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

Christopher R. Cox[1], Matthew Cooper Borkenhagen[2] and Mark S. Seidenberg[2]

1. Louisiana State University, 2. University of Wisconsin-Madison

Introduction

- **Generalization**—the ability to apply existing knowledge to novel cases—is essential for language acquisition and learning to read.
- Nonce words that do not occur in natural speech like NUST or GLORP can be read aloud [1].
- Children's vocabulary development depends on their exposure to spoken language, which varies considerably [2, 3] with enormous consequences for learning to read and other aspects of schooling.
- Knowledge gaps cannot be closed solely through explicit instruction because there isn't sufficient classroom time.
- **Our research examined the relation between generalization and efficiency of learning.**

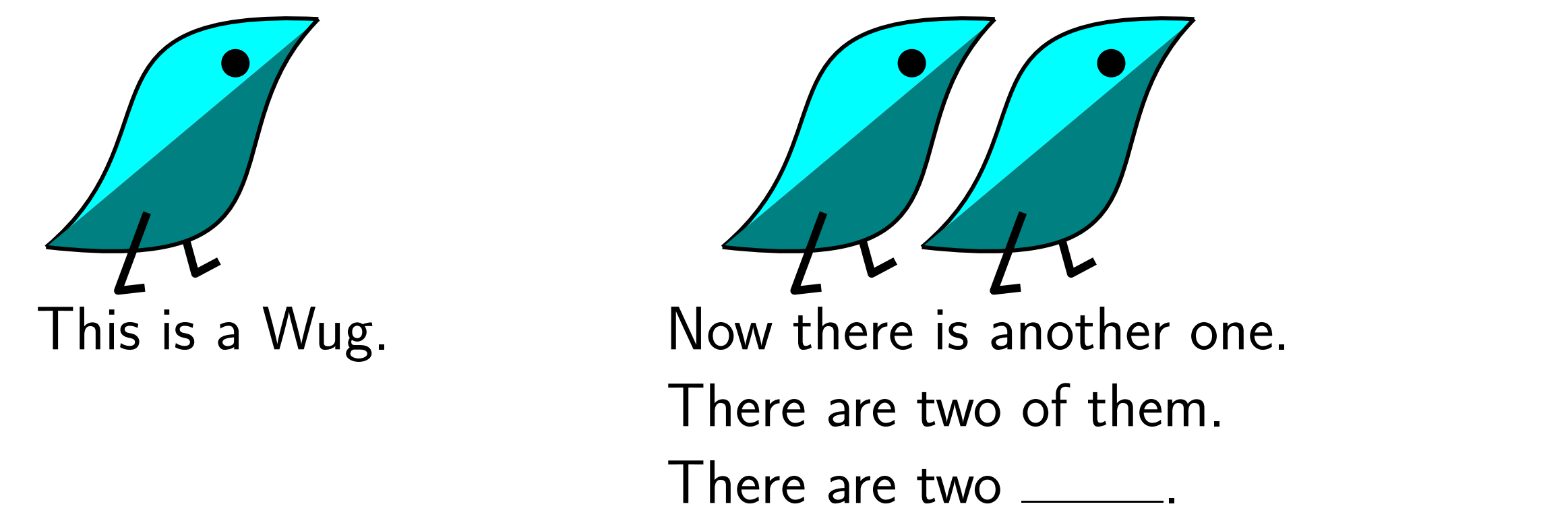

Figure 1:The WUG Test [4]. When presented with a novel noun, even young children will generalize the plural "s" when speaking about more than one of them.

Approach

- Children need 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.**

We examined efficiency of learning as a function of training set size using well-studied models of learning orthography-phonology correspondences [1, 5].

Model training and evaluation

One million models were run, 100k with each of 10 training set sizes $(100, 200, \ldots, 1000)$ comprised of words sampled randomly without replacement from the 2881 word target set. Each model was trained for 3000 weight updates with a constant learning rate $(\eta = 0.1)$. 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.

Conclusion

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.

Model Performance

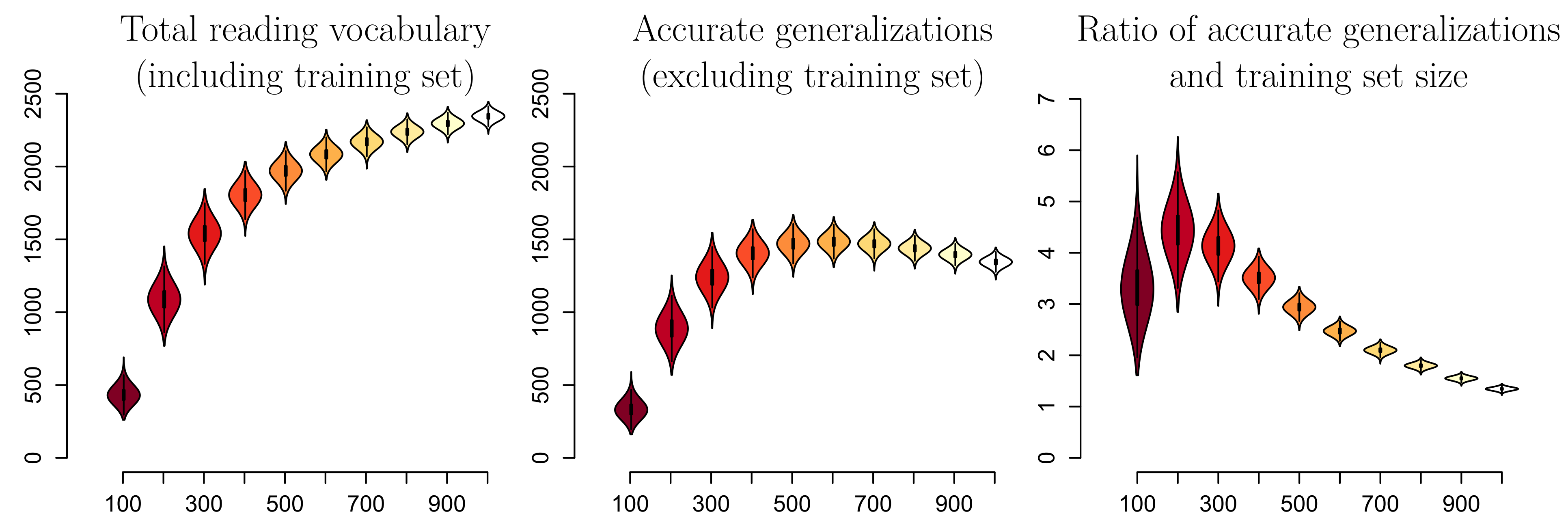

Figure 2:Summary of model performance as a function of the size of the training set.

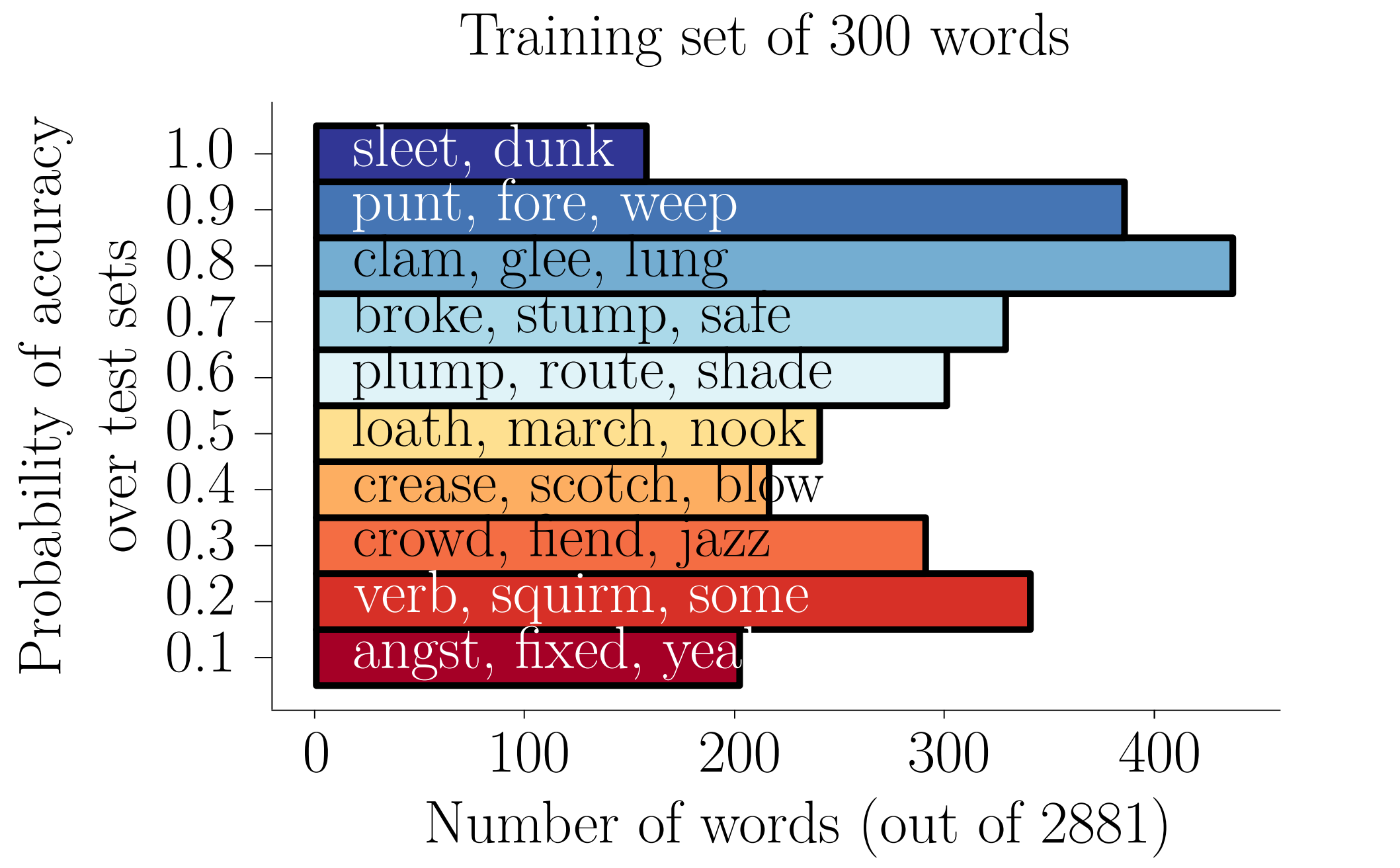

Figure 3:The probability that each word would be accurately generalized to, given that it was not trained on, summarized in a histogram. A random selection of words in each bin are printed.

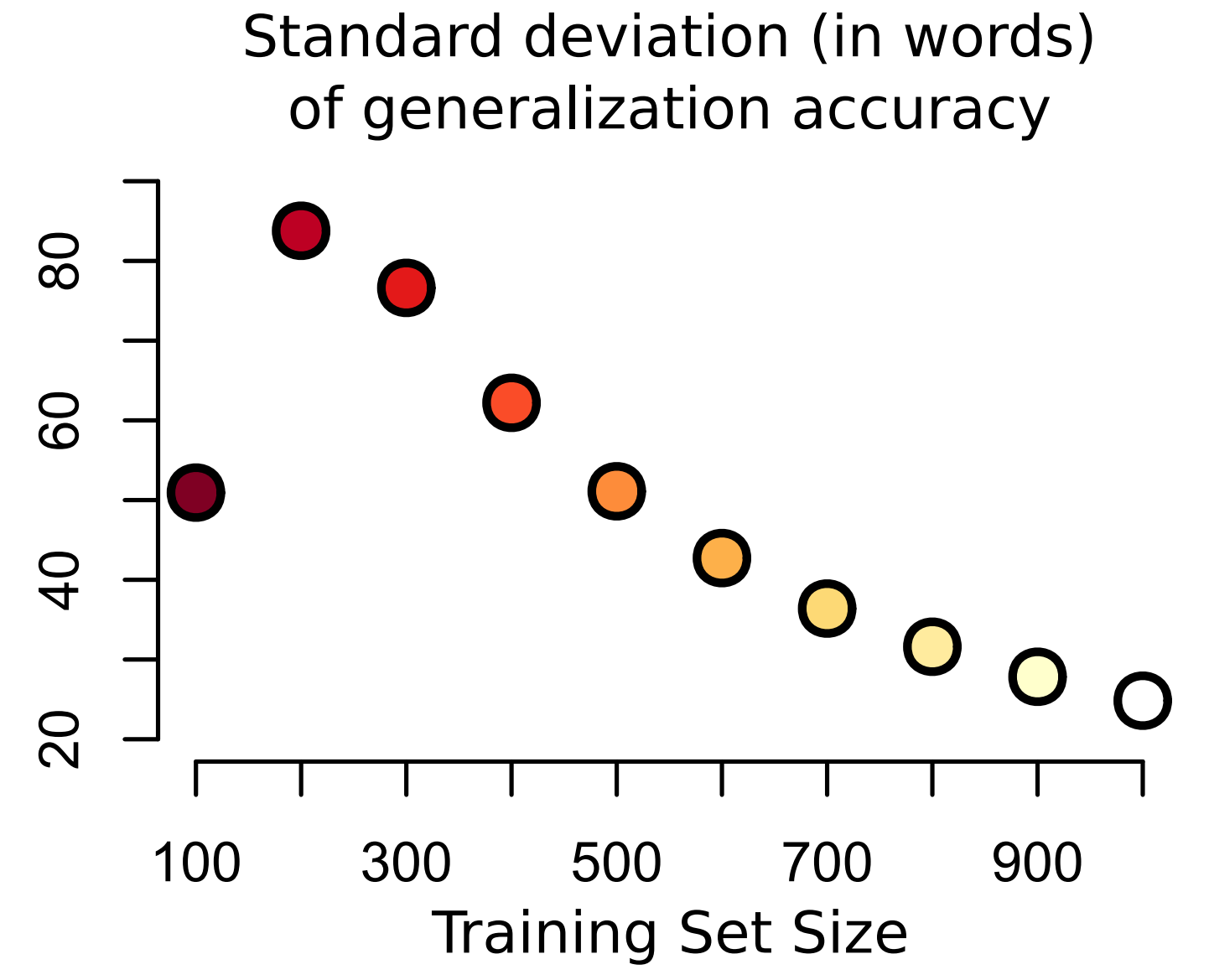

Figure 4:Standard deviation of generalization accuracy over the 100k models fit at each training set size.

Predicting generalizability

	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 1:Correlation among lexical variables. Bottom row shows each variable correlated with probability of accuracy.

Word Level	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 2:Effect sizes describing the variance explained by regressing the probability of accurately generalizing to an untrained word on several variables.

Model Level	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 3:Effect sizes describing the variance explained by regressing generalization accuracy on aggregate training set stats.

Target words

The simulations used a set of **2881 monosyllabic English** words employed in previous research [5]. Word length ranged from 2–8 letters and 1–7 phonemes. Words were presented uniformly, ignoring frequency in natural language.

Model architecture and word representations

Feedforward network with an input orthographic layer (102 units), an output phonological layer (66 units) and a single hidden layer (100 units). Weights updated with gradient descent and backpropagation after accumulating cross-entropy error over all words in the training set. The model was implemented using scikit-learn in Python 3.6 using a multilayer perceptron, and training was executed in parallel using HTCondor [6] and computational resources maintained by the Center for High Throughput Computing at UW Madison.

Orthographic representations were generated as follows:

1. Words were vowel-centered and padded with "empty letters" to establish uniform length (14 letters).
2. The letter *y* was treated as a consonant when it began a word and a vowel otherwise.
3. Each letter was represented by one unit in a 26 element vector.
4. Vectors were condensed by eliminating nodes for unused features, resulting in an input layer with 102 features.

Phonological word forms were represented using 41 phonemes (26 consonants, 15 vowels):

1. 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).
2. Each phoneme was defined by 25 phonetic features [5].
3. Vectors were condensed by eliminating nodes for unused features, resulting in an output layer with 66 features.

References

[1] Mark S. Seidenberg and James L. McClelland.
A distributed, developmental model of word recognition and naming.
Psychological Review, 96(4):523 – 568, 1989.

[2] Betty Hart and Todd R. Risley.
Meaningful differences in the everyday experience of young American children.
Baltimore, MD: Paul H. Brookes, 1995.

[3] Jill Gilkerson, Jeffrey A. Richards, Steven F. Warren, Judith K. Montgomery, Charles R. Greenwood, D. Kimbrough Oller, John H. L. Hansen, and Terrance D. Paul.
Mapping the early language environment using all-day recordings and automated analysis.
American Journal of Speech-Language Pathology, 26:248–265, 2017.

[4] Jean Berko.
The child's learning of english morphology.
Word, 14:150 – 177, 1958.

[5] M W Harm and M S Seidenberg.
Phonology, reading acquisition, and dyslexia: insights from connectionist models.
Psychological review, 106:491–528, 1999.

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

Acknowledgements

This work has been supported by the Vilas Trust at UW-Madison and by the Institute of Education Sciences, US Department of Education, through Award R305B150003 to UW-Madison. The opinions expressed are those of the authors and do not represent views of the US Department of Education.