September 18, 2023
This is a preprint that has not undergone peer review yet.

Convolutional networks can model the functional modulation of MEG responses during reading

Marijn van Vliet[1*], Oona Rinkinen[1], Takao Shimizu[1], Anni-Mari Niskanen[1], Barry Devereux[3], and Riitta Salmelin[1,2]

[1]Department of Neuroscience and Biomedical Engineering, Aalto University, Finland
[2]Aalto NeuroImaging, Aalto University, Finland
[3]School of Electronics, Electrical Engineering and Computer Science, Queen's University Belfast, United Kingdom
[*]Corresponding author: marijn.vanvliet@aalto.fi

Abstract

Neuroimaging studies have provided a wealth of information about when and where changes in brain activity might be expected during reading. We sought to better understand the computational steps that give rise to such task-related modulations of neural activity by using a convolutional neural network to model the macro-scale computations necessary to perform single-word recognition. We presented the model with stimuli that had been shown to human volunteers in an earlier magnetoencephalography (MEG) experiment and evaluated whether the same experimental effects could be observed in both brain activity and model. In a direct comparison between model and MEG recordings, the model accurately predicted the amplitude changes of three evoked MEG response components commonly observed during single-word reading. In contrast to traditional models of reading, our model directly operates on the pixel values of an image containing text. This allowed us to simulate the whole gamut of processing from the detection and segmentation of letter shapes to word-form identification, with the deep learning architecture facilitating inclusion of a large vocabulary of 10k Finnish words. Interestingly, the key to achieving the desired behavior was to use a noisy activation function for the units in the model as well as to obey word frequency statistics when repeating stimuli during training. We conclude that the deep learning techniques that revolutionized models of object recognition can also create models of reading that can be straightforwardly compared to neuroimaging data, which will greatly facilitate testing and refining theories on language processing in the brain.

1 Introduction

What computational operations is the brain performing when it recognizes some splotches of ink on a piece of paper as a meaningful word? This question has been the focus of a large number of neuroimaging studies that examine brain activity during reading. Noninvasive measurement techniques such as electroencephalography (EEG),[1] magnetoencephalography (MEG)[2] and functional magnetic resonance imaging (fMRI)[3] have provided a wealth of information about when and where changes in activity might be expected during various tasks involving orthographic processing.[4]

In the case of visual word recognition, a lot of these results come in the form of changes in the amplitude of specific components of the neural response evoked by stimuli that are designed to create interesting experimental contrasts.[5] Such evoked components reflect macro-level computations — that is, the net result of thousands of individual biological neurons working together. Taken together, the results indicate the presence of a processing pipeline, starting with the extraction of low-level visual features (e.g., edges, line segments), which are subsequently refined into more complex features (e.g., letter shapes) and further into lexical features (e.g., bigrams, words).[6] While neuroimaging studies provide us with

[1] Grainger and Holcomb, 2009
[2] Salmelin, 2007
[3] Price, 2012
[4] Carreiras et al., 2014
[5] Luck, 2014
[6] Dehaene et al., 2005; Grainger and Holcomb, 2009

information about what processing steps are performed where and when, the observed data alone yield little information as to what kind of computations are performed during these steps.[7] To develop such an understanding, we need to make these computations explicit, model them, and test and refine the model against the data provided by imaging studies.[8]

[7] Poeppel, 2012

[8] Barber and Kutas, 2007; Doerig et al., 2023; Price, 2018

In this study, we aimed to computationally reproduce the results of Vartiainen et al. (2011), which is a representative MEG study that employed experimental contrasts designed to study key processing steps throughout the entire visual word recognition pipeline. The authors catalogued the effects of the experimental contrasts on the amplitudes of all major evoked MEG responses found in the data, and concluded that the significant effects could be attributed to three components that dominate the early MEG time course during visual word recognition,[9] namely:

[9] Jobard et al., 2003; Salmelin, 2007

type-I This component peaks occipitally around 100 ms after stimulus onset, is modulated by the visual complexity of the stimulus and hence thought to reflect the processing of low-level visual features.[10] Vartiainen et al. (2011) used a contrast between stimuli with and without added visual noise to highlight this processing stage.

[10] Tarkiainen et al., 1999

type-II This component peaks occipito-temporally around 150 ms after stimulus onset, and is modulated by whether the stimulus contains letters of the participant's native alphabet.[11] Hence, this component is colloquially referred to as the "letter string" response and thought to reflect letter shape detection. Vartiainen et al. (2011) used a contrast between letter strings and symbol strings (visually similar to letters but not part of the Finnish alphabet) to highlight this processing stage.

[11] Parviainen et al., 2006; Tarkiainen et al., 1999

N400m This component peaks at around 400 ms after stimulus onset, and is often studied in the context of priming,[12] but experimental effects can also be observed when using isolated stimuli. For example, this component is modulated by the lexicality of the stimulus, hence associated with more lexical processing.[13] Vartiainen et al. (2011) used a contrast between consonant strings, valid Finnish words and pseudowords to highlight this processing stage.

[12] Kutas and Federmeier, 2011

[13] Halgren et al., 2002; Helenius et al., 1998; Salmelin, 2007; Service et al., 2007

These three components are also observed in EEG studies (referred to as the P1, N1/N170 and N400 components respectively[14]). Together, they are indicative of three processing stages during visual word recognition: basic visual analysis, orthographic analysis and lexical analysis. To explore possible cognitive computations during these stages, we sought to create a computational model that performs the macro-level computations needed to achieve visual word recognition in such a manner as to reproduce the behavior of the three evoked components when presented with the same stimuli as a human participant.

[14] Brem et al., 2009

In past decades, several computational models have been created that successfully capture some aspects of visual word recognition. For the purposes of this study, we draw a distinction between the older "traditional-style" models[15] and the large artificial neural networks trained through deep learning that have been gaining popularity as models of brain function.[16] We will show that a model will need to combine elements from both approaches if it is to be able to reproduce the stimulus-dependent modulation of all evoked components mentioned above.

[15] Norris, 2013

[16] LeCun et al., 2015; Richards et al., 2019

Amongst the first models of visual word recognition, the interactive activation and competition (IAC) model of letter perception by McClelland and Rumelhart,[17] showed how the interplay of bottom-up and top-down connections results in a system capable of "filling in the blanks" when faced with a partially obscured word. This model was later extended to

[17] McClelland and Rumelhart, 1981; Rumelhart and McClelland, 1982

model semantics as well, showing how the activation of some semantic features ("is alive", "is a bird") leads to the subsequent activation of more semantic features ("can fly", "lays eggs"), in what became known as the parallel distributed processing (PDP) framework.[18] Note that while models such as these consist of many interconnected units, those units and their connections do not aim to mimic a biological connectome, but are an abstract representation of macro-level computations performed by one. Coltheart et al. (2001) pointed out the benefits of explicitly defined connections in their dual-route cascaded (DRC) model of visual word recognition and reading out loud, as they grant the researcher exact control over the macro-level computations. However, as the scope of the models increases, "hand-wiring" the connections in the model becomes increasingly difficult. Therefore, most current models employ back-propagation to learn the connection weights between units based on a large training dataset. Together, IAC, PDP and DRC models have been shown to account for many behavioral findings, such as reaction times and recognition errors, in both healthy volunteers and patients.[19] Furthermore, Laszlo and Plaut (2012) demonstrated how a PDP-style model can produce an N400-like signal by summing the activity of the computational units in specific layers of the model.

[18] McClelland and Rogers, 2003

[19] McClelland and Rogers, 2003; McLeod et al., 2000; Perry et al., 2007

However, none of the aforementioned models can produce detailed quantitative data that can be directly compared with neuroimaging data, leaving researchers to focus on finding indirect evidence that the same macro-level computations occur in both the model and the brain, with varying levels of success.[20] Due to the computational constrains at the time, the simulated environments in these models are extremely simplified (e.g. the Laszlo and Plaut (2012) model operated exclusively on 3-letter words), whereas the brain data will, by nature, reflect the reading process in more varied and realistic visual settings. An even bigger limitation is the manner in which the initial visual input is presented to the model. These models make use of "letter banks", where each letter of a written word is encoded in as a separate group of inputs. The letters themselves are encoded as either a collection of 16 predefined line segments (Figure 1)[21] or a binary code indicating the letter.[22] This rather high level of visual representation sidesteps having to deal with issues such as visual noise, letters with different scales, rotations and fonts, segmentation of the individual letters, and so on. More importantly, it makes it impossible to create the visual noise and symbol string conditions used in the MEG study we were trying to reproduce[23] to modulate the type-I and type-II components. In order to model the process of visual word recognition to the extent where one may reproduce neuroimaging studies such as Vartiainen et al. (2011), we need to start with a model of vision that is able to directly operate on the pixels of a stimulus.

[20] Barber and Kutas, 2007; Jobard et al., 2003; Protopapas et al., 2016

Figure 1: The 16 predefined line segments of McClelland and Rumelhart (1981) and five example letters

[21] Coltheart et al., 2001; McClelland and Rumelhart, 1981

[22] Laszlo and Armstrong, 2014; Laszlo and Plaut, 2012

[23] Vartiainen et al., 2011

Models of vision seem to have converged on a sequence of convolution-and-pooling operations to simulate the macro-level behavior of the visual cortex.[24] Such models have become much more powerful as advances in deep learning and its software ecosystem are rapidly changing our notion of what is computationally tractable to model.[25] Convolutional neural networks (CNNS) have emerged that perform scale- and rotation-invariant visual object recognition at very high levels of accuracy.[26] Furthermore, they model some functions of the visual cortex well enough that a direct comparison between network state and neuroimaging data has become possible.[27]

[24] Lindsay, 2020; Serre et al., 2007

[25] LeCun et al., 2015; Richards et al., 2019

[26] Dai et al., 2021; Krizhevsky et al., 2017; Simonyan and Zisserman, 2015

[27] Devereux et al., 2018; Schrimpf et al., 2020; Yamins and DiCarlo, 2016

Visual word recognition can be seen as a specialized form of object recognition.[28] Developmental studies suggest that orthographic-specific neuroimaging findings such as the visual word form area (VWFA) and the type-II evoked components emerge after part of our established vision system is reconfigured as we learn to read.[29] Studies on the visual cortex of non-human primates have shown that visual systems even without extensive exposure to

[28] Grainger, 2018

[29] Cohen and Dehaene, 2004; Dehaene-Lambertz et al., 2018a; Parviainen et al., 2006

written language have neural representations that carry enough information to distinguish letter and word shapes,[30] even accounting for distortions,[31] and this finding was reproduced in CNNs. Therefore, CNNs may very well be suitable tools for increasing the scale of traditional connectionist models of reading, thus allowing for a much more realistic modeling of the early orthographic processing steps than what has been possible so far. Indeed, earlier work has established that training a CNN to classify written words can coerce it to form a VWFA-like region on top of an existing visual system,[32] and cause it to form high-level orthographic representations[33] which can be used by a linear regression algorithm to predict human MEG activity in the visual cortex.[34]

[30] Rajalingham et al., 2020
[31] Katti and Arun, 2022
[32] Hannagan et al., 2021
[33] Katti and Arun, 2022; Testolin et al., 2017
[34] Caucheteux and King, 2022

Whereas traditional-style models are typically designed to replicate a specific experimental effect in a biologically feasible manner, deep learning models are designed to perform in a more naturalistic setting, regardless of biological plausibility. Consequently, traditional-style models have been evaluated by comparing the timing and amount of the simulated activity to the amplitude of an evoked component[35] or human reaction times[36] in a tightly controlled experimental setting, and deep learning models have been evaluated using data from participants who are operating in a more challenging setting, such as watching movies,[37] reading sentences[38] or listening to stories.[39] For the latter type of comparisons, information-based metrics are used. For example, representational similarity analysis (RSA)[40] quantifies the extent to which the inner state of a model discriminates between the same types of stimuli as a neural response, and BrainScore[41] quantifies the extent to which a (usually linear) regressor or classifier can use the inner state of a model to predict brain activity. However, information-based metrics by themselves provide at best only a rough indication of how closely a model mimics the computations performed by the brain. It would be desirable to establish a more direct link between a model and the experimental results obtained in neuroscience studies.[42] Therefore, in the current study, we evaluated the performance of a CNN in the spirit of the traditional models, namely by its ability to accurately reproduce the behavior of the type-I, type-II and N400m responses in the tightly controlled experimental setting of Vartiainen et al. (2011). The resulting model combines the ability of deep learning models to operate directly on the pixels of a stimulus and to have a large vocabulary, with the ability of traditional models to replicate specific experimental results.

[35] Laszlo and Armstrong, 2014; Laszlo and Plaut, 2012
[36] Agrawal et al., 2020; Norris and Kinoshita, 2012
[37] Huth et al., 2012
[38] Caucheteux and King, 2022
[39] Huth et al., 2016
[40] Kriegeskorte et al., 2008
[41] Schrimpf et al., 2020
[42] Bowers et al., 2022

We started with a basic CNN architecture, namely VGG-11 ,[43] that had been pre-trained to classify images and evolved it into a model of visual word recognition by re-training it to classify written words. Different versions of the model were created by introducing slight modifications to the base architecture and using different training regimens. The performance of the trained models was evaluated by presenting them with the same set of stimuli that had been used by Vartiainen et al. (2011) and comparing the amount of activity within its layers with the amplitude of the type-I, type-II, and N400m components measured from the human participants.

[43] Simonyan and Zisserman, 2015

2 Results

2.1 The three MEG responses have unique response profiles

The computational models in this study were evaluated by their success in replicating the functional effects observed in the evoked MEG components as recorded by Vartiainen et al. (2011) during the presentation of written words, pseudowords, consonant strings, symbol strings and noise-embedded words (Figure 2A) to 15 participants. Performing minimum norm estimate, dynamic statistical parametric mapping (MNE-dSPM)[44] source

[44] Dale et al., 2000

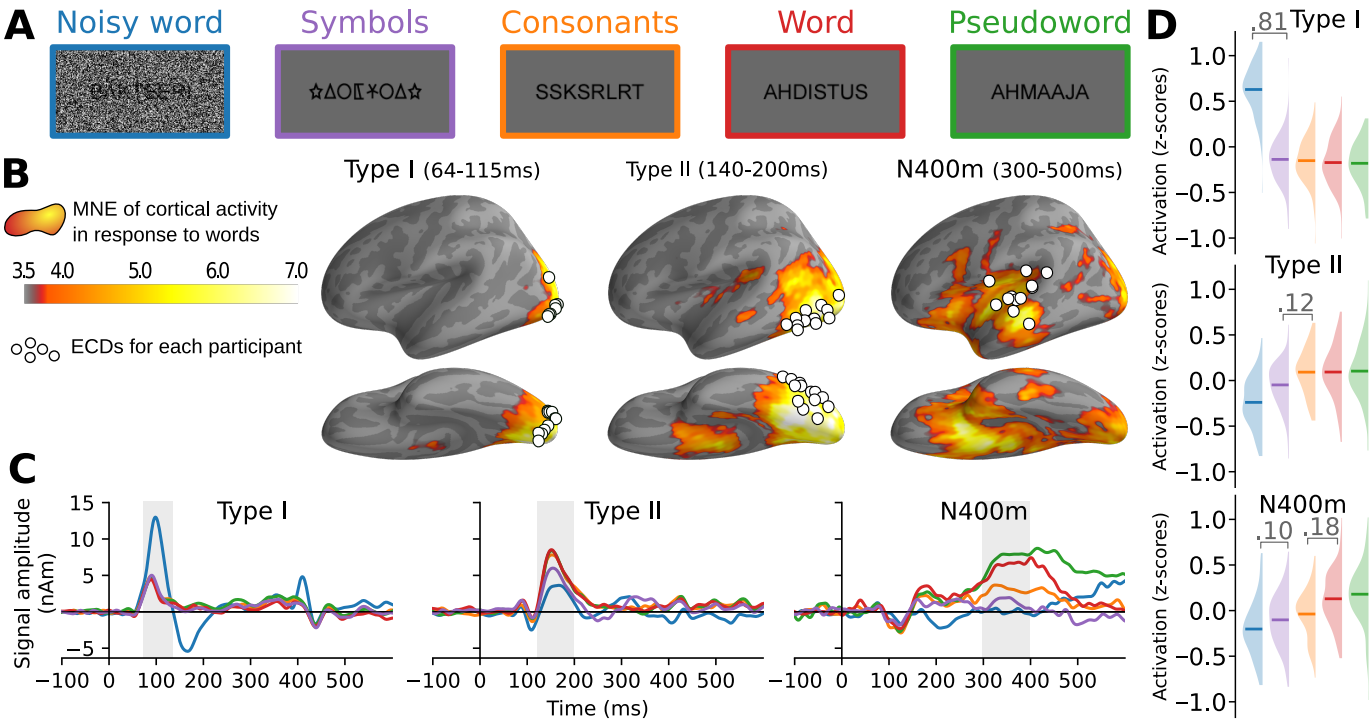

Figure 2: Summary of the MEG results obtained by Vartiainen et al. (2011).
A: Examples of stimuli used in the MEG experiment. Each stimulus contained 7–8 letters or symbols.
B: Source estimate of the evoked MEG activity, using MNE-dSPM. The grand-average activity to word stimuli, averaged for three time intervals, is shown in orange hues. For each time interval, white circles indicate the location of the most representative left-hemisphere ECD for each participant, as determined by Vartiainen et al. (2011).
C: Grand-average time course of signal strength for each group of ECDs in response to the different stimulus types. The traces are color-coded to indicate the stimulus type as shown in A. Shaded regions indicate time periods over which statistical analysis was performed.
D: For each group of ECDs shown in B, and separately for each stimulus type (different colors, see A), the distribution (and mean) of the grand-average response amplitudes to the different stimulus types, obtained by integrating the ECD signal strength over the time intervals highlighted in C. Whenever there is a significant difference (linear mixed effects (LME) model, $p < 0.05$, false discovery rate (FDR) corrected) between two adjacent distributions, the corresponding difference in means is shown.

analysis yielded a comprehensive map of where activity can be found (Figure 2B), which is used in this study mostly to provide context for a deeper investigation into three peaks of activity that can be observed along the ventral stream. Through guided equivalent current dipole (ECD) modeling,[45] the high-dimensional data was summarized as a sparse set of ECDs, each one capturing an isolated spatiotemporal component, allowing us to study selected components in more detail. The original study[46] found three ECD groups along the ventral stream that correspond to the type-I, type-II and N400m responses, based on their location (Figure 2B), timing (Figure 2C) and responses to the different stimulus types (Figure 2D). Similar ECD groups were identified in the right hemisphere (Supplementary Figure 1), but the effects were clearer in the left. Therefore, in this study, we re-used the three ECD groups on the left hemisphere to serve as ground-truth for our modeling efforts.

[45] Salmelin, 2010

[46] Vartiainen et al., 2011

To compare the evoked MEG responses and model layer activity, we collected summary statistics of both, referred to as "response profiles". For the ECD time courses, the response profiles were formed by computing for each stimulus the average signal amplitude in the time intervals indicated in Figure 2C, z-transformed across the stimuli, and averaged across participants to obtain grand-average response profiles (Figure 2D). The response profile for the type-I response demonstrates how this occipital component is driven by the visual complexity of the stimulus and is characterized by a large response to noise-embedded

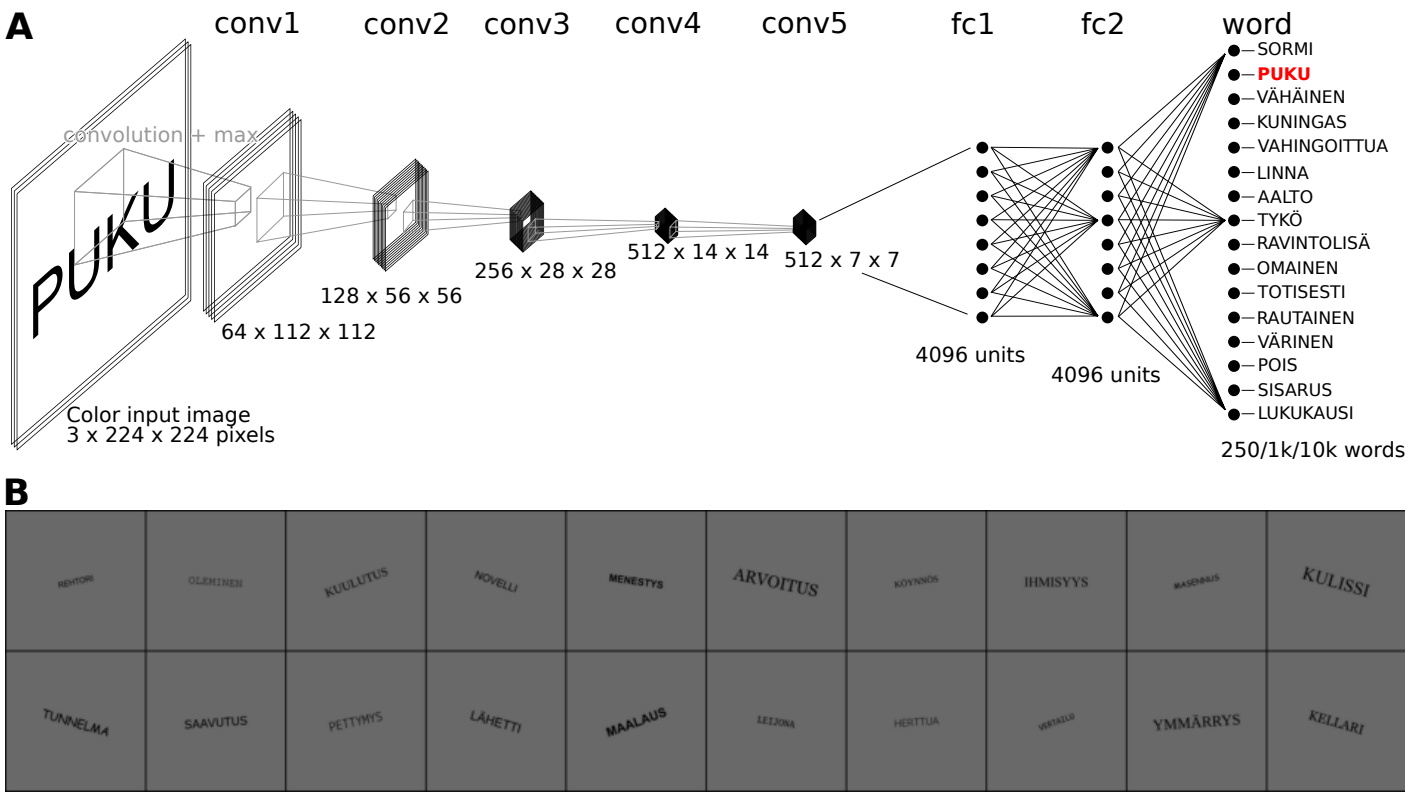

Figure 3: Overview of the proposed computational model of visual word recognition in the brain.
A: The VGG-11 model architecture, consisting of five convolution layers, two fully connected layers and one output layer.
B: Examples of the images used to train the model.

words relative to all other stimulus types. The type-II response, located further along the fusiform gyrus, exhibits sensitivity to whether the stimulus contains letters that are part of the participant's native alphabet and, in contrast to the type-I, has a reduced response to the noise-embedded words. The N400m response is located in the temporal cortex and is modulated by the lexical content of the stimulus, manifested in this study as a larger response to the word-like (i.e., words and pseudowords) versus the non-word-like stimuli.

2.2 CNNs produce responses that are qualitatively similar to the MEG evoked components

As a model of the computations underlying the brain activity observed during the MEG experiment, we used a VGG-11 [47] network architecture, pretrained on ImageNet.[48] This architecture consists of five convolution layers (three of which perform convolution twice), followed by two densely connected layers, terminating in an output layer (Figure 3A). Variations of the model were trained to perform visual word recognition using training sets consisting of images of written Finnish words, rendered in varying fonts, sizes and rotations (Figure 3B). In each case, training lasted for 20 epochs, which was enough for each model to obtain a near perfect accuracy (> 99%) on an independent validation set, as well as being able to correctly identify the original 118 word stimuli used in the MEG experiment. During training, the models were never exposed to any of the other stimulus types used in the MEG experiment (noisy words, symbols, consonant strings, and pseudowords) nor to the exact word stimuli used in that experiment. After training was complete, response profiles for all layers were created by computing the $\ell 2$ norm of the rectified linear unit (RELU) activations of the units in each layer in response to each stimulus, followed by a z-transformation (Figure 4). Finally, the response profiles of the CNN models were compared

[47] Szegedy et al., 2015

[48] Russakovsky et al., 2015

to those of the three MEG evoked components, both qualitatively by comparing the effects of the experimental contrasts and quantitatively through correlation. This allowed for an iterative design approach, where the network architecture and training diet of the model were manipulated to address shortcomings in the response profiles, until a model was reached with the response profiles of the contained layers matching those of the type-I, type-II and N400m evoked MEG components.

The design process started with an evaluation of the standard VGG-11 model, trained only on ImageNet. The layer response profiles of this model showed a distinctive pattern in response to the different types of stimuli used in Vartiainen et al. (2011) (Figure 4, top row), where the stimuli containing a large amount of noise evoke a very high level of activation in the first convolution layer, which carries over into all other layers. Thus, this model simulated the type-I component well, but it failed to simulate the type-II and N400m components. This is not surprising given that this model is illiterate.

Next, we trained the model on a training set containing a million examples of 250 possible written words (Figure 4, second row). After the model had learned to recognize written words (regardless of size, font family and some rotation), the convolution layers still showed a response profile similar to the type-I. However, the later linear layers now had a less extreme response to the noisy stimuli and also showed a dissociation between symbols and letters, making their response profiles more similar to that of the type-II and N400m components of the evoked response. Nonetheless, in this version of the model, the response to noisy stimuli is still too large to be a good fit with the type-II and N400m response profiles.

In vision research, having unit activations be noisy has been shown to make a system more robust against image perturbations, both in the brain and in CNNs.[49] In our case, the addition of a small amount of Gaussian noise to all unit activations in the model reduced markedly the response of the later layers to the noisy stimuli, resulting in less activity to the noisy stimuli than the other stimulus types (Figure 4, third row), while the recognition accuracy of the model was not degraded. We found that in order for the model to display this behavior, both noisy unit activations and batch normalization were necessary (Supplementary Figure 2). For this model, the response profiles of the first four convolution layers match that of the type-I component, and that of the first connected layer is similar to the type-II component in that both the noisy word stimuli and symbol stimuli evoked smaller responses than stimuli containing letters. However, this model does not contain any layers with response profiles matching that of the N400m.

[49] Carandini et al., 1997; Dapello et al., 2020

The N400m component of the evoked MEG response is modulated by the lexicality of the stimulus, as demonstrated through the contrast between consonant strings, proper words and pseudowords. Whereas consonant strings evoked a smaller N400m response than proper words, pseudowords produced a response that was equal to (or even larger than) that to proper words (Figure 2D). In this version of the model however, the response to pseudowords is less than that of proper words. This can be attributed to the relatively small vocabulary (250 words) of this model. The pseudowords used in the MEG experiment were selected to be orthographically similar to words in the extended vocabulary of a native Finnish speaker, but not necessarily similar to any of the 250 words in the vocabulary of the model. Increasing the vocabulary size to 1000 words was enough to solve the problem regarding pseudowords and yielded a model where the response profiles of the early convolution layers matched that of the type-I, those of the two fully connected layers matched that of the type-II, and that of the output layer matched that of the N400m (Figure 4, fourth row).

Further increasing the vocabulary size to 10 000 words initially had a detrimental effect

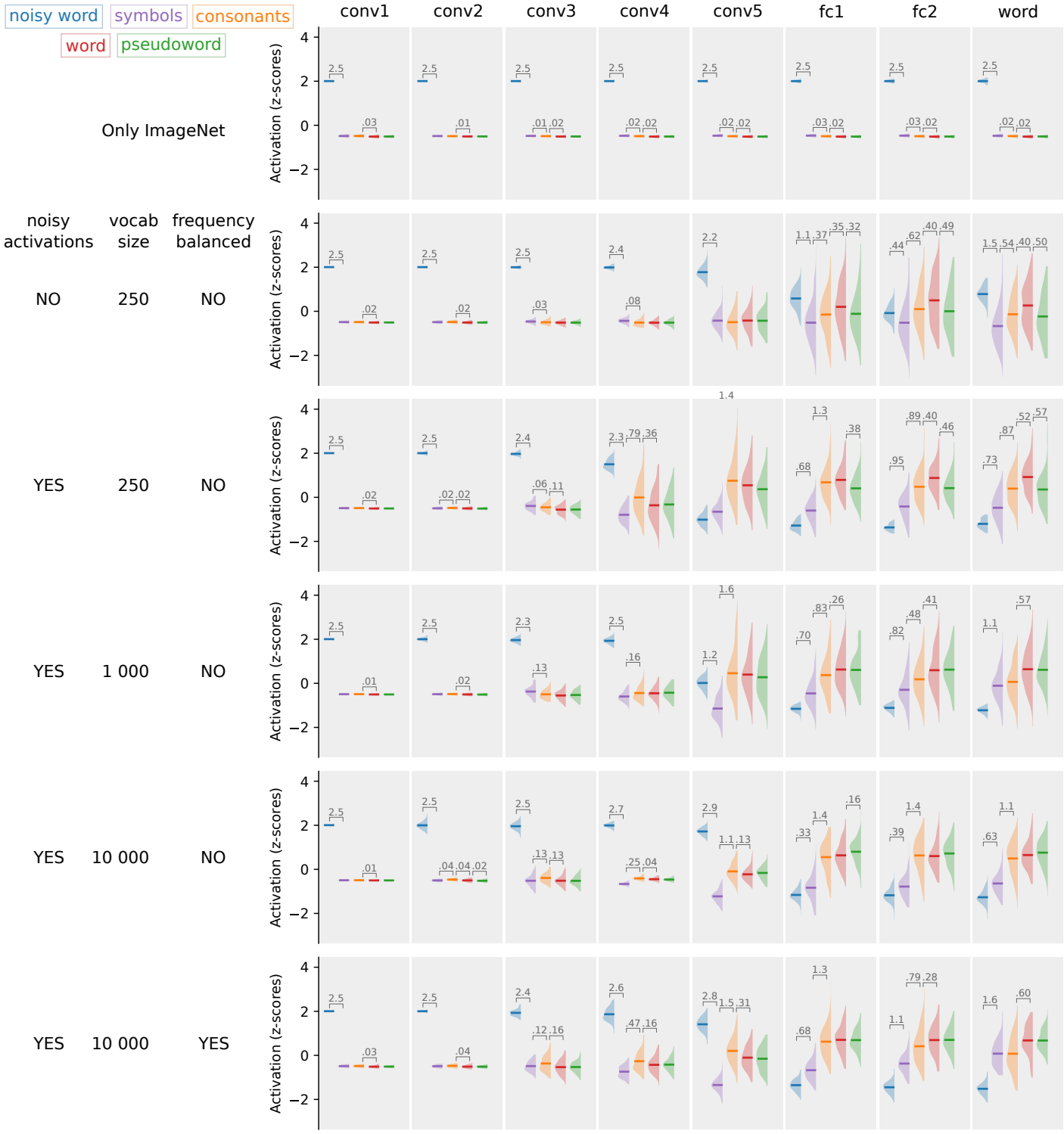

Figure 4: response profiles obtained from the models. For each layer, the response profile, i.e. the z-scored magnitude of ReLU activations in response to the same stimuli as used in the MEG experiment, is shown. Whenever there is a significant difference (t-test, $p < 0.05$, FDR corrected) between two adjacent distributions, the corresponding difference in means is shown.

on the model's ability to simulate the N400m component, as the response to consonant strings was nearly as strong as that to words and pseudowords (Figure 4, fifth row). In order to reproduce the response profile of the N400m during all experimental conditions of the MEG experiment, a model needs to hit a sweet spot regarding specificity of activation to in-vocabulary versus out-of-vocabulary letter strings. We found that this specificity can be controlled by manipulating the frequency in which words are presented to the model during training. By scaling the number of examples of a word in the training set in accordance with the frequency of occurrence of the word in a large text corpus,[50] a model was obtained with a large vocabulary of 10 000 and the response profiles of the convolution layers matching that of the type-I, response profiles of the fully connected layers matching that of the type-II, and the response profiles of the output layer matching that of the N400m (Figure 4, bottom row). These response profiles were stable across different random initializations when training the model (Supplementary Figure 3).

[50] Kanerva et al., 2014

Various variations in model architecture and training procedure were evaluated (Supplementary Figures 4–6). We found that reducing the number of layers by either removing convolution layers or fully connected layers, resulted in a model that did not exhibit the desired change in response to noisy stimuli from an enlarged response in the early layers to a diminished response in the later layers (Supplementary Figure 4). Increasing the number of layers resulted in a model that is very similar to the final model. As the vocabulary of the final model (10 000) exceeds the number of units in its fully-connected layers, a bottleneck is created in which a sub-lexical representation is formed. The number of units in the fully-connected layers, i.e. the width of the bottleneck, did not greatly affect the response profiles of the model (Supplementary Figure 5).

A model with either randomly initialized weights, or trained only on ImageNet, have response profiles in which each layer resembled that of the type-I response (Supplementary Figure 6, top rows). Training on word stimuli was required to simulate the type-II and N400m components. While starting from a pre-trained model on ImageNet reduced the number of epochs necessary for the model to reach >99 % accuracy, pre-training was not found to be strictly necessary in order to arrive at a good model that can simulate all three components (Supplementary Figure 6, bottom rows).

2.3 CNNs produce responses that correlate well with the MEG evoked components

Since both model layer activation and MEG evoked response amplitudes are described by a single number per stimulus in the response profiles, comparison between response profiles can be performed through straightforward correlation analysis (Figure 5, Figure 6). The models were designed to simulate the response amplitudes of an idealized average human participant, consistently producing the same response to a given stimulus, without attempting to model inter-participant variability and variability across repetitions. These sources of variability are present in the MEG data and place a boundary on the maximal obtainable correlation score, which we will refer to as the noise ceiling, following Kriegeskorte et al. (2008). Evoked MEG response amplitudes can vary substantially between repetitions of the same stimulus (either when presenting it multiple times to the same participant or presenting it to multiple participants) and are hence typically studied by averaging across repetitions.[51] The need for averaging becomes apparent when looking at the estimated noise ceilings for single-participant data (Figure 5). The pessimistic estimate for the single-participant noise ceiling is very low for all three MEG components (computed as the average correlation between the response amplitudes of a single participant versus those averaged across all

[51] Luck, 2014

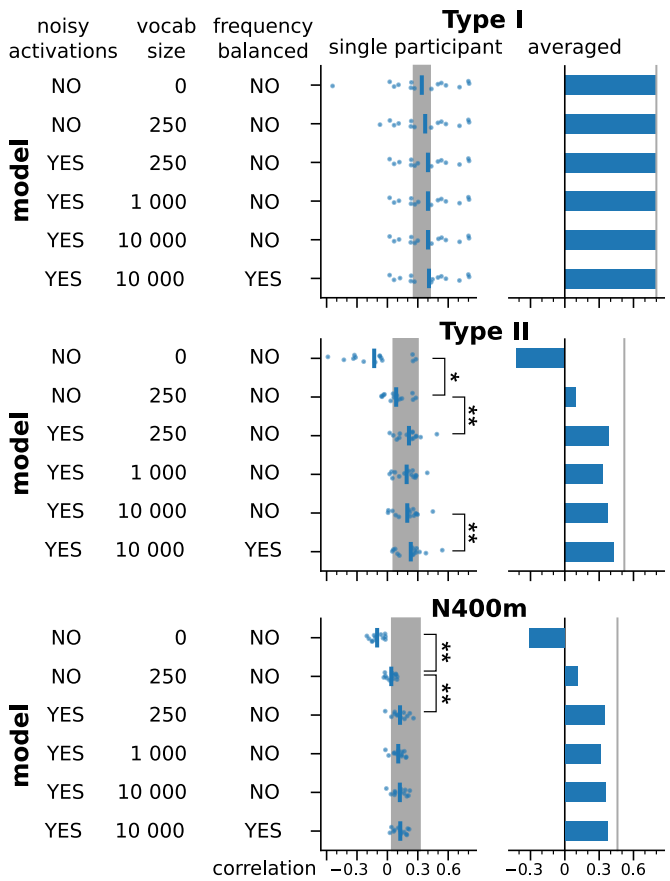

Figure 5: Correlations between all models and MEG responses. For each model, the maximum correlation between the response profiles obtained from all layers of the model and the response profiles of the type-I, type-II and N400m evoked MEG components. This was evaluated on both the grand-average level and for each individual subject. Estimated noise ceilings are indicated with vertical lines. The grand average analysis only has a single estimate, for the single subject analysis the shaded area indicates the range between pessimistic estimate and the optimistic estimate, in between which the true noise ceiling should lie. Significant differences between neighboring distributions are indicated with stars (paired t-test, $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$, FDR corrected).

remaining participants[52]), indicating a large amount of inter-participant variability. When averaging across participants, the estimated noise ceilings are much higher (estimated through a procedure proposed by Lage-Castellanos et al. (2019)), which is why we will focus on the comparison between the models and the average MEG response amplitudes.

[52] Kriegeskorte et al., 2008

The amount of activity in the first convolution layer of each model we evaluated correlates with the amplitude of the type-I component up to the estimated noise ceiling, which is in line with our qualitative finding that all models produce similar response profiles in the first convolution layers. For the type-II and N400m components, the illiterate model (trained only on ImageNet, Figure 5 top row) shows a strong negative correlation, which can be attributed to its invalid response to the noisy word condition. Training the model on orthographic stimuli greatly improves its correlation with these components (Figure 5, second row), but also here we see that the model without noisy unit activations, which failed to produce the proper response to the noisy word condition, has a weak correlation to the type-II and N400m components. Among the models with noisy unit activations (Figure 5, bottom four rows), which produced the proper response to the noisy word condition, the differences in correlation with the type-II and N400m components are much smaller. Notably, the models that produced qualitatively incorrect N400m response profiles (Figure 5, rows three and five) have slightly higher correlation with the N400m component than a model that produced a qualitatively correct responses (Figure 5, fourth row). This shows how a poor simulation of a brain response during one experimental condition can be compensated for by a better simulation of the brain response during another. However, the final model that produces correct responses to all experimental conditions and has the largest vocabulary (Figure 5, bottom row) also has the strongest correlation with the components of the evoked response. Still, there remains a gap between the model-brain correlations and the estimated noise ceiling, indicating there might remain some reproducible variance in the type-II and N400m

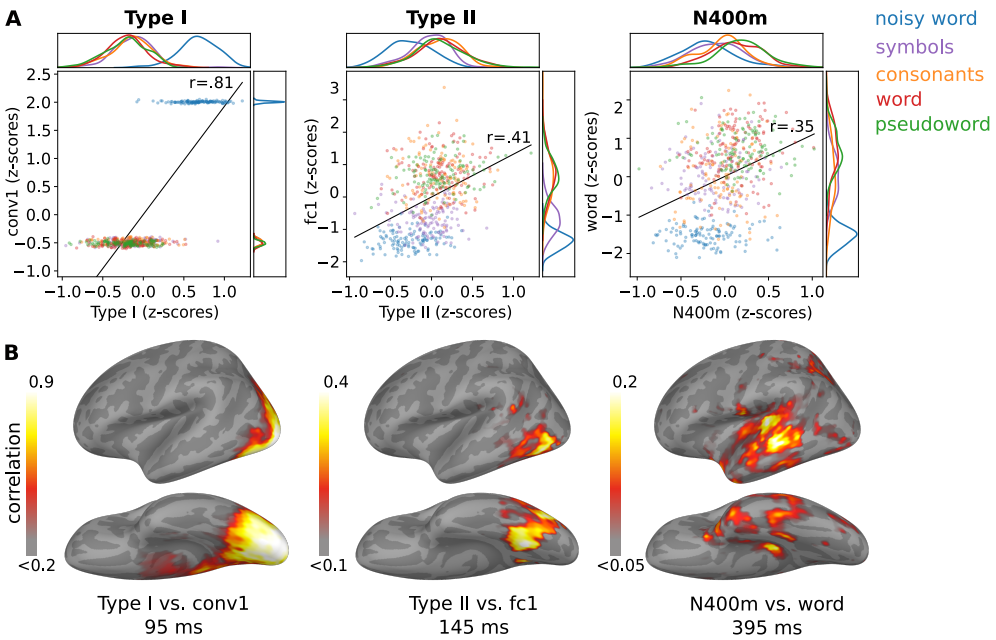

Figure 6: A closer look at the relationship between the final model and MEG responses.
A: Relationship between the response profiles obtained from evoked MEG activity, quantified using three ECD groups, and the response profiles obtained from three layers of the model. Kernel density distributions are shown at the borders.
B: Correlation between the MNE-dSPM source estimate and the model. Grand-average source estimates were obtained in response to each stimulus. The correlation map was obtained by correlating the activity at each source point with that for three layers of the model. The correlation map is shown at the time of peak correlation (within the time windows indicated in Figure 2C). Only positive correlations are shown.

components that the model does not capture.

Upon closer examination of the relationship between the response profiles of the final model and the MEG evoked components (Figure 6A), we see that the correlation is mostly driven by differences in response strength between the experimental conditions. Within each stimulus type, MEG evoked component amplitude and model layer activation are not significantly uncorrelated ($p > 0.05$ for all components), which is not surprising as the stimuli were designed to minimize within-condition variation. To verify that the three ECD groups capture the most important correlations between the brain activity and the model, we compared the model to cortex-wide MNE-dSPM source estimates (Figure 6B). The areas with maximum correlation appear around the locations of the ECDs and at a time that is close to the peak activation of the ECDs.

To test whether a multivariate analysis would yield additional insights, we trained a linear ridge regression model to predict brain activity given the activity within the layers of a model, what Schrimpf et al. (2020) refer to as BrainScore. BrainScores where computed for both the illiterate model, trained only on ImageNet (cf. Figure 4, top row), and the final model (cf. Figure 4, bottom row). We found that for both models, every layer could be used to predict the brain activity along key locations along the ventral stream (Supplementary Figure 7). Furthermore, there was very little difference in BrainScore between the illiterate and the final model. Hence, this information-based metric was in this case not helpful for evaluating model-brain fit.

2.4 CNNs can reproduce experimental results beyond the original MEG study

The experimental contrasts employed in the MEG experiment were designed to distinguish the type-I, type-II and N400m components from one another. However, many more experimental contrasts have been used to study these components in the past. We explored the effect of some of these on our final model (Figure 7) in order to ascertain the extent to which the model can simulate the behavior of the three components under conditions extending beyond the original MEG data that guided its design.

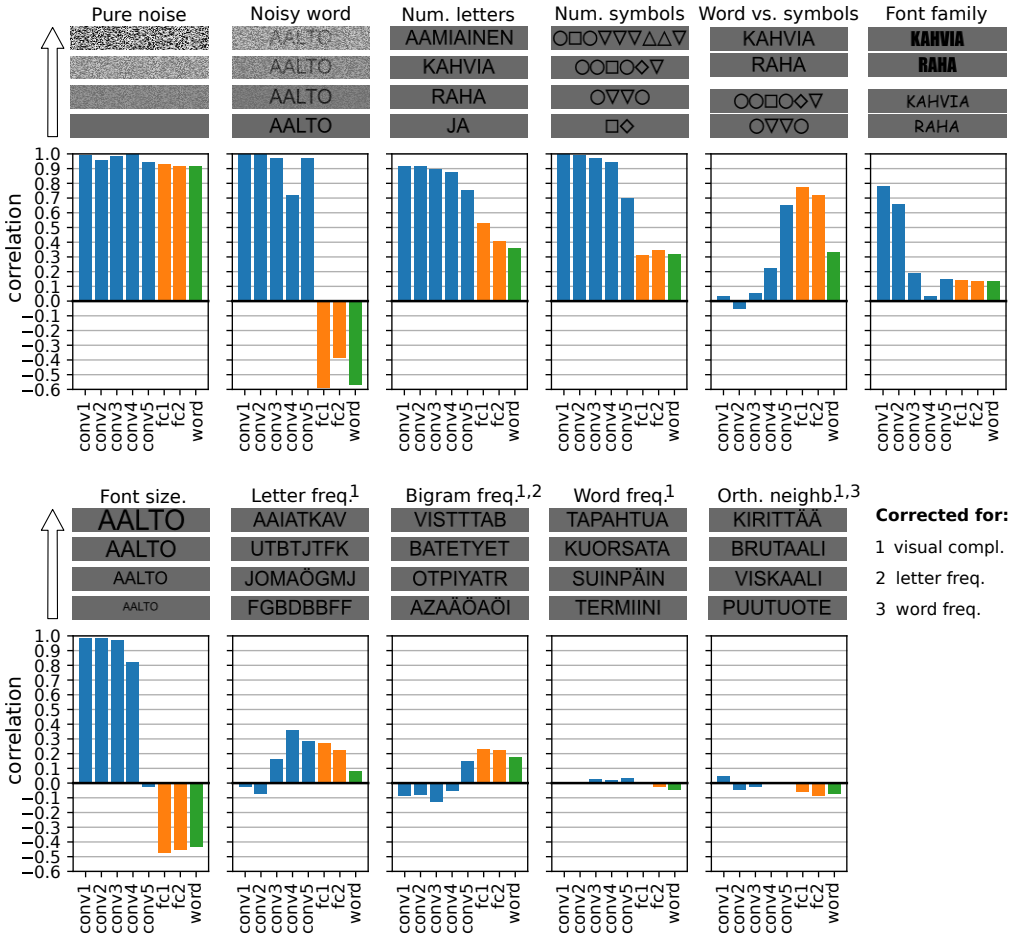

Figure 7: Post-hoc exploration of various experimental contrasts. For each contrast, four sample stimuli are shown to demonstrate the effect of the manipulated stimulus property and below are the correlation between the manipulation and the amount of activity in each layer of the final model. For the experimental contrasts indicated with a number, one or more confounding factors were corrected for (partial correlation). Different colors indicate convolution layers (blue), fully connected layers (orange) and the output layer (green).

In Tarkiainen et al. (1999), the difference between the type-I and type-II component was highlighted by contrasting stimuli that contained only noise with stimuli that contained both noise and letters. Our model successfully replicates the finding that the amplitude of both the type-I and type-II components increase with the amount of visual noise, except when there are letters present, in which case the amplitude of the type-II component decreases with the amount of noise while that of the type-I component is unaffected. This striking change in behavior occurs in the model when the layer type switches from convolution to fully connected linear. Furthermore, our model reproduces the finding by Tarkiainen et al. (1999) that the amplitudes of both the type-I and type-II component increases when a stimulus contains more letters or symbols, however given two strings of equal length, the amplitude of the type-II component is larger for the string containing letters versus a string containing symbols, while the amplitude of the type-I component is similar for both strings.

Basic visual features such as font family and font size have a large effect on the activity in the initial convolution layers of the model, and less so on the activity in the later layers. This roughly corresponds to findings in the masked repetition priming EEG literature, where they find that early components such as the N/P150 are affected by such properties, whereas later components are not.[53] However, our results may not be directly comparable to that literature, because our model cannot do repetition priming. Even so, it is surprising that the fully connected linear layers in the model show a decrease in activity with increasing font size, rather than being unaffected.

[53] Chauncey et al., 2008; Grainger and Holcomb, 2009; Hauk and Pulvermüller, 2004

Using invasive recordings, Woolnough et al. (2021) found effects of letter and bigram frequency in the occipitotemporal cortex, where we find the type-II component in MEG, but

not in occipital cortex, where we find the type-I component. This result has been shown in EEG studies as well.[54] Our model shows similar sensitivity to letter and bigram frequency in the later convolution and fully connected layers. However, we found that our model did not show any sensitivity to word frequency or orthographic neighbourhood size, even though these are known to affect later components in EEG studies.[55]

[54] Grainger, 2018; Laszlo and Federmeier, 2014

[55] Dambacher et al., 2006; Holcomb et al., 2002

3 Discussion

In this study, we demonstrated how training a CNN for visual word classification results in a model that can explain key functional effects of the evoked MEG response during visual word recognition, provided some conditions to enhance biological realism are met. Necessary attributes of the CNN and its training diet were found to be that unit activations should be noisy, the vocabulary should be large enough, and for very large vocabularies, the frequency of occurrence of words in the training set should follow their frequency in the language environment.

In contrast to traditional models of reading in the brain that operate on collections of pre-defined lines or letter banks,[56] a CNN stimulates the entire processing pipeline starting from raw pixel values of a bitmap image. This allowed us to evaluate model-brain correspondence using the same experimental contrasts as have been used in neuroimaging studies of visual word recognition. Following a neuroconnectionist approach,[57] we started with an illiterate CNN, trained to perform image classification, and iteratively made changes to its architecture and training diet, until it could replicate the effects of manipulating low-level visual properties such as noise levels, fonts and sizes, to mid-level orthographic properties such as symbols versus letters, and to high-level lexical properties such as consonants strings versus pseudowords versus real words. Some of these contrasts explore what we consider to be "edge cases" of visual word processing that explore what happens when the brain tries to recognize a word and fails. Symbol strings, consonant strings and pseudowords do not occur often in our natural environment and we don't explicitly train ourselves to distinguish them from real words when we learn to read. Following this logic, we included only real words in the training set for the model, so that responses to the nonword stimuli were simulations of a computational process failing to properly recognize a word.

[56] for example: Coltheart et al., 2001; Laszlo and Armstrong, 2014; McClelland and Rumelhart, 1981

[57] Doerig et al., 2023

The resulting models bridge an important gap in modeling visual word recognition, by establishing a direct link between computational processes and key findings in experimental neuroimaging studies.

3.1 A computational process exhibiting type-I, type-II and N400m-like effects

In this study, the computational process of visual word recognition was taken to be a transformation from a bitmap image containing a written word (with some variation in letter shapes, size and rotation) to the identity of the word. Many network architectures can be found in the deep-learning literature that could by used to implement such a process. For this study, we chose the VGG architecture (Figure 3), where high accuracy is achieved by performing the same few operations many times, rather than many different operations or clever tricks. The operations performed by the network are: convolution, pooling, batch normalization, and linear transforms. Our rationale was to start with these operations, which already have links to the neuroscience of vision[58] and only when they prove insufficient to explain a desired experimental effect, introduce others. This turned out to be unnecessary for explaining the results of Vartiainen et al. (2011), but one can imagine the necessity of

[58] Carandini et al., 1997; Lindsay, 2020; Serre et al., 2007; Yamins and DiCarlo, 2016

introducing recurrence[59] or attention mechanisms[60] in future extensions of the model into for example semantic processing.

[59] Kubilius et al., 2019

[60] Vaswani et al., 2017

The visual word recognition task itself did not prove particularly difficult for a CNN, as all variations of the network architecture and training diet that were tried resulted in a model that achieved near perfect accuracy. However, not all variations produced response profiles matching those of the evoked components (Figure 4, Supplementary Figure 4). This validates a criticism sometimes heard concerning whether deep learning models are brain-like, namely that the ability of a model to perform a task that once could only be done by a brain is not enough evidence that a model computes like a brain.[61] Indeed, we found that task performance alone, i.e. the identification of the correct stimulus word, was a poor indicator for judging a model's suitability as computational hypothesis for the brain. More compelling evidence was obtained by comparing the model's internal state to measurements of brain activity. Two variations of the model, with respective vocabulary sizes of 1k and 10k words, implement the computational process in such a way that they produce amounts of activity within their layers that resemble the amplitudes of the type-I, type-II and N400m components of the MEG evoked response. The model with the larger vocabulary was chosen to be the final iteration of the model and subjected to a more extensive range of experimental contrasts (Figure 7).

[61] Bowers et al., 2022

The computational process implemented by the final model starts with a series of convolution-and-pooling operations. Individual convolution units have a restricted receptive field, which is gradually widened through pooling operations. The units in the first layer see a patch of 3×3 pixels and those in the final fifth convolution layer combine information from patches of 18×18 pixels. This setup enforces a hierarchical representation forming in those layers, from edges to line segments and corners to (parts of) letter shapes.[62] Hence, increasing the visual complexity, i.e. adding more edges to an image by for example introducing visual noise or adding more symbols, will increase the activity of the convolution units. This behavior corresponds closely to that of the type-I component, which is in line with the growing body of evidence of convolution-and-pooling being performed in visual cortex.[63]

[62] Dehaene et al., 2005; Dehaene-Lambertz et al., 2018b

[63] Lindsay, 2020; Serre et al., 2007

The kind of information selected for in the later layers of the model changes through the introduction of non-relevant input signals during training, here done through noisy unit activations. Without noisy activations, the task of the computational units is limited to finding features to distinguish one letter from another. With noisy activations, their task also entails isolating letter-shape information from non-relevant information. The property of not merely having a preference for, but actively seeking to isolate certain types of information is key for producing the response profiles of the type-II and N400m components. When receptive field sizes were 8×8 pixels and larger (fourth and fifth layers), the convolution units showed a preference for letters over symbols and a preference for more frequently used letters. This orthographic tuning is also the hallmark of the type-II component. However, while the amplitude of this component increases with increasing visual complexity through the addition of more symbols or letters to a stimulus,[64] an important aspect of the type-II is that manipulations that reduce readability, such as the addition of visual noise, decrease its amplitude, even though the overall visual complexity of the stimulus is increased. It is in this aspect that the convolution units in our models are not necessarily a good fit for the type-II response, since their tight receptive field strongly ties their activation to that of their counterparts in the previous layer. Increased activation in the early convolution layers tends to carry over, with the later convolution layers having only limited capabilities for filtering out noise. With small vocabularies, we did observe noise-filtering behavior in the

[64] Tarkiainen et al., 1999

fifth convolution layer, but less so when vocabulary size was increased from 250 to 1k and the effect was gone with a vocabulary of 10k.

In the model, convolution units are followed by pooling units, which serve the purpose of stratifying the response across changes in position, size and rotation within the receptive field of the pooling unit. Hence, the effect of small differences in letter shape, such as the usage of different fonts, was only present in the early convolution layers, in line with findings in the EEG literature.[65] However, the ability of pooling units to stratify such differences depends on the size of their receptive field, which might explain why, in later layers, we observed a decreased response to stimuli where text was rendered with a font size exceeding the receptive field of the pooling units (Figure 7).

[65] Chauncey et al., 2008; Grainger and Holcomb, 2009; Hauk and Pulvermüller, 2004

In the final model, the convolution-and-pooling phase is followed by three fully connected layers, which pool information across the entire input image. Their large receptive fields make them more effective in isolating orthographic information from noise than the convolution layers (Figure 4) and their response profile matches that of the type-II component well, also when the model has a large vocabulary. Moreover, units in the fully connected layers can see multiple letters, making them sensitive not only to letter frequency, but also bigram frequency, allowing them to replicate these effects on the type-II component.[66]

[66] Woolnough et al., 2021

The final fully connected layer, the output layer, represents the model's notion of a mental lexicon: a collection of known word-forms, not yet linked to their semantic meaning (which we didn't model at all). Since orthographic word-form and semantic meaning are generally only very loosely coupled,[67] the lexicon often appears as a separate unit in computational models.[68] At face value, since we employed cross-entropy as loss function during training, the output layer simulates the lexicon following a localist structural approach, which harks back to the earliest models of visual word recognition.[69] In the localist approach, each word is represented by a dedicated unit that should only be active when the sensory version of that word (in this study, a bitmap image containing the word as rendered text) is provided as input to the model, to the exclusion of all other words in the vocabulary. This is known in the machine learning literature as "one-hot" encoding. The localist approach is often contrasted with the distributed approach, wherein each word is represented by a unique combination of multiple output units that respond to subword parts (e.g., n-grams, pairings of letters in the first and final position, etc.), which is a more efficient encoding scheme.[70] Both the localist and distributed approaches have their strengths and weaknesses when it comes to explaining experimental results, and the debate on how exactly the human lexicon is encoded in the brain is far from settled.[71] Interestingly, because our model covers the entire process from pixel to word-form, its simulation of the lexicon is not strictly localist, but combines both localist and distributed approaches.

[67] Monaghan et al., 2014

[68] Woollams, 2015

[69] McClelland and Rumelhart, 1981; Morton, 1969

[70] Seidenberg and McClelland, 1989

[71] Woollams, 2015

When the input image contains a word that is present in the model's vocabulary, it is true that the corresponding word-specific unit in the output layer will have the highest amount of activation of all units in the output layer. However, given the connectionist nature of our model, all units tuned to orthographically similar words will co-activate to some degree. This inaccuracy is important when it comes to simulating the amplitude of the N400m component in response to pseudowords, which is as large as (and sometimes even larger than) the response to real words. In the model, a pseudoword does not have a dedicated output unit, but given a large enough vocabulary, there will be enough units that respond to orthographically similar words to produce a large amount of activity nonetheless. Crucially, letter strings that are not word-like, such as consonant strings, will not activate many units in the output layer, which is an important difference between the response profiles of the

N400m and the type-II. Furthermore, in our final model, the first two fully connected layers have fewer units (4096) than the number of words in the vocabulary (10 000), which produces a distributed representation, where one unit is involved in the encoding of many words. The response profiles of both the second fully connected layer (distributed representation) and the output layer (localist representation) resemble that of the N400m response. Hence, our model does not provide direct cause for preferring a localist or distributed approach when modeling the lexicon, but shows how both approaches interact when creating a model of the entire process of sensory input to lexicon: the localist approach can function as training objective and a distributed representation emerges in the hidden layers.

3.2 Limitations of the current model

The VGG-11 architecture was originally designed to achieve high image classification accuracy on the ImageNet challenge.[72] Although we have introduced some modifications that make the model more biologically plausible, the final model is still incomplete in many ways as a complete model of brain function during reading.

[72] Simonyan and Zisserman, 2015

One important limitation of the current model is the lack of an explicit mapping from the units inside its layers to specific locations in the brain at specific times. Multiple layers have a maximum correlation with the source estimate at the exact same location and time (Figure 6B). Furthermore, there is no clear relationship between the number of layers the signal needs to traverse in the model to the processing time in the brain. The signal needs to propagate through five convolution layers to reach the point where it matches the type-II response at 140 ms to 200 ms, but only through one more additional layer to reach the point where it starts to match the N400m response at 300 ms to 500 ms. This also raises the question what the brain is doing during this time that seems to take so long. Still, cutting down on the number of times convolution is performed in the model seems to make it unable to achieve the desired suppression of noise (Supplementary Figure 4). A promising way forward may be to use a network architecture like CORNet,[73] that performs convolution multiple times in a recurrent fashion, yet simultaneously propagates activity forward after each pass. The introduction of recursion into the model will furthermore align it better with traditional-style models, since it can cause a model to exhibit attractor behavior,[74] which will be especially important when extending the model into the semantic domain.

[73] Kubilius et al., 2019

[74] McLeod et al., 2000

Despite its limitations, our model is an important milestone for computational models of reading that leverages deep learning techniques to extend the scope of traditional-style models to encompass the entire computational process starting from raw pixels values to representations of word-forms in the mental lexicon. This allowed us to directly compare activity within the model to MEG neuroimaging results and show how the stimulus-dependent amplitudes of the type-I, type-II and N400m components of the MEG evoked response can be computationally derived from a series of convolution-and-pooling steps, followed by several fully connected linear operations. Overall, we found that a qualitative evaluation of the response profiles (Figure 4) gave a most direct path to finding ways to improve the model. It was more informative than the correlation scores (Figure 5, Supplementary Figure 7), which were mostly driven by the large differences between the noise-embedded words and symbol strings versus the other stimulus types. This underlines the importance of finding ways to directly link the activity simulated within a model to measurements obtained from the brain and reason about them in both a quantitative and qualitative fashion.

4 Materials and Methods

4.1 MEG study design and data analysis

The MEG data that had been collected by Vartiainen et al. (2011) was re-analyzed for the current study using MNE-python[75] and FreeSurfer.[76] We refer to the original publication for additional details on the study protocol and data collection process.

[75] Gramfort et al., 2013

[76] Dale et al., 1999

Simultaneous MEG and EEG was recorded from 15 participants (who gave their informed consent, in agreement with the prior approval of the Ethics Committee of the Helsinki and Uusimaa Hospital District). The stimuli consisted of visually presented Finnish words, pseudowords (pronounceable but meaningless), consonant strings (random consonants), symbol strings (randomly formed from 10 possible shapes), and words embedded in high-frequency visual noise (Figure 2A). Each stimulus category contained 112 different stimuli and each stimulus contained 7–8 letters/symbols. The stimuli were presented sequentially in a block design. Each block started with a random period of rest (0 ms to 600 ms), followed by 7 stimuli of the same category. Each stimulus was shown for 300 ms with an inter-stimulus interval of 1500 ms of gray background. In addition to the five stimulus conditions (16 blocks each), there were 16 blocks of rest (12 s). An additional 5 target blocks were added, in which one stimulus appeared twice in a row, that the participants were asked to detect and respond to with a button press. The target blocks were not included in the analysis. The same paradigm was applied in an fMRI-EEG recording of the same participants as well (data not used here).

The data were recorded with a Vector View device (Elekta Neuromag, now MEGIN Oy, Finland) in a magnetically shielded room. In addition, vertical and horizontal electro-oculography (EOG) were recorded. The signals were bandpass filtered at 0.03 Hz to 200 Hz and digitized at 600 Hz. For analysis, the data were further processed using the maxfilter software (MEGIN Oy, Finland) and bandpass filtered at 0.1 Hz to 40 Hz. Eye-movement and heart-beat artifacts were reduced using independent component analysis (ICA): the signal was high-passed at 1 Hz, and components with correlations larger than 5 standard deviations with any of the EOG channels or an estimated electro-cardiogram (based on the magnetometers without the maxfilter applied) were removed (3–7 out of 86–108 components). Epochs were cut –200 ms to 1000 ms relative to the onset of the stimulus and baseline-corrected using the pre-stimulus interval. Participant-specific noise covariance matrices of the MEG sensors were estimated using the pre-stimulus interval.

Source estimation of the sensor-level signals was performed using two methods. The first, MNE-dSPM,[77] yields a comprehensive map of where activity can be found, which is used in this study mostly to provide context for a deeper investigation into three peaks of activity that can be observed along the ventral stream. The second, guided ECD modeling,[78] summarizes the high-dimensional data as a sparse set of ECDs, each one capturing an isolated spatiotemporal component, allowing us to study selected components in more detail. The advantage of this sparse method over distributed source estimation techniques such as the minimum norm estimate (MNE) or beamformers, is that in cases like Vartiainen et al. (2011), a single ECD may conveniently capture an isolated component of the evoked response. Furthermore, this sparsity makes it straightforward to take individual differences regarding the cortical origin of the evoked component into account, since the location and orientation of the ECDs were fitted to optimally explain the data of each individual. For each individual, the matching between ECDs and the type-I, type-II and N400m responses was performed based on their approximate location, timing of peak activity, and behavior across

[77] Dale et al., 2000

[78] Salmelin, 2010

the different stimulus types. The groups do not necessarily contain an ECD for each of the 15 participants, as some participants did not have an ECD to contribute that matched well enough in position and/or timing. The number of participants contributing an ECD to each group were: type-I: 14, type-II: 14, and N400m: 15.

For both ECD and MNE-dSPM source estimation, 3-layer boundary element method (BEM) forward models were used. These models were based on T1-weighted anatomical images that were acquired using a 3 T Signa EXCITE magnetic resonance imaging (MRI) scanner. FreeSurfer[79] was used to obtain the surface reconstructions required to compute the BEM models. Vartiainen et al. (2011) defines three groups of ECDs in the left-hemisphere, associated with the type-I, type-II and N400m responses. The locations and orientations of the ECDs in these groups were re-used in the current study. For each individual epoch, the signal at each of the three ECDs and a cortex-wide MNE-dSPM source estimate was computed using the noise covariance matrices and BEM forward models. Finally, the locations of the ECDs and MNE-dSPM source points were morphed to FreeSurfer's template brain.

[79] Dale et al., 1999

4.2 Computational models

The models of the computational process underlying the brain activity observed during the MEG experiment used a VGG-11 [80] network architecture, pretrained on ImageNet,[81] as defined in the TorchVision[82] module of the PyTorch[83] package. This architecture consists of five convolution layers (with the final three performing convolution twice), followed by two fully connected linear layers, terminating in a fully connected linear output layer (Figure 3A). Each individual convolution step is followed by batch-normalization[84] and max-pooling. Every unit, including the output units, used the following non-linear activation function:

[80] Szegedy et al., 2015

[81] Russakovsky et al., 2015

[82] Marcel and Rodriguez, 2010

[83] Paszke et al., 2019

[84] Ioffe and Szegedy, 2015

$$\text{noisyReLU}(x) = \max(x + N(0, \sigma_{\text{noise}})), \tag{1}$$

where N denotes Gaussian distributed random noise with zero mean and standard deviation σ_{noise}. For version of the model with noisy activations $\sigma_{\text{noise}} = 0.1$, and for versions without $\sigma_{\text{noise}} = 0$.

The model was trained to perform visual word recognition using a training set of approximately 1 000 000 images, where each image depicted a word, rendered in varying fonts, sizes and rotations (Figure 3B). An independent test set containing 100 000 images was used to track the performance of the model during training. None of the images in the training or test set were an exact duplicate of the images used as stimuli in the MEG experiment. Therefore, the stimulus set can be treated as an independent validation set. The task for the model was to identify the correct word (regardless of the font, size and rotation used to render the text) by setting the corresponding unit in the output layer to a high value.

Different versions of the model were created using vocabulary sizes of 250, 1 000 and 10 000 Finnish words. The vocabulary was compiled using the 112 words used in the MEG, extended with the most frequently used Finnish words, as determined by Kanerva et al. (2014), having a length of 3–9, excluding proper names and using their lemma form. No symbol strings, consonant strings, nor pseudowords were present in the training set.

For the training sets that do not take word frequency into account, each word was rendered 100 times in various fonts, sizes and rotations. When word frequency was taken into account, the number of times each word occurred in the training set c was scaled by its frequency f in a large Finnish text corpus[85] in the following manner:

[85] Kanerva et al., 2014

$$r = \sqrt{\frac{f}{f_0}}, \tag{2}$$

$$c = \left\lfloor N \frac{r}{\sum \mathbf{r}} \right\rfloor, \tag{3}$$

where f_0 is the frequency of the least commonly occurring word in the training vocabulary, r is the frequency ratio, $\sum \mathbf{r}$ is the total frequency ratio of all words in the training vocabulary, N is the desired number of items in the training set, in our case 1 000 000. The actual number of items in the training set will deviate a little from N, as the floor operation causes $\sum \mathbf{c} \approx N$. The number of times a word occurred in the test set was always 10.

Images were assembled by overlaying the rendered text onto an image of visual noise. Visual noise was rendered by randomly selecting a gray-scale value (0 % to 100 %) for each pixel. Next, if the training image contained text, the word was rendered in uppercase letters, using a randomly chosen font (15 possible fonts), font size (14 pt to 32 pt), and rotation (-20° to 20°). The list of fonts consisted of: Ubuntu Mono, Courier, Luxi Mono Regular, Lucida Console, Lekton, Dejavu Sans Mono, Times New Roman, Arial, Arial Black, Verdana, Comic Sans MS, Georgia, Liberation Serif, Impact, and Roboto Condensed.

For every variation of the model, the training procedure was the same. The training objective was to minimize the cross-entropy loss between the model output and a one-hot encoded vector of a length equal to the vocabulary size, indicating the target word. Optimization of the model's parameters was performed using stochastic gradient descend with an initial learning rate of 0.01, which was reduced to 0.001 after 10 epochs (an "epoch" in this context refers to one sequence of all 1 000 000 training images), and a momentum of 0.9. During training, the noisy-ReLU activation function was temporarily disabled for the output units, and re-activated during evaluation and normal operation of the model. In total, training was performed for 20 epochs, by which point the performance on the test set, for all variations of the model, had plateaued at around 99.6 %.

4.3 Comparison between model and MEG data

To compare the model with the MEG responses, statistical summaries were made of the high-dimensional data, that we refer to as "response profiles", that could be compared either qualitatively by examining the pattern of activation across stimulus types, and quantitatively by computing Pearson correlation between them.

To construct the response profiles for a model, it was applied to the bitmap images that were used during the MEG experiment (padded to be 256 × 256 pixels with 50 % grayscale pixels). The activity at each layer of the model was quantified by recording the ReLU activation right before the maxpool operation, and computing the ℓ_2 norm across the units. This yielded, for each stimulus image, a single number for each layer in the model representing the amount of activity within that layer. For each layer, the response profile was z-scored across the stimulus images.

To construct the response profiles for the type-I, type-II and N400m components, the neural activity represented by the corresponding ECD was quantified by taking the mean of its timecourse during an appropriate time window (indicated in Figure 2C, 64 ms to 115 ms for the type-I, 140 ms to 200 ms for the type-II, and 300 ms to 400 ms for the N400m, after stimulus onset). This yielded for each stimulus image a single number that indicates the amount of signal at the ECD. Collected across all stimuli, these numbers form the response pattern of each ECD. The response patterns were first z-scored across the stimulus images

independently for each participant and later averaged across participants for the grand-average analysis.

The MNE-dSPM source estimates used 4000 source points, distributed evenly across the cortex. The response profile at each source point was computed by taking the maximum signal value of the source point's timecourse within the same time window as the ECD analysis. This yielded for each stimulus image, a single number for each source point for each participant, indicating the amount of signal at the source point. The response patterns were first z-scored across the stimulus images independently for each participant and later averaged across participants during the grand-average analysis.

4.4 Statistics

The comparisons between response profiles obtained from the models and the MEG data were performed through Pearson correlation (Figure 5, Figure 6, Figure 7). In cases where distributions were tested for a significant difference in mean (Figure 4, Figure 5), this was done through two-tailed paired t-tests, as implemented in SciPy.[86] In Figure 2D, multiple participants contribute to each data point in the distributions, hence a linear mixed effects (LME) model was used to compare distributions, with participants modeled as random effect (both slope and intercept), as implemented in the Julia MixedModels package.[87]

[86] Virtanen et al., 2020

[87] Bates et al., 2023

The noise ceilings, indicating the maximum obtainable correlation between the response profile of a model and those of the evoked components (Figure 5), were estimated using two different methods. A noise ceiling depends on the amount of variation in repeated measurements of the same response profile. When estimating the noise ceiling for a response profile obtained from a single participant, repeated measurements were available since the same response profile could be obtained from multiple participants. Hence, the estimation procedure outlined in Kriegeskorte et al. (2008) could be used, which yields an optimistic and pessimistic estimate, which act as upper and lower bounds for the true noise ceiling. For each evoked component, the noise estimate was computed as the mean correlation between the response profile obtained from each individual participant and the grand-average response profile. For the optimistic estimate, the individual under consideration was included in the grand-average (hence contributing data twice) and for the pessimistic estimate, the individual was not. When estimating the noise ceiling for the grand-average response profile, no repeated measurements were directly available. In this case, Lage-Castellanos et al. (2019) propose estimating the noise ceiling n based on the between-participant variation:

$$n = \frac{\sqrt{\sigma_\mu^2 - \frac{\mu_{\sigma^2}}{N}}}{\sigma_\mu}, \tag{4}$$

Where σ_μ is the standard deviation (across the stimuli) of the grand-average response profile, μ_{σ^2} is the mean (across the stimuli) variation across the participant-specific response profiles, and N is the number of participants.

4.5 Post-hoc experimental contrasts

The final model was not only tested with the stimuli of Vartiainen et al. (2011), but also sets of stimuli designed with additional experimental contrasts (Figure 7). For each contrast, a different set of stimuli was generated. Unless mentioned otherwise, stimuli were text strings rendered on a gray background in the Arial font at a fontsize of 23 pixels.

Pure noise 1 000 images were generated containing a gray background, which was blended with Gaussian noise with varying levels between 0 % to 100 %, with no rendered text.

Noisy word All words of length 8 were selected from the original 10k vocabulary and rendered. They were blended with Gaussian noise at random levels between 0 % to 50 %.

Num. letters All words in the 10k vocabulary were rendered.

Num. symbols For each word in the 10k vocabulary, a symbol string of matching length was generated using the unicode characters "square", "circle", "triangle up", "triangle down", "diamond" and rendered.

Word vs. symbols The stimulus lists used for the "Num. letters" and "Num. symbols" contrasts were combined. For this contrast, the model response profile was correlated with a vector containing zeros indicating symbols and ones indicating words.

Font family All words in the 10k vocabulary were rendered in both the Impact and Comic Sans fonts. For this contrast, the model response profile was correlated with a vector containing zeros indicating stimuli rendered in the Impact font, and ones indicating stimuli rendered in the Comic Sans font.

Font size All words of length 8 were selected from the original 10k vocabulary. Each selected word was rendered at 8 fontsizes ranging from 14–32 pixels.

Letter freq. 10 000 random letter strings of length 8 were generated and rendered. In this contrast, the model response profile was correlated with the total letter frequency of each random letter string in the 10k vocabulary.

Bigram freq. 10 000 strings consisting of 4 bigrams were generated and rendered. Since there is generally a strong correlation between letter and bigram frequency, the random bigrams were selected in a fashion designed to minimize this confounding factor. Bigrams were randomly selected from a set of 50 common bigrams (freq >= 30 000) with the lowest letter frequency, and a set of 50 uncommon bigrams (freq < 10 000) with the highest letter frequency. Bigram and letter frequencies were computed on the 10k vocabulary.

Word freq. All words in the 10k vocabulary were rendered. The word frequency was computed as the square root of the number of times the word occured in a large corpus of Finnish text.[88]

[88] Kanerva et al., 2014

Orth. neighb. All words in the 10k vocabulary were rendered. The orthographic neighbourhood size was computed as the number of words in the 10k vocabulary with a Levenshtein distance of 1 to the target word.

4.6 Data and materials availability

The trained model and MEG derivatives necessary for performing the comparison is available at https://osf.io/nu2ep. As the consent obtained from the participants prohibits sharing of personal data, only MEG data that has been aggregated across participants is available. The code for training the model and reproducing the results of this paper is available at https://github.com/wmvanvliet/viswordrec-baseline.

5 Acknowledgements

We acknowledge the computational resources provided by the Aalto Science-IT project. This research was funded by the Academy of Finland (grants #310988 and #343385 to M.v.V, #315553 and #355407 to R.S.) and the Sigrid Jusélius Foundation (to R.S.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

References

Agrawal, A., Hari, K., & Arun, S. (2020). A compositional neural code in high-level visual cortex can explain jumbled word reading. *eLife*, *9*, e54846. https://doi.org/10.7554/eLife.54846

Barber, H. A., & Kutas, M. (2007). Interplay between computational models and cognitive electrophysiology in visual word recognition. *Brain Research Reviews*, *53*(1), 98–123. https://doi.org/10.1016/j.brainresrev.2006.07.002

Bates, D., Alday, P., Kleinschmidt, D., José Bayoán Santiago Calderón, P., Zhan, L., Noack, A., Bouchet-Valat, M., Arslan, A., Kelman, T., Baldassari, A., Ehinger, B., Karrasch, D., Saba, E., Quinn, J., Hatherly, M., Piibeleht, M., Mogensen, P. K., Babayan, S., Holy, T., & Gagnon, Y. L. (2023). *Juliastats/mixedmodels.jl: V4.14.1* (Version v4.14.1). Zenodo. https://doi.org/10.5281/zenodo.7979497

Bowers, J. S., Malhotra, G., Dujmović, M., Montero, M. L., Tsvetkov, C., Biscione, V., Puebla, G., Adolfi, F., Hummel, J. E., Heaton, R. F., Evans, B. D., Mitchell, J., & Blything, R. (2022). Deep Problems with Neural Network Models of Human Vision. *Behavioral and Brain Sciences*, 1–74. https://doi.org/10.1017/S0140525X22002813

Brem, S., Halder, P., Bucher, K., Summers, P., Martin, E., & Brandeis, D. (2009). Tuning of the visual word processing system: Distinct developmental ERP and fMRI effects. *Human Brain Mapping*, *30*(6), 1833–1844. https://doi.org/10.1002/hbm.20751

Carandini, M., Heeger, D. J., & Movshon, J. A. (1997). Linearity and normalization in simple cells of the macaque primary visual cortex. *The Journal of Neuroscience: The Official Journal of the Society for Neuroscience*, *17*(21), 8621–8644. https://doi.org/10.1523/JNEUROSCI.17-21-08621.1997

Carreiras, M., Armstrong, B. C., Perea, M., & Frost, R. (2014). The what, when, where, and how of visual word recognition. *Trends in Cognitive Sciences*, *18*(2), 90–98. https://doi.org/10.1016/j.tics.2013.11.005

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Nature Communications Biology*, *5*(1), 1–10. https://doi.org/10.1038/s42003-022-03036-1

Chauncey, K., Holcomb, P. J., & Grainger, J. (2008). Effects of stimulus font and size on masked repetition priming: An event-related potentials (ERP) investigation. *Language and Cognitive Processes*, *23*(1), 183–200. https://doi.org/10.1080/01690960701579839

Cohen, L., & Dehaene, S. (2004). Specialization within the ventral stream: The case for the visual word form area. *NeuroImage*, *22*(1), 466–476. https://doi.org/10.1016/j.neuroimage.2003.12.049

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), 204–56. https://doi.org/10.1037/0033-295X.108.1.204

Dai, Z., Liu, H., Le, Q. V., & Tan, M. (2021). CoAtNet: Marrying convolution and attention for all data sizes. *arXiv:2106.04803 [cs]*arxiv 2106.04803.

Dale, A. M., Fischl, B. R., & Sereno, M. I. (1999). Cortical surface-based analysis: I. segmentation and surface reconstruction. *NeuroImage*, *9*(2), 179–194. https://doi.org/10.1006/nimg.1998.0395

Dale, A. M., Liu, A. K., Fischl, B. R., Buckner, R. L., Belliveau, J. W., Lewine, J. D., & Halgren, E. (2000). Dynamic statistical parametric mapping: Combining fMRI and MEG for high-resolution imaging of cortical activity. *Neuron*, *26*(1), 55–67.

Dambacher, M., Kliegl, R., Hofmann, M., & Jacobs, A. M. (2006). Frequency and predictability effects on event-related potentials during reading. *Brain Research*, *1084*(1), 89–103. https://doi.org/10.1016/j.brainres.2006.02.010

Dapello, J., Marques, T., Schrimpf, M., Geiger, F., Cox, D. D., & DiCarlo, J. J. (2020). Simulating a Primary Visual Cortex at the Front of CNNs Improves Robustness to Image Perturbations. *bioRxiv*, 2020.06.16.154542. https://doi.org/10.1101/2020.06.16.154542

Dehaene, S., Cohen, L., Sigman, M., & Vinckier, F. (2005). The neural code for written words: A proposal. *Trends in Cognitive Sciences*, *9*(7). https://doi.org/10.1016/j.tics.2005.05.004

Dehaene-Lambertz, G., Monzalvo, K., & Dehaene, S. (2018a). The emergence of the visual word form: Longitudinal evolution of category-specific ventral visual areas during reading acquisition. *PLOS Biology*, *16*(3), e2004103. https://doi.org/10.1371/journal.pbio.2004103

Dehaene-Lambertz, G., Monzalvo, K., & Dehaene, S. (2018b). The emergence of the visual word form: Longitudinal evolution of category-specific ventral visual areas during reading acquisition. *PLOS Biology*, *16*(3), e2004103. https://doi.org/10.1371/journal.pbio.2004103

Devereux, B. J., Clarke, A., & Tyler, L. K. (2018). Integrated deep visual and semantic attractor neural networks predict fMRI pattern-information along the ventral object processing pathway. *Scientific Reports*, *8*(1). https://doi.org/10.1038/s41598-018-28865-1

Doerig, A., Sommers, R. P., Seeliger, K., Richards, B., Ismael, J., Lindsay, G. W., Kording, K. P., Konkle, T., van Gerven, M. A. J., Kriegeskorte, N., & Kietzmann, T. C. (2023). The neuroconnectionist research programme. *Nature Reviews Neuroscience*, 1–20. https://doi.org/10.1038/s41583-023-00705-w

Grainger, J. (2018). Orthographic processing: A 'mid-level' vision of reading: The 44th Sir Frederic Bartlett Lecture. *Quarterly Journal of Experimental Psychology*, *71*(2), 335–359. https://doi.org/10.1080/17470218.2017.1314515

Grainger, J., & Holcomb, P. J. (2009). Watching the word go by: On the time-course of component processes in visual word recognition. *Linguistics and Language Compass*, *3*(1), 128–156. https://doi.org/10.1111/j.1749-818X.2008.00121.x

Gramfort, A., Luessi, M., Larson, E., Engemann, D. A., Strohmeier, D., Brodbeck, C., Goj, R., Jas, M., Brooks, T., Parkkonen, L., & Hämäläinen, M. S. (2013). MEG and EEG data analysis with MNE-Python. *Frontiers in Neuroscience*, *7*(December), 1–13. https://doi.org/10.3389/fnins.2013.00267

Halgren, E., Dhond, R. P., Christensen, N., Van Petten, C., Marinkovic, K., Lewine, J. D., & Dale, A. M. (2002). N400-like magnetoencephalography responses modulated by semantic context, word frequency, and lexical class in sentences. *NeuroImage*, *17*(3), 1101–1116. https://doi.org/10.1006/nimg.2002.1268

Hannagan, T., Agrawal, A., Cohen, L., & Dehaene, S. (2021). Emergence of a compositional neural code for written words: Recycling of a convolutional neural network for reading. *Proceedings of the National Academy of Sciences*, *118*(46). https://doi.org/10.1073/pnas.2104779118

Hauk, O., & Pulvermüller, F. (2004). Effects of word length and frequency on the human event-related potential. *Clinical Neurophysiology*, *115*(5), 1090–1103. https://doi.org/10.1016/j.clinph.2003.12.020

Helenius, P., Salmelin, R., Service, E., & Connolly, J. F. (1998). Distinct time courses of word and context comprehension in the left temporal cortex. *Brain*, *121*(6), 1133–1142. https://doi.org/10.1093/brain/121.6.1133

Holcomb, P. J., Grainger, J., & O'Rourke, T. (2002). An electrophysiological study of the effects of orthographic neighborhood size on printed word perception. *Journal of Cognitive Neuroscience*, *14*(6), 938–950. https://doi.org/10.1162/089892902760191153

Huth, A. G., Heer, W. A. D., Griffiths, T. L., Theunissen, F. E., & Jack, L. (2016). Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, *532*(7600), 453–458. https://doi.org/10.1038/nature17637

Huth, A. G., Nishimoto, S., Vu, A. T., & Gallant, J. L. (2012). A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, *76*(6), 1210–1224. https://doi.org/10.1016/j.neuron.2012.10.014

Ioffe, S., & Szegedy, C. (2015). Batch normalization: Accelerating deep network training by reducing internal covariate shift, In *Proceedings of the 32nd international conference on international conference on machine learning - volume 37*, JMLR.org.

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

Kanerva, J., Luotolahti, J., Laippala, V., & Ginter, F. (2014). Syntactic N-gram collection from a large-scale corpus of internet Finnish. *Frontiers in Artificial Intelligence and Applications*, *268*, 184–191. https://doi.org/10.3233/978-1-61499-442-8-184

Katti, H., & Arun, S. P. (2022). A separable neural code in monkey IT enables perfect CAPTCHA decoding. *Journal of Neurophysiology*, *127*(4), 869–884. https://doi.org/10.1152/jn.00160.2021

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

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2017). ImageNet classification with deep convolutional neural networks. *Communications of the ACM*, *60*(6), 84–90. https://doi.org/10.1145/3065386

Kubilius, J., Schrimpf, M., Kar, K., Rajalingham, R., Hong, H., Majaj, N., Issa, E., Bashivan, P., Prescott-Roy, J., Schmidt, K., Nayebi, A., Bear, D., Yamins, D. L., & DiCarlo, J. J. (2019). Brain-like object recognition with high-performing shallow recurrent ANNs, In *Advances in neural information processing systems*, Curran Associates, Inc.

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (erp). *Annual Review of Psychology*, *62*, 621. https://doi.org/10.1146/annurev.psych.093008.131123

Lage-Castellanos, A., Valente, G., Formisano, E., & Martino, F. D. (2019). Methods for computing the maximum performance of computational models of fMRI responses. *PLOS Computational Biology*, *15*(3), e1006397. https://doi.org/10.1371/journal.pcbi.1006397

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., & Federmeier, K. D. (2014). Never seem to find the time: Evaluating the physiological time course of visual word recognition with regression analysis of single-item event-related potentials. *Language, Cognition and Neuroscience*, *29*(5), 642–661. https://doi.org/10.1080/01690965.2013.866259

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

Lindsay, G. W. (2020). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of Cognitive Neuroscience*, *33*(10), 1–15. https://doi.org/10.1162/jocn_a_01544

Luck, S. J. (2014). *An introduction to the event-related potential technique* (Second). Cambridge, MA, USA, A Bradford Book.

Marcel, S., & Rodriguez, Y. (2010). TorchVision the machine-vision package of torch, In *Proceedings of the 18th ACM international conference on multimedia*, Association for Computing Machinery. https://doi.org/10.1145/1873951.1874254

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nature Reviews Neuroscience*, *4*(4), 310–322. https://doi.org/10.1038/nrn1076

McClelland, J. L., & Rumelhart, D. E. (1981). An interactive activation model of context effects in letter perception: Part I. an account of basic findings. *Psychological Review*, *88*(5), 580–596. https://doi.org/10.1016/B978-1-4832-1446-7.50048-0

McLeod, P., Shallice, T., & Plaut, D. C. (2000). Attractor dynamics in word recognition: Converging evidence from errors by normal subjects, dyslexic patients and a connectionist model. *Cognition*, *74*(1), 91–114. https://doi.org/10.1016/S0010-0277(99)00067-0

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *369*, 20130299. https://doi.org/10.1098/rstb.2013.0299

Morton, J. (1969). Interaction of information in word recognition. *Psychological Review*, *76*(2), 165–178. https://doi.org/10.1037/h0027366

Norris, D. (2013). Models of visual word recognition. *Trends in Cognitive Sciences*, *17*(10), 517–524. https://doi.org/10.1016/j.tics.2013.08.003

Norris, D., & Kinoshita, S. (2012). Reading through a noisy channel: Why there's nothing special about the perception of orthography. *Psychological Review*, *119*, 517–545. https://doi.org/10.1037/a0028450

Parviainen, T., Helenius, P., Poskiparta, E., Niemi, P., & Salmelin, R. (2006). Cortical sequence of word perception in beginning readers. *Journal of Neuroscience*, *26*(22), 6052–6061. https://doi.org/10.1523/JNEUROSCI.0673-06.2006

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). PyTorch: An imperative style, high-performance deep learning library. In *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc. http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf

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

Poeppel, D. (2012). The maps problem and the mapping problem: two challenges for a cognitive neuroscience of speech and language. *Cognitive Neuropsycholy*, *29*(1-2), 34–55. https://doi.org/10.1080/02643294.2012.710600

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. (2018). The evolution of cognitive models: From neuropsychology to neuroimaging and back. *Cortex*, *107*, 37–49. https://doi.org/10.1016/j.cortex.2017.12.020

Protopapas, A., Orfanidou, E., Taylor, J. S. H., Karavasilis, E., Kapnoula, E. C., Panagiotaropoulou, G., Velonakis, G., Poulou, L. S., Smyrnis, N., & Kelekis, D. (2016). Evaluating cognitive models of visual word recognition using fMRI: Effects of lexical and sublexical variables. *NeuroImage*, *128*, 328–341. https://doi.org/10.1016/j.neuroimage.2016.01.013

Rajalingham, R., Kar, K., Sanghavi, S., Dehaene, S., & DiCarlo, J. J. (2020). The inferior temporal cortex is a potential cortical precursor of orthographic processing in untrained monkeys. *Nature Communications*, *11*(1), 3886. https://doi.org/10.1038/s41467-020-17714-3

Richards, B. A., Lillicrap, T. P., Beaudoin, P., Bengio, Y., Bogacz, R., Christensen, A., Clopath, C., Costa, R. P., de Berker, A., Ganguli, S., Gillon, C. J., Hafner, D., Kepecs, A., Kriegeskorte, N., Latham, P., Lindsay, G. W., Miller, K. D., Naud, R., Pack, C. C., … Kording, K. P. (2019). A deep learning framework for neuroscience. *Nature Neuroscience*, *22*(11), 1761–1770. https://doi.org/10.1038/s41593-019-0520-2

Rumelhart, D. E., & McClelland, J. L. (1982). An interactive activation model of context effects in letter perception: Part II. the contextual enhancement effect and some tests and extensions of the model. *Psychological Review*, *89*(1), 60–94. https://doi.org/10.1037/0033-295X.89.1.60

Russakovsky, O., Deng, J., Su, H., Krause, J., Satheesh, S., Ma, S., Huang, Z., Karpathy, A., Khosla, A., Bernstein, M., Berg, A. C., & Fei-Fei, L. (2015). ImageNet large scale visual recognition challenge. *International Journal of Computer Vision*, *115*(3), 211–252. https://doi.org/10.1007/s11263-015-0816-y

Salmelin, R. (2007). Clinical neurophysiology of language: The MEG approach. *Clinical Neurophysiology*, *118*(2), 237–254. https://doi.org/10.1016/j.clinph.2006.07.316

Salmelin, R. (2010). Multi-dipole modeling in MEG. In P. Hansen, M. Kringelbach, & R. Salmelin (Eds.), *MEG: An introduction to methods* (pp. 124–155). New York, Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195307238.001.0001

Schrimpf, M., Kubilius, J., Hong, H., Majaj, N. J., Rajalingham, R., Issa, E. B., Kar, K., Bashivan, P., Prescott-Roy, J., Schmidt, K., Yamins, D. L. K., & DiCarlo, J. J. (2020). Brain-Score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, 407007. https://doi.org/10.1101/407007

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523–568. https://doi.org/10.1037/0033-295X.96.4.523

Serre, T., Wolf, L., Bileschi, S., Riesenhuber, M., & Poggio, T. (2007). Robust object recognition with cortex-like mechanisms. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *29*(3), 411–426. https://doi.org/10.1109/TPAMI.2007.56

Service, E., Helenius, P., Maury, S., & Salmelin, R. (2007). Localization of syntactic and semantic brain responses using magnetoencephalography. *Journal of Cognitive Neuroscience*, *19*(7), 1193–1205. https://doi.org/10.1162/jocn.2007.19.7.1193

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. *arXiv:1409.1556 [cs]*arxiv 1409.1556.

Szegedy, C., Liu, W., Jia, Y., Sermanet, P., Reed, S., Anguelov, D., Erhan, D., Vanhoucke, V., & Rabinovich, A. (2015). Going deeper with convolutions. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, *07-12-June*, 1–9. https://doi.org/10.1109/CVPR.2015.7298594

Tarkiainen, A., Helenius, P., Hansen, P. C., Cornelissen, P. L., & Salmelin, R. (1999). Dynamics of letter string perception in the human occipitotemporal cortex. *Brain*, *122*(11), 2119–2132. https://doi.org/10.1093/brain/122.11.2119

Testolin, A., Stoianov, I., & Zorzi, M. (2017). Letter perception emerges from unsupervised deep learning and recycling of natural image features. *Nature Human Behaviour*, *1*(9), 657–664. https://doi.org/10.1038/s41562-017-0186-2

Vartiainen, J., Liljeström, M., Koskinen, M., Renvall, H., & Salmelin, R. (2011). Functional magnetic resonance imaging blood oxygenation level-dependent signal and magnetoencephalography evoked responses yield different neural functionality in reading. *The Journal of Neuroscience*, *31*(3), 1048–1058. https://doi.org/10.1523/jneurosci.3113-10.2011

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. arXiv. https://doi.org/10.48550/arXiv.1706.03762

Virtanen, P., Gommers, R., Oliphant, T. E., Haberland, M., Reddy, T., Cournapeau, D., Burovski, E., Peterson, P., Weckesser, W., Bright, J., van der Walt, S. J., Brett, M., Wilson, J., Millman, K. J., Mayorov, N., Nelson, A. R. J., Jones, E., Kern, R., Larson, E., … SciPy 1.0 Contributors. (2020). SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, *17*, 261–272. https://doi.org/10.1038/s41592-019-0686-2

Woollams, A. M. (2015). Lexical is as lexical does: Computational approaches to lexical representation. *Language, Cognition and Neuroscience*, *30*(4), 395–408. https://doi.org/10.1080/23273798.2015.1005637

Woolnough, O., Donos, C., Rollo, P. S., Forseth, K. J., Lakretz, Y., Crone, N. E., Fischer-Baum, S., Dehaene, S., & Tandon, N. (2021). Spatiotemporal dynamics of orthographic and lexical processing in the ventral visual pathway. *Nature Human Behaviour*, *5*(3), 389–398. https://doi.org/10.1038/s41562-020-00982-w

Yamins, D. L. K., & DiCarlo, J. J. (2016). Using goal-driven deep learning models to understand sensory cortex. *Nature Neuroscience*, *19*(3), 356–365. https://doi.org/10.1038/nn.4244