December 9, 2020
Draft version intended for future publication.

A large scale computational model of word recognition and its comparison with MEG data

Marijn van Vliet[1*], Oona Rinkinen[1], Takao Shimizu[1], Barry Devereux[2], and Riitta Salmelin[1]

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

Abstract

1 Introduction

What computational steps is the brain performing when it recognizes some lines on a piece of paper as a specific 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] However, it is rarely straightforward to translate observations of brain activity into a mechanistic understanding of the computational process being performed by the brain.[5]

[1] Grainger and Holcomb, 2009
[2] Salmelin, 2007
[3] Price, 2012
[4] Carreiras et al., 2014
[5] Poeppel, 2012

Computational models facilitate the development of cognitive theories by allowing us to reason about conceptual "box and arrow" ideas in a qualitative and quantitative manner.[6] However, the predictions made by existing models of reading are not directly comparable to actual neuroimaging data, and it is an often repeated sentiment that there should be more contact between the two.[7]

[6] Barber and Kutas, 2007; Price, 2018
[7] Carreiras et al., 2014; Laszlo and Armstrong, 2014; Laszlo and Plaut, 2012; Poeppel, 2012; Taylor et al., 2013

In the domain of models of reading in the brain, connectionist models using parallel distributed processing (PDP)[8] and "dual route" approaches[9] have been shown to account for many observational findings in both healthy volunteers and patients.[10] Furthermore, Laszlo and Plaut (2012) have shown that by summing the activity of the computational units in specific layers of a connectionist model, the resulting time varying signal resembles a well known component, observed in EEG and MEG studies, known as the N400 potential.[11] This result has later been extended to model more of such components.[12] However, the signal produced by the model of Laszlo and Plaut (2012) cannot be directly compared with neuroimaging data, because the simulated environment was extremely simplified to reduce the complexity of the model, whereas the brain data will by nature reflect the reading process in a realistic setting. For example, the model operates on 5-letter words with an alphabet of only 3 possible letters. Nevertheless, they demonstrated how a computational model can both perform a simplified reading task and produce neuroimaging-like data.

[8] McClelland and Rogers, 2003
[9] Perry et al., 2007
[10] McClelland and Rogers, 2003; McLeod et al., 2000; Perry et al., 2007
[11] Kutas and Federmeier, 2011
[12] Laszlo and Armstrong, 2014

Recent advances in deep learning and its software ecosystem are rapidly changing our notion of what is computationally tractable to model.[13] Convolutional neural networks (CNNs) have emerged that perform visual object recognition at a large enough scale to enable a direct comparison between network state and neuroimaging data[14] and consequently our understanding of basic visual processing has increased tremendously.[15] Since the first stages of reading, namely visual word recognition, can be seen as a specialized form of object recognition, CNNs may very well be a suitable tool for increasing the scale of traditional connectionist models of

[13] Richards et al., 2019
[14] Devereux et al., 2018; Schrimpf et al., 2018; Yamins and DiCarlo, 2016
[15] Lindsay, 2020

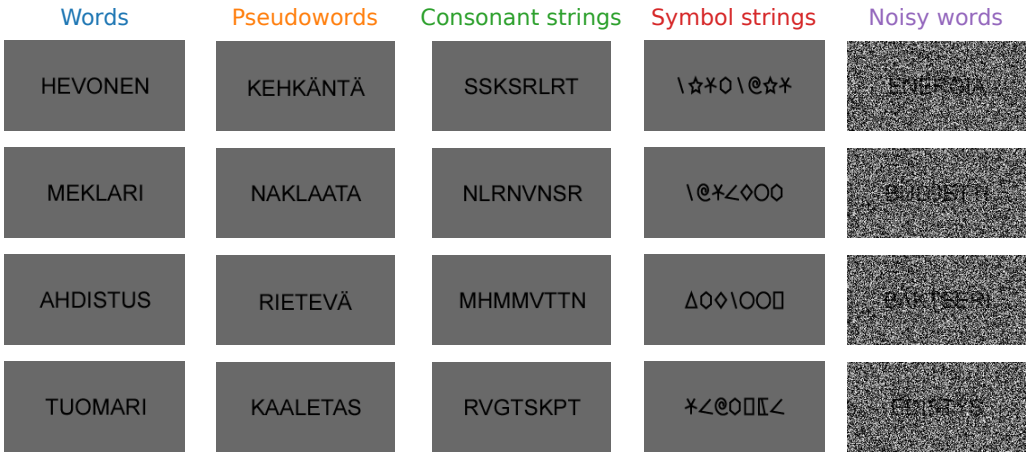

Figure 1: Examples of stimuli used in the MEG experiment. Each stimulus contained 7–8 letters or symbols.

reading.

In this study, we trained a CNN to perform visual word recognition on bitmap images of rendered text. The same set of stimulus images was presented to both the model and human volunteers in order to directly compare the activation inside the model to the amplitude of MEG evoked responses recorded from the study participants. Whereas the training set of the model consisted only of images of either valid Finnish words or only visual noise, the stimulus set used in the experiment contained images of valid Finnish words, which were similar to the ones present in the training data for the model, but also consonant strings, symbol strings and pseudo-words. We show how various layers in the model behave similarly to several well studied evoked responses and evaluate this similarity both qualitatively and, for the first time, quantitatively.

2 Results

2.1 The brain

For this study, we reused the MEG data that was collected as part of an earlier study by Vartiainen et al. (2011). During the recording session, 15 participants (who gave their informed consent, in agreement with the prior approval of the Helsinki and Uusimaa Ethics Committee) were presented with 560 orthographic stimuli (silent reading task), designed to form a series of experimental contrasts that highlight three processing stages during single word reading. Stimuli included valid Finnish words, pseudowords, consonant strings, letter-like symbol strings and Finnish words embedded in visual noise (Figure 1).

To summarize the high-dimensional MEG data, the sensor level signals were segregated into cortical-level spatiotemporal components by means of guided equivalent current dipole (ECD) modeling.[16] This yielded for each participant, a collection of 11–15 ECDs that together account for at least 75 % of the signal. The ECDs where then grouped based on their location and the timing of peak activation. The current study re-uses three of such dipole groups defined in the original study[17] along the ventral stream (Figure 3A), that capture the different processing stages which the experiment sought to highlight. For each dipole, the response to each stimulus was summarized by integrating the activity over the time window used in Vartiainen et al. (2011) for statistical analysis. To obtain the group-level response to each stimulus, the activity integrals for the dipoles in the group were z-transformed across the stimuli, and averaged.

The first group of ECDs is occipitally located and is characterized by early onset activity in the visual cortex, peaking 65 ms to 115 ms after stimulus onset. The activity at these dipoles is driven by the visual complexity of the stimulus and is characterized in this study by a large response to noise embedded words relative to all other stimulus types. The second group is

[16] Hämäläinen et al., 1993

[17] Vartiainen et al., 2011

located further along the fusiform gyrus, sometimes referred to as the visual word form area (VWFA)[18] and their activity peaks slightly later at 140 ms to 200 ms. Dipoles belonging this group exhibit sensitivity to whether the stimulus contains letters that are part of the participant's native alphabet,[19] and is characterized in this study by a smaller response to stimuli containing symbol strings and noise embedded words, relative to stimuli that contain letters. The third and final group is located temporally, peaking much later at 300 ms to 500 ms, corresponding to the N400 component of the evoked response.[20] In this group, activity is modulated by the lexical content of the stimulus, and is characterized in this study by a larger response to the word-like (i.e., words and pseudowords) versus the non-word-like stimuli.

[18] Cohen and Dehaene, 2004

[19] Tarkiainen et al., 1999

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

2.2 The model

As a model of the computations underlying the brain activity observed during the MEG experiment, we used a VGG-11[21] network architecture, pretrained on ImageNet,[22] as provided by the TorchVision package.[23] This architecture consists of five convolution layers (three of which perform convolution twice), followed by two densely connected layers, terminating in an output layer. The model was trained to perform visual word recognition using a training set that contained 1 000 000 images, where each image depicted one of 10 000 possible Finnish words, rendered in varying fonts, sizes and rotations, with varying degrees of visual background noise (Figure 2). The task for the model was to identify the correct word (by setting the corresponding unit in the output layer to a high value), irregardless of the font, size and rotation used to render the text. The vocabulary size (10 000) was chosen to exceed the number of units in the densely connected layers (4 096), forcing the model to construct a sub-lexical representation.

[21] Szegedy et al., 2015

[22] Russakovsky et al., 2015

[23] Marcel and Rodriguez, 2010

In addition to the images of valid Finnish words, 50 000 images consisting of only visual noise were added to the training set. The inclusion of this "no word present" condition introduces a detection element to what is otherwise a discrimination task. If no word was present in the image, all output units should have a low value. During training, the performance of the model was evaluated on an independent test set of 100 000 images that contained words and 5 000 that contained only noise. Training was stopped when the model's performance plateaued, at which point the accuracy on the test set was 99.3 %.

After training, the model was used to classify the stimulus images used during the MEG experiment (see Supplementary Table 1). The model classified 113 of the 118 word stimuli correctly (accuracy 95.8%). The 5 incorrectly classified stimuli were instead classified as close orthographic neighbours (e.g. LUOMINEN→TUOMINEN). All noise embedded words were misclassified, since the amount of noise made them unrecognizable, with 99/118 being classified as the word METSÄTEOLLISUUS, which is one of the words in the training set with the highest visual complexity. Of course, all the non-word stimuli were misclassified, as the model was not trained on any of these types of stimuli (i.e. pseudowords, consonant strings, symbol strings). In this case, it was more difficult to find a pattern in the way they were classified. In many cases, similarity in letter shape seemed to play a role (e.g. SKKNTMT→SUKUNIMI, ÄHKÄÄJÄ→VARAAJA), not not always (e.g. TTNRNHR→UUTINENKIRJE, INKRIHTI→EMOLEVY).

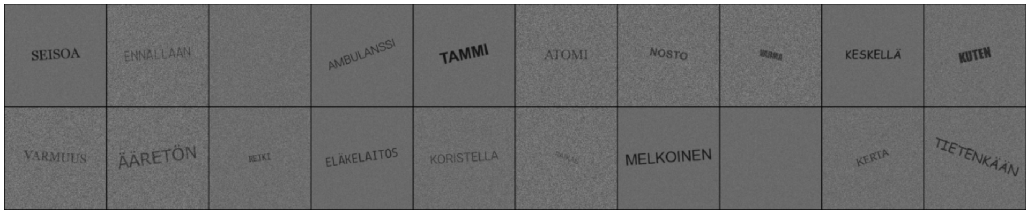

Figure 2: Examples of images in the training set for the model.

2.3 Comparing model and brain

To directly compare layer activations in the model to dipole responses in the brain (Figure 3, middle), the sum activation in each layer of the model was recorded for each stimulus as it passed through the model (Figure 3B). Since the same stimuli were processed by both the model and the brain, we can measure the model-brain correspondence qualitatively by examining the responses to each stimulus type, and quantitatively by computing the correlation between the summarized activity in the model layers and dipole groups (Table 1).

In the first three convolution layers of the model, we see a much larger response to the "noisy word" stimuli, relative to the other stimulus types, likely driven by the much higher visual complexity of these stimuli. Convolution filters that are detecting edges and corners output high values for visual noise and low values for stretches of flat gray background. This corresponds well to the response of the first dipole group (Pearson correlations: 0.79, 0.79, 0.68), localized in the visual cortex of the brain, which shows a similar sensitivity to noise, and does not distinguish between the other stimulus types.

We see a change in the response pattern of the model in the next two convolution layers: layers 4 and 5. In these layers, the activity in response to symbol strings is now lower than that of the other types. This indicates the presence of convolution filters that are sensitive to the specific line arrangements that are found in letters of the Finnish alphabet, which are not present in the symbol strings. In this regard, these layers correspond to the responses of dipole group two in the brain. However, many of the filters are still sensitive to noise, which does not correspond to the responses of the second dipole group, resulting in a negative correlation (Pearson correlations: -0.03, -0.32). Overall, layers 4 and 5 correspond best to dipole group one (Pearson correlations: 0.15, 0.66).

The selectivity for letter shapes becomes more pronounced in the two fully connected layers of the model. In these layers, we also see less activity in response to the noise stimuli. However, no distinction is made between consonant strings and word-like stimuli, indicating that letters are detected in isolation. This means the response patterns of these layers correspond to that of the second dipole group (Pearson correlations 0.39, 0.47).

Finally, at the output layer of the model, we observe that consonant strings produce less activity, indicating that the shapes of multiple letters are combined to produce the output. It is noteworthy that pseudowords, which were not present in the training set, produce a roughly equal amount of activation in the output layer as valid Finnish words. This means that the response pattern of the output layer corresponds to that of the third dipole group (Pearson correlation 0.44).

3 Discussion

One may ask why the lexicon layer of the model is a one-hot encoded output vector. This is plainly incompatible with how the brain works. Some abstract semantic representation, such

Layer	Type	Dipole group 1	Dipole group 2	Dipole group 3
1	Convolution	0.79	-0.48	-0.35
2	Convolution	0.79	-0.48	-0.35
3	Convolution	0.68	-0.42	-0.33
4	Convolution	0.15	-0.03	-0.01
5	Convolution	0.66	-0.32	-0.19
6	Fully conn.	-0.32	0.39	0.39
7	Fully conn.	-0.53	0.47	0.43
8	Output	-0.61	0.45	0.44

Table 1: Pearson correlations between each layer of the model and the grand-average of each dipole group.

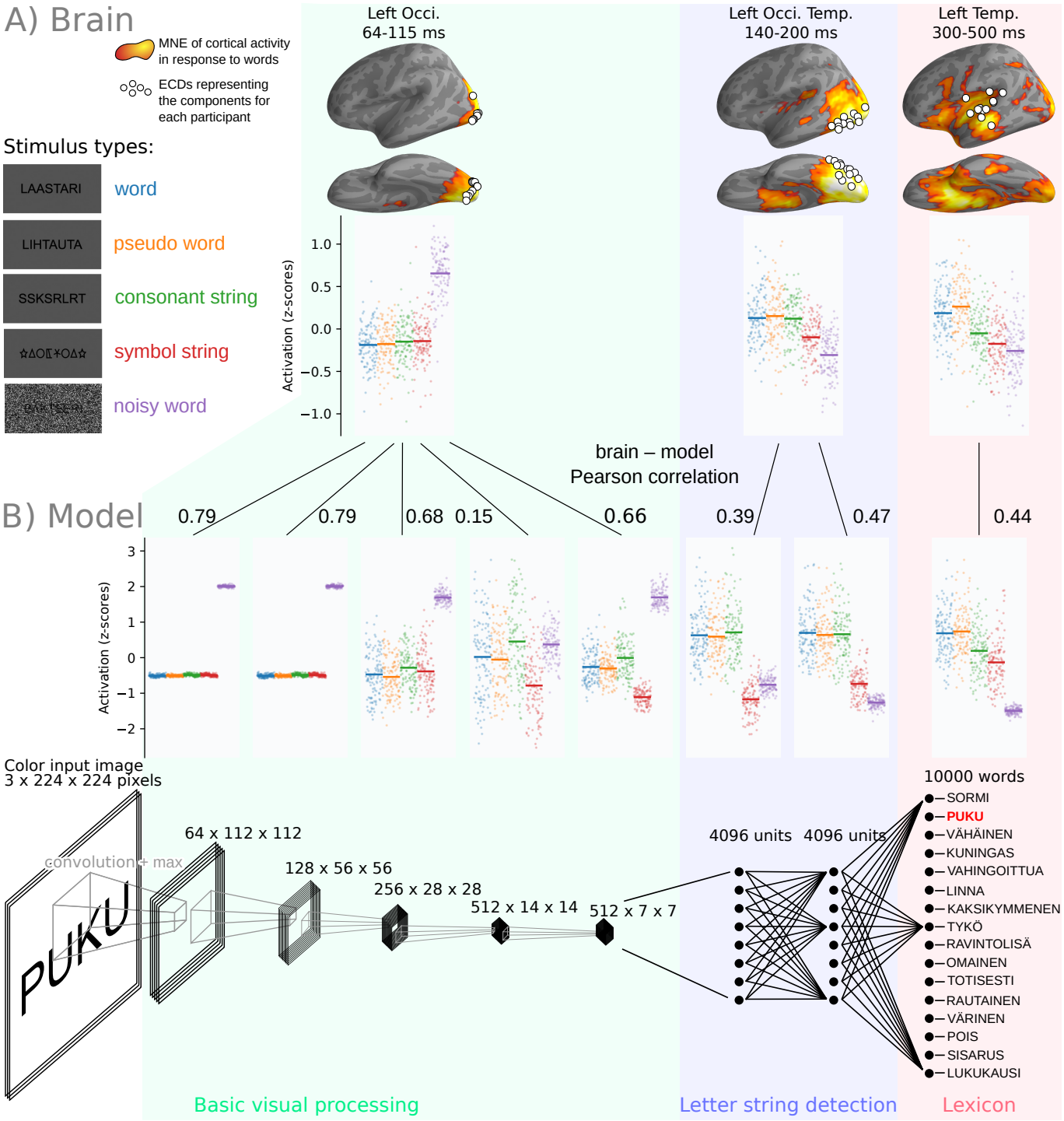

Figure 3: Comparison between MEG evoked activity and sum activity in each layer of the model. Based on the response pattern to each stimulus type, three processing stages where identified that correspond to different time windows in the MEG activity and different layer types in the model.

A) Evoked MEG activity, quantified during three time intervals. The grand-average minimum norm estimate (MNE) source activity to valid Finnish words is shown in orange hues. Overlaid are the positions of the most representative ECD for each participant during the indicated time interval, as determined by Vartiainen et al. (2011). Below is shown for each stimulus, the grand-average activity at the ECDs, integrated over the indicated time interval.

B) For each layer of the model, the sum rectified linear unit (ReLu) activation in each layer in response to each stimulus. The network architecture is shown below.

as word2vec or semantic features would clearly be better candidates. The reason why the final layer of the model is the way it is, is because this is the point where a hard 90 degree turn needs to be made from orthographic similarity to semantic similarity.

The model used in this study is a standard convolutional design and has many shortcomings as a model of the brain. Nevertheless, the fact that the model performs well despite these shortcomings shows the power of using deep learning models to implement cognitive theories.

4 Methods

5 Acknowledgements

We acknowledge the computational resources provided by the Aalto Science-IT project. This research was funded by the Academy of Finland (grant #310988 to M.v.V, #255349, #256459, #283071 and #315553 to R.S.).

References

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

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. doi:10.1016/j.tics.2013.11.005

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

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). doi:10.1038/s41598-018-28865-1

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. doi:10.1111/j.1749-818X.2008.00121.x

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. doi:10.1006/nimg.2002.1268

Hämäläinen, M. S., Hari, R., Ilmoniemi, R. J., Knuutila, J., & Lounasmaa, O. V. (1993). Magnetoencephalography - theory, instrumentation, and applications to noninvasive studies of the working human brain. *Reviews of Modern Physics*, *65*(2), 414–507. doi:10.1103/revmodphys.65.413

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. doi:10.1093/brain/121.6.1133

Kutas, M., & Federmeier, K. D. (2011). Thirty years and counting: Finding meaning in the N400 component of the event related brain potential (ERP). *Annu. Rev. Psychol.*, *62*, 621. doi:10.1146/annurev.psych.093008.131123

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 Lang.*, *132*, 22–27. doi:10.1016/j.bandl.2014.03.002

Laszlo, S., & Plaut, D. C. (2012). A neurally plausible Parallel Distributed Processing model of Event-Related Potential word reading data. *Brain Lang.*, *120*(3)arXiv NIHMS150003, 271–281. doi:10.1016/j.bandl.2011.09.001

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

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. doi:10.1145/1873951.1874254

McClelland, J. L., & Rogers, T. T. (2003). The parallel distributed processing approach to semantic cognition. *Nat. Rev. Neurosci.*, *4*(4), 310–322. doi:10.1038/nrn1076

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. doi:10.1016/S0010-0277(99)00067-0

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. doi: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. *Cogn. Neuropsychol.*, *29*(1-2), NIHMS150003, 34–55. doi:10.1080/02643294.2012.710600

Price, C. J. (2012). A review and synthesis of the first 20years of PET and fMRI studies of heard speech, spoken language and reading. *NeuroImage*, *62*(2), 816–847. doi: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. doi:10.1016/j.cortex.2017.12.020

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. doi:10.1038/s41593-019-0520-2

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. doi:10.1007/s11263-015-0816-y

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

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. (2018). Brain-Score: Which artificial neural network for object recognition is most brain-like? *bioRxiv*, 407007. doi:10.1101/407007

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. doi:10.1162/jocn.2007.19.7.1193

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*arxiv 1409.4842, 1–9. doi: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. doi:10.1093/brain/122.11.2119

Taylor, J. S. H., Rastle, K., & Davis, M. H. (2013). Can Cognitive Models Explain Brain Activation During Word and Pseudoword Reading? A Meta-Analysis of 36 Neuroimaging Studies. *Psychological Bulletin*, *139*(4), 766–791. doi:10.1037/a0030266

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. doi:10.1523/JNEUROSCI.3113-10.2011

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