplementary Table 3 in the online supporting information). Importantly, the phonological searchlight did not reveal any significant clusters in the primary visual cortex. This points to the external validity of the current approach, where correspondence between model and neural representations is able to clearly distinguish between orthographic inputs and phonological outputs in expected brain areas.

Intermediate Representations from Models

The hidden layer of the ANN was used to model information related to the transformation of orthography to phonology. The RDM constructed from model activations in this hidden layer correlated with left hemisphere neural representations in the anterior temporal pole, the orbital and triangular IFG, the precentral sulcus, the FG, the intraparietal sulcus, the cuneus, and the MTG (Figure 5 and Supplementary Table 4 in the online supporting information). Additionally,

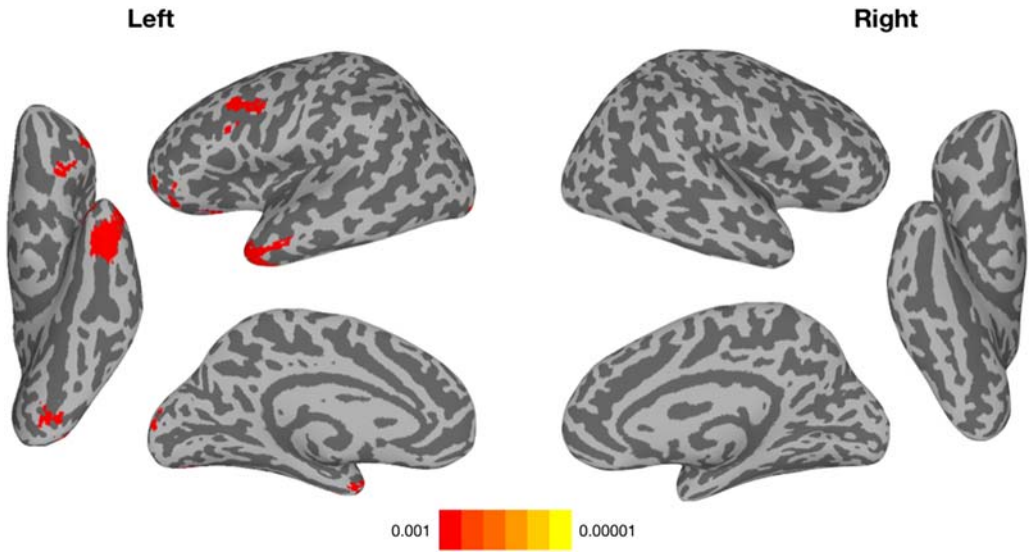

Figure 3. Correspondence between the orthographic input similarity structure and the neural similarity structure.

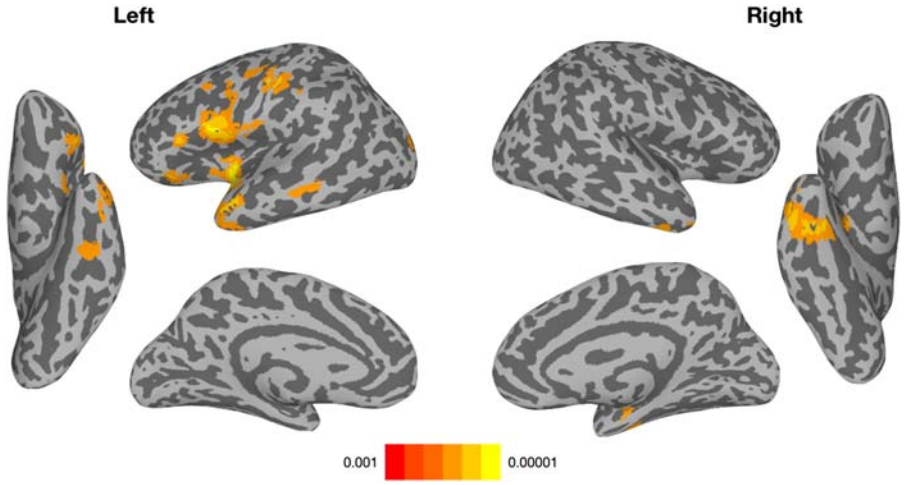

Figure 4. Correspondence between the phonological output similarity structure and the neural similarity structure.

the hidden layer RDM correlated with right hemisphere representations in the parahippocampal gyrus. This pattern of results largely agrees with a standard neuropsychological view, in which reading aloud is supported by a combination of visual-related areas in the occipito-temporal cortex, language production areas in the left IFG, and intermediate areas along the left MTG (Binder et al., 2009; Fiez & Petersen, 1998; Price, 2012; Taylor et al., 2013; Turkeltaub et al., 2002). At the same time, our approach lends a new level of specificity to the functions being carried out in these areas during reading.

As our aim was to model the mapping between orthography and phonology, we elected to use representations generated from the DRC model layer that specifically implemented this process: the GPC rule route (Figure 2). This route also provides the best match to the function of the hidden layer of the ANN model (Figure 1). The RDM based on GPC representations from the DRC model was correlated with a more restricted set of largely left hemisphere representations in the superior temporal gyrus and sulcus, the postcentral gyrus, the straight gyrus, and the

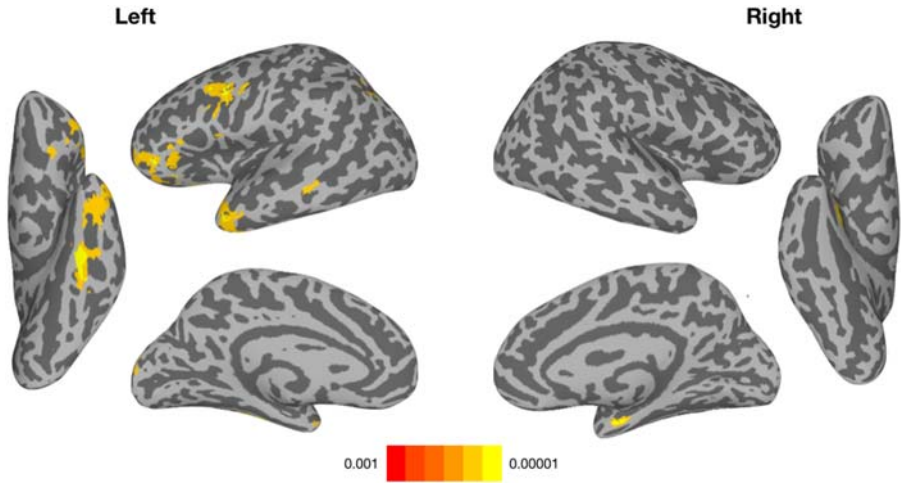

Figure 5. Correspondence between the artificial neural network hidden layer similarity structure and the neural similarity structure.

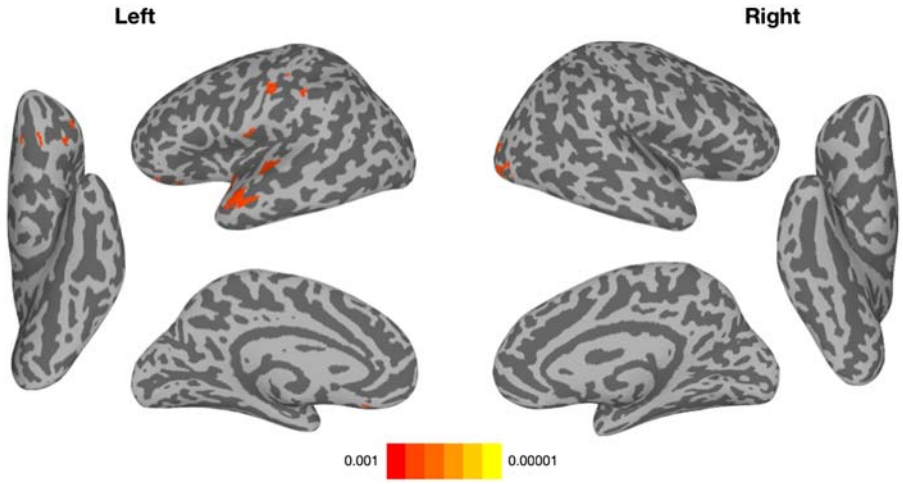

Figure 6. Correspondence between the dual-route cascaded intermediate graphene-phoneme correspondence route similarity structure and the neural similarity structure.

IFG pars orbitalis (Figure 6 and Supplementary Table 5 in the online supporting information). In the right hemisphere, the DRC RDM correlated with the neural representation in the inferior occipital gyrus. These neural correspondences with the intermediate GPC level of the DRC are spatially restricted relative to those found for the ANN, and they appear in areas less typically related to language and reading. This suggests that the ANN model might provide a better fit to the neural data than the DRC model.

Partial Correlations between Models

To formally test the ability of each model to account for the neural data after accounting for the other, we used partial correlations. This approach also accommodates the moderate correlation between the ANN and DRC intermediate layer representations. A searchlight analysis correlating the intermediate layers of the DRC model to the neural data with the variance due to the ANN hidden layer RDM partialled out revealed no results. However, when the reversed analysis was

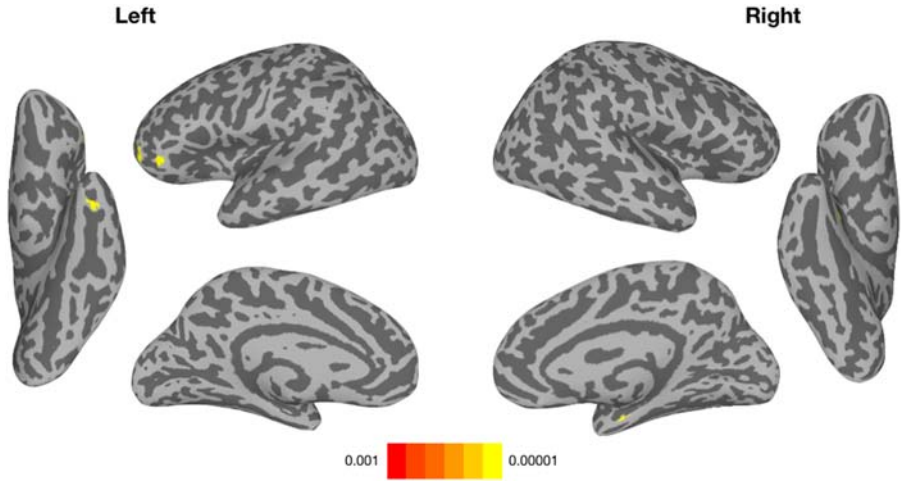

Figure 7. Correspondence between the artificial neural network hidden layer similarity structure and the neural similarity structure with variance due to the dual-route cascaded model partialled out.

conducted—the ANN hidden layer RDM correlated to the neural data with the variance due to the DRC model partialled out—a primarily left-lateralized set of areas emerged. These include the middle frontal gyrus, the orbital IFG, and the anterior inferior temporal gyrus. The ANN representations with variance due to the DRC model removed also correlated to neural activity in the right parahippocampal gyrus (Figure 7 and Supplementary Table 6 in the online supporting information). These results are consistent with the hypothesis that the distributed, nonsymbolic nature of the representations generated by the ANN model of reading reflects neural information processing to a greater degree than the symbolic representations of the DRC model.

DISCUSSION

While there is broad agreement as to where in the brain written language is processed, what is being coded or represented in these regions has remained largely unclear. Existing computational cognitive models that account for reading behavior use theories containing very different kinds of representations. ANN models use distributed nonsymbolic representations, while the DRC model uses local symbolic representations. We used RSA to relate neural data obtained during reading to the representations of orthography-to-phonology transforms generated by these two distinct computational models. The results suggest that both the DRC and the ANN models correspond to neural activity related to transformations from orthographic to phonological representations. Consistent with our predictions, we also show that the distributed ANN model of reading reflects the neural activity involved in reading aloud to a greater degree than the symbolic DRC model.

Importantly, the orthographic and phonological representations used as inputs to and outputs from the computational models correlate with neural activity in relevant cortex. The orthographic representations correspond to a network of cortical regions commonly associated with reading (Price, 2012; Turkeltaub et al., 2002). This network spans frontal, temporal, and occipital cortex. Unexpectedly, orthographic inputs did not correlate with neural representations in the posterior occipito-temporal sulcus typically associated with orthographic processing. We used grapheme-level orthographic representations, consistent with previous ANN models (Plaut et al., 1996; Woollams et al., 2007). However, graphemes contain clusters of letters such as "th," meant to more closely correspond to phonology than would purely orthographic inputs. Hence, graphemic similarity structure across words may reflect higher-order associations beyond orthography.

Phonological representations also correlate with brain structures that have been shown to be sensitive to phonological structure, including the left hemisphere middle and superior temporal gyri, and the ATL (Hoffman et al., 2015; Richlan et al., 2009). The phonological similarity structure was also related to neural activity in the pre- and postcentral gyri, most likely reflecting word production (Guenther et al., 2006; Indefrey & Levelt, 2004).

Both models account for how orthographic inputs are transformed to phonological outputs. This transform is reflected in the hidden layer of the ANN model and the GPC route of the DRC model. These intermediate representations corresponded to neural representation in areas of the brain involved in reading. In particular, the neural correspondences with the hidden layer of the ANN model overlap heavily with those for the orthographic representations, showing clusters in IFG, MTG, and ATL. Many of these regions have been previously implicated in orthography-to-phonology transformation (Jobard et al., 2003). The ANN representations also corresponded with neural representation in the FG. While the exact role of the FG has been debated (Dehaene & Cohen, 2011; Price & Devlin, 2011), it has been consistently implicated in the reading of words. A recent study suggests that the middle (mFG) and anterior fusiform gyrus (aFG), but not the posterior FG, contain neural representations of both orthographic and phonological information (Zhao et al., 2017). Consistent with this finding, we also found partially overlapping representation

of orthographic and phonological features in both the mFG and aFG. We also found correspondence between the ANN and the neural similarity structures in the mFG and aFG, further supporting the validity of the ANN hidden layer representations as an intermediate step between orthography and phonology. Abstractness of orthographic information, as indexed by invariance to case, font, and letter position (Dehaene et al., 2004), and representation of larger orthographic as opposed to low-level visual features, have been shown to increase along a posterior-to-anterior axis (Vinckier et al., 2007). Whereas the orthographic RSA identified clusters of correlation along the length of the FG, the hidden layer RSA reveals clusters in only the mFG and aFG, suggesting that the distributed representations generated from the model may indeed reflect larger and more abstract orthographic units. The ANN representations were also correlated with neural representations in the motor cortex, potentially reflecting a final transformation from the orthographic input to the motor code associated with a phonological output (Guenther et al., 2006).

The DRC model intermediate representations showed neural correspondences that spatially overlapped with the phonological representations used as target outputs in the ANN model in anterior temporal and Rolandic cortices. The DRC model intermediate representations also showed some overlap with the standard reading network (Binder et al., 2009; Cattinelli et al., 2013; Fiez & Petersen, 1998; Murphy et al., 2019; Price, 2012; Taylor et al., 2013; Turkeltaub et al., 2002) in the inferior frontal, the superior temporal, and the anterior temporal cortex. The ATL and the IFG pars orbitalis have been implicated in semantic processing (Binder et al., 2009; Hoffman et al., 2015) and the superior temporal gyrus in phonological processing (Fiez & Petersen, 1998; Price, 2012; Turkeltaub et al., 2002). This correspondence with reading-related areas was, however, more spatially limited than was found for the ANN model. Counter to our expectations, the DRC model did not correlate with activity in the FG, a putative locus of orthographic processing. The DRC model's lack of correlation with cortex generally associated with orthographic processing may be due to the different, more localist form of the input, or it may point to the distributed nature of the neural representations reflected in the intermediate representations of the ANN but not the DRC model.

Despite the significant correlation between the representations generated by the models, partial correlation analysis revealed that the distributed representations generated by the ANN model reflect neural representations to a greater degree. The network of neural activity resulting from this analysis is spatially limited, but partially corresponds to the standard brain reading network. In particular, both the IFG pars orbitalis and the inferior ATL are implicated in semantic processing during reading (Binder et al., 2009; Hoffman et al., 2015). Additionally, lesions of these same cortical regions produce regularization errors, in which a regular spelling-to-sound correspondence is improperly applied to an irregular word, as in pronouncing "sew" to rhyme with "do" (Binder et al., 2016). Overall, partial correlation results suggest that the ANN model better captures the neural representations involved in transforming orthography to phonology.

A possible explanation for the ANN-to-neural correspondence being greater than the DRC-to-neural correspondence is the sparser, binary nature of the DRC representations. ANN representations are dense and take rational number values, and thus are natively more like fMRI data. However, in order to assess ANNs as a model of neural representation, it must first be established that (1) ANNs actually do learn structured representations analogous to those in the brain and (2) the fit of a representation derived from an ANN model does in fact correspond to neural data above and beyond a representation derived from a symbolic model. Our results suggest that the ANN model does learn representations similar to neural representations, and that those representations do fit neural data better than representations derived from the symbolic DRC model. In short, brain activity does not tend to resemble local, symbolic rules, but instead computations over distributed elements.

The similarity structure of neural activity in the left ATL was related to every representational similarity structure we tested. The exact role of the left ATL in reading is unclear, although it has been linked to both semantic and phonological processing. There is fMRI evidence in healthy participants that it maps between semantic and phonological representations, particularly when reading exception words (Hoffman et al., 2015). Evidence from lesion deficit studies, however, suggests a more direct role in producing phonology. Rudrauf et al. (2008) found that lesions to the left ATL produced deficits in naming, but not recognition, of pictures across a range of categories. The co-presence of sensitivity to orthographic and phonological information in the left ATL was unexpected. This result may suggest that the role of the ATL involves integration across multiple levels and types of reading-related representation. Further investigation is warranted to identify the exact role of the ATL in reading.

An interesting possibility arises from the DRC model's conceptualization of spelling–sound knowledge as GPC rules. The similarity structures generated by the DRC will be sensitive to word regularity, or how well they correspond to the GPC rules (Coltheart et al., 2001). The ANN model instead exploits consistency, the statistical regularities between letter combinations and phonology, relative to the frequency of that spelling–to–sound correspondence in the entire corpus of words on which it is trained (Plaut et al., 1996). Evidence from reaction times suggest that consistency has a larger effect on reading performance (Cortese & Simpson, 2000; Jared et al., 1990), but independent effects of regularity have also been demonstrated (Glushko, 1979). It is possible that the correlation between the DRC model's representations and neural activity reflects cortical sensitivity to regularity, whereas the correlation between the ANN model and neural activity reflects sensitivity to consistency. Future research should seek to disambiguate the neural correlates of consistency and regularity.

The internal representations generated by the two models were moderately and significantly correlated ($r = 0.52$, $p < 0.0001$). The strength of correlation was surprising, as the theoretical bases for the models are highly divergent. ANN models learn representations over the repeated presentation of stimuli, and they assume that many units are able to contribute to the representation of a given word and, furthermore, that these units reflect generalized lexical consistency. No single unit codes for a particular spelling-to-sound correspondence. Instead, the pronunciation common to a letter string shared across words is coded by multiple units (Seidenberg & McClelland, 1989). The DRC model, in contrast, assumes that once the model finishes cycling, only the units corresponding to the relevant symbolic rules and the corresponding lexical entry for a given input word will be activated (Coltheart et al., 2001). Thus, the correlation between the intermediate representations in these models may reflect that both types are successful at accounting for reading aloud, in that they ultimately represent valid phonological outputs for the input words. By analogy to Marr's framework for levels of analysis in cognitive systems (Marr, 1982), these models differ at the level of their representations and algorithms, but their correlation may arise from solving the same specific computational problem of reading aloud. For example, the lexicons and grapheme–phoneme rules in the DRC model may represent a high-level interpretation of the more basic, neurally inspired mechanism instantiated in the ANN model. The presumably more basic mechanisms of the ANN model, on the other hand, may reflect areas of feature convergence, where intermediate blends of orthographic features are remapped to accommodate the statistical regularities of phonological output.

One potential reason that the ANN and DRC models showed fairly similar correspondence to neural representations is the steps taken to equate the model representations. The DRC model representations were derived from the GPC route, with the lexical route allowed to contribute to the final outputs but not considered in our main analysis (but see Supplementary Figure 3 and Supplementary Table 1 in the online supporting information for near identical results including

the lexical route in DRC representations). To ensure fair treatment of both models, we limited the representations derived from the ANN model to a single hidden layer representing the orthography-to-phonology pathway, allowing external frequency-based sematic input to contribute to the model's output but not directly affecting the hidden layer. There are clear extensions that can be made to the ANN model that are likely to improve correspondence to neural representations. Using deeper networks (networks with more hidden layers), potentially with neurobiologically inspired patterns of intra- versus interlayer connectivity (see Laszlo & Armstrong, 2014; Laszlo & Plaut, 2012), and adding recurrence are natural ways to improve the ANN model's fidelity with neural processing. It is less clear how one would extend the DRC model to emulate the complexity of the brain. Furthermore, a plethora of work in the connectionist tradition has examined potential semantic route implementations for ANN models of reading (Harm & Seidenberg, 2004; Monaghan et al., 2017; Ueno et al., 2011); we are unaware of any work that develops the dual-route model's semantic pathway. In this sense, enforcing fair comparison between the models has actually limited the fair treatment of the ANN model with respect to its correspondence with neural representation.

These results support ANNs with distributed representations as a preferred model for the neural processing of reading. While the symbolic representations from the DRC model do correlate with neural activity, the effect is spatially restricted compared with the RSA map from the ANN representations. More stringently, removing the variance due the ANN model reveals that the DRC model corresponds to neural representations in no additional brain areas. This suggests that word representations in the brain are coded as distributed patterns of activation and are more sensitive to spelling–sound consistency than to rule-based regularity. The fact that the DRC model explains no neural variance above and beyond that of the ANN model also may suggest that the lower-level features of the ANN account for the GPC rules of the DRC model to some degree.

More generally, our results demonstrate the promise of directly comparing and adjudicating between computational models using neural data. While behavioral evidence alone has been insufficient for arbitration, RSA allowed us to compare the neural plausibility of predictions made by the ANN and DRC models. The correlation of the ANN model with a large number of relevant brain areas, and its ability to explain neural variance beyond that of the DRC model, demonstrate that the ANN model holds particular promise for connecting cognitive and neural approaches to reading in a computationally principled way. This method is simple to extend to any computational model that produces explicit, quantitative predictions.

One limitation to this study is the approximation of semantics used in the ANN model. While using frequency as a proxy for semantics reflects the intuitive idea that more common words are more likely to be understood, frequency is not a true semantic variable. Most notably, it ignores any effects that might vary based on the actual semantic content of words. For example, imageability has been shown to affect word reading speed and accuracy (Strain et al., 1995) and has distinct neural correlates compared to consistency and frequency (Graves et al., 2010). However, this frequency manipulation has been successfully used as an approximation for semantic input in multiple previous studies using the same model architecture used in this study (Plaut et al., 1996; Woollams et al., 2007). We plan to expand the ANN model to include a fuller implementation of semantics in future studies.

Exactly what the nature of these semantic representations should be remains to be determined. Optimization of hardware and advances in machine learning techniques have permitted modelling language at a larger scale than ever before. The recent success of transformer models suggests that computations over word co-occurrence and position are sufficient to generate relatively comprehensible and syntactically correct sentences (Vaswani et al., 2017). Indeed, co-occurrence

based models are capable of predicting human behavior across a wide range of semantic tasks (Hofmann et al., 2018; Mandera et al., 2017; Pereira et al., 2016). However, a growing body of evidence from the embodied or grounded cognition literature suggests that neural representations of body states, emotions, and somatosensory experience play a role in how the brain computes word meaning (Barsalou, 2008). Both the co-occurrence and embodied feature approaches appear to be relevant to how the brain represents semantics. This was shown in a study using a word familiarity task in fMRI, which demonstrated distinct neural correlates of co-occurrence based and abstract semantic feature based (including embodied feature) representations of abstract words (Wang et al., 2017). Furthermore, context is known to have an effect on the processing of semantics, such as when disambiguating the meaning of homonyms (Rodd, 2020). As such, an online semantic subnetwork accounting for embodied features, co-occurrence, and context is a likely candidate for producing the most cognitively relevant semantic representations.

A second limitation is the nature of the phonological representations. Phonology is coded here as binary units, representing the presence or absence of a phoneme. Spatially, the phonological vector representation is subdivided into three sections corresponding to the onset, vowel, or coda of a word. This coding scheme has been used productively in previous modelling work (Plaut et al., 1996; Woollams et al., 2007), but is by no means a perfect all-purpose representation. It captures phoneme-level similarity (e.g., "cat" and "hat" are more similar than "cat" and "dog") but cannot capture the more fine-grained differences that phonetic features might (e.g., "/p/" and "/b/" are exactly as similar as "/p/" and "/a/"). While our results agree with neurobiological models of phonology output, showing correspondence between the phonological representations and neural activity in the pre- and postcentral gyri (Indefrey & Levelt, 2004), future research could examine different phonetic or articulatory feature representations to determine which best fits the observed neural activity. In particular, it seems likely that more detailed phonological representations would improve correspondence with the middle and superior temporal gyri, which are sensitive to familiar phoneme patterns (Indefrey & Levelt, 2004; Price, 2012).

A final limitation is the performance of the ANN model on pronounceable nonwords. While the model did quite well, achieving a human-level accuracy of 97.2% on the tested real words, its 84.9% accuracy (77.4% by exact-match criterion) on the nonword test list remains somewhat below human-level performance. This issue may be partially addressed by extending the ANN model to more than one hidden layer. A single hidden layer allows for blends and partial combinations of features to be mapped to outputs, but multiple layers allow for combinations of those feature blends to be processed (LeCun et al., 2015). Deeper neural networks may not only improve performance on the nonword test set but may also better model the neural processes involved in reading.

Here we have directly compared two major computational cognitive models of reading in terms of their correspondence with neural data acquired during reading. The distributed ANN model accounted for unique variance beyond that of the DRC, revealing correspondence with neural representations in areas previously shown to be related to orthography or orthography–phonology mapping. Critically, correspondence of model representations with neural representations allows for the direct interpretation of neural patterns in terms of the computational function being implemented.

ACKNOWLEDGMENTS

We thank Dr. Blair Armstrong for providing a version of the LENS neural network simulator that could be complied with later versions of Linux.

FUNDING INFORMATION

William W. Graves, Eunice Kennedy Shriver National Institute of Child Health and Human Development (http://dx.doi.org/10.13039/100009633), Award ID: HD065839.

AUTHOR CONTRIBUTIONS

William W. Graves developed the study concept and collected the data. Ryan Staples analyzed the data and drafted the manuscript. Both authors contributed to study design and the interpretation of the results. Both authors approved the final manuscript for submission.

REFERENCES

Alexander, M. P., Friedman, R. B., Loverso, F., & Fischer, R. S. (1992). Lesion localization of phonological aphasia. *Brain and Language*, *43*, 83–95. https://doi.org/10.1016/0093-934X(92)90022-7

Barsalou, L. W. (2008). Grounded cognition. *Annual Review of Psychology*, *59*, 617–645. https://doi.org/10.1146/annurev.psych.59.103006.093639

Binder, J. R., & Desai, R. H. (2011). The neurobiology of semantic memory. *Trends in Cognitive Sciences*, *15*(11), 527–536. https://doi.org/10.1016/j.tics.2011.10.001

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. https://doi.org/10.1093/cercor/bhp055

Binder, J. R., Medler, D. A., Desai, R., Conant, L. L., & Liebenthal, E. (2005). Some neurophysiological constraints on models of word naming. *NeuroImage*, *27*(3), 677–693. https://doi.org/10.1016/j.neuroimage.2005.04.029

Binder, J. R., Medler, D. A., Westbury, C. F., Liebenthal, E., & Buchanan, L. (2006). Tuning of the human left fusiform gyrus to sublexical orthographic structure. *NeuroImage*, *33*(2), 739–748. https://doi.org/10.1016/j.neuroimage.2006.06.053

Binder, J. R., & Mohr, J. P. (1992). The topography of callosal reading pathways: A case control analysis. *Brain*, *115*, 1807–1826.

Binder, J. R., Pillay, S. B., Humphries, C. J., Gross, W. L., Graves, W. W., & Book, D. S. (2016). Surface errors without semantic impairment in acquired dyslexia: A voxel-based lesion–symptom mapping study. *Brain*, *139*(5), 1517–1526. https://doi.org/10.1093/brain/aww029

Brem, S., Bach, S., Kucian, K., Guttorm, T. K., Martin, E., Lyytinen, H., Brandeis, D., & Richardson, U. (2010). Brain sensitivity to print emerges when children learn letter–speech sound correspondences. *Proceedings of the National Academy of Sciences*, *107*(17), 7939–7944. https://doi.org/10.1073/pnas.0904402107

Buchsbaum, B. R., Baldo, J., Okada, K., Berman, K. F., Dronkers, N., D'Esposito, M., & Hickok, G. (2011). Conduction aphasia, sensory-motor integration, and phonological short-term memory—An aggregate analysis of lesion and fMRI data. *Brain and Language*, *119*(3), 119–128. https://doi.org/10.1016/j.bandl.2010.12.001

Buckner, R. L., Andrews-Hanna, J. R., & Schacter, D. L. (2008). The brain's default network: Anatomy, function, and relevance to disease. *Annals of the NewYork Academy of Sciences*, *1124*, 1–38. https://doi.org/10.1196/annals.1440.011

Cattinelli, I., Borghese, N. A., Gallucci, M., & Paulesu, E. (2013). Reading the reading brain: A new meta-analysis of functional imaging data on reading. *Journal of Neurolinguistics*, *26*(1), 214–238. https://doi.org/10.1016/j.jneuroling.2012.08.001

Chao, L. L., Haxby, J. V., & Martin, A. (1999). Attribute-based neural substrates in temporal cortrex for perceiving and knowing about objects. *Nature Neuroscience*, *2*(10), 913–919.

Cheyette, S. J., & Plaut, D. C. (2017). Modeling the N400 ERP component as transient semantic over-activation within a neural network model of word comprehension. *Cognition*, *162*, 153–166. https://doi.org/10.1016/j.cognition.2016.10.016

Cohen, L., Dehaene, S., Naccache, L., Lehéricy, S., Dehaene-Lambertz, G., Hénaff, M.-A., & Michel, F. (2000). The visual word form area: Spatial and temporal characterization of an initial stage of reading in normal subjects and posterior split-brain patients. *Brain*, *123*(2), 291–307.

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*, 204–256.

Cortese, M. J., & Simpson, G. B. (2000). Regularity effects in word naming: What are they? *Memory & Cognition*, *28*, 1269–1276.

Cox, R. W. (1996). AFNI: Software for analysis and visualization of functional magnetic resonance neuroimages. *Computers and Biomedical Research*, *29*(3), 162–173.

Damasio, A. R., & Damasio, H. (1983). The anatomic basis of pure alexia. *Neurology*, *33*(12), 1573–1583.

Damasio, H., & Damasio, A. R. (1980). The anatomical basis of conduction aphasia. *Brain*, *103*, 337–350.

Damasio, H., Tranel, D., Grabowski, T., Adolphs, R., & Damasio, A. (2004). Neural systems behind word and concept retrieval. *Cognition*, *92*(1–2), 179–229. https://doi.org/10.1016/j.cognition.2002.07.001

Dehaene, S., & Cohen, L. (2011). The unique role of the visual word form area in reading. *Trends in Cognitive Science*, *15*(6), 254–262. https://doi.org/10.1016/j.tics.2011.04.003

Dehaene, S., Jobert, A., Naccache, P., Ciuciu, P., Poline, J.-B., Le Bihan, D., & Cohen, L. (2004). Letter binding and invariant recognition of masked words. *Psychological Science*, *15*, 307–313.

Fiez, J. A., Balota, D. A., Raichle, M. E., & Petersen, S. E. (1999). Effects of lexicality, frequency, and spelling-to-sound correspondence on the functional anatomy of reading. *Neuron*, *24*, 205–218.

Fiez, J. A., & Petersen, S. E. (1998). Neuroimaging studies of word reading. *Proceedings of the National Academy of Sciences*, *95*, 914–921.

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. https://doi.org/10.1016/j.cortex.2017.02.017

Forstmann, B. U., Wagenmakers, E. J., Eichele, T., Brown, S., & Serences, J. T. (2011). Reciprocal relations between cognitive neuroscience and formal cognitive models: Opposites attract? *Trends in Cognitive Sciences*, *15*(6), 272–279. https://doi.org/10.1016/j.tics.2011.04.002

Glushko, R. J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance*, *5*(4), 674–691. https://doi.org/10.1037/0096-1523.5.4.674

Graves, W. W., Binder, J. R., Desai, R. H., Humphries, C., Stengel, B. C., & Seidenberg, M. S. (2014). Anatomy is strategy: Skilled reading differences associated with structural connectivity differences in the reading network. *Brain and Language*, *133*, 1–13. https://doi.org/10.1016/j.bandl.2014.03.005

Graves, W. W., Desai, R., Humphries, C., Seidenberg, M. S., & Binder, J. R. (2010). Neural systems for reading aloud: A multiparametric approach. *Cerebral Cortex*, *20*(8), 1799–1815. https://doi.org/10.1093/cercor/bhp245

Guenther, F. H., Ghosh, S. S., & Tourville, J. A. (2006). Neural modeling and imaging of the cortical interactions underlying syllable production. *Brain and Language*, *96*(3), 280–301. https://doi.org/10.1016/j.bandl.2005.06.001

Hanke, M., Halchenko, Y. O., Sederberg, P. B., Hanson, S. J., Haxby, J. V., & Pollmann, S. (2009). PyMVPA: A python toolbox for multivariate pattern analysis of fMRI data. *Neuroinformatics*, *7*(1), 37–53. https://doi.org/10.1007/s12021-008-9041-y

Harm, M. W., & Seidenberg, M. S. (2004). Phonology, reading acquisition, and dyslexia: Insights from connectionist models. *Psychological Review*, *106*, 491–528.

Hashimoto, R., & Sakai, K. L. (2004). Learning letters in adulthood: Direct visualization of cortical plasticity for forming a new link between orthography and phonology. *Neuron*, *42*, 311–322.

Haxby, J. V., Connolly, A. C., & Guntupalli, J. S. (2014). Decoding neural representational spaces using multivariate pattern analysis. *Annual Review of Neuroscience*, *37*, 435–456. https://doi.org/10.1146/annurev-neuro-062012-170325

Hillis, A. E., & Caramazza, A. (1991). Category-specific naming and comprehension impairment: A double dissociation. *Brain*, *114*, 2081–2094.

Hinton, G. E. (1989). Connectionist learning procedures. *Artificial Intelligence*, *40*, 185–234.

Hoffman, P., Lambon Ralph, M. A., & Woollams, A. M. (2015). Triangulation of the neurocomputational architecture underpinning reading aloud. *Proceedings of the National Academy of Sciences*, *112*(28), E3719–3728. https://doi.org/10.1073/pnas.1502032112

Hofmann, M. J., Biemann, C., Westbury, C., Murusidze, M., Conrad, M., & Jacobs, A. M. (2018). Simple co-occurrence statistics reproducibly predict association ratings. *Cognitive Science*, *42*(7), 2287–2312. https://doi.org/10.1111/cogs.12662

Humphreys, G. F., Hoffman, P., Visser, M., Binney, R. J., & Lambon Ralph, M. A. (2015). Establishing task- and modality-dependent dissociations between the semantic and default mode networks. *Proceedings of the National Academy of Sciences*, *112*(25), 7857–7862. https://doi.org/10.1073/pnas.1422760112

Humphries, C., Binder, J. R., Medler, D. A., & Liebenthal, E. (2007). Time course of semantic processes during sentence comprehension: An fMRI study. *NeuroImage*, *36*(3), 924–932. https://doi.org/10.1016/j.neuroimage.2007.03.059

Indefrey, P., & Levelt, W. J. (2004). The spatial and temporal signatures of word production components. *Cognition*, *92*(1–2), 101–144. https://doi.org/10.1016/j.cognition.2002.06.001

Jared, D. (2002). Spelling–sound consistency and regularity effects in word naming. *Journal of Memory and Language*, *46*(4), 723–750. https://doi.org/10.1006/jmla.2001.2827

Jared, D., McRae, K., & Seidenberg, M. (1990). The basis of consistency effects in word reading. *Journal of Memory and Language*, *29*(6), 687–715.

Jobard, G., Crivello, F., & Tzourio-Mazoyer, N. (2003). Evaluation of the dual route theory of reading: A metanalysis of 35 neuroimaging studies. *NeuroImage*, *20*(2), 693–712. https://doi.org/10.1016/s1053-8119(03)00343-4

Kable, J. W., Kan, I. P., Wilson, A., Thompson-Schill, S. L., & Chatterjee, A. (2005). Conceptual representations of action in the lateral temporal cortex. *Journal of Cognitive Neuroscience*, *17*(12), 1855–1870.

Kriegeskorte, N., & Kievit, R. A. (2013). Representational geometry: Integrating cognition, computation, and the brain. *Trends in Cognitive Sciences*, *17*(8), 401–412. https://doi.org/10.1016/j.tics.2013.06.007

Kriegeskorte, N., Mur, M., & Bandettini, P. (2008). Representational similarity analysis—Connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, *2*, 4. https://doi.org/10.3389/neuro.06.004.2008

Lambon Ralph, M. A., Jefferies, E., Patterson, K., & Rogers, T. T. (2017). The neural and computational bases of semantic cognition. *Nature Reviews Neuroscience*, *18*(1), 42–55. https://doi.org/10.1038/nrn.2016.150

Lambon Ralph, M. A., Lowe, C., & Rogers, T. T. (2007). Neural basis of category-specific semantic deficits for living things: Evidence from semantic dementia, HSVE and a neural network model. *Brain*, *130*(4), 1127–1137. https://doi.org/10.1093/brain/awm025

Lancaster, J. L., Woldorff, M. G., Parsons, L. M., Liotti, M., Freitas, C. S., Rainey, L., Kochunov, P. V., Nickerson, D., Mikiten, S. A., & Fox, P. T. (2000). Automated Talairach atlas labels for functional brain mapping. *Human Brain Mapping*, *10*, 120–131.

Laszlo, S., & Armstrong, B. C. (2014). PSPs and ERPs: Applying the dynamics of post-synaptic potentials to individual units in simulation of temporally extended event-related potential reading data. *Brain and Language*, *132*, 22–27. https://doi.org/10.1016/j.bandl.2014.03.002

Laszlo, S., & Plaut, D. C. (2012). A neurally plausible parallel distributed processing model of event-related potential word reading data. *Brain and Language*, *120*(3), 271–281. https://doi.org/10.1016/j.bandl.2011.09.001

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. https://doi.org/10.1038/nature14539

Leff, A. P., Crewes, H., Plant, G. T., Scott, S. K., Kennard, C., & Wise, R. J. S. (2001). The functional anatomy of single-word reading in patients with hemianopic and pure alexia. *Brain*, *124*, 510–521.

Levandowsky, M., & Winter, D. (1971). Distance between sets. *Nature*, *234*, 34–35.

Liuzzi, A. G., Bruffaerts, R., Peeters, R., Adamczuk, K., Keuleers, E., De Deyne, S., Storms, G., Dupont, P., & Vandenberghe, R. (2017). Cross-modal representation of spoken and written word meaning in left pars triangularis. *NeuroImage*, *150*, 292–307. https://doi.org/10.1016/j.neuroimage.2017.02.032

Malone, P. S., Glezer, L. S., Kim, J., Jiang, X., & Riesenhuber, M. (2016). Multivariate pattern analysis reveals category-related organization of semantic representations in anterior temporal cortex. *Journal of Neuroscience*, *36*(39), 10089–10096. https://doi.org/10.1523/JNEUROSCI.1599-16.2016

Mandera, P., Keuleers, E., & Brysbaert, M. (2017). Explaining human performance in psycholinguistic tasks with models of semantic similarity based on prediction and counting: A review and

empirical validation. *Journal of Memory and Language*, *92*, 57–78. https://doi.org/10.1016/j.jml.2016.04.001
Mano, Q. R., Humphries, C., Desai, R. H., Seidenberg, M. S., Osmon, D. C., Stengel, B. C., & Binder, J. R. (2013). The role of left occipitotemporal cortex in reading: Reconciling stimulus, task, and lexicality effects. *Cerebral Cortex*, *23*(4), 988–1001. https://doi.org/10.1093/cercor/bhs093
Marr, D. (1982). *Vision*. New York: W. H. Freeman.
Mattheiss, S. R., Levinson, H., & Graves, W. W. (2018). Duality of Function: Activation for meaningless nonwords and semantic codes in the same brain areas. *Cerebral Cortex*, *28*(7), 2516–2524. https://doi.org/10.1093/cercor/bhy053
McCandliss, B. D., Cohen, L., & Dehaene, S. (2003). The visual word form area: Expertise for reading in the fusiform gyrus. *Trends in Cognitive Sciences*, *7*(7), 293–299. https://doi.org/10.1016/s1364-6613(03)00134-7
McCann, R. S., & Besner, D. (1987). Reading pseudohomophones: Implications for models of pronunciation assembly and the locus of word frequency effects in naming. *Journal of Experimental Psychology: Human Perception and Performance*, *13*, 13–24.
Monaghan, P., Chang, Y.-N., Welbourne, S., & Brysbaert, M. (2017). Exploring the relations between word frequency, language exposure, and bilingualism in a computational model of reading. *Journal of Memory and Language*, *93*, 1–21. https://doi.org/10.1016/j.jml.2016.08.003
Mumford, J. A., Turner, B. O., Ashby, F. G., & Poldrack, R. A. (2012). Deconvolving BOLD activation in event-related designs for multivoxel pattern classification analyses. *NeuroImage*, *59*(3), 2636–2643. https://doi.org/10.1016/j.neuroimage.2011.08.076
Murphy, K. A., Jogia, J., & Talcott, J. B. (2019). On the neural basis of word reading: A meta-analysis of fMRI evidence using activation likelihood estimation. *Journal of Neurolinguistics*, *49*, 71–83. https://doi.org/10.1016/j.jneuroling.2018.08.005
Oosterhof, N. N., Connolly, A. C., & Haxby, J. V. (2016). CoSMoMVPA: Multi-modal multivariate pattern analysis of neuroimaging data in Matlab/GNU Octave. *Frontiers in Neuroinformatics*, *10*, 27. https://doi.org/10.3389/fninf.2016.00027
Pereira, F., Gershman, S., Ritter, S., & Botvinick, M. (2016). A comparative evaluation of off-the-shelf distributed semantic representations for modelling behavioural data. *Cognitive Neuropsychology*, *33*(3–4), 175–190. https://doi.org/10.1080/02643294.2016.1176907
Perry, C., Ziegler, J. C., & Zorzi, M. (2007). Nested incremental modeling in the development of computational theories: The CDP+ model of reading aloud. *Psychological Review*, *114*(2), 273–315. https://doi.org/10.1037/0033-295X.114.2.273
Perry, C., Ziegler, J. C., & Zorzi, M. (2010). Beyond single syllables: Large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. *Cognitive Psychology*, *61*(2), 106–151. https://doi.org/10.1016/j.cogpsych.2010.04.001
Pillay, S. B., Stengel, B. C., Humphries, C., Book, D. S., & Binder, J. R. (2014). Cerebral localization of impaired phonological retrieval during rhyme judgment. *Annals of Neurology*, *76*(5), 738–746. https://doi.org/10.1002/ana.24266
Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*, 56–115.
Plaut, D. C., & Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. *Cognitive Neuropsychology*, *10*(5), 377–500. https://doi.org/10.1080/02643299308253469
Popov, V., Ostarek, M., & Tenison, C. (2018). Practices and pitfalls in inferring neural representations. *NeuroImage*, *174*, 340–351. https://doi.org/10.1016/j.neuroimage.2018.03.041
Price, C. J. (2012). A review and synthesis of the first 20 years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*, *62*(2), 816–847. https://doi.org/10.1016/j.neuroimage.2012.04.062
Price, C. J., & Devlin, J. T. (2011). The interactive account of ventral occipitotemporal contributions to reading. *Trends in Cognitive Science*, *15*(6), 246–253. https://doi.org/10.1016/j.tics.2011.04.001
Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the N400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behavior*, *2*(9), 693–705. https://doi.org/10.1038/s41562-018-0406-4
Raichle, M. E., MacLeod, A. N., Snyder, A. Z., Powers, W. J., Gusnard, D. A., & Schulman, G. L. (2001). A default mode of brain function. *Proceedings of the National Academy of Sciences*, *98*(2), 676–682.
Richlan, F., Kronbichler, M., & Wimmer, H. (2009). Functional abnormalities in the dyslexic brain: A quantitative meta-analysis of neuroimaging studies. *Human Brain Mapping*, *30*(10), 3299–3308. https://doi.org/10.1002/hbm.20752
Rodd, J. (2020). Settling into semantic space: An ambiguity-focused account of word-meaning access. *Perspectives on Psychological Science*, *15*(2), 411–427.
Rothlein, D., & Rapp, B. (2014). The similarity structure of distributed neural responses reveals the multiple representations of letters. *NeuroImage*, *89*, 331–344. https://doi.org/10.1016/j.neuroimage.2013.11.054
Rudrauf, D., Mehta, S., Bruss, J., Tranel, D., Damasio, H., & Grabowski, T. J. (2008). Thresholding lesion overlap difference maps: Application to category-related naming and recognition deficits. *NeuroImage*, *41*(3), 970–984. https://doi.org/10.1016/j.neuroimage.2007.12.033
Sandak, R., Mencl, W. E., Frost, S. J., & Pugh, K. R. (2004). The neurobiological basis of skilled and impaired reading: Recent findings and new directions. *Scientific Studies of Reading*, *8*(3), 273–292. https://doi.org/10.1207/s1532799xssr0803_6
Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*, 523–568.
Shaywitz, B. A., Shaywitz, S. E., Blachman, B. A., Pugh, K. R., Fulbright, R. K., Skudlarski, P., Mencl, W. E., Constable, R. T., Holahan, J. M., Marchione, K. E., Fletcher, J. M., Lyon, G. R., & Gore, J. C. (2004). Development of left occipitotemporal systems for skilled reading in children after a phonologically-based intervention. *Biological Psychiatry*, *55*(9), 926–933. https://doi.org/10.1016/j.biopsych.2003.12.019
Shaywitz, B. A., Skudlarski, P., Holahan, J. M., Marchione, K. E., Constable, R. T., Fulbright, R. K., Zelterman, D., Lacadie, C., & Shaywitz, S. E. (2007). Age-related changes in reading systems of dyslexic children. *Annals of Neurology*, *61*(4), 363–370. https://doi.org/10.1002/ana.21093
Strain, E., & Herdman, C. M. (1999). Imageability effects in word naming: An individual differences analysis. *Canadian Journal of Experimental Psychology*, *53*, 347–359. https://doi.org/10.1037/h0087322
Strain, E., Patterson, K., & Seidenberg, M. S. (1995). Semantic effects in single-word naming. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *21*(5), 1140–1154. https://doi.org/10.1037/0278-7393.21.5.1140
Taylor, J. S., Rastle, K., & Davis, M. H. (2013). Can cognitive models explain brain activation during word and pseudoword reading? A meta-analysis of 36 neuroimaging studies. *Psychological Bulletin*, *139*(4), 766–791. https://doi.org/10.1037/a0030266

Taylor, J. S. H., Davis, M. H., & Rastle, K. (2017). Comparing and validating methods of reading instruction using behavioural and neural findings in an artificial orthography. *Journal of Experimental Psychology: General*, *146*(6), 826–858. https://doi.org/10.1037/xge0000301

Turkeltaub, P. E., Eden, G. F., Jones, K. M., & Zeffiro, T. A. (2002). Meta-analysis of the functional neuroanatomy of single-word reading: Method and validation. *NeuroImage*, *16*(3), 765–780. https://doi.org/10.1006/nimg.2002.1131

Twomey, T., Kawabata Duncan, K. J., Price, C. J., & Devlin, J. T. (2011). Top-down modulation of ventral occipito-temporal responses during visual word recognition. *NeuroImage*, *55*(3), 1242–1251. https://doi.org/10.1016/j.neuroimage.2011.01.001

Ueno, T., Saito, S., Rogers, T. T., & Lambon Ralph, M. A. (2011). Lichtheim 2: Synthesizing aphasia and the neural basis of language in a neurocomputational model of the dual dorsal-ventral language pathways. *Neuron*, *72*(2), 385–396. https://doi.org/10.1016/j.neuron.2011.09.013

van der Mark, S., Bucher, K., Maurer, U., Schulz, E., Brem, S., Buckelmuller, J., Kronbichler, M., Loenneker, T., Klaver, P., Martin, E., & Brandeis, D. (2009). Children with dyslexia lack multiple specializations along the visual word-form (VWF) system. *NeuroImage*, *47*(4), 1940–1949. https://doi.org/10.1016/j.neuroimage.2009.05.021

Vaswani, A., Shazeer, N., Parmer, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Informatrion Processing Systems*, 6000–6010. https://arxiv.org/abs/1706.03762

Vinckier, F., Dehaene, S., Jobert, A., Dubus, J. P., Sigman, M., & Cohen, L. (2007). Hierarchical coding of letter strings in the ventral stream: Dissecting the inner organization of the visual word-form system. *Neuron*, *55*(1), 143–156. https://doi.org/10.1016/j.neuron.2007.05.031

Wang, X., Wu, W., Ling, Z., Xu, Y., Fang, Y., Wang, X., Binder, J. R., Men, W., Gao, J. H., & Bi, Y. (2017). Organizational principles of abstract words in the human brain. *Cereberal Cortex*, *28*(12), 4305–4318. https://doi.org/10.1093/cercor/bhx283

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), 3047. https://doi.org/10.1038/s41598-018-21062-0

Welbourne, S. R., & Lambon Ralph, M. A. (2005). Exploring the impact of plasticity-related recovery after brain damage in a connectionist model of single-word reading. *Cognitive, Affective, & Behavioral Neuroscience*, *5*(1), 77–92.

Woollams, A. M., Lambon Ralph, M. A., Plaut, D. C., & Patterson, K. (2007). SD-squared: On the association between semantic dementia and surface dyslexia. *Psychological Review*, *114*(2), 316–339.

Zevin, J., & Seidenberg, M. (2006). Simulating consistency effects and individual differences in nonword naming: A comparison of current models. *Journal of Memory and Language*, *54*(2), 145–160. https://doi.org/10.1016/j.jml.2005.08.002

Zhao, L., Chen, C., Shao, L., Wang, Y., Xiao, X., Chen, C., Yang, J., Zevin, J., & Xue, G. (2017). Orthographic and phonological representations in the fusiform cortex. *Cerebral Cortex 27*(11), 5197–5210. https://doi.org/10.1093/cercor/bhw300