The Words Matter in Early Reading Acquisition

Matthew J. Cooper Borkenhagen[1, 2], Devin M. Kearns[3], & Christopher R. Cox[4]

[1] Florida State University
[2] The Florida Center for Reading Research
[3] North Carolina State University
[4] Louisiana State University

Author Note

Correspondence concerning this article should be addressed to Matthew J. Cooper Borkenhagen, 1114 W. Call Street, Tallahassee, FL 32306. E-mail: mcooperborkenhagen@gmail.com

Abstract

...

Keywords: Reading development · connectionism

Word count: – words

The Words Matter in Early Reading Acquisition

Introduction

On the way to developing reading skill children encounter many different words. Each encounter brings with it exposure to properties of print that provide the basis for learning and generalization, where early in development the critical properties primarily concern letters and sounds and their patterns of co-occurrence across words (Castles, Rastle, & Nation, 2019; Seidenberg & MacDonald, 2019). Due to a variety of factors children vary in their word recognition skills. These include genetic factors, disability, instruction, and exposure to print, among others (Connor, Morrison, & Katch, 2004; Cunningham & Stanovich, 1993; Cunningham & Stanovich, 2013; Gayan & Olson, 2003; Peterson & Pennington, 2015). It is thought that an additional factor in outcomes originates in the set of printed words to which the child is exposed in their learning environment (e.g., through reading aloud printed words independently or with others). Indirect evidence for this can be found in experimental studies in highly controlled training environments: Children perform differently on words with different properties (e.g., length, regularity, consistency; Seidenberg et al., 1984) and differ in their choice of pronunciation for ambiguous test items (e.g., zead) after rehearsal on different practice items (e.g., heat and zeal or bread and head; Apfelbaum, Hazeltine, & McMurray, 2013). Learning studies have shown that even with limited exposure (even in absence of instruction) children are able to bootstrap their existing knowledge of letters and sounds to retain information about the identities of specific novel words to which they were previously exposed (Cunningham, Perry, Stanovich, & Share, 2002; Nation, Angell, & Castles, 2007). However, in studies that involve learning, less consideration is given to the variability in children's performance vis-a-vis the exposure to words over a longer developmental span (i.e., over the course of early reading acquisition) and in generalization performance. In educational practice, the words children encounter are primarily selected based on their teachers' implicit theories about how children learn. This often involves the assumption that children need to read most words quickly and accurately and can self-teach unfamiliar ones. In very early reading acquisition in an alphabetic language (e.g., English, Spanish, Dutch), one theory is that children achieve automaticity fastest through direct teaching on a large number of letter-sound patterns apparent in the language (e.g., the sound of ai in main). This is far less straightforward in writing systems like English where the mappings between letters and phonemes are less structured, a property known as quasiregularity (see Seidenberg, Cooper Borkenhagen, & Kearns, 2020 for a discussion). In quasiregular writing systems the ensembles of experiences with printed words are particularly important in early development given the lack of generativity of rules that relate individual letters and phonemes, though we know little about how learning experiences at the level of word ensemble relate to outcomes in word recognition skills (Compton, Miller, Elleman & Steacy, 2014). As a proxy for learning about the properties of words that support learning and generalization across ensembles of words, children are typically taught letter-sound patterns prioritized by their frequency in the language alongside being provided texts containing many instances of these patterns (i.e., phonics instruction). A common alternative teaching approach to phonics focuses on sight reading. In this approach, children are exposed to the most common words in the language and learn to pronounce them using methods that leverage rote memorization rather than emphasizing the letter-sound patterns they contain. Importantly, these (and other) instructional theories employ different ideas about the types of experiences young readers should have with printed words on their way to skilled reading. Is it important to organize experiences in a way that privileges some words over others? This question motivates the current study. We consider whether word-recognition performance varies depending on the particular words students learn, i.e., the ensemble of words that comprises the child's print learning environment.

Difficult Phenomena to Study

Answering this question has consequences for theories of reading acquisition and approaches to early reading instruction. Unfortunately, measuring the relative impacts of different ensembles of printed words is difficult due to the diffuse nature of early print experiences. Learning to read words aloud happens gradually over many learning trials that are distributed across many experiences with different words, and over wide developmental timescales. This gradual process is difficult to capture with traditional experimentation given the challenges associated with controlling the learning environment for developing children at a large scale. Learning studies like those mentioned previously are designed to examine learning at a causal level using narrowly constructed learning experiments in a laboratory setting where factors can be carefully controlled (typically using nonwords to avoid confounds of familiarity with stimuli). Additionally, typical behavioral assessments are not capable of measuring subtle differences in word reading performance (e.g., whether a given pronunciation includes minor imprecisions) or variation beyond the few items they include.

The Connectionist Approach

One way to assess acquisition and performance of this kind is to use simulations. Connectionism offers a robust set of tools to investigate these learning questions given the established success of connectionist models of human cognition (McClelland, Botvinick, Noelle, Plaut, Rogers, Seidenberg, & Smith, 2011). The paradigm has been used to model learning to read (Plaut, 2005; Seidenberg, Farry-Thorn, & Zevin, 2022); simulated learners have similar levels of performance accuracy and response time as humans on word-reading measures. An important advantage (especially related to the current project) is the large scale at which connectionist models can be deployed in simulating environment effects on learning (see Cox, Cooper Borkenhagen, & Seidenberg, 2019 for an example). In the approach taken here, a series of models that simulate early print learning (naming a word based on its printed form) represents a large sample of learners, each trained with a different ensemble of words, each of which comprises 300 words randomly selected from a text corpus. Given that the models differ only with respect to their learning environment (i.e., the ensemble on which it is trained), differences in learning outcomes can be causally attributed to variation in the environment. Thus, observed differences in model performance after training provide evidence that the specific words encountered influence learning outcomes. This approach also allows us to explore the degree to which reading experience matters relative to the characteristics of the learner, that is, capabilities of the learner that will affect their reading success. For example, Monaghan and Wollams (2017) simulated dyslexia in a reading model by adding noise to the part of the model responsible for pronunciation of the printed form. Patterson, Seidenberg, & McClelland (1989) created a model of dyslexia by reducing the system's capacity by removing hidden units from the network, which resulted in a pattern similar to what is seen in children with a learning difficulty. In addition to studying the effects of print environment on outcomes, our method here involves manipulating models in a similar way to Patterson et al. (1989) by manipulating the number of hidden units in order to introduce variation in representational capacity. Measuring the relative impact of words on later reading performance is challenging when using real-world samples because human learners vary in countless ways, both observable and unobservable, which can confound analyses. While statistical methods, such as regression, can control for some of these factors in examining human performance, these methods are limited in their ability to account for all sources of variability. This makes it difficult to isolate the specific influence of the words themselves among other covarying factors. The simulation approach used here addresses this challenge by enabling precise experimental control around the capacities of the (simulated) learners and the learning environment to which learners are exposed. It allows us to (a) examine the performance of an identical learner across different word ensembles and (b) compare learners with systematically varied characteristics (e.g., representational capacity) on the same word ensemble. In the experiments reported here, we do both simultaneously in order to examine the extent to which the words matter for early print development outcomes for learners that vary in terms of their ability to represent print-speech mappings.

Questions

Our general question is whether the different ensembles of words learned during early reading acquisition contributed to different word reading outcomes. We investigate this with three specific questions. First, to what extent is there variation in learning as a function of exposure to ensembles of different training items? Second, to what extent is there variation in learning as evidenced in transfer effects to untaught words? Finally, is the effect of the training environment modulated by the capacity of the learner to represent the mappings between phonology and orthography? Regarding questions one and two, we anticipate that there will be wide variation in performance on the words learned, and that this variation would also be evidenced in generalization items at the end of learning. With regard to question three, variability in the ensemble effect based on learning capacity, in general we expected for models with reduced representational capacity to perform more poorly and with greater variability than their high-capacity counterparts. We expected this increase in variation because we anticipate that learners with diminished capacity to map print to speech would benefit more from ensembles that, in general, conferred better (more generalizable) print-speech structure.

Table 1
Descriptive Statistics for Words Included in Sampling Process

	Holdout Items				Full Dataset			
Variable	Mean	SD	Min	Max	Mean	SD	Min	Max
Letters	4.41	1.06	2.00	8.00	4.52	1.08	2.00	9.00
Phonemes	3.48	0.82	2.00	6.00	3.57	0.83	1.00	9.00
Frequency	8.76	2.15	3.87	16.19	8.88	2.08	0.00	16.96

Note. 300 holdout items were selected from 2,869 items in the full dataset. Frequency = log transformed frequency of words from Hyperspace Analogue of Language (Lund & Burgess, 1996).

Methods

Word Data

Monosyllabic words (N = 2,869) were collected from popular children's books, taken from the corpus developed in Lewis et al. (2021). All monosyllabic words from the Lewis et al. (2021) corpus comprised the pool of words from which training and generalization items were sampled. Learning was weighted by frequency such that during each learning trial the error signal associated with any word presented to the model was proportional to its frequency as established in the Hypertext Analog of Language corpus (Lund & Burgess, 1996; accessed through data from English Lexicon Project, see Balota et al., 2007)[1]. All words participating in the simulations were monosyllabic. Table 1 presents descriptive statistics for both the full dataset and the subset of the data that were only used to evaluate generalization (in a holdout set never used for model training). The values for the descriptive statistics did not show a significant difference between the holdout items and the others: letters, F(1, 2868) = 3.38, p = .07; phonemes, F(1, 2774) = 3.25, p = .07; frequency F(1, 2773) = 1.08, p = .30.

Training and Holdout Item Sets

Models learned 300 words across 50 learning trials (epochs) and were all tested on the same independent set of 300 holdout words in order to examine generalization. This

[1] See appendix for a technical description of how the scaling is accomplished.

number of epochs was chosen because testing different configurations revealed this to be a reasonable number of weight updates to achieve asymptotic performance across a large number of pilot runs and to do so in a way that could be accomplished on a short timeline (in terms of learning). Other values are also possible but yielded similar results during a pilot phase. The testing procedure is described more fully below in the “Testing” section.

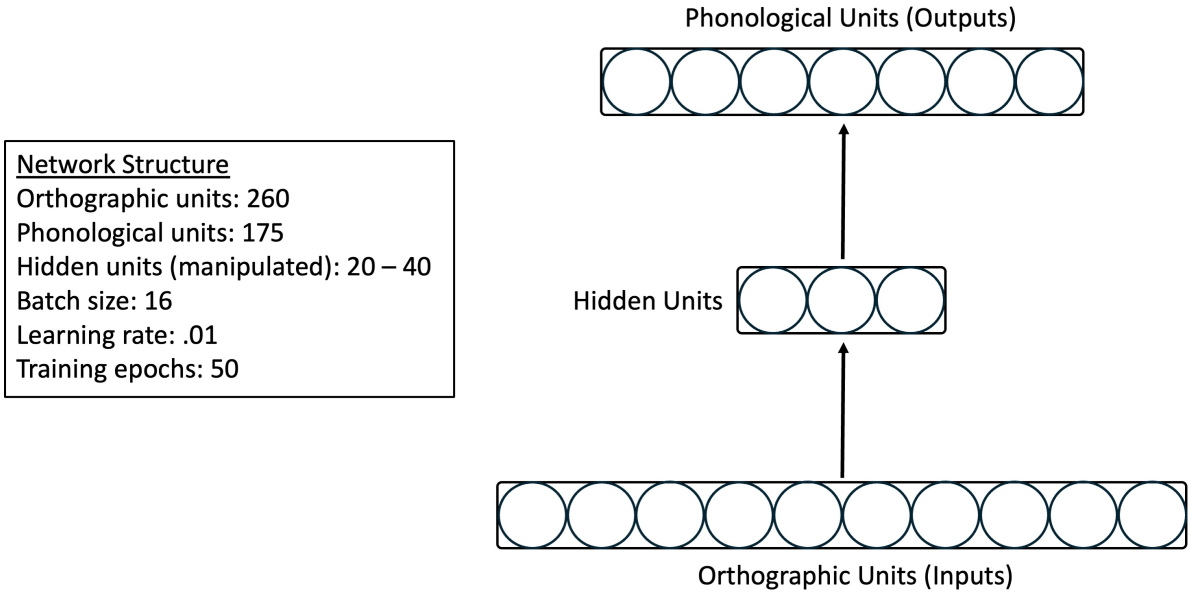
Modeling Procedure

Computational models that map print to speech were used, following the procedures in other simulation work (Brown et al., 2015; Cox et al., 2019; Harm & Seidenberg, 1999; Seidenberg & McClelland, 1989). The model architecture used here was a feedforward network with 260 orthographic input units, 175 phonological output units, and an experimentally manipulated number of hidden units (see description of representational capacity below). Each orthographic unit has a directed, weighted connection to every hidden unit, and each hidden unit has a directed, weighted connection to every phonological unit. Every hidden and phonological unit also receives weighted input from a special unit that always has activation equal to 1 (i.e., a bias unit), which serves a similar function as the intercept term in a linear regression. The networks were all trained in ways that were consistent with other connectionist implementations of feedforward reading models (Seidenberg & McClelland, 1989; Plaut et al., 1996). During learning weights were updated using gradient descent through the use of backpropagation. Error was calculated using binary cross entropy as the loss function (see Plaut et al., 1996 for similar simulations). Except for the number of hidden units, all other hyperparameters were decided prior to experimentation and held constant for all models. Notably, we used a batch size of 16, a learning rate of .01, and 50 training epochs. Models were implemented in Python 3.11.5 using Keras 2.12.0 with Tensorflow 2.12.0 as the backend.

Input and output patterns were generated by concatenating representations for letters and phonemes, respectively. Words were vowel-centered such that leftmost vowel appeared in the 4th slot for both orthography and phonology. The methods for creating representations were taken from prior connectionist reading simulations (Brown et al, 2015; Cox et al., 2019; Harm & Seidenberg, 1999). First, each letter in a string is assigned to a sequential orthographic “slot”, such that the first vowel in each string occupies the same slot across words; this was the fourth slot for our set of words. For words where two vowel letters correspond to a single vowel phoneme, the orthographic slot following the first vowel letter is occupied by the second vowel letter. For all other words, this slot is left blank, so that every slot handles only consonants or only vowels. For example, “eat” was represented “______eat_**” and “at” was represented “**_*a*_*t*___”[2]. Representations are configured in this way such that all orthographic inputs have the same number of slots (10 total). Padding representations in this way ensures that all representations are the same length and “vowel centered”. The same padding logic is followed to assign phonemes to slots; the vowel occupied the fourth slot and each representation has seven slots. The value in each slot (including the “empty” value) is translated into binary vectors that the model can utilize as either an input or output (in these simulations the input is always orthography and the output is always phonology). Each orthographic slot is replaced with a 26-element vector that orthogonally codes each of the 26 letters of the English alphabet (i.e., “a” has 1 in the first position and 0 elsewhere, “b” has 1 in the second position and 0 elsewhere, etc.). Empty slots are all-zero vectors of length 26. Likewise, each phonological slot is replaced by a 33-element binary vector. However, rather than being orthogonal like the letter representations, phonemes are coded as distributed representations (see Harm & Seidenberg, 1999), where each position in the vector corresponds to distinct phonological features. Distributed representations for phonemes were used to represent each phoneme for a word situated in a seven slot output representation. Features were defined by standard distinctive features used to uniquely identify the sounds of English (Chomsky & Halle, 1968; see Harm & Seidenberg, 1999 for a description of the features used).

Finally, the vectors within each slot are concatenated to form a single pair of orthographic and phonological vectors for each word. Orthographic and phonological vectors are then arranged into separate matrices in order to be utilized during simulations. Empty phonological slots are filled with all-zero vectors of length 33. A depiction of the network is shown in Figure 1.

[2] The notation here includes empty slots as “_”, which in the implemented model are represented as vectors where all features are off (i.e., are zeros).



Network Structure
Orthographic units: 260
Phonological units: 175
Hidden units (manipulated): 20 – 40
Batch size: 16
Learning rate: .01
Training epochs: 50

Representational Capacity (i.e., Hidden Units) The representational capacity of the model was manipulated by changing the number of hidden units in the network. This was done in such a way that each ensemble in one level of hidden unit has an identical "twin" in each other level such that the only thing that differed across them was the number of hidden units. This manipulation allowed us to test whether the effect of word set (the ensemble effect) differed based on characteristics of different models (the learner effect). The manipulation of hidden units is associated with variation found across readers given that hidden units provide the underlying capacity with which print and speech can be integrated in the model (Seidenberg & McClelland, 1989). Models with very few hidden units fail to encode generalizable structure in the training corpus. This is particularly problematic in writing systems like English where letters and sound map together in quasiregular ways (Seidenberg et al., 2020).

Number of Training Runs

The words selected for training and the representational capacity of each model were both manipulated. At each level of three capacity conditions (ranging from 20 to 40 hidden units) 10,000 models were run (N = 30,000). Simulations were conducted such that for each of the 10,000 models within each level of representational capacity there was a "twin" model in each of the other levels. Each set of twins differed only in terms of the representational capacity they possessed, and were matched in terms of the ensemble of words and all other hyperparameters of the simulation other than hidden units (i.e., learning rate, training epochs, batch size). The capacity conditions were static across models but training items were sampled randomly from the 2,869 words in the full monosyllabic word corpus. For each training ensemble a random set of 300 words was drawn from the corpus. No two ensembles were identical. See supplement for identities of words included in the sampling process. Words were presented in the same order throughout training for all models.

Testing

Error Metric. A primary objective of learning is a high level of achievement of the subject material. In order to assess each model's learning at the end of the training period we analyze the error associated with each model's performance on the training and holdout sets (300 words each) using mean squared error. This measure of error is common among simulation studies using connectionist models of reading because of the direct way in which error reflects processing difficulties associated with the model's ability to produce a phonological output from orthographic input (e.g., Chang, Monaghan, & Welbourne, 2019; see also Seidenberg & McClelland, 1989, where summed squared error was used). Mean squared error is calculated over output phonological features relative to the features for the word's target. The formula for mean squared error is shown in (1). Here, n is the number of output units ($n = 175$), y_i is the output of the ith unit, and $\hat{y}_i$ is the target state for that same (ith) unit.

$$\text{Mean Squared Error} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \quad (1)$$

In the results section mean squared error is reported in its standardized form (standardized using all the training and test data) in order to represent variation in performance in a more concrete way (i.e. in standard deviation units). Raw mean squared error is also provided. Given our specific interest in the effect of word ensemble during learning and generalization, error for each model is averaged across all of its training words (when measuring training outcomes) and test words (when measuring generalization).

Table 2

Hypothetical Distribution of Pronunciation Errors Relative to the Target Word "breathed"

Spelling	Pronunciation	Diff. features	Diff. MSE Raw	Diff. MSE Z-score	Accuracy
breathed	b iðd	–	0.00	0.00	1.00
breathg	b iðg	1	0.00	0.01	1.00
breazg	b izg	2	0.01	0.02	0.99
broozg	b uzg	3	0.01	0.03	0.99
bruised	b uzd	4	0.02	0.04	0.98

Note. The difference between the pronunciation for "breathed" and "bruised" is four features (also sometimes referred to as "units"). Pronunciations are transcribed using the International Phonetic Alphabet. The change in number of features (Diff. features) is relative to the top row ("breathed"). Spellings for nonwords are hypothetical given the pronunciation. The estimates of change in standardized MSE (Diff. MSE Standardized) are taken from standardized data for training and generalization. Accuracy is provided for reference, though results in the paper are explained in terms of mean squared error. Note that the differences across successive rows are defined by differences in individual features, though they also result in differences in phonemes. It is not always the case that a difference of one feature results in two different legal phonemes in English.

Understanding the Error Metric

For additional context when interpreting error results we will say a few things about the relevance of error as a measure of performance in the models reported here (and other connectionist models of reading). The output for each word is a sequence of 175 features, each representing a binary phonological feature (e.g., plosive, nasal, bilabial, etc.) within a phonological slot associated with pronunciation. We analyze results in terms of performance (error) in the output features because it allows for continuously graded data on pronunciation. Table 2 provides examples of minimal pairs of alternative pronunciations and the corresponding impact on error (raw and standardized) on output features.

References

PLACE REFERENCES HERE.

Data Availability Statement