# Specifying the orthographic prediction error for a better understanding of efficient visual word recognition in humans and machines

**Wanlu Fu (wanlu.fu@uni-koeln.de) & Benjamin Gagl (benjamin.gagl@uni-koeln.de)**
Self Learning Systems Lab at the Department of Special Education and Rehabilitation
Frangenheimstraße 4, Cologne, Germany

## Abstract

**Recent evidence suggests that readers optimize low-level visual information following the principles of predictive coding. Based on a transparent neurocognitive model, we postulated that readers optimize their percept by removing redundant visual signals, which allows them to focus on the informative aspects of the sensory input, i.e., the orthographic prediction error (oPE). Here, we test alternative oPE implementations by assuming all-or-nothing signaling units based on multiple thresholds and compare them to the original oPE implementation. For model evaluation, we implemented the comparison based on behavioral and electrophysiological data (EEG at 230, 430 ms). We found the highest model fit for the oPE with a 50% threshold integrating multiple prediction units for behavior and the late EEG component. The early EEG component was still explained best by the original hypothesis. In the final evaluation, we used image representations of both oPE implementations as input to a deep-neuronal network model (DNN). We compared the lexical decision performance of the DNN in two tasks (words vs. consonant strings; words vs. pseudowords) to the performance after training with unaltered word images and found better DNN performance when trained with the 50% oPE representations in both tasks. Thus, the new formulation is adequate for late but not early neuronal signals and lexical decision behavior in humans and machines. The change from early to late neuronal processing likely reflects a transformation in the representational structure over time that relates to accessing the meaning of words.**

**Keywords:** Visual word recognition; Visual-orthographic processing; Predictive coding; EEG; Behavior; Computer vision model

## Introduction

Predictive processing hypotheses assume that the brain uses learned patterns in the environment to optimize information processing (Clark, 2013; De Lange, Heilbron, & Kok, 2018). By making predictions based on past experiences, the brain constructs probabilistic inferences about upcoming sensory information. The predominant paradigm to investigate predictive processing relies on manipulating predictability from the immediate or recently learned context of the sensory input (Hawelka, Schuster, Gagl, & Hutzler, 2015; Heilbron, Richter, Ekman, Hagoort, & de Lange, 2020; Hofmann, Remus, Biemann, Radach, & Kuchinke, 2022; Eisenhauer, Gagl, & Fiebach, 2022; Yan, de Lange, & Richter, 2023; Alink, Schwiedrzik, Kohler, Singer, & Muckli, 2010). In contrast, we recently showed that humans implement predictive processing based on patterns stored in our long-term memory (Gagl et al., 2020). Here, we investigate these knowledge-based predictions by implementing multiple models that inform about the structure of the neuronal implementation, which allows the formation of prediction error representations that underly efficient human reading. Central investigation is the precision in predictive processing in visual word recognition (de Lange, Schmitt, & Heilbron, 2022). We implement alternative hypotheses based on transparent computational models (Guest & Martin, 2021) and evaluate them based on human behavior and brain responses but also with computer vision models.

The principles of predictive coding (Rao & Ballard, 1999) provide a framework for investigating the interaction between sensory processing (i.e., vision) and higher levels of processing essential to reading (i.e., phonology, semantics). The Prediction Error Model of Reading (PEMoR, Gagl et al., 2020) postulates that the reader removes redundant visual signals to focus on the informative aspects of the percept (i.e., orthographic prediction error; oPE). The oPE is essential for efficient word perception and correlates with early brain activation in visual-perceptual systems of the occipital cortex.

A central characteristic of the PEMoR is its simplicity, and the implementation is fully transparent, without free parameters based only on a single image computation. Here, we explore two hypotheses that implement a more precise binary all-or-nothing signal processing on the level of the prediction and the prediction error. This hypothesis originated from the finding that neuronal signaling likely implementers all-or-nothing signaling (i.e., spikes; Qin et al., 2020; Saszik & DeVries, 2012). In the original PEMoR, the prediction results from a pixel-by-pixel mean integrating

the visual appearance of all words stored in the mental lexicon. Thus, the resulting prediction and prediction error representations comprised graded values for each pixel. When assuming binary all-or-nothing coding, the value of a pixel can be either 0 or 1 (see Fig. 1). For the implementation of the prediction, we chose two options: One that integrated information from multiple prediction units (Fig. 1A) and one that integrated information from one prediction unit (Fig. 1B). We use multiple thresholds to explore different implementations for the binary units (see examples of $g$ in Fig. 1A, B). Here, we contrast the original $oPE$ implementation with the two variants of the new implementation for each threshold based on behavioral and EEG data.

In the final part, we use a computer vision model based on a deep neural network (DNN) to investigate the informativeness of the winning oPE representations in comparison to unaltered images. Here, we train the DNN with either image representations of the oPE or word images and then compare the performance of the trained models. This evaluation is interesting as DNNs have demonstrated good capabilities of learning from experience (He, Zhang, Ren, & Sun, 2016; LeCun, Bengio, & Hinton, 2015; Krizhevsky, Sutskever, & Hinton, 2012). The inherent ability of DNNs to overcome the drawback of traditional algorithms dependent on hand-designed features comes at the cost of interpretability of the inner workings of the models (Samek, Montavon, Lapuschkin, Anders, & Müller, 2021; Alvarez Melis & Jaakkola, 2018). The high flexibility of DNN architectures often prevents concrete insights into the processes underlying its decisions, thus creating a new researchable artifact (Ma & Peters, 2020). This characteristic stands in steep contrast to the transparency of the implementations of the oPE models described above (van Rooij et al., 2023).

## Method

***New Implementation of the prediction error model of reading.*** As assumed in PEMoR, the orthographic prediction error was implemented by image-based computations, indicating that for each pixel of the images (e.g., 5-letter words: 140x40 pixels in size), we estimate a knowledge-based prediction and stimulus-specific prediction errors (see Eq. (1)). So, the resulting prediction error representation is an image containing binary values for each pixel. The prediction matrix was calculated based on two algorithms: (i) mean and (ii) binary predictions (i.e., shown in Fig. 1 A and B). The consequence of this differentiation is that if we implement one unit, the prediction input is also binary. If we implement multiple units, we assume a graded signal again. This results from the case that when multiple binary inputs are integrated, the result will be graded in every situation in which no consensus is present (i.e., only 0 when all prediction units signal 0; only 1 when all prediction signals signal 1; for all cases in which we have a combination of 0s and 1s we get a graded signal, e.g., when assuming four units and only one signals a one the result would be 0.25). Note that the words used to determine the prediction for each pixel were taken from the German SUBTLEX database (i.e., all five-letter words; Brysbaert et al., 2011). We implement the estimation of the prediction error of both variants using thresholds from 10-90% of the signal in 10% steps (see Eq. (1) and examples in Fig. 1).

$$oPE_{i,j} = \begin{cases} 1 & \text{if } (SensoryInput_{i,j} - Prediction_{i,j}) > Threshold \\ 0 & \text{otherwise} \end{cases}$$

$$i, j = 140, 40$$

$$Threshold = \begin{bmatrix} .1 & .2 & .3 & .4 & .5 & .6 & .7 & .8 & .9 \end{bmatrix} \tag{1}$$

Here, $SensoryInput_{i,j}$ indicates the image input that is presented at a given trial, including all pixels, represented by gray values between $[0,1]$. $Prediction_{i,j}$ represents the prediction matrix, calculated as an average across the $SensoryInput_{i,j}$ related to all words in the lexicon (i.e., again represented by gray values between $[0,1]$). The pixel value of the orthographic prediction error, $oPE$, was only set to $1$ when the difference between the $SensoryInput$ and the $Prediction$ exceeded a $Threshold$ (e.g., at a 50% threshold, when the sensory input is 1, and the prediction is 0.2, the prediction error is 1 because the difference, i.e., 0.8 is higher than the 50% threshold, i.e., 0.5; if the prediction is 0.6, the prediction error is 0 as the difference value, i.e., .2 is lower than the value of the threshold).

We implemented the main difference between the two new formulations in the algorithm that determined the $Prediction$. For the mean prediction algorithm, we used the same formulation described in (Gagl et al., 2020). Here, we calculated the mean over the pixel values from all words from the lexicon. For the binary algorithm, in addition, the same $Threshold$ as for $oPE$ calculation was applied to the prediction so that the grayscale representation of the prediction became binary (i.e., only including $[0,1]$).

$$Prediction_{i,j} = \begin{cases} 1 & \text{if } (\frac{\sum_1^n StoredWords_{i,j}}{n}) > Threshold \\ 0 & \text{otherwise} \end{cases}$$

$$i, j = 140, 40$$

$$Threshold = \begin{bmatrix} .1 & .2 & .3 & .4 & .5 & .6 & .7 & .8 & .9 \end{bmatrix} \tag{2}$$

The pixel values of the binary $Prediction$ (see Eq. (2)) are $1$ when the average across the pixel values of the image representation of the $StoredWords$ in the lexicon (i.e., $StoredWords_{i,j}$) exceeded the $Threshold$. The pixel values of the binary $Prediction$ can have any value between $[0, 1]$ as no threshold is applied (i.e., cp. the prediction images shown in Fig. 1)

For the model evaluations, we generated an aggregated $oPE$ for each stimulus presented in the studies. We summed gray values from the $oPE$ images for each stimulus to achieve this. So, as shown in the examples of Figure 1, the gray values represent the amount of error per pixel (i.e., a white pixel represents a high prediction error; see Gagl et al., 2020 for more details).

***Model evaluations.*** In this study, we evaluate the new oPE implementations by model comparisons. In the first part, we compared nineteen new oPE assumptions (binary vs. mean prediction for nine threshold levels each) with each other and against the original implementation of the oPE described in Gagl et al. (2020). We use linear mixed effect models (Bates, Mächler, Bolker, & Walker, 2014) and the Akaike Information Criterion (AIC, Akaike, 1973) to find the best fitting model using behavioral lexical decision data. In the second part, we implement the same model comparison logic but use electrophysiological data measured by EEG (find detailed descriptions of the dataset in Gagl et al., 2020). Here, we had the chance to investigate the highly time-resolved neuronal processes underlying the behavioral responses investigated in the first evaluations.

In the third part, we evaluate the winning models from the behavioral and neuronal evaluations based on the performance of a computer vision model (i.e., ResNet18, He et al., 2016; see Ma & Peters, 2020 for a similar logic). We used the ResNet architecture as it has a good image classification performance and integrated batch normalization, which is essential to DNN performance (Laurent, Pereyra, Brakel, Zhang, & Bengio, 2016; Santurkar, Tsipras, Ilyas, & Madry, 2018; Ba, Kiros, & Hinton, 2016). The latter is interesting as it performs a process similar to that assumed for estimating the oPE representations. Here, the input is subtracted by the input of the mean based on a $batch$ of data used for training and then transformed based on two free parameters set during model training. The area size under each learning curve (AUC) has been proposed as a precise measure for evaluating the predictive ability of learning algorithms (Huang & Ling, 2005), the AUC scores comparison in this study allows the test of the $oPE$ representations compare to unaltered images of letter strings. So, we can use the AUC score interpreted as a summary measure of batch normalization applied to unaltered word images results in a similar performance compared to the oPE representations that more transparently implement the optimization of the sensory input. This evaluation is not based on the aggregated $oPE$ but on the image representations shown in Figure 1. We compared the performance in two lexical decision tasks where we changed the type of the non-words. We had an easy task in which we used consonant strings and a more challenging task in which we used pseudowords as non-words. Lexical decisions involving consonant strings are easy, as one only has to identify if the letter involves vowels to identify the words in the task. The task is more difficult when we include pseudowords, as here, a more general difference must be detected as both words and non-words include vowels. This evaluation indicates how informative the $oPE$ representations are compared to unaltered images of letter strings. Overall, we provide two evaluations based on data from human data and one utilizing the performance of a computer vision model.

*Word Recognition Behavior.* We adopt the model comparison analysis procedure and three datasets from Gagl et al. (2020) to evaluate the new orthographic prediction error (oPE) variants (behavioral lexical decision data; find here: https://osf.io/d8yjc/). All stimuli had five letters (Behaviour: 800 Words & non-words; EEG: 200 Words & non-words), and participants were typically reading native speakers (Behaviour: N = 35; EEG: N = 31).

We estimated linear mixed models (Bates et al., 2014), including the oPEs as predictors to describe logarithmic transformed response times. Inspired by the previous analysis (see Gagl et al., 2020) estimating the interaction of the oPE and word lexicality (i.e., word or non-word), we also added the number of pixels, word frequency, and decision accuracy as covariates of no interest to the models. For the model fit comparison, we used the Akaike Information Criterion (AIC) concerning the null model without an oPE (Akaike, 1973).

*Electrophysiological data.* To test the timing and effect changes of the new oPE on the neuronal level, we focused on the two datasets previously identified as relevant: (i) posterior electrodes at 230 ms after stimulus onset showing an oPE effect and (ii) frontal electrodes at 430 ms after stimulus onset showing an interaction of the oPE with lexicality. EEG epochs were pre-processed (see Gagl et al., 2020), and for the first LMM-based analysis, in addition to the new oPEs, we added lexicality and the number of pixels as covariates of no interest. Again, the AIC is used to compare the model fit.

As an additional comparison, we again ran the linear regression analysis described in Gagl et al. (2020) for the entire EEG epochs (from 200 ms before and 800 ms after word onset). We used the multiple regression analysis with the same parameters as in the time-point-specific EEG evaluations and the winning oPE model (i.e., number of pixels, word/non-word, and the interactions with word/non-word distinction). Inference was made based on a cluster-based permutation test (Maris & Oostenveld, 2007). All clusters with a probability of less than an assumed alpha value of 0.05 under this simulated null hypothesis were considered statistically significant.

*Computer vision model performance.* Here, we used a computer vision model based on a deep-network architecture, the ResNet18 (He et al., 2016), with a multi-layered structure including batch normalization from the *pytorch* framework `https://github.com/pytorch`. The latter is especially interesting, as it, similar to PEMoR, is implemented to optimize the model's input for image recognition tasks (Bjorck, Gomes, Selman, & Weinberger, 2018). This process optimizes the sensory input based on the previously learned input (i.e., a *batch*; Santurkar et al., 2018). Here, one could assume that the ResNet implements a similar representation as the oPE. So, if this is the case, the critical test will compare the performance of a DNN trained with oPE representations and a DNN trained on unaltered letter string images. If the oPE representation is more informative than a word input, we expect better performance with less data during training and potentially better overall lexical decision performance.

Here, we trained the ResNet18 multiple runs (50 simulations) on training data and tested the performance based on left-out training data. This procedure was repeated for the winning oPE, the original oPE, and letter string images (baseline). In detail, to investigate the learning curve, we trained the model on parts of the data (increases by 1% each time between 1% and 5%; increases by 10% each time between 10% and 50%; 100% of the training dataset. Total: 9506 images for each condition, i.e., words, pseudowords, and consonant strings). The letter strings had 3-5 letters projected to a 64x64 pixels image, and the words were taken from the SUBTLEX-DE database (Brysbaert et al., 2011). Non-words were created by randomly replacing vowels from the words with other vowels to create pseudowords and replacing vowels with random consonants to create consonant strings. Also, we checked whether a created letter string was accidentally a word. When this was the case, we repeated the procedure. To account for the very different nature of the two types of non-words, we implemented two tasks: (i) words against pseudowords and (ii) words against consonant strings.

## Results

***oPE new formulations in Word Recognition Behavior.*** Correlations between the new and the original oPEs in both stimulus sets showed low similarity for the low percentage thresholds (10-20% thresholds, range: r = .07 to .32) and high similarity for the other implementations (30-90% thresholds range: r = .52 to .92).

***Behavioral results.*** Behavioral evaluations indicated the highest model fit at a threshold of 50% for the Mean prediction with only one AIC point difference to the Binary prediction (see Fig. 2A). Generally, higher model fits are found for the Mean prediction. Also, compared to the original oPE, all but two oPEs (10 & 90% thresholds; 2 of 9) based on the Mean prediction and only three for the Binary prediction (lower AICs for 7 of 9) showed higher model fits.

In detail, we found, for the mean oPE with a 50% threshold, a significant interaction with lexicality (Interaction Estimate: 31.74, SE = 3.29, t = 9.64) and a significant fixed effect of oPE (Estimate: -26.69, SE = 2.49, t = -10.70; Covariates of no interest: Estimate: -0.03, SE = 0.00, t = -5.94). When inspecting the response time patterns (see Fig. 3), we find, again, that for words, we can identify a positive effect and, for non-words, a negative effect.

***Electrophysiological results.*** The original oPE had the highest model fit at amplitudes at 230 ms. Still, all but one new oPE formulation (10% threshold, binary prediction) increased the model fit significantly in contrast to the null model (i.e., AIC difference $> 3$; see Fig. 2B). Again, oPEs based on the Mean prediction had higher model fits in all but one case (40%).

At 430 ms, the model comparison pattern was highly similar to the behavioral results (cp. Binary: r = .97; Mean: r = .93 vs. Binary: r = .76; Mean: r = .75 for the 230 ms time window). Although the general AIC differences have been smaller than in the early time window, we found a clear peak for the 50% threshold oPE of both prediction assumptions that was higher cp. to the original oPE. Again, we found higher AIC differences for oPEs based on the Mean prediction assumptions (6 of 9).

In addition, we investigated the effect of the 50% threshold mean oPE with multiple regression analysis for the full epochs. The models included the 50% threshold oPE, number of pixels, lexicality, and the interactions of the oPE with lexicality. In comparison to the original oPE, we again find a significant main effect of the new oPE, but the cluster was only from 210 ms to 250 ms in the similar sensors (cp. dashed lines vs. yellow area in Fig. 4A). This

finding is in line with the 230 ms model comparison analysis. A significant of lexicality was found from 360 ms to 800 ms, started later but lasted longer compared to the original oPE (see 4B). In contrast, the interaction between 50% threshold oPE and lexicality started earlier and lasted longer for the new (from 300 ms to 800 ms) compared to the original oPE implementation (see Fig. 4C).

***Computer vision model performance.*** Based on the previous analysis, we focused here on the 50 % oPE (winning model for behavior and EEG at 430 ms), the original oPE (winning model for EEG at 230 ms), and unaltered letter string images as input to the ResNet18. Lexical decision performance increased with the training dataset size for all three inputs and the two task versions in both the training and test datasets (see Fig. 5). Distinguishing words from consonant strings was much easier than pseudowords (see Fig. 5A). Welch's ANOVA Comparisons of the AUC scores showed significant differences between oPE models and baseline images in the easier task (words vs. consonant strings, $W(2,97.61)=12.27$, $p<0.00$, (see Fig. 5B), post hoc comparisons revealed ResNet18 performance was better after training with oPE representations (all $p<0.02$). Similarly results found in Words vs. Pseudowords task ($W(2,89.2)=285.95$, $p<0.00$, see Fig. 5C), a larger AUC score for the new oPE formulation was found (i.e., mean prediction and a 50% threshold) compared to both original oPE model and the baseline condition (all $p<0.00$).

## Discussion

The Prediction error model of reading (PEMoR) highlights the role of orthographic prediction errors, representations that integrate visual-orthographic knowledge with sensory inputs, for efficient word perception (Gagl et al., 2020). Here, we revisited the original PEMoR and investigated if prediction error precision is central to conserving critical visual information for word identification. Adapting the formulation of the PEMoR allowed us to test several new hypotheses that we implemented based on all-or-nothing (i.e., binary) predictions and prediction errors, facilitating transparent theoretical development (i.e., as suggested by Guest & Martin, 2021). Our main finding was that all-or-nothing (i.e., binary; Qin et al., 2020) orthographic prediction errors (oPE) implemented with a 50% threshold that results from a mean prediction best explained human behavioral performance. Also, we found that late brain potentials (after 300 ms) and, when these prediction errors are used as input for computer vision models, model performance was best. In contrast, the original oPE implementation (i.e., mean prediction and prediction error) better described an electrophysiological component around 200 ms. The finding that one likely implements the new formulation that assumes precise prediction error signals in late processes (N400, behavior) but not necessarily in early processing is intriguing. The finding that the original formulation best describes the early component highlights the only difference between the formulations: The implementation of a graded prediction error in the original and a binary prediction error in the new formulation. The difference between early, graded prediction error and late binary prediction error could indicate a transformation of the oPE representation over time, settling on a solution with less nuance (i.e., only binary signaling). In other words, we potentially examine a case of precision increase in a prediction error representation from early perceptual to later meaning retrieval processes.

Revisiting the formulation of oPE allowed a detailed investigation of how integrating sensory input and knowledge could be implemented. We found that implementing a more precise, all-or-noting oPE representation is better suited for meaning extraction. We found consistent evidence for implementing a more precise prediction error in human and model behavior, potentially representing the end product of a perceptual process that reduces sensory noise in reading (Norris & Kinoshita, 2012). The association with the extraction of meaning comes from the repeated finding that the oPE effect is positive in words and negative in non-words (i.e., for words: higher response times and larger EEG amplitudes when the oPE is high; non-words: inverted pattern). This dissociation of the oPE effect for meaningful and meaningless letter strings could stem from a process in preparation for lexical access filtering out letter strings without meaning (e.g., the lexical categorization process, Gagl et al., 2022). In line with the lexical categorization proposal is the implementation of the original and the new PEMoR that assumes the integration of multiple units on the level of the knowledge-based prediction from long-term memory (e.g., generally in line with most proposals in visual word recognition Carreiras, Armstrong, Perea, & Frost, 2014; Price & Devlin, 2011; Twomey, Duncan, Price, & Devlin, 2011; Taylor, Rastle, & Davis, 2013; Coltheart, 2005). New implementations for oPE used multiple thresholds to detect optimal neural coding of informative visual features and found that the 50% threshold oPE leads to better word recognition behavior. However, graded signals better describe early neuronal signatures of efficient word recognition.

The evaluation of new formulations of oPE on the EEG time course showed that around 200 ms after stimulus onset, posterior electrodes showed an oPE effect, which was more substantial when estimated with the original oPE. This effect is similar to the previously identified occipitotemporal N170 component, which was interpreted as an earlier maker of visual-orthographic processes in visual word recognition (i.e., differentiates between orthographic

stimuli other stimuli such as symbols and faces Carreiras et al. (2014); Dehaene, Cohen, Sigman, and Vinckier (2005); Solomyak and Marantz (2009)). Studies focusing on the N170 component (see Amora et al., 2022; Pickering & Gambi, 2018 for a systematic review) also reported that N170 activity may correspond to logographic processing where a word is processed as a whole during the reading of very frequent words or after a massive repetition of a word (Simon, Petit, Bernard, & Rebaï, 2007). However, such findings are hardly found consistently across studies (Nieuwland, 2019). Thus, our and previous findings contribute to a deeper understanding of the neural mechanisms underlying word perception and align with prior research on N170's role in discriminating and processing orthographic features in the visual-orthographic domain (Maurer, Zevin, & McCandliss, 2008). Still, overall, as found here and described previously, processing at this stage precedes lexical processing, as the oPE effect is similar in letter strings with and without meaning, also preceding lexicality effects, typically found 400 ms after stimulus onset.

The N400 component is typically sensitive to lexical and semantic processing (Kutas & Federmeier, 2011; Kutas & Hillyard, 1980). Previous studies showed strong effects when extracting meaning from perceived words (Eisenhauer, Fiebach, & Gagl, 2019; Eisenhauer et al., 2022; Almeida & Poeppel, 2013; Fiebach, Gruber, & Supp, 2005; Dufau, Grainger, Midgley, & Holcomb, 2015) and capturing an implicit, probabilistic representation of sentence comprehension (Rabovsky, Hansen, & McClelland, 2018). The observation that the new oPE formulation, including more precise representations, was the adequate predictor for brain potentials in this time window may indicate the role of the representation as a key to word meaning (Gagl et al., 2022). Central here is that the oPE is part of an interaction term with the lexicality of the word, showing inverted effects for meaningful and meaningless letter strings. This interaction is also present in behavioral responses in lexical decision tasks. Furthermore, when comparing the interaction effects based on the original and the new oPE formulation, we find that the main effect of lexicality becomes significant at a later point in time when implementing the analysis with the new, more precise prediction error. Also, the oPE by lexicality interaction clusters starts before the lexicality cluster, which is the opposite when one utilizes the original oPE formulation. Here, we might speculate that the high precision oPE could be a better key to word meaning. Not only does the interaction of oPE and lexicality precede the lexicality main effect, but also the finding that Deep neuronal network models can extract relevant information for lexical decision tasks better when trained with the precise oPE representation. Thus, combined, these findings indicate a potential role of the oPE when extracting word meaning in visual word recognition.

Using deep neural networks to evaluate our oPE assumptions follows the idea of using computer vision models as research artifacts (Ma & Peters, 2020). This class of models was successfully used to investigate potential object recognition architectures potentially implemented in human object recognition (Geirhos et al., 2017; Ma & Peters, 2020; Lindsay, 2021) or more specific visual word and letter recognition (Hannagan, Agrawal, Cohen, & Dehaene, 2021; Testolin, Stoianov, & Zorzi, 2017; LeCun et al., 1989; Yin, Biscione, & Bowers, 2023). Here, we are only interested in the model architectures so far as we wanted to use a model with recurrent connections (i.e., top-down connections as assumed in predictive coding accounts, e.g., see Rao & Ballard, 1999; Gagl et al., 2020) and batch-normalization (Laurent et al., 2016; Cooijmans, Ballas, Laurent, Gülçehre, & Courville, 2016; Lu, Sindhwani, & Sainath, 2016). The latter implements a process similar to estimating the oPE, as the sensory input to the model is normalized based on the mean of the previous model input. This model configuration of the ResNet model family showed that the automatically generated internal representation of the neuronal network model generated after learning with unaltered word images could not catch up with theoretically defined oPE representations. This finding was present in an easy (Word. vs. Consonant strings) and hard (Word. vs. Pseudowords) version of a lexical decision task. Architecture is essential, but understanding the representations implemented in visual word recognition might not only allow us to come up with ways to increase human reading abilities (e.g., see Gagl & Gregorova, 2023) but also allow better performance for object recognition models (e.g., see Sexton & Love, 2022 for a similar approach but using brain response data as an input).

***Limitations.*** While revisiting the oPE implementation, we still lack neuronal plausibility in our theoretical accounts. Here, the implementation is minimalistic, lacking important neurophysiological details (Bastos et al., 2012). Thus, future work should not only increase the neuronal plausibility of the PEMoR. Also, several aspects of reading are currently not covered so, including prediction error representations on other levels of implementation (as partly outlined here (Gagl, Weyers, & Mueller, 2021)) that are likely part of access to meaning (e.g., see Coltheart, 2005). Derived from the findings of this study, implementing a mechanism conducting prediction error precision increase over time will be essential. One proposal could be a version of the PEMoR that involves timing such that one implements a precision increase before or while the oPE representation could be passed on to help access word meaning (see Fig. 6).

***Conclusion.*** Here, we revisited the PEMoR and identified a process indicating an evolvement of efficient representations over time. We found that implementing a graded prediction better describes early processing, and a more precise binary prediction better describes late processes and word recognition behavior in humans and machines. Thus, our results provide a view on developing efficient visual-orthographic representations such that these prediction error-based representations could hold the key to accessing word meaning. The image computable nature of the PEMoR representations based on images and domain-general assumptions of neuronal processing (e.g., Rao & Ballard, 1999) could allow future studies to investigate representations in higher vision outside of the domain of visual word recognition and reading.

***Data and materials availability.*** The script and data for model evaluations are available at https://github.com/Wanlu-koeln/oPE-new-formulation-behavior-EEG.git. The Resnet18 lexical decision training code and oPE images are made publicly available for research purpose at https://github.com/Wanlu-koeln/Resnet-lexical-decision-performance.git.

## References


Akaike, H. (1973). Maximum likelihood identification of gaussian autoregressive moving average models. *Biometrika*, *60*(2), 255–265.

Alink, A., Schwiedrzik, C. M., Kohler, A., Singer, W., & Muckli, L. (2010). Stimulus predictability reduces responses in primary visual cortex. *Journal of Neuroscience*, *30*(8), 2960–2966.

Almeida, D., & Poeppel, D. (2013). Word-specific repetition effects revealed by meg and the implications for lexical access. *Brain and language*, *127*(3), 497–509.

Alvarez Melis, D., & Jaakkola, T. (2018). Towards robust interpretability with self-explaining neural networks. *Advances in neural information processing systems*, *31*.

Amora, K. K., Tretow, A., Verwimp, C., Tijms, J., Leppänen, P. H., & Csépe, V. (2022). Typical and atypical development of visual expertise for print as indexed by the visual word n1 (n170w): a systematic review. *Frontiers in Neuroscience*, *16*, 898800.

Ba, J. L., Kiros, J. R., & Hinton, G. E. (2016). Layer normalization. *arXiv preprint arXiv:1607.06450*.

Bastos, A., Usrey, W., Adams, R., Mangun, G., Fries, P., & Friston, K. (2012). Canonical microcircuits for predictive coding. *Neuron*, *76*(4), 695-711. doi: https://doi.org/10.1016/j.neuron.2012.10.038

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2014). *Fitting linear mixed-effects models using lme4.* arXiv. doi: 10.48550/ARXIV.1406.5823

Bjorck, N., Gomes, C. P., Selman, B., & Weinberger, K. Q. (2018). Understanding batch normalization. *Advances in neural information processing systems*, *31*.

Brysbaert, M., Buchmeier, M., Conrad, M., Jacobs, A. M., Bölte, J., & Böhl, A. (2011). The word frequency effect. *Experimental psychology*.

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.

Clark, A. (2013). Whatever next? predictive brains, situated agents, and the future of cognitive science. *Behavioral and brain sciences*, *36*(3), 181–204.

Coltheart, M. (2005). Modeling reading: The dual-route approach. *The science of reading: A handbook*, 6–23.

Cooijmans, T., Ballas, N., Laurent, C., Gülçehre, Ç., & Courville, A. (2016). Recurrent batch normalization. *arXiv preprint arXiv:1603.09025*.

Dehaene, S., Cohen, L., Sigman, M., & Vinckier, F. (2005). The neural code for written words: a proposal. *Trends in cognitive sciences*, *9*(7), 335–341.

De Lange, F. P., Heilbron, M., & Kok, P. (2018). How do expectations shape perception? *Trends in cognitive sciences*, *22*(9), 764–779.

de Lange, F. P., Schmitt, L.-M., & Heilbron, M. (2022). Reconstructing the predictive architecture of the mind and brain. *Trends in Cognitive Sciences*.

Dufau, S., Grainger, J., Midgley, K. J., & Holcomb, P. J. (2015). A thousand words are worth a picture: Snapshots of printed-word processing in an event-related potential megastudy. *Psychological science*, *26*(12), 1887–1897.

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).

Eisenhauer, S., Gagl, B., & Fiebach, C. J. (2022). Predictive pre-activation of orthographic and lexical-semantic representations facilitates visual word recognition. *Psychophysiology*, *59*(3), e13970.

Fiebach, C. J., Gruber, T., & Supp, G. G. (2005). Neuronal mechanisms of repetition priming in occipitotemporal cortex: spatiotemporal evidence from functional magnetic resonance imaging and electroencephalography. *Journal of Neuroscience*, *25*(13), 3414–3422.

Gagl, B., & Gregorova, K. (2023). Investigating lexical categorization in visual word recognition based on a joint diagnostic and training approach for language learners.

Gagl, B., Richlan, F., Ludersdorfer, P., Sassenhagen, J., Eisenhauer, S., Gregorova, K., & Fiebach, C. J. (2022). The lexical categorization model: A computational model of left ventral occipito-temporal cortex activation in visual word recognition. *Plos Computational Biology*, *18*(6), e1009995.

Gagl, B., Sassenhagen, J., Haan, S., Gregorova, K., Richlan, F., & Fiebach, C. J. (2020). An orthographic prediction error as the basis for efficient visual word recognition. *NeuroImage*, *214*, 116727. doi: https://doi.org/10.1016/j.neuroimage.2020.116727

Gagl, B., Weyers, I., & Mueller, J. L. (2021). Speechless reader model: A neurocognitive model for human reading reveals cognitive underpinnings of baboon lexical decision behavior. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 43).

Geirhos, R., Janssen, D. H., Schütt, H. H., Rauber, J., Bethge, M., & Wichmann, F. A. (2017). Comparing deep neural networks against humans: object recognition when the signal gets weaker. *arXiv preprint arXiv:1706.06969*.

Guest, O., & Martin, A. E. (2021). How computational modeling can force theory building in psychological science. *Perspectives on Psychological Science*, *16*(4), 789–802.

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), e2104779118.

Hawelka, S., Schuster, S., Gagl, B., & Hutzler, F. (2015). On forward inferences of fast and slow readers. an eye movement study. *Scientific reports*, *5*(1), 1–8.

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the ieee conference on computer vision and pattern recognition* (pp. 770–778).

Heilbron, M., Richter, D., Ekman, M., Hagoort, P., & de Lange, F. P. (2020, jan). Word contexts enhance the neural representation of individual letters in early visual cortex. *Nature Communications*, *11*(1). doi: 10.1038/s41467-019-13996-4

Hofmann, M. J., Remus, S., Biemann, C., Radach, R., & Kuchinke, L. (2022). Language models explain word reading times better than empirical predictability. *Frontiers in Artificial Intelligence*, *4*, 214.

Huang, J., & Ling, C. X. (2005). Using auc and accuracy in evaluating learning algorithms. *IEEE Transactions on knowledge and Data Engineering*, *17*(3), 299–310.

Krizhevsky, A., Sutskever, I., & Hinton, G. E. (2012). Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, *25*.

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*, 621–647.

Kutas, M., & Hillyard, S. A. (1980). Reading senseless sentences: Brain potentials reflect semantic incongruity. *Science*, *207*(4427), 203–205.

Laurent, C., Pereyra, G., Brakel, P., Zhang, Y., & Bengio, Y. (2016). Batch normalized recurrent neural networks. In *2016 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 2657–2661).

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436–444.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L. D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural computation*, *1*(4), 541–551.

Lindsay, G. W. (2021). Convolutional neural networks as a model of the visual system: Past, present, and future. *Journal of cognitive neuroscience*, *33*(10), 2017–2031.

Lu, Z., Sindhwani, V., & Sainath, T. N. (2016). Learning compact recurrent neural networks. In *2016 ieee international conference on acoustics, speech and signal processing (icassp)* (pp. 5960–5964).

Ma, W. J., & Peters, B. (2020). A neural network walks into a lab: towards using deep nets as models for human behavior. *arXiv preprint arXiv:2005.02181*.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of eeg-and meg-data. *Journal of neuroscience methods*, *164*(1), 177–190.

Maurer, U., Zevin, J. D., & McCandliss, B. D. (2008). Left-lateralized n170 effects of visual expertise in reading: evidence from japanese syllabic and logographic scripts. *Journal of Cognitive Neuroscience*, *20*(10), 1878–1891.

Nieuwland, M. S. (2019). Do 'early'brain responses reveal word form prediction during language comprehension? a critical review. *Neuroscience & Biobehavioral Reviews*, *96*, 367–400.

Norris, D., & Kinoshita, S. (2012). Reading through a noisy channel: why there's nothing special about the perception of orthography. *Psychological review*, *119*(3), 517.

Pickering, M. J., & Gambi, C. (2018). Predicting while comprehending language: A theory and review. *Psychological bulletin*, *144*(10), 1002.

Price, C. J., & Devlin, J. T. (2011). The interactive account of ventral occipitotemporal contributions to reading. *Trends in cognitive sciences*, *15*(6), 246–253.

Qin, H., Gong, R., Liu, X., Bai, X., Song, J., & Sebe, N. (2020). Binary neural networks: A survey. *Pattern Recognition*, *105*, 107281.

Rabovsky, M., Hansen, S. S., & McClelland, J. L. (2018). Modelling the n400 brain potential as change in a probabilistic representation of meaning. *Nature Human Behaviour*, *2*(9), 693–705.

Rao, R. P., & Ballard, D. H. (1999). Predictive coding in the visual cortex: a functional interpretation of some extra-classical receptive-field effects. *Nature neuroscience*, *2*(1), 79–87.

Samek, W., Montavon, G., Lapuschkin, S., Anders, C. J., & Müller, K.-R. (2021). Explaining deep neural networks and beyond: A review of methods and applications. *Proceedings of the IEEE*, *109*(3), 247–278.

Santurkar, S., Tsipras, D., Ilyas, A., & Madry, A. (2018). How does batch normalization help optimization? *Advances in neural information processing systems*, *31*.

Saszik, S., & DeVries, S. H. (2012). A mammalian retinal bipolar cell uses both graded changes in membrane voltage and all-or-nothing na+ spikes to encode light. *Journal of Neuroscience*, *32*(1), 297–307.

Sexton, N. J., & Love, B. C. (2022). Reassessing hierarchical correspondences between brain and deep networks through direct interface. *Science Advances*, *8*(28), eabm2219.

Simon, G., Petit, L., Bernard, C., & Rebaï, M. (2007). N170 erps could represent a logographic processing strategy in visual word recognition. *Behavioral and Brain Functions*, *3*(1), 1–11.

Solomyak, O., & Marantz, A. (2009). Lexical access in early stages of visual word processing: A single-trial correlational meg study of heteronym recognition. *Brain and language*, *108*(3), 191–196.

Taylor, J., 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.

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.

Twomey, T., Duncan, K. J. K., Price, C. J., & Devlin, J. T. (2011). Top-down modulation of ventral occipito-temporal responses during visual word recognition. *Neuroimage*, *55*(3), 1242–1251.

van Rooij, I., Guest, O., Adolfi, F. G., de Haan, R., Kolokolova, A., & Rich, P. (2023). Reclaiming ai as a theoretical tool for cognitive science. *ArXiv preprint*.

Yan, C., de Lange, F. P., & Richter, D. (2023). Conceptual associations generate sensory predictions. *Journal of Neuroscience*, *43*(20), 3733–3742.

Yin, D., Biscione, V., & Bowers, J. S. (2023). Convolutional neural networks trained to identify words provide a surprisingly good account of visual form priming effects. *Computational Brain & Behavior*, 1–16.

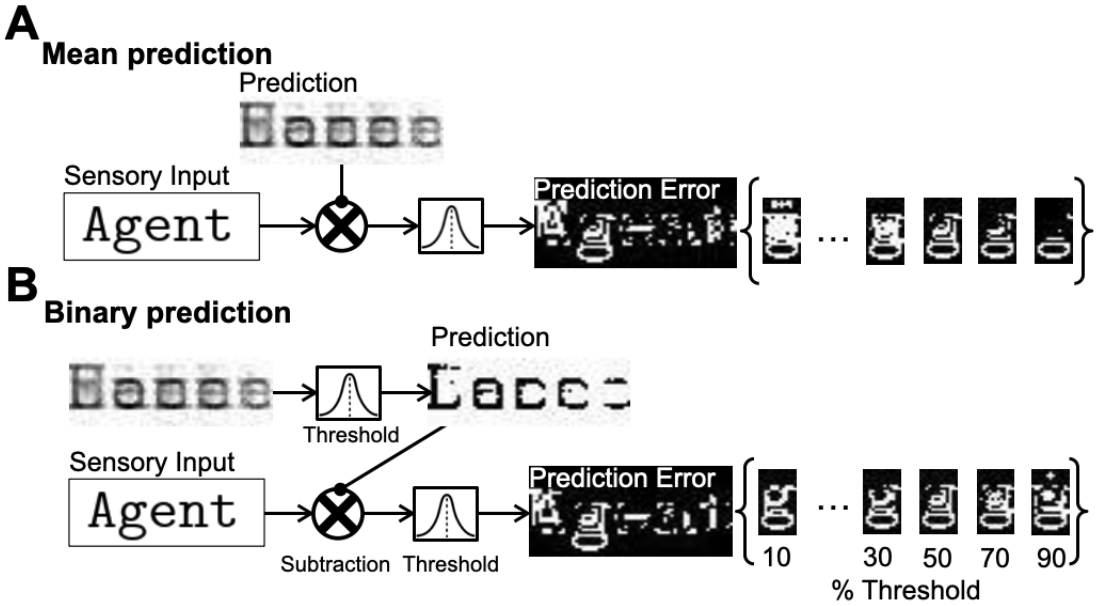


Figure 1: Schematic depiction of the new PEMoR implementations. (A) Structure assuming multiple prediction units with a binary output. Integrating multiple binary prediction units results in a graded prediction, which we approximate by a pixel-by-pixel mean of the pixel values from all words in the lexicon (see Gagl et al., 2020). (B) The structure assumes a single prediction unit that has a binary output. We implemented both binary prediction and prediction error units with the same threshold. We show orthographic prediction error examples for both assumptions for the letter "g" based on multiple thresholds.

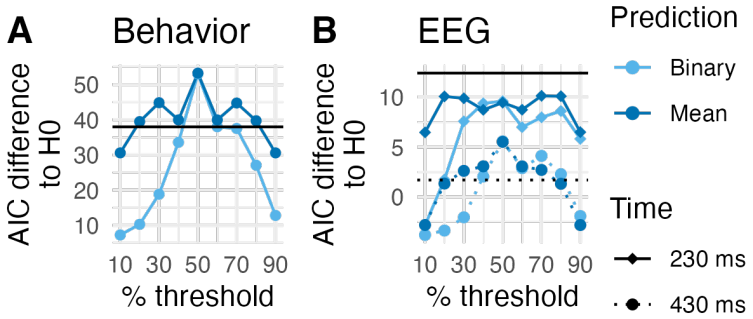


Figure 2: Model comparisons results based on (A) lexical decision response times for the original oPE (solid black line) and new oPE implementations, including all thresholds and both prediction assumptions. All AICs presented here are against a version of the linear mixed model without an oPE predictor (Null hypothesis, H0). (B) Same but for EEG amplitudes at 230 ms and 430 ms. The AIC difference for the original oPE is presented as solid (230 ms) and dotted black line (430 ms).

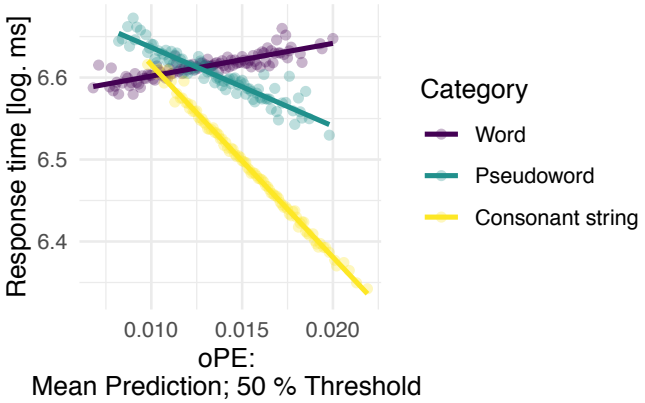


Figure 3: Orthographic prediction error (oPE) effect on Lexical (i.e., word/non-word) decision task behavior based on the oPE with a mean prediction and a 50% threshold for the prediction error representation (German nouns, five letters length). Purple lines show the effects for words, petrol lines for pseudowords (pronounceable non-words), and yellow lines for consonant strings (unpronounceable non-words). Dots represent mean reaction time estimates across all participants, separated into bins of oPE and stimulus category after excluding confounding effects.

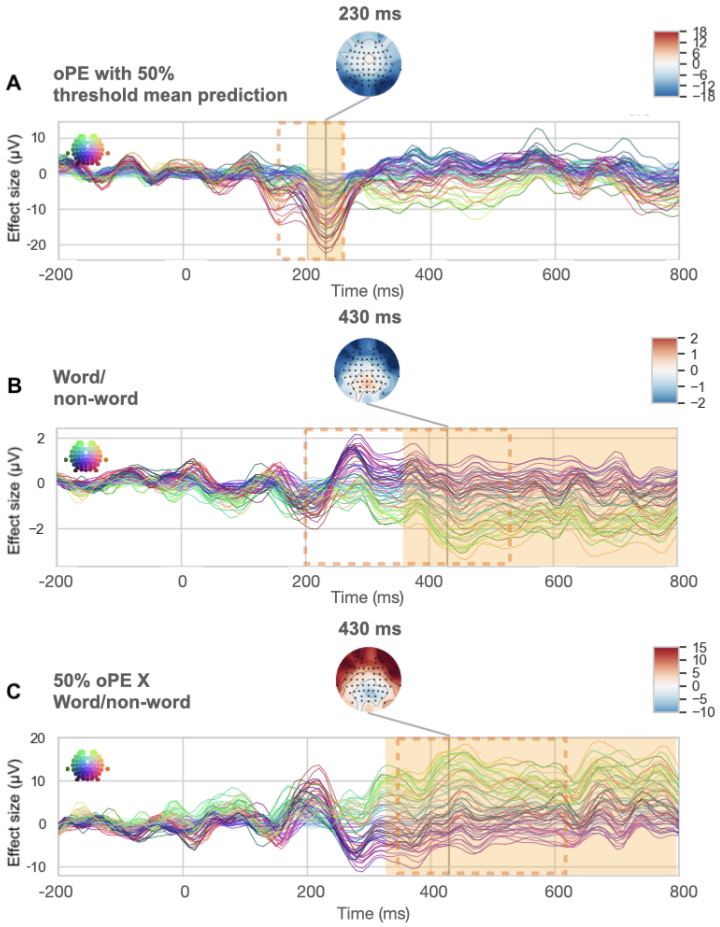


Figure 4: EEG results for lexical decision task of 200 words and 200 non-words (100 pseudowords, 100 consonant strings): Timing of 50% threshold orthographic prediction error effects. Effect sizes from multiple regression ERPs are presented as time courses for each sensor and time-point with orange shadow areas marking time windows with significant activation clusters (dashed rectangles mark the significant activation time windows from the multiple regression analysis Fig. 5sis for original oPE). ERP results are shown for (A) the main effect of the 50 % oPE, (B) the word/non-word effect, and (C) the 50% oPE by word/non-word effects.

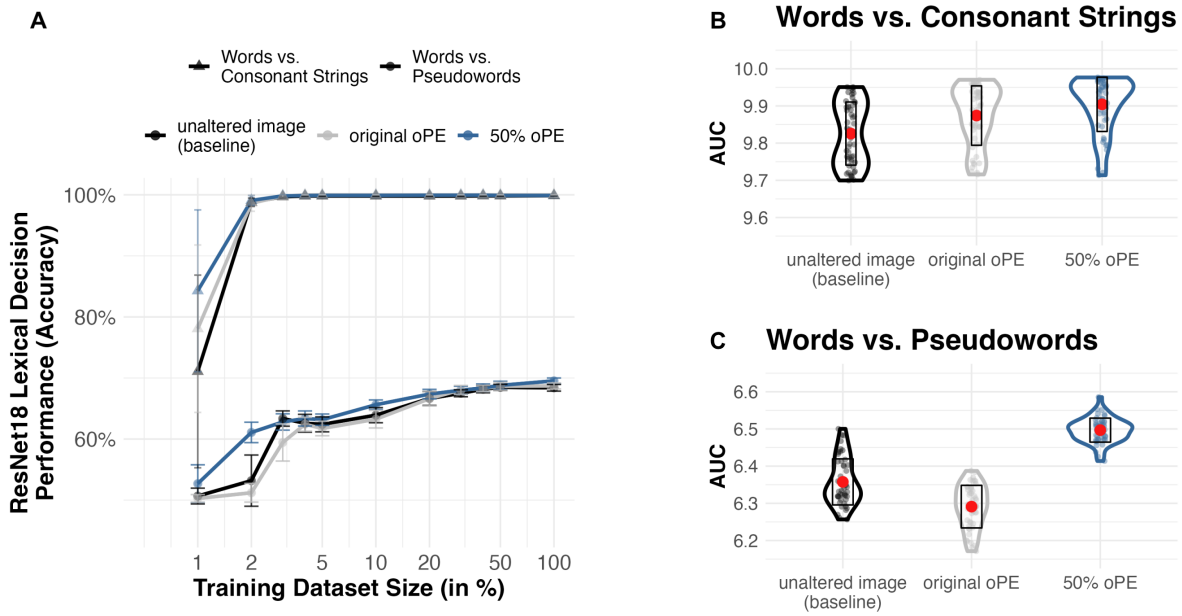


Figure 5: ResNet18 Lexical decision performance (Accuracy) for 50 % oPE, original oPE, and standardized letter string images as input. (A) Lexical decision accuracy for the test environment over a percentage of training data (i.e., 50 simulations per point and condition), including the standard deviation, separated for two types of non-words. Area size under each simulation of ResNet18 performance for 50 % oPE, original oPE, and standardized letter string images as input, separated for Words vs. Cosonant Strings (B) and Words vs. Pseudowords tasks. Each dot in the violin plot represents one simulation per condition from different percentages of training data. Under each task, the violin plot of each oPE model contains 50 simulations.

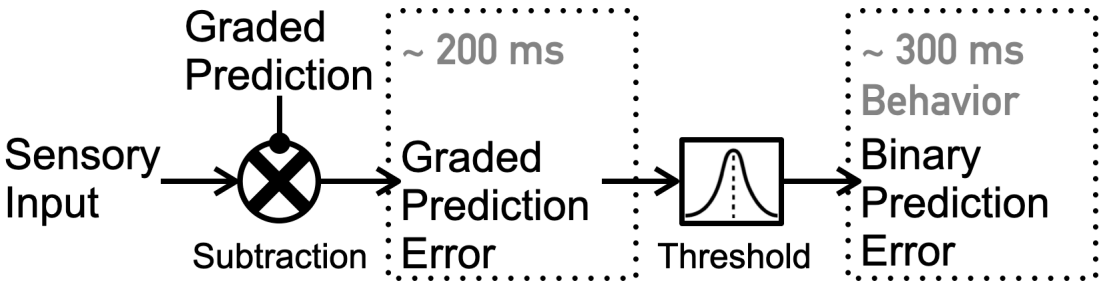


Figure 6: Schematic summary of results. The presented findings suggest efficient reading involves subtracting a graded prediction to generate a graded prediction error representation utilized in pre-lexical processing around 200 ms after stimulus onset. Next, we find evidence for thresholding the graded prediction error to increase the precision of the prediction error representation in processes associated with the extraction of word meaning.