Understanding the Neuro-Cognitive Mechanisms of Orthographic Learning in Humans and Baboons: A Comparison of Mechanistic and Connectionist Models

Janos Pauli and Benjamin Gagl
Self Learning Systems Lab, University of Cologne, Cologne, Germany

Abstract

Learning to read is essential for social participation. Here, we investigate how humans and baboons learn orthographic information. We use a neuro-cognitive mechanistic model—the Speechless Reader (SLR) and two connectionist models (CORnet-Z and ResNet-18) to investigate a human and a baboon dataset. The connectionist models employ neuronally plausible CNN architectures, while the SLR provides transparent implementations of orthographic decision behavior using pixel, letter, and letter sequence level prediction errors as representations. To align models and data, we train the models using identical trial sequences for each human and baboon. The SLR outperforms the CNNs across both species, especially on trial-wise metrics. While CNN responses diverge from individual behavioral patterns, the SLR's interpretable errors reveal that the complexity of orthographic representations increases with training. This finding suggests that domain-specific mechanistic models offer valuable insight into learned visual behavior across species.

Keywords: Reading; Computer Vision; Orthographic Decisions; Neuro-Cognitive Phenotyping; Humans; Baboons

Introduction

Efficient reading is critical for success in modern societies (Huettig & Pickering, 2019). Reading research is dominated by two types of models: mechanistic and connectionist models. Mechanistic models provide a transparent, handcrafted implementation of cognitive processes in reading (Coltheart et al., 2001). This line of research was particularly successful in generating effective remediation programs (i.e., phonics; Galuschka et al.,2014) and descriptions of individual differences in reading behavior (i.e., computational phenotypes; Perry et al.,2019). In contrast to these domain-specific models, data-driven connectionist models successfully described benchmark effects in reading behavior (i.e., Seidenberg & McClelland,1989; see Norris (2013) for a review). Recent work increasingly integrates CNN-based vision models to understand behavior and neural dynamics. Here, we assess whether learning orthographic stimuli—letter strings—can be better understood through a mechanistic, neuro-cognitively grounded approach.

Methods

We use two orthographic learning datasets. Baboons (Grainger et al., 2012) and humans (Eisenhauer et al., 2019) participants and models learned to classify known and novel letter strings over multiple trials. We used three models for simulation: (i) Speechless Reader (SLR, Gagl et al., 2024): A mechanistic model allowing the inspection of how participants learned. (ii) CORnet-Z: A shallow recurrent CNN designed to mimic cortical processing (Kubilius et al., 2018). (iii) ResNet-18: A deeper CNN with skip connections (He et al., 2016).

We trained CORnet-Z and ResNet-18 models using PyTorch (Paszke et al., 2019) with a binary output layer (familiar vs. novel) and cross-entropy loss. One model was trained for each human and baboon (learning rate of 0.0001; Adam optimizer (Kingma & Ba, 2014)). Input stimuli were grayscale letter strings (228×228 pixels, black Arial on white). To simulate prior visual experience, early layers were frozen in ImageNet-pretrained models (ResNet: "conv1", "bn1", "layer1"; CORnet: "V1", "V2"). Human models were further pre-trained on a lexical decision task using 1,074 German words and 1,074 pseudowords. Baboons' models were trained for one epoch on the same stimuli they saw, without validation or data augmentation, mirroring the experimental setup (see Hannagan et al.,2014,Linke et al.,2017 for a similar approach). Human models were fine-tuned across four training epochs and validated on font-switched data (Times New Roman), which matched the experimental sessions (N = 960; n = 240 per session). We measured model accuracy and used mean-squared errors and trial-wise similarity (see Geirhos et al., 2020) to compare the model to participant responses.

Results

Baboon Data. Baboons improved their performance gradually, with the first noticeable gains after about 1,000 trials (Fig. 1A). In contrast, both SLR models achieved higher performance much earlier—around 60% for the best-fitting and 70% for the best-performing variant. ResNet performed nearly perfectly, and CORnet showed a performance of around 85%. These differences in the behavior resulted in higher MSE values, reflecting the difference between model and baboon performance, for connectionist models than the mechanistic SLR implementations (Fig. 1B). After approximately 10,000 trials, all models reached similar accuracy levels. Still, the best-fit SLR model had the lowest MSE and the highest trial-wise similarity (Fig. 1C). While connectionist models showed a lower trial-wise similarity than both SLR models, they still had a higher similarity (i.e., κ) than the average between-baboon similarity.

Human Data. Mean human and model accuracies show an increase with learning for all except the CORnet model (see Fig. 1D). Human performance was unmatched, with only the best fitting SLR achieving 80% accuracy after four training

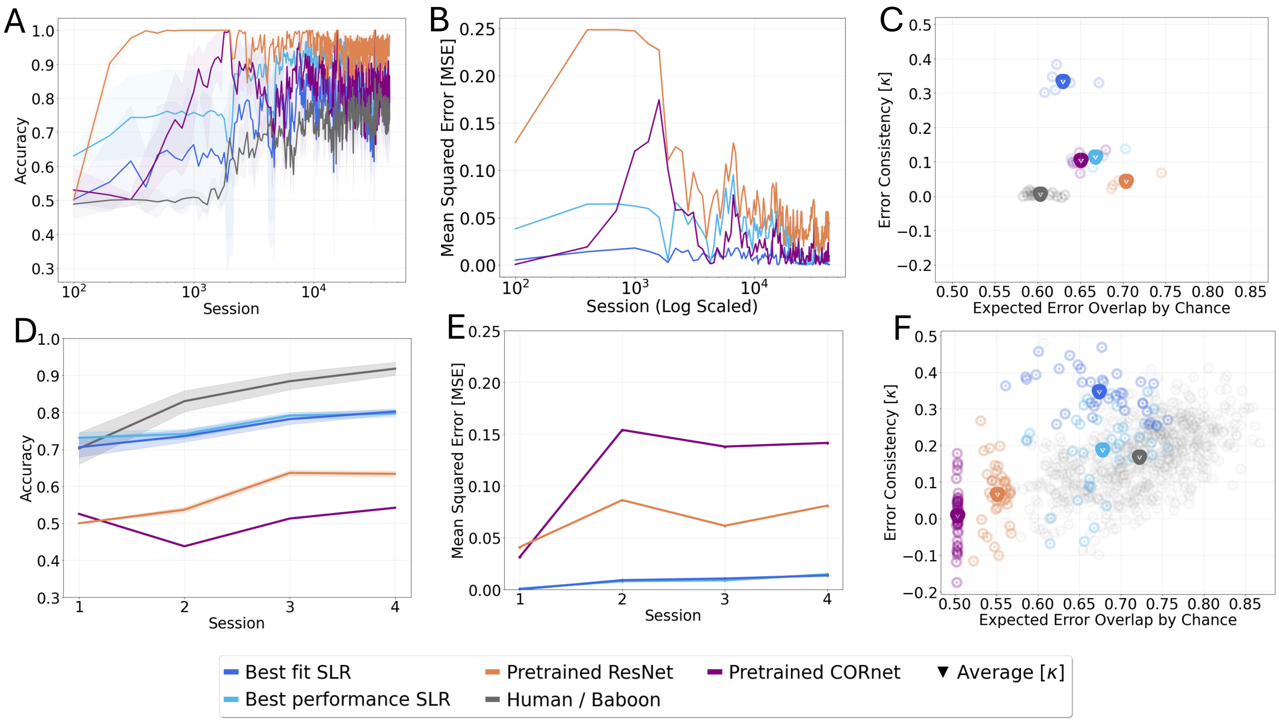

Figure 1: Baboon, human and model performance from the ResNet, CORnet, and two Speechless Reader model implementations (Best-fit and Best-performance variant). (A) Baboon, SLR, and CNN training session-level accuracy, including the 95% confidence intervals. (B) Model fit based on session-level mean squared errors comparing model with baboon performance. (C) Error consistency values (Cohens κ) plotted against the expected error overlap by chance, comparing model and baboon behavior on the level of single trials. Higher κ values indicate stronger item-level behavioral agreement that tends to increase with the expected error overlap (i.e., at high accuracies, the expected overlap is typically higher). The expected error overlap is based on the accuracies of two models (e.g., best-fit SLR and human). The higher the two model accuracies are, the more likely it is that they made the same item-level decisions by chance. In (D), we show session-level accuracy and CNN validation accuracy (E), session-level model fit, and (F) trial-level error consistency for the human dataset.

sessions. Again, connectionist models had lower overall and trial-wise similarity than both SLR variants (see Fig. 1E/F). Only ResNet showed a relatively high κ of 0.14 in the first session (i.e., highest κ of all connectionist models, Range: -0.02 - .10). Thus, both SLR models simulated the learning trajectories more accurately in both datasets, suggesting that we can utilize them for computational phenotyping.

Neuro-cognitive Phenotypes. The best-fitting SLR model reveals that, early in learning, all three representations—visual-pixel-level (oPE), letter-level (LPE), and sequence-level (sPE)—are engaged in both humans and baboons (Humans/Baboons, % oPE: 19/59, % LPE: 97/66, % sPE: 92/67). As experience increases and oPE becomes less relevant, reliance shifts toward LPE and sPE, which support orthographic processing (Humans/Baboons, % oPE: 0/31, % LPE: 100/97, % sPE: 100/88). This shift coincides with growing differences in prediction errors between learned and novel letter strings, leading to a decline in the informativeness of oPE (oPE difference learned/novel: early: -.7/-.4; late: -.6/-.1). Eventually, oPE shows the smallest error differences, explaining its reduced contribution to orthographic decisions.

Discussion

Here, we demonstrate that simple, domain-specific models, such as SLR, can effectively capture the learning dynamics of both baboons and humans in orthographic learning, outperforming general-purpose connectionist models (ResNet, CORnet) with increased interpretability. In baboons, connectionist models rapidly reached ceiling performance due to task simplicity (i.e., frequent stimulus repetition). In contrast, in the human dataset, connectionist models struggle as task difficulty increases (i.e., more stimuli are learned in fewer trials). In contrast, the SLR models revealed a consistent learning progression. From the used representation, we find that with learning, representations change from low-level pixel representations to higher-level orthographic units, aligning with theories of reading development (Gagl et al., 2015).

SLR's strength lies in its parsimony (only one free parameter), resilience to overfitting, and ability to offer interpretable, individual-level cognitive insights. These qualities make it well-suited to model reading, learning, and identifying precursors of reading difficulties. Ultimately, the results support the value of mechanistic, task-specific models in cognitive neuroscience and reading research.

References

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.

Eisenhauer, S., Fiebach, C. J., & Gagl, B. (2019). Context-based facilitation in visual word recognition: Evidence for visual and lexical but not pre-lexical contributions. *eNeuro*, *6*(2). Retrieved from https://www.eneuro.org/content/6/2/ENEURO.0321-18.2019 doi: 10.1523/ENEURO.0321-18.2019

Gagl, B., Hawelka, S., & Wimmer, H. (2015, apr). On sources of the word length effect in young readers. *Scientific Studies of Reading*, *19*(4), 289–306. Retrieved from https://doi.org/10.1080%2F10888438.2015.1026969 doi: 10.1080/10888438.2015.1026969

Gagl, B., Weyers, I., Eisenhauer, S., Fiebach, C. J., Colombo, M., Scarf, D., ... Mueller, J. L. (2024). Non-human recognition of orthography: How is it implemented and how does it differ from human orthographic processing. *bioRxiv*. Retrieved from https://www.biorxiv.org/content/early/2024/08/13/2024.06.25.600635 doi: 10.1101/2024.06.25.600635

Galuschka, K., Ise, E., Krick, K., & Schulte-Körne, G. (2014). Effectiveness of treatment approaches for children and adolescents with reading disabilities: A meta-analysis of randomized controlled trials. *PloS one*, *9*(2), e89900.

Geirhos, R., Meding, K., & Wichmann, F. A. (2020). Beyond accuracy: quantifying trial-by-trial behaviour of cnns and humans by measuring error consistency. *CoRR*, *abs/2006.16736*. Retrieved from https://arxiv.org/abs/2006.16736

Grainger, J., Dufau, S., Montant, M., Ziegler, J. C., & Fagot, J. (2012). Orthographic processing in baboons (papio papio). *Science*, *336*(6078), 245–248. Retrieved from https://www.science.org/doi/abs/10.1126/science.1218152 doi: 10.1126/science.1218152

Hannagan, T., Ziegler, J. C., Dufau, S., Fagot, J., & Grainger, J. (2014, Jan). Deep learning of orthographic representations in baboons. *PLoS ONE*, *9*(1), e84843. Retrieved from https://doi.org/10.1371/journal.pone.0084843 doi: 10.1371/journal.pone.0084843

He, K., Zhang, X., Ren, S., & Sun, J. (2016, June). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition (cvpr).*

Huettig, F., & Pickering, M. J. (2019, Jun). Literacy advantages beyond reading: Prediction of spoken language. *Trends in Cognitive Sciences*, *23*(6), 464–475. Retrieved from https://doi.org/10.1016/j.tics.2019.03.008 doi: 10.1016/j.tics.2019.03.008

Kingma, D. P., & Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980* .

Kubilius, J., Schrimpf, M., Nayebi, A., Bear, D., Yamins, D. L. K., & DiCarlo, J. J. (2018). Cornet: Modeling the neural mechanisms of core object recognition. *bioRxiv*. Retrieved from https://www.biorxiv.org/content/early/2018/09/04/408385 doi: 10.1101/408385

Linke, M., Bröker, F., Ramscar, M., & Baayen, H. (2017, Aug). Are baboons learning ”orthographic” representations? probably not. *PLOS ONE*, *12*(8), e0183876. Retrieved from https://doi.org/10.1371/journal.pone.0183876 doi: 10.1371/journal.pone.0183876

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

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., ... Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems* (Vol. 32). Curran Associates, Inc. Retrieved from https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf

Perry, C., Zorzi, M., & Ziegler, J. C. (2019). Understanding dyslexia through personalized large-scale computational models. *Psychological Science*, *30*(3), 386–395. Retrieved from https://doi.org/10.1177/0956797618823540 (PMID: 30730792) doi: 10.1177/0956797618823540

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological review*, *96*(4), 523.