Context variability promotes generalization in reading aloud: Insight from a neural network simulation

Ian D. Miller (i.miller@mail.utoronto.ca)
Department of Psychology, University of Toronto Scarborough, 1265 Military Trail, Toronto, ON, M1C 1A4, Canada

Nicolas Dumay (n.dumay@exeter.ac.uk)
Department of Psychology, University of Exeter, Perry Road, Exeter, EX4 4QG, UK

Mark Pitt (pitt.2@osu.edu)
Department of Psychology, Ohio State University, 1835 Neil Ave., Columbus, OH, 43210, USA

Brian Lam (brianps.lam@mail.utoronto.ca)
Division of Engineering Science, University of Toronto, 35 St. George Street, Toronto, ON, M5S 1A4, Canada

Blair C. Armstrong (blair.armstrong@utoronto.ca)
Department of Psychology, University of Toronto Scarborough, 1265 Military Trail, Toronto, ON, M1C 1A4, Canada

Abstract

How do neural network models of quasiregular domains learn to represent knowledge that varies in its consistency with the domain, and generalize this knowledge appropriately? Recent work focusing on spelling-to-sound correspondences in English proposes that a graded "warping" mechanism determines the extent to which the pronunciation of a newly learned word should generalize to its orthographic neighbors. We explored the micro-structure of this proposal by training a network to pronounce new made-up words that were consistent with the dominant pronunciation (regulars), were comprised of a completely unfamiliar pronunciation (exceptions), or were consistent with a subordinate pronunciation in English (ambiguous). Crucially, by training the same spelling-to-sound mapping with either one or multiple items, we tested whether variation in adjacent, within-item context made a given pronunciation more able to generalize. This is exactly what we found. Context variability, therefore, appears to act as a modulator of the warping in quasiregular domains.

Keywords: quasiregularity, neural network models, context variability, read aloud, spelling-to-sound correspondences, reading acquisition.

Introduction

In many domains, typically referred to as "quasiregular" domains, knowledge acquisition often entails learning about typical patterns and regularities alongside exceptions and violations. For instance, birds *typically* have wings and can fly. However, bats, which do have wings and can fly, are not birds. Meanwhile, penguins, which cannot fly, are still birds (Rogers & McClelland, 2004). Similarly, in the domain of learning the tenses of English verbs, the past tense can usually be guessed correctly by adding -ed to the present tense of the verb, as in walk-ed and talk-ed. However, violations of consistency must also be learned, such as the past tense of go being *went*, not *go-ed* (Seidenberg & Plaut, 2014).

The same kind of tension between regular patterns and exceptions characterizes the reading system of many languages, including English, in which relationships between spelling and sound are quite opaque. Learning how to translate spelling into sound requires that regularities, such as the English "i" typically pronounced in *mint*, *hint*, or *tint*, be able to coexist alongside exceptions, such as the "i" of *pint* being pronounced as in *eye*. Learning regularities leads to an efficient representation of knowledge, as storing a repeated pattern means that it can be applied to most instances, including new ones. In the illustration above, generalizing the typical pronunciation to the plausible novel word *kint*, for example, would in all likelihood produce a correct response. However, stimuli that violate the prototypical pattern should not be allowed to generalize, and this needs to be learned as well. How a computational model should deal with these two competing pressures—to generalize and not to generalize—has been a challenge for any account of how the brain's architecture learns to represent quasiregular domains.

One class of accounts that attempts to address these competing pressures are "rules-and-exceptions" models, which assume that rules and exceptions are coded by distinct pathways characterized by specific computation abilities and underlying neuro-anatomical substrates. In the domain of reading aloud, the Dual Route Cascaded model (DRC; Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001), as well as closely related models, such as CDP++ (Zorzi, Houghton, & Butterworth, 1998), have one route via which the most typical pronunciation can be retrieved and applied to most words as well as novel, unknown words. These models, therefore, must also have another separate route in order to code exceptions and ensure that these do not generalize. This type of account is appealing to psychologists due to its high-level theoretical transparency.

However, the limitations of this type of account become more prominent as soon as one tries to understand what makes a given word an exception to the rule. In particular, not all exceptions to the rules are equal, and some may be more extreme than others. For example, the pronunciation of "i" in *pint* may not be shared by any other rhyming word. In other rhyme neighbourhoods, however, the pronunciation

of "i" is more *ambiguous* and it is less clear which case is the *exception to the rule*, as in *give* and *live* versus *hive*, and *drive*. Thus, addressing the full quasiregularity continuum is more challenging for such dual route accounts, often involving assumptions about the speed and probability of completing processing via one route or the other under race conditions.

To address this challenge, a competing class of accounts, often exemplified by neural network models, has been developed. These models rely upon a single set of simple neuro-biologically inspired learning, processing, and representation principles that allow for information to be summated, nonlinearly transformed, and transmitted between "units" that are akin to neurons. However, despite the simplicity of the underlying mechanisms, these models can learn internal connectivity structures that encode the diversity of cases encountered on a quasiregular continuum, such as English spelling-to-sound mappings (e.g., Plaut, McClelland, Seidenberg, & Patterson, 1996; McClelland, 2015).

Work by Kim, Pitt, and Myung (2013) has shed light onto exactly how these neural networks accomplish this feat through *representational warping*, as illustrated in Figure 1. Warping is a hypothesized mechanism underlying quasiregularity that can be understood as a simplified two-dimensional schematic of the high-dimensional mapping between spelling and sound in the trained network. The representational space is relatively flat (minimally warped), so that consistent pronunciations can easily generalize to neighbouring areas (e.g. the "i" in *mint*, *lint*). Encoding a word that violates this consistency will entail *warping* a local area of the representational space where a different spelling-sound mapping can apply (e.g., the "i" in *pint*). Warped representations should minimally bleed out to impact neighbouring areas of the space, including to regions where newly-learned words (and nonwords) would be represented. However, the more exemplars that follow an atypical pronunciation (i.e., the more ambiguous the string is in terms of its pronunciation), the wider the area occupied by the alternate pronunciation and the shallower the slope. On that basis, neighboring strings are more likely to slip into that area and adopt the alternate pronunciation, thereby increasing the likelihood of generalizing ambiguous pronunciations.

These core tenets of representational warping have received direct empirical support through a series of coordinated neural network simulations and behavioural investigations (Armstrong, Dumay, Kim, & Pitt, 2017). Simulations and college-aged human participants were assessed on how well they learned and generalized the pronunciation of new words that were either consistent or inconsistent with their prior knowledge (that is, regular or exceptional), or somewhere in between (ambiguous). The behavioural results paralleled those of the simulations, showing differential rates of pronunciation generalization of new words even though these words had been learned equally well.

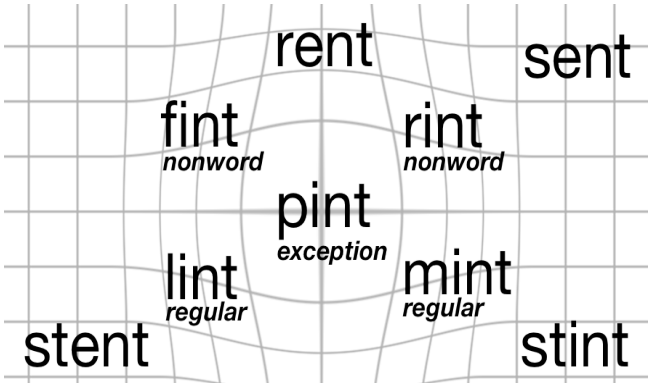

Figure 1: Warped space enabling the representation of the pronunciation of exception word *pint* with minimal spill-over to neighbouring words (lint, mint) and nonwords (fint, rint).

Present Aims

Although promising, the initial studies of warping leave much unanswered about exactly which properties of a word drive the formation of warped representations, and how warping could be modulated to impact generalization while preserving the learnability of items. To advance this line of reasoning, the current work looks at context variability, which has previously been shown to promote learning of the input structure (e.g., Lively, Logan, & Pisoni, 1993; Rost & McMurray, 2009). Any dimension (even across modalities) that covaries with the target domain can in theory assist with its partitioning into meaningful perceptual categories (e.g., Thiessen, 2007). Previous work in non-linguistic domains has established that context variability can enhance learning of initial items as well as generalization to novel items (e.g., Finch, Carvalho, & Goldstone, 2016). Therefore, the aim of the simulations we report herein is to identify how context variability—the number of words that share a rhyming pronunciation—could impact warping and generalization.

The interaction between context variability and warping is especially important because it allows an exploration of the micro-structure of representational warping. For example, learning a new exception word such as *suff*, rhyming with *roof* not *cuff*, should generate maximal warping. However, if instead, two new exception words sharing that rhyming word are learned, such as *chuff* and *vuff*, this "exception" should be slightly less exceptional, thereby requiring less warping to store both exceptions, ultimately increasing generalization for this new pronunciation. This should be true even when holding the total frequency of each rhyme mapping constant, such as by presenting the pair of exceptions at half the frequency as the single exception.

Taken together, the results of the current simulations will further our understanding of the warping mechanism, contributing to an explicit mechanistic understanding of how and why context variability modulates learning and generalization in a number of experimental studies, without recourse to qualitatively distinct representations of "rules" versus "exceptions."

Neural Network Simulation

We explored how the number of examples of a new pronunciation that varied in its consistency with prior knowledge impacted the ease of learning the new pronunciation, and how well this pronunciation generalized to previously unseen nonword neighbors in the surrounding representational space.

As in Armstrong, Dumay, et al. (2017), we extended a popular connectionist model of reading English words aloud (Plaut et al., 1996). First, we trained this model on a base vocabulary of English, and then introduced new items that varied in their consistencies with the established spelling-sound regularities. The task of reading aloud in English is particularly well-suited for the present purpose because of its quasiregularity: most words follow common sets of regularities to determine how spelling maps to sound—but there are also some exceptions.

Once the model was trained, we then introduced new regular, ambiguous, and exception words into the vocabulary, which we refer to as *anchors*. Holding rhyme frequency constant, we manipulated context variability by varying whether there was one, two, or three new words that contained each new pronunciation—we refer to this manipulation as *dilution*.

Finally, we examined how well these new words were learned and the degree to which their representations generalized to neighbouring nonword *probes*. Nonwords are an ideal test of generalization because unlike neighbouring words, which may have been explicitly trained to some particular pronunciation, nonwords have not been explicitly represented, and their performance is determined by how the pronunciation of known words is extended to them.

At a macro-structural level, we predicted that when all three anchor types were equally well-learned, there should be substantial generalization of the ambiguous anchors to their corresponding probes, and minimal generalization of the exception anchors. This is because learning the exceptions will entail more warping than learning the ambiguous anchors. As we dilute each pronunciation, this macro-structural prediction should be recapitulated at a micro-structural level: having three words sharing an exceptional (or ambiguous) rhyme should make that pronunciation less exceptional, thereby leading to reduced warping and more generalization of the new pronunciation. Regular words serve as a basic control condition in all cases; they should already be pronounced correctly even before training via generalization of existing pronunciations, as should regular probes.

Methods

Architecture

The architecture of the model is presented in Figure 2. The model consists of an input layer coding for a word's *orthography*, which feeds forward to an intermediate hidden layer, which in turn feeds into the *phonology* output layer.

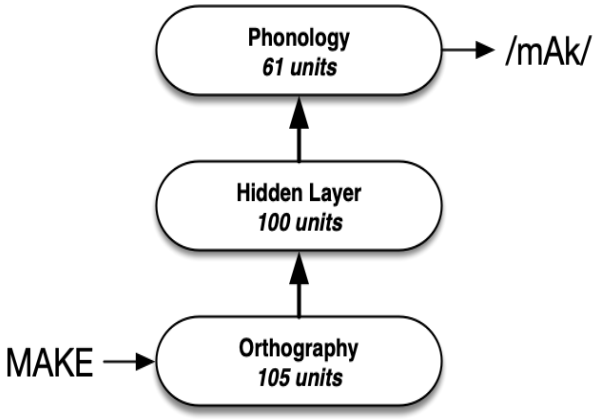

Figure 2: Model architecture. An example word, *MAKE*, is presented to the orthographic layer. The corresponding pronunciation, */mAk/*, is produced by the phonology layer.

Both the input and output layers use a slot-based coding to denote the position of elements in a word, thereby enabling the representation of the spelling and sound of monosyllabic English words. A sequence of slots in the orthographic layer codes for each grapheme in a word, in sequence. Similarly, a sequence of slots in the phonological layer codes for the onset, vowel, and coda of words (e.g., the "h", "i" and "nt" in *hint*).

Representations

Base Vocabulary The training corpus consisted of 2998 monosyllabic words, representing the bulk of common English monosyllabic words. Their written word frequencies were also used during training to make the simulations more realistic given the importance of frequency effects in language processing. For consistency with prior simulations, we used the frequency data from Kučera and Francis (1967).

New Word Anchors After training on the baseline vocabulary, the model learned three sets of new anchors:

1. *Regular* anchors, whose pronunciation was consistent with the regularities in the baseline vocabulary (e.g., *bint*, rhyming with *mint, hint, tint*).

2. *Exception* anchors, whose pronunciation was inconsistent with the regularities in the baseline vocabulary (e.g., *suff*, rhyming with *roof*, not *cuff*).

3. *Ambiguous* anchors, whose pronunciation was consistent with a subordinate pronunciation, but not the dominant regularity, in the baseline vocabulary (e.g., *bive*, rhyming with *give, live*, not *drive, hive*).

Each set consisted of ten triplets, with each triplet coding for three words sharing the same rhyme (vowel + coda), using the stimuli in Armstrong, Dumay, et al. (2017) as the basis

for the first word in each triplet. For example, one exception triplet consisted of {*suff, chuff, vuff*}.

To study the effects of *diluting* the exposure to each word (onset + vowel + coda) while holding the frequency of the rhymes (vowel + coda) in each triplet constant, we created three dilution sets presenting either the first anchor in a triplet, the first and second anchors in a triplet, or the entire triplet. The summed word frequency of each pronunciation was held constant using a dilution scaling factor, such that the word frequency of presenting a single anchor in a triplet would be scaled by 1, presenting two anchors in a triplet would be scaled by $1/2$, and presenting all three anchors would be scaled by $1/3$. We refer to each of these sets as the the *low, moderate, and high* dilution sets, respectively. We trained models using two different orderings of the triplets to test the robustness of the model's performance to the effects of dilution. The results were highly similar, so for brevity we report the average results across orderings.

To address other key psycholinguistic properties that can behaviourally modulate the naming aloud task, we also matched the words on several other properties—to the extent possible given the constraints of English—including length in letters, length in phonemes, and orthographic neighbourhood size.

Probes (nonwords to test generalization)

Probes consisted of four rhyme neighbours for each of the anchor triplets (e.g., *vlit* as a probe for anchor *blit*, rhyming with *slit*). Put differently, the probes and their corresponding anchors all share the same rhyme. The only difference between these two types of stimuli is that the model learned to read the anchors through supervised learning, whereas the model never learned (i.e., never had its weights adjusted) based on exposure to the probes. Thus, the probes reflect a test of the generalization of the pronunciation learned from the anchors to these novel (nonword) items.

Training

Model Initialization

Prior to training, the weights in the simulation were instantiated to small random values in proportion to N, the number of weights between layers, as (mean = 0, range = $\pm$ $1/\sqrt{N}$). All bias values were initially set to -1.85, which reduced the mean activity in the hidden and output units and facilitated the learning of the sparse output patterns. To gain insight into the systematicity of the results, two variants of each set of random weights were run, each using a different random seed to initialize the weights. The results of each individual simulation were very similar, so we report the average across the simulations.

Base Vocabulary For the first 350 epochs, the model was trained on the base vocabulary only. Weight adjustments for the initial 10 epochs were calculated using Steepest Gradient Descent, with a learning rate of 0.0001. For the next 340 epochs, weight adjustments were performed using Adaptive Moment Estimation (Adam) with a learning rate of 0.01 and the default algorithm parameters ($betas = (0.9, 0.99)$; $epsilon = 1 \times 10^8$) (Kingma & Ba, 2014). The effects of word frequency, F were simulated by scaling the cross-entropy error between the target unit activations and the actual unit activations in the output layer by the log-transformed word frequency, $ln(F+2)$. We completed one sweep through every word in the training corpus before updating the weights (i.e., batch learning). The updated weights were also subject to a small amount of weight decay (1×10^{-6}). After training on the base vocabulary for a total of 350 epochs, error had reduced to a small, asymptotic state, and 96.2% percent of words were pronounced correctly. Here, a "correct" response was defined as whether the most active vowel in the output pattern matched the target vowel in the training example.

Learning new words After epoch 350, the base vocabulary was expanded to also include the training anchors for each dilution condition. The word frequency for the anchors was specified for three dilution scaling amounts, $D = \{1, \frac{1}{2}, \frac{1}{3}\}$, as $ln((10 * D) + 2)$, in order to maintain equal exposure to each rhyme regardless of how many anchors shared that rhyme. After training on the base vocabulary and prior to training on the expanded vocabulary, the Adam optimizer was reset. The model was then trained for an additional 350 epochs until overall performance in the base vocabulary and for each anchor type reached a stable state and all anchors were pronounced correctly.

Results

The predicted behavioural markers of warping and dilution are as follows: first, whereas the network will not need to warp the space to represent regular anchors, it will need to increasingly warp the space for ambiguous and exception anchors. More warping should prevent generalization for exception probes compared to ambiguous probes. The dilution manipulation should recapitulate these macro-structural changes at a micro-structural level: diluting an exception anchor (or an ambiguous anchor) will make it slightly less exceptional, leading to increased generalization. To test for these predictions in a simple and straightforward manner, our analyses focused on how often the model produced the pronunciation of the vowel that was consistent with the anchors learned during training. Warping was expected to be necessary to accommodate the atypical pronunciations associated with exceptions, and to a lesser extent, with the ambiguous words.

In the vast majority of cases when the model was not producing this training-consistent pronunciation, it produced a regularized pronunciation consistent with the regularities of the baseline vocabulary and not some third other type of unexpected ("erronous") response. Note that for regular anchors, the regularized response is also the training-consistent response.

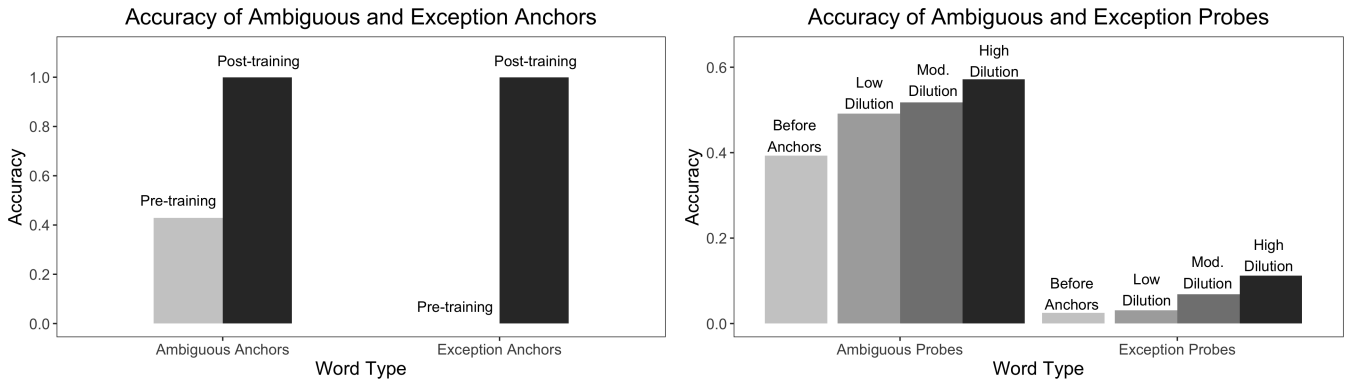

Figure 3: Proportion of Anchors and Probes producing training-consistent responses. The figure in the left panel depicts the effects of training with anchors, in which all anchors were trained to perfect accuracy for all word types. The figure in the right panel depicts the final accuracy of probes after training upon anchors, as well as the pre-training baseline. Regulars are not depicted in the figure because they begin with perfect accuracy due to their rule-consistent pronunciations. Note the different y-axis ranges were used across the two figures to better highlight the effects of interest.

Learning Anchors

The proportion of training-consistent responses for ambiguous and exception anchors is presented in the left panel of Figure 3. The "Pre-training" data was sampled from training sweep 350, after the learning of the base vocabulary, and the "Post-training" data were sampled from training sweep 700, following an additional 350 sweeps of training for both the anchors and the base vocabulary.

The "Pre-training" data establishes how well the base vocabulary can be used as grounds for generalizing to the untrained anchors, and as a basic validation of the core characteristics of our anchors sets and of the reimplementation of the Plaut et al. (1996) model. The results show that the model never produces the correct (training-consistent) pronunciation of an exception anchor, sometimes produces a correct response for an ambiguous anchor, and always produces a correct pronunciation of a regular anchor, as expected.

After training on the anchors, the model learns to pronounce all anchors perfectly for all dilution levels, demonstrating that all item types can be represented in the network. The critical question, then, is how well these representations generalize to neighbouring nonword probes.

Generalizing to nonword probes

The proportion of training-consistent responses for the probes is presented in the right panel of Figure 3. First, at a macro-structural level, the results clearly show that the rates of generalization are higher for ambiguous words than for exception words, as predicted by the warping mechanism. The predicted effects of dilution within each item type also recapitulate these macro-structural effects at the micro-structural level: the proportion of generalized anchor pronunciations increases as a function of increased dilution. That is, generalization of a newly learned pronunciation increases as a function of increased context variability, here defined as varying word onsets. Collectively, these results provide important insights and explicit quantitative evidence of how different aspects of a word's representation influence the warping of representational space, and the resulting changes in generalization of the pronunciation of newly learned words.

Discussion

Understanding how quasiregular domains are represented so as to enable the generalization of regularities but not inconsistencies is a fundamental issue in the cognitive sciences. The present work expands past work exploring warping to understand its sensitivity to context variables that could affect word acquisition and generalization. In particular, we probed whether macro-structural differences in generalization rates for newly learned regular, ambiguous, and exception anchors would be recapitulated at the micro-structural level when the rhyming portion of a newly learned word was diluted by presenting it in more than one new word. We observed that this was indeed the case, such that increased context variability in word onsets caused the network to infer that there was stronger evidence that this new pronunciation should be generalized, both for ambiguous and exception items. Thus, context variability can play a critical role in modulating the formation of warped representations. Notably, these results were observed in the context of equal ceiling performance for the explicitly trained anchor words of each item type.

The results of our simulation suggest that generalization is driven by the sublexical components of a word's representation, and that increased context variability in word onsets for a same-rhyme ending helps the network infer that a new sublexical regularity is emerging. Statistical cues to segmentation a la Saffran, Aslin, and Newport (1996) and

exemplar based models of segmentation (Perruchet & Vinter, 1998) work in similar ways. These results provide strong predictions regarding how humans should perform when learning these types of items which, if confirmed empirically, could offer a new avenue for contrasting neural network and dual route models. In particular, our results were observed when the sublexical rhyme portion of the representation was presented equally frequently and it was only the onset + rhyme conjunction (i.e., the lexical representation) whose frequency was varied across dilutions. This makes specific quantitative predictions regarding the relative importance of sublexical and lexical representations for generalization, as well as for how and why rule-like behaviour emerges for new consistencies.

Empirical data consistent with these claims has been reported previously in a number of studies, wherein increased context variability (broadly construed) leads to learning benefits, although the exact magnitude of these benefits has varied across studies (for discussion, see Roembke, Freedberg, Hazeltine, & McMurray, 2020.) Whereas this past work has typically focused primarily on the learnability of new explicitly taught words (anchors, in present parlance), the work we report highlights how such learning is likely simultaneously impacting generalization rates. This possibility is well-illustrated by considering how our present findings might offer a slightly different interpretation of the results reported by Roembke et al. (2020). In their study, the authors tested children and observed a benefit for variability in a consonontal frame similar to our dilution manipulation. However, accuracy for their "anchors" did not reach 100% at training. Our results suggest that, insofar as some anchors sharing a same pronunciation had not been learned, the benefits from context variability may have arisen not just from explicit *learning* of the anchors, but also from from *generalization* of the knowledge of the successfully learned anchors to the unsuccessfully learned anchors (effectively, probes, in our nomenclature). Indeed, in additional simulations not reported here for the sake of brevity, we have observed that the likelihood of generalization of anchor knowledge is modulated by how much experience the model has had for learning the anchors in the lead up to the "steady state" performance that we reported in detail in the results section, wherein accuracy for the probes and anchors had remained relatively stable for hundreds of training sweeps. A more detailed examination of these transient effects is therefore a clear direction for future work.

Our results may also relate to other aspects of word learning. Several studies have reported how rule-like generalizations can be enhanced in infants by providing broader evidence of context variability in processing. For example, Gerken (2006) noted that infants learning stimuli generated using an AAB rule (e.g., *leledi, dededi, wiwije, jijili*) were more likely to learn an abstract representation of this rule if a more diverse set of syllables was used to demonstrate the abstract rule (e.g., *leledi, wiwije, jijili*). In contrast, if only a single syllable was used to code for the "B" portion of the rule (e.g., *leledi, wiwidi, jijidi*), generalizations appeared restricted to novel items that shared this B syllable. For instance, in the former but not the latter case, infants were more likely to generalize the rule to novel stimuli that had no syllable overlap with the training stimuli, such as *kokoba*.

These results are consistent with our own findings that experience with a diversity of onsets and a consistent word ending lead to generalizations of the word ending to other novel onsets. They also suggest that the warping principle could be extended even further to explain more abstract rule-like behaviour if we varied not only the onset of a word, but also its rhyme. As such, the warping mechanism may be relevant for understanding broad cross-sections of the statistical learning literature focused on learning both linguistic and non-linguistic stimuli (Armstrong, Frost, & Christiansen, 2017). Thus, warping may have major implications not only for the domain of reading aloud, but also for learning and generalizing other aspects of language - and beyond.

Conclusion

The present simulations clearly outline important predictions regarding the micro-structure of how new representations are learned and generalized, and how these processes can be modulated by context variability that dilutes the degree to which an inconsistent word's onset occurs with a particular rhyme. These results have clear connections and implications for what and how representations are learned and generalized when reading aloud, as well as for other linguistic domains and domains in the cognitive sciences more generally. Our explicit computational framework also has close links to potential analogous behavioural experiments that examine learning and generalization. Collectively, such coordinated computational and behavioural studies should also allow for the efficient exploration of how representational warping is modulated by other key psycholinguistic factors, yielding further insights into the operation of this powerful and flexible mechanism.

Acknowledgements

The authors would like to thank the anonymous reviewers. ND is supported by the Economic and Social Research Council (UK) standard grant #ES/R006288/1. BCA is supported by NSERC DG 502584 and CFI JELF/ORF 36578.

References

Armstrong, B. C., Dumay, N., Kim, W., & Pitt, M. A. (2017). Generalization from newly learned words reveals structural properties of the human reading system. *Journal of Experimental Psychology: General*, *146*(2), 227–249. doi: 10.1037/xge0000257

Armstrong, B. C., Frost, R., & Christiansen, M. H. (2017, January). The long road of statistical learning research:

Past, present and future. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *372*(1711), 20160047. doi: 10.1098/rstb.2016.0047

Coltheart, M., Rastle, K., Perry, C., Langdon, R., & Ziegler, J. (2001). DRC: A Dual Route Cascaded Model of Visual Word Recognition and Reading Aloud. *Psychological Review*, *108*(1), 204–256.

Finch, D., Carvalho, P. F., & Goldstone, R. L. (2016). Variability in category learning: The Effect of Context Change and Item Variation on Knowledge Generalization. In *CogSci.*

Gerken, L. (2006). Decisions, decisions: Infant language learning when multiple generalizations are possible. *Cognition*, *98*(3), B67-B74. doi: 10.1016/j.cognition.2005.03.003

Kim, W., Pitt, M. A., & Myung, J. I. (2013, October). How Do PDP Models Learn Quasiregularity? *Psychological review*, *120*(4), 903–916. doi: 10.1037/a0034195

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

Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Dartmouth Publishing Group.

Lively, S. E., Logan, J. S., & Pisoni, D. B. (1993, September). Training Japanese listeners to identify English /r/ and /l/. II: The role of phonetic environment and talker variability in learning new perceptual categories. *The Journal of the Acoustical Society of America*, *94*(3), 1242–1255. doi: 10.1121/1.408177

McClelland, J. L. (2015). Capturing gradience, continuous change, and quasi-regularity in sound, word, phrase, and meaning. *The handbook of language emergence*, 53–80.

Perruchet, P., & Vinter, A. (1998). PARSER: A model for word segmentation. *Journal of memory and language*, *39*(2), 246–263.

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.

Roembke, T. C., Freedberg, M. V., Hazeltine, E., & McMurray, B. (2020). Simultaneous training on overlapping grapheme phoneme correspondences augments learning and retention. *Journal of Experimental Child Psychology*, *191*, 1–19. doi: 10.1016/j.jecp.2019.104731

Rogers, T. T., & McClelland, J. L. (2004). *Semantic Cognition: A Parallel Distributed Processing Approach*. MIT Press.

Rost, G. C., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental science*, *12*(2), 339–349.

Saffran, J. R., Aslin, R. N., & Newport, E. L. (1996, December). Statistical learning by 8-month-old infants. *Science*, *274*(5294), 1926–1928. doi: 10.1126/science.274.5294.1926

Seidenberg, M. S., & Plaut, D. C. (2014). Quasiregularity and Its Discontents: The Legacy of the Past Tense Debate. *Cognitive Science*, *38*(6), 1190–1228. doi: 10.1111/cogs.12147

Thiessen, E. D. (2007). The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language*, *56*(1), 16–34.

Zorzi, M., Houghton, G., & Butterworth, B. (1998). Two routes or one in reading aloud? A connectionist dual-process model. *Journal of Experimental Psychology: Human Perception and Performance*, *24*(4), 1131–1161. doi: 10.1037/0096-1523.24.4.1131