Efficiency of Learning in Experience-Limited Domains: Generalization Beyond the WUG Test

Christopher R. Cox (chriscox@lsu.edu)
Department of Psychology, Louisiana State University
1005 Field House Dr, Baton Rouge, LA 70802 USA

Matthew Cooper Borkenhagen and **Mark S. Seidenberg**
Department of Psychology, University of Wisconsin-Madison
1202 W. Johnson Street, Madison, WI 53706 USA

Abstract

Learning to read English requires learning the complex statistical dependencies between orthography and phonology. Previous research has focused on how these statistics are learned in neural network models provided with as much training as needed. Children, however, are expected to acquire this knowledge in a few years of school with only limited instruction. We examined how these mappings can be learned efficiently, defined by tradeoffs between the number of words that are explicitly trained and the number that are correct by generalization. A million models were trained, varying the sizes of randomly-selected training sets. For a target corpus of about 3000 words, training sets of 200–300 words were most efficient, producing generalization to as many as 1800 untrained words. Composition of the 300 word training sets also greatly affected generalization. The results suggest directions for designing curricula that promote efficient learning of complex material.

Keywords: reading; efficient learning; generalization; computational modeling; human and machine learning

Introduction

Generalization—the ability to apply existing knowledge to novel cases—is an important capacity observed, with varying complexity, in many species (Santolin & Saffran, 2018). Human generalization encompasses a broad range of behaviors, ranging from generalizations about the properties of three dimensional space to ones based on physical appearance. The behavioral and neurobiological bases of generalization are a focus of much research (e.g., Goldberg, 2009; Zhang, Bengio, Hardt, Recht, & Vinyals, 2016).

Generalization is especially important in language acquisition and learning to read. Children rapidly acquire knowledge that allows them to generalize beyond the limited sample of utterances they experience (Chomsky, 1965). The classic demonstration is the WUG Test (Berko, 1958). A child who has learned about plural formation can generalize to novel cases: one wug, two wugs. Similarly, a beginning reader who has learned correspondences between spelling and pronunciation can read aloud nonce words such as NUST and GLORP (Seidenberg & McClelland, 1989). Generalization has traditionally been taken as evidence for symbolic rules, but it is also observed in neural networks of varying complexity (Seidenberg & Plaut, 2014; LeCun, Bengio, & Hinton, 2015).

Our research examined generalization from a different perspective, efficiency of learning. Efficiency is a concern in real-world contexts in which, unlike most machine learning applications, learning opportunities are constrained. For example, children's vocabulary development depends on their time- and context-limited exposure to spoken language, which varies considerably (Hart & Risley, 1995; Gilkerson et al., 2017). The resulting differences in vocabulary size and quality have an enormous impact on learning to read and other aspects of schooling (Seidenberg, 2017). Knowledge gaps cannot be closed solely through explicit instruction because there isn't sufficient classroom time. The same holds for learning mappings between written and spoken language. Instruction ("phonics") is helpful, but only a small subset of patterns can be taught. In these and other knowledge domains, children learn from relatively limited data and generalization is paramount.

In the classic WUG test generalization is assessed by performance on nonce forms or, in machine learning, withheld words. The exact composition of the examples that support generalization is not the focus of attention, but is critical in experience-limited domains. We therefore re-formulated the generalization question as follows, using spelling-sound knowledge as a test case:

- Children need to acquire the ability to generate pronunciations for many written words (the target set);
- They are explicitly taught the correspondences between orthography and phonology for a much smaller subset of words (the training set);
- Generalization is assessed in terms of correct performance on untrained items from the target set, rather than nonce forms. This shifts the focus of generalization to acquiring real-world knowledge.

The research question is then how the size and composition of the training set affects generalization to untrained items. Learning is efficient if the ratio between the number of trained items and the number of generalization items is low. We examined efficiency of learning as a function of the size of the training set using simple, well-studied models of learning orthography-phonology correspondences (Seidenberg & McClelland, 1989; Harm & Seidenberg, 1999). We also examined how efficiency was affected by the composition of a training set of a given size. The results suggest that it may be possible to structure children's reading experiences in ways that promote more efficient learning.

Materials and Methods

Words

The simulations used a set of 2881 monosyllabic English words employed in previous research (Harm & Seidenberg, 1999). Word length ranged from 2–8 letters and 1–7 phonemes.

Model architecture

The model was a simple feedforward network with an input orthographic layer (102 units), an output phonological layer (66 units) and a single hidden layer (100 units). It was structured and trained in standard ways, with weights updated with gradient descent and backpropagation after accumulating cross-entropy error over all words in the training set.

Orthographic representations were generated as follows. Words were centered on the vowel (or the first vowel in a digraph), adding empty letters to the onset as necessary. If the first vowel was followed immediately by a consonant, an empty letter was also added between them, except in cases where the consonant is voiced as part of the vowel (e.g., the letter *w* in SAW). The letter *y* was treated as a consonant when it began a word and a vowel otherwise. Finally, empty letters were added to the end of each word, resulting in orthographic codes of uniform length (14 letters including empty ones).

Each letter was represented by one unit in a 26 element vector, with no units activated for the empty letter. The 14 vectors were concatenated to represent each word. To make these representations more concise, they were stacked to create a 2881×364 matrix, and all-zero columns were dropped, leaving 102 units.

Phonological word forms were represented using 41 phonemes (26 consonants, 15 vowels). They were aligned on the first vowel, adding empty phonemes at the beginning or end to produce phonological representations of equal length (10 phonemes including empty phonemes). Each phoneme was defined by 25 phonetic features (Harm & Seidenberg, 1999). The 10 phoneme by 25 feature vectors were condensed by eliminating nodes for unused features, resulting in an output layer with 66 features.

The model was implemented using scikit-learn in Python 3.6 using a multilayer perceptron, and training was executed in parallel using HTCondor (Thain, Tannenbaum, & Livny, 2005) and computational resources maintained by the Center for High Throughput Computing at UW Madison.

Model training

One million models were run, each using a set of words sampled randomly without replacement from the 2881 word target set. Training sets ranged from 100 to 1000 words in increments of 100, with an equal number of each size.

Each model was trained for 3000 weight updates with a constant learning rate (0.1). The model was exposed to the whole training set before each update. Each model was then tested on the untrained remainder of the target corpus to evaluate generalization. An output pattern was scored as correct if all unit activations were within 0.5 of their target state.

Model evaluation

Using all untrained words as the holdout set to evaluate generalization performance for each model means that the holdout set is not held constant. This is a deliberate design decision: when a word is explicitly trained, it no longer needs to be generalized to. Training on exceptional, irregular words may be the only way to accurately produce them—that explicit training not only develops the model to encode that orth-phon relationship, but also removes that exceptional word from the generalization set. On the other hand, this exceptional word may not teach the model anything generally useful. The give and take between what is in the training set or test set is central to the research question.

An alternative approach is possible, where a single test set is constructed a priori and used for all generalization. This has the advantage of serving as a true benchmark, but poses a critical challenge. It requires composing a representative test set that expresses all relational othographic and phonological structure. Our attempts at dimensionality reduction on the model representations that map between orthography and phonology for the full corpus indicate that 50 dimensions are necessary to express 80% of the variance in that structure. Sampling representatively from that high dimensional space would be necessary for constructing a useful benchmark test set. The problem of constructing this test set is the same as the problem of constructing a representative and efficient training set, and does not have a simple solution.

Results

Training set size and generalization

Figure 1A shows generalization to untrained items as a function of training set size. Smaller training sets afford more opportunities for generalization, but are less able to provide representative coverage of the corpus. Increasing the size of the training set produced diminishing generalization returns. Increasing training sets beyond 500 words did not yield greater

Size	Mean	(Ratio)	Max	(Ratio)
100	333	(3.33)	590	(5.90)
200	889	(4.45)	1252	(6.26)
300	1240	(4.13)	1546	(5.15)
400	1404	(3.51)	1634	(4.08)
500	1469	(2.94)	1668	(3.34)
600	1484	(2.47)	1654	(2.76)
700	1470	(2.10)	1618	(2.31)
800	1438	(1.80)	1566	(1.96)
900	1395	(1.55)	1510	(1.68)
1000	1344	(1.34)	1444	(1.44)

Table 1: Mean and maximum generalization performance over 100k models fit with each training set size. Ratios divide the previous descriptive statistic by the training set size.

generalization.

Figure 1B shows total number of words correct (trained and generalized). No model produced correct performance for all words. Some words were only learned if they were included in the training set; they were never produced correctly by generalization. These include words with highly atypical spellings and pronunciations such as SIXTH, DRAUGHT, SCHEME, COUPS, and JINX.

Figure 1C shows an index of *training set efficiency*, defined as the number of words correct by generalization divided by the number of words trained. Training sets with 100 words are less efficient than those with 200 words on average and in the limit, indicating that the larger set captures more of the structure relevant to untrained words. Training sets of 300 words are somewhat less efficient than those with 200, but after 300 words efficiency drops rapidly. Taking all three metrics into account, 300 words appears to be a sweet spot (see also Table 1).

Analyses of training environments containing 300 words show that they yielded reading vocabularies of 1540 words on average ($SD = 76.62$) and 1846 words at best (failing to decode 1035). Given that efficiency is a primary concern for early reading curricula, it is noteworthy that this is 75.5% of the largest reading vocabulary achieved with any training set (2444 words, achieved after training on 1000 words). Note that this 598 word increase required growing the training set by 700 words. If we subtract the training set from all reading vocabularies and just focus on words that were generalized to, the best model trained on 300 words (1546) achieves 92.7% of the maximum amount of generalization achieved with any training set (1668, achieved after training on 500 words).

These results indicate that nearly all systematic structure relating English orthography and phonology within our corpus of 2881 monosyllabic words can be learned from an appropriately constructed 300 word subset. It is possible to establish a reading vocabulary of over 1800 words based on explicit training on only 300 words, a 6-fold return on instructional investment. However, achieving this level of performance is highly dependent on the composition of the training set: the best and worst models trained with 300 words are separated in performance by over 600 accurate generalizations (min: 906; max: 1546). Thus, in future work it will be important to understand how properties of training sets are related to generalization.

What makes a word likely to be correct by generalization?

The rates at which individual words were correct by generalization across training sets varied greatly, forming a roughly bimodal distribution (Figure 2).

At one extreme are words that are correct by generalization with almost any random selection of training words; at the other are words that for which generalization is highly sensitive to training set composition. The former contain spelling patterns and orthography-phonology mappings that occur more often in this corpus; the latter words have less common patterns and more idiosyncratic mappings.

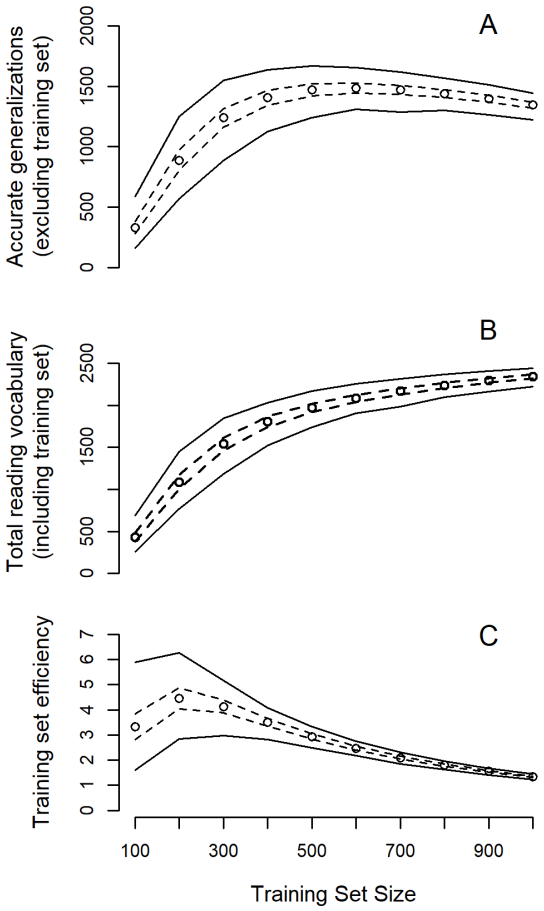

Figure 1: Reading vocabulary size and generalization ability for increasing training set sizes. A) The number of accurate generalization peaks at lower training set sizes and B) the rate of reading vocabulary growth slows. No model trained on a subset of words is capable of reading all words. C) The ratio of generalization performance and training set size, efficiency, is highest with training sets with 200–300 words. Dots indicate the mean; dotted lines are $\pm 1SD$; solid lines are minimum and maximum values.

Whether a word was likely to be generalized to was related to quantifiable measures of orthographic, phonological, and relational (mapping) typicality. We examined several lexical factors that have been employed in previous research:

- Word length: number of letters
- Orthographic neighborhood: number of words whose spelling differs from a word by a one letter substitution, deletion, or addition ($D_{Levenshtein} < 1$).
- Phonological neighborhood: number of words with the same rime (e.g., for "must", the "ust" words like "dust" and "lust").

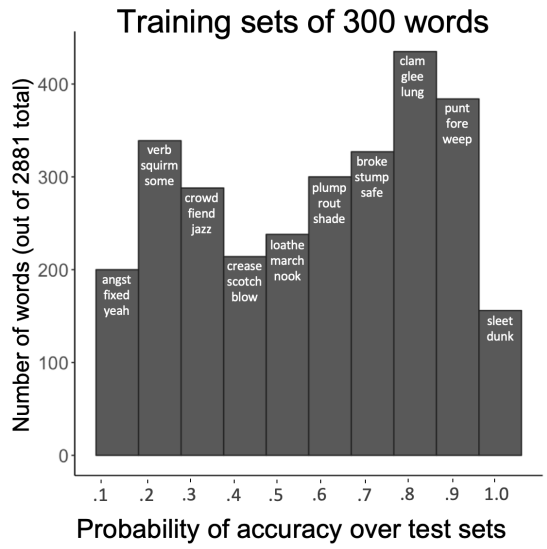

Figure 2: When aggregating over the 100k 300-word model training environments, each word occurs in many test sets. The proportion of times a word occurs in the test set and is accurately generalized to corresponds to how difficult that word is to learn. Representative words belonging to each bin are displayed.

- Consistency: the proportion of words with a given word body (the orthographic equivalent of the rime) and the same phonological rime (e.g., for GAVE, the proportion of -AVE words pronounced "ave"; (Plaut, McClelland, Seidenberg, & Patterson, 1996).

The correlations among these variables, and between these variables and the probability of accurate generalization, are reported in Table 2. The number of orthographic neighbors tends to decrease as word length increases ($r = -0.65$); a similar but weaker trend applies to the size of phonological neighborhoods ($r = -0.28$). This is representative of the English language in general. There is also a moderate relationship between neighborhood size across modalities, such that words that belong to large othographic neighborhoods are expected to belong to large phonological neighborhoods ($r = 0.35$). That this correlation is not higher demonstrates the asymmetry of structure across the modalities. The consistency of a word's pronunciation given its orthography, however, is uncorrelated with the modality-specific metrics. Words are more likely to be generalized to if they are short, belong to large phonological and (especially) orthographic neighborhoods, and have consistent pronunciation given their spelling (Table 2, bottom row).

Given the high correlations among variables, and to gain perspective on how jointly-predictive these factors are of the probability of accurate generalization, we regressed the probability of accuracy over test sets on all four variables in an additive linear model (no interaction terms). This simple model accounts for 39% of the variance in generalization accuracy. Of the variables we considered, the consistency metric accounted for the most unique variance ($\Delta R^2 = 0.15$), but orthographic neighborhood size was a close second ($\Delta R^2 = 0.13$). Once accounting for other variables, phonological neighborhood size and word length did not appreciably improve the model.

These results are broadly consistent with previous research. Effects of spelling-sound consistency have been observed in many behavioral studies of skilled and beginning readers (Jared, McRae, & Seidenberg, 1990), and simulated in earlier models that examined performance over the course of learning many words (Seidenberg & McClelland, 1989; Plaut et al., 1996). Our results suggest that factors that affected ease of learning in the earlier models also affect probability of generalization as studied in the present work.

Out of the variables we considered, phonological neighborhood size is the most studied in the context of word acquisition, where it is understood to influence the order in which words are acquired (Storkel, 2003). Orthographic neighborhood size is often studied in terms of performance, specifically visual word recognition and lexical access (Andrews, 1997). It is also negatively correlated with age of acquisition norms, which indicates that words with more dense orthographic neighborhoods tend to be learned earlier (Cameirão & Vicente, 2010). Words with consistent orthographic to phonology relationships are also processed more efficiently (Ziegler, Ferrand, & Montant, 2004).

	WL	ON	PN	Con.
Word Length	1.00			
Orth. Neighbors	-0.65	1.00		
Phon. Neighbors	-0.28	0.35	1.00	
Consistency	-0.03	-0.02	-0.02	1.00
P(accuracy)	-0.27	0.47	0.28	0.38

Table 2: Correlation among lexical measures. The bottom row reports the pairwise correlation of each variable with the probability of generalization accuracy for each word, defined as the number of times accurately generalized to divided by the number of test sets a word appeared in.

regressor	η_p^2	$\Delta\mathbf{R^2}$
Word length	0.01	0.00
Orth. Neighbors	0.17	0.13
Phon. Neighbors	0.03	0.02
Consistency	0.20	0.15

Table 3: Effect sizes for the regressors that account for variance in the probability of accurately generalizing each word. These effect size metrics are perspectives on the unique variance explained by each variable. Because of collinearity among the regressors, the sum of the ΔR^2 values will be less than total $R^2 = 0.39$.

What makes a good training set?

The word-level features reviewed above give some insight into which words will tend to be generalized to, and which will not, in the context of any given training set. The deeper question pertains to the qualities of the training set foster the most efficient generalization to untrained words in the language. One angle on this question is to consider that the word-level features are in fact reflective of how the word is situated relative to the broader linguistic environment. While we did not test this directly, it is plausible to assume that neighborhood size predicts how likely a word is to be generalized to. Good training sets are *representative* of the broader environment. If a neighborhood is split across training and test sets, the consequence is that the neighbors in the test set have representation within the training set. Given that we randomly split our corpus into training and test sets, there is no guarantee that neighborhoods are efficiently split in this way. However, words that belong to larger neighborhoods are more likely to be split across training and test sets by chance, so we might expect that training sets with larger orthographic and phonological neighborhoods on average will foster more generalization. It is clear that words with no orthographic neighbors ($n = 271$) are generalized to far less often (median probability 0.10) than words with at least one neighbor (median probability 0.56).

Such a crude metric, however, would be largely insensitive to the relative composition of the two sets. For instance, training sets that contain many words with large neighborhoods may simply contain all the words belonging to those large neighborhoods. Such a training set would be unrepresentative of the test set, and unlikely to foster generalization. What we would rather know is each word's neighborhood size relative to the number of its neighbors that also belong to the training set.

On the other hand, orthographic and phonological neighborhood structure is only helpful to the extent that they are aligned. An orthographic neighborhood populated with words with irregular and idiosyncratic pronunciations is not likely to foster generalization on a reading-aloud task. Thus, training sets that have a large and varied collection of words with consistent pronunciations may be expected to generalize well. While it is easy to determine the mean consistency of a training set, it is less clear how to account for the variability across consistent relationships and determine the representativeness of such relationships to the target environment.

We regressed the generalization performance of the 100,000 models trained on 300 word training sets on the mean word length, orthographic and phonological neighborhood sizes, and consistency over all 300 words in each set. The effect sizes are reported in Table 4. We see that, despite being a very crude measure, mean orth-phon consistency accounts for about 13.6% of the variance unexplained by the other variables, indicating that item level characteristics may provide insight on how to construct efficient training sets. However, the vast majority of variance remains unexplained and provides fertile ground for continued research.

regressor	η_p^2	$\Delta\mathbf{R^2}$
Word length	0.002	0.001
Orth. Neighbors	0.006	0.005
Phon. Neighbors	0.000	0.000
Consistency	0.137	0.136

Table 4: Effect sizes for the regressors used to account for variance in generalization accuracy over the 100,000 models fit to random 300 word training sets. Generalization was to all untrained words in the corpus. Because of collinearity among the regressors, the sum of the ΔR^2 values will be less than total $R^2 = 0.14$.

Discussion

We have established a computational procedure for investigating two aspects of generalization in learning basic reading skills: how many words need to be learned to generalize to real English words yet to be learned, and what aspects of reading vocabulary promote this transfer. Our findings indicate that while printed vocabulary continues to grow along with the number of words taught, the efficiency of learning does not grow along with it.

These findings are relevant to real-world learning conditions. As a human teacher grows the number of words they would like to teach, the amount of learning time needed grows along with it. Our findings suggest a trade-off where a smaller number of words could be taught, increasing efficiency of learning and teaching for sake of near-optimal generalization capacity. This has potentially important implications for reading education where there is a need to teach spelling-sound patterns (phonics) but only enough time to sample from the large set of patterns. Many educators oppose teaching phonics because it is seen as requiring "drill and kill" amounts of instruction and practice. This may be less of a concern if, as our results suggest, patterns can be selected in a way that maximizes generalization.

The problem of maximizing generalization with the smallest possible training set can be formalized as a *machine teaching* optimization problem (Zhu, 2015). We have drawn on this literature by manipulating the learning environment while holding the abilities of the learner constant, and then performing careful analyses of the outcomes to identify the factors that contribute to training the most proficient models. In doing so we have demonstrated systematic relationships between the composition of the training set and generalization performance that machine teachers may be able to discover and exploit.

These results are empirical; our next step will be to identify properties of words and word-sets responsible for better generalization both at the word- and set-level. As indicated in our regression model reported, item-wise measures of phonology, orthography, and especially orth-phon consistency account

for non-trivial amounts of generalization error. Next steps will be oriented towards accounting for more of the variance in generalization accuracy, and to scale up analyses to model-wise characteristics that promote generalization. It may also be possible to improve efficiency even further by using training sets attuned to children's vocabulary development, and by optimizing the sequence of learning experiences. Ultimately the aim is to discover the principle axes of the orth-phon mapping space, and exploit that structure in a theory-driven way to construct idealized training environments.

The reported models were trained on representations of the orthography with 14 "slots" for letters and tested on phonology with 10 "slots" for phonemes. This has consequences for learning that are artificial relative to how a child learns to decode orthography. Most salient is that each slot has an independent set of weights that project to the hidden layer. This means that what is learned about letters in one slot is not necessarily transferred to other slots—once the model has learned to pronounce the consonant *K* in the third slot, it will fail to generalize that knowledge when presented with a *K* in the fifth slot. This and other limitations of the slot based representation scheme contribute to our focus (and the focus of the modeling literature, generally) on monosyllabic words. Monosyllabic words are short and fairly consistent in length with a single vowel phoneme. After vowel-centering, the limits of using slots are effectively attenuated in the monosyllabic context, but it is not a solution that scales up. Models of reading that attempt to reflect more plausible visual processes and accommodate disyllabic words are needed. The slot-based approach may add some complexity to the decoding problem while simplifying the "visual" experience of our models.

Though preliminary, these simulations demonstrate that it is possible to be more efficient with curricula that attend to the number of words taught and the words that are prioritized in teaching.

References

Andrews, S. (1997). The effect of orthographic similarity on lexical retrieval: Resolving neighborhood conflicts. *Psychonomic Bulletin & Review*, *4*(4), 439–461. doi: 10.3758/bf03214334

Berko, J. (1958). The child's learning of english morphology. *Word*, *14*, 150 - 177.

Cameirão, M. L., & Vicente, S. G. (2010). Age-of-acquisition norms for a set of 1,749 portuguese words. *Behavior Research Methods*, *42*(2), 474–480. doi: 10.3758/brm.42.2.474

Chomsky, N. (1965). *Aspects of the theory of syntax.* Cambridge, M.I.T. Press.

Gilkerson, J., Richards, J. A., Warren, S. F., Montgomery, J. K., Greenwood, C. R., Oller, D. K., ... Paul, T. D. (2017). Mapping the early language environment using all-day recordings and automated analysis. *American Journal of Speech-Language Pathology*, *26*, 248–265.

Goldberg, A. E. (2009). The nature of generalization in language. *Cognitive Linguistics*, *20*(1), 93 - 127.

Harm, M. W., & Seidenberg, M. S. (1999). Phonology, reading acquisition, and dyslexia: insights from connectionist models. *Psychological review*, *106*, 491–528.

Hart, B., & Risley, T. R. (1995). *Meaningful differences in the everyday experience of young american children.* Baltimore, MD: Paul H. Brookes.

Jared, D., McRae, K., & Seidenberg, M. S. (1990). The basis of consistency effects in word naming. *Journal of Memory and Language*, *29*(6), 687–715. doi: 10.1016/0749-596x(90)90044-z

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436–444. doi: 10.1038/nature14539

Plaut, D. C., McClelland, J. L., Seidenberg, M. S., & Patterson, K. (1996). Understanding normal and impaired word reading: Computational principles in quasi-regular domains. *Psychological Review*, *103*(1), 56 - 115.

Santolin, C., & Saffran, J. R. (2018). Constraints on statistical learning across species. *Trends in Cognitive Sciences*, *22*(1), 52 - 63.

Seidenberg, M. S. (2017). *Language at the speed of sight: How we read, why so many can't, and what can be done about it.* New York : Basic Books.

Seidenberg, M. S., & McClelland, J. L. (1989). A distributed, developmental model of word recognition and naming. *Psychological Review*, *96*(4), 523 - 568.

Seidenberg, M. S., & Plaut, D. C. (2014). Quasiregularity and its discontents: The legacy of the past tense debate. *Cognitive Science*, *38*(6), 1190 - 1228.

Storkel, H. L. (2003). Learning new words II. *Journal of Speech, Language, and Hearing Research*, *46*(6), 1312–1323. doi: 10.1044/1092-4388(2003/102)

Thain, D., Tannenbaum, T., & Livny, M. (2005). Distributed computing in practice: the condor experience. *Concurrency and Computation-Practice and Experience*, *17*(2-4), 323 - 356.

Zhang, C., Bengio, S., Hardt, M., Recht, B., & Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *arXiv*.

Zhu, X. (2015). Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In *Proceedings of the twenty-ninth AAAI conference on artificial intelligence* (pp. 4083–4087). AAAI Press. Retrieved from `http://dl.acm.org/citation.cfm?id=2888116.2888288`

Ziegler, J. C., Ferrand, L., & Montant, M. (2004). Visual phonology: The effects of orthographic consistency on different auditory word recognition tasks. *Memory & Cognition*, *32*(5), 732–741. doi: 10.3758/bf03195863