Generalization from newly learned words reveals structural properties of the human reading system

Blair C. Armstrong

Department of Psychology & Centre for French & Linguistics at Scarborough,

University of Toronto, Canada

BCBL. Basque Center on Cognition, Brain and Language, Spain

Nicolas Dumay

Department of Psychology, University of Exeter, UK

BCBL. Basque Center on Cognition, Brain and Language, Spain

Woojae Kim

Department of Psychology, Howard University, DC

Mark A. Pitt

Department of Psychology, Ohio State University, OH

Corresponding author:
Blair C. Armstrong (blair.armstrong@utoronto.ca)
Department of Psychology & Centre for French & Linguistics
1265 Military Trail
Toronto, Ontario, Canada
M1C 1A4
416-287-7146

WORD COUNT: 16 982

Abstract

Connectionist accounts of quasiregular domains, such as spelling-sound correspondences in English, represent exception words (e.g., *pint*) amidst regular words (e.g., *mint*) via a graded "warping" mechanism. Warping allows the model to extend the dominant pronunciation to nonwords (regularization) with minimal interference (spillover) from the exceptions. We tested for a behavioral marker of warping by investigating the degree to which participants generalized from newly learned made-up words, which ranged from sharing the dominant pronunciation (regulars), a subordinate pronunciation (ambiguous), or a previously non-existent (exception) pronunciation. The new words were learned over two days, and generalization was assessed 48 hours later using nonword neighbors of the new words in a tempo naming task. The frequency of regularization (a measure of generalization) was directly related to degree of warping required to learn the pronunciation of the new word. Simulations using the Plaut et al. (1996) model further support a warping interpretation. Findings highlight the need to develop theories of representation that are integrally tied to how those representations are learned and generalized.

Keywords: quasiregularity; connectionist models; word learning; tempo naming.

Generalization from newly learned words reveals structural properties of the human reading system

Mastery of reading in alphabetic languages, such as English, requires an individual to learn correspondences between strings of letters and their pronunciations. Typically, this mapping is consistent, whereby a string of letters is pronounced in one way across many words (*int* in *mint, hint, print*). Although such regularities can simplify reading acquisition, there are also exceptions that are inconsistent with these regularities (e.g., *pint*) that must also be learned, and these clearly pose a challenge. Nonetheless, most people become proficient readers. Why does learning an exception word like *pint* not disrupt reading orthographically similar words like *mint*? This paper investigates how skilled readers represent both regular words and exceptions to the rule, and yet exhibit only minimal interference between the two categories.

Considerable computational works has been devoted to understanding how quasiregularity is represented in memory in a way that enables accurate reading of both regulars and exceptions. For example, the DRC model (Coltheart, Rastle, Perry, Langdon, & Ziegler, 2001) has maintained that two qualitatively different mechanisms are required to do so: a rule-based system to deal with regulars and new words, and a memory-based system to accommodate exceptions. Others have argued that it is possible for a single mechanism to represent both regularities and the various degrees of inconsistency characteristic of natural language (e.g., Plaut, Seidenberg, Patterson, & McClelland, 1996). A connectionist, parallel distributed processing (henceforth PDP) network that maps between spelling and sound via an intermediate pool of hidden units has been shown to learn how to pronounce both regulars and exceptions with a performance comparable to that of skilled readers. One impressive but puzzling finding of this simulation work is that learning exceptions minimally disrupts the ability to pronounce words the network had never seen (i.e., nonwords; Plaut et al., 1996). This emergent property of how PDP models learn an internal representation between orthography and phonology highlights how learning to represent new words is not necessarily a separate, independent component of a theory of reading, but one that is fundamentally intertwined with representation formation.

The current project explores this inter-dependence, which is so central to PDP models across the board (McClelland, 1998; 2015). We test predictions of the PDP mechanism that enables learning quasiregularity and present behavioural and computational evidence for its plausibility. Although we use reading behaviour to test our predictions, our findings have implications not just for theories of reading (e.g., Coltheart et al., 2001; Perry, Ziegler, & Zorzi, 2010; Plaut et al., 1996), but also speak to issues related to language acquisition, such as why the properties of some words generalize but others do not (Apfelbaum, Hazeltine, & McMurray, 2013; Treiman, Kessler, Zevin, Bick, & Davis, 2006). They should also have implications for learning in other quasiregular domains, such as second language learning and bilingualism (e.g., how partially overlapping regularities in English and Spanish can co-exist in the bilingual brain; Ijalba & Obler, 2015), grammar acquisition (e.g., how regular and exceptional past tense forms such as learn/learned vs. go/went are represented; Pinker & Ullmann, 2002; Seidenberg & Plaut, 2014), and semantic cognition (e.g., how a penguin can be classified as a bird, not as a fish; McClelland, McNaughton & O'Reilly, 1995), just to cite a few. In addition, because PDP models are statistical learning systems, our data provide insight into an intriguing discrepancy in the statistical learning literature: why learning in different domains yields sometimes a high degree of generalization and sometimes stimulus-specific learning and minimal generalization (Frost, Armstrong, Seigelman, & Christiansen, 2015).

To better understand how internal representations in the Plaut et al. (1996) model made it both generalize appropriately and learn exceptions, Kim, Pitt, and Myung (2013) performed a series of simulations to flesh out how representations of regularities and exceptions were organized in that model. Intuitively, the characteristics of the representations of regular and exception pronunciations would seem to be at odds: learning an exception like *pint* should hinder the model's ability to generalize its knowledge of how *int* should be pronounced when encountering new words (e.g., *kint, bint, gint*). Kim et al.'s simulations revealed that the crux for reconciling these contrasting pressures lies in how exceptions are accommodated in a representational system that is designed to ensure that new words are pronounced using the dominant pronunciation (short I, as in *mint*). Insertion of an exception word requires a warping of the representational space to permit the mapping of a letter (e.g., i) onto an *additional* phoneme (e.g., long dipthong /aI/, as in *pint*). Warping is confined essentially to the specific context of the exception (*p* onset and *nt* coda). Hence, it is not about modifying some context-free, small-grain size association, akin to a grapheme-phoneme correspondence affecting all words containing a particular letter or group of letters. (see Treiman, Kessler, Zevin, Bick, & Davis, 2006, for a similar developmental view).

Intriguingly, however, and as depicted in Figure 1, there was some significant spillover which resulted in the model generalizing the new pronunciation to (untrained) nonword neighbours (e.g., *kint, bint*, g*int*); that is, on a significant proportion of trials, these items which the model had never seen were mispronounced using the exception (long) vowel, instead of the short vowel. Kim et al.'s simulations also showed graded effects in the violation of established consistencies: whereas the spillover from single exceptions to the rule (e.g., *pint*) was minimal, it was more pronounced the more there were instances in which the inconsistent pronunciation was the correct one. In these ambiguous cases (e.g., *ive* typically pronounced with a long vowel as in *hive*, except in *give* and *live*), the model effectively assumes that there is a newly forming regularity and it therefore increases the likelihood that the subordinate (short-vowel) pronunciation is chosen.

Kim et al.'s (2013) findings provide an opportunity to test new predictions of the single-route / single mechanism model of reading and the connectionist architecture more broadly, and to explore the link between representation and generalization. If when asked to name nonword neighbours of the exception, readers mimic what is observed in the simulations and thus exhibit some spillover from the newly learned exception (i.e., occasionally naming *kint* as *pint*, not *mint*), then warping as the core representational mechanism instantiated in the single-route model will be supported. In particular, the frequency of regularization of nonword probes should be (inversely) related to the amount of warping required to learn the pronunciation of a new irregular word. For instance, ambiguous cases, for which the relevant pronunciation is already available (though in a subordinate position), require less warping compared to exceptions introducing a completely new spelling-sound mapping. This inverse relationship between warping and extent of generalization is schematized in Figure 2.

In the present study, participants were taught a set of made-up words (coined *anchors*) that varied in the extent to which pronunciation of their vowels agreed with English spelling-sound correspondences, as reflected in other rhyming words: (1) *Regular* anchors, pronounced exactly as in all other words of the language, except for a few loan exceptions (e.g., *blit*, rhyming with *wit*); (2) *Exception* anchors, breaking a rule for which there are no exceptions so far (e.g., *suff*, rhyming with *roof*, and not with *cuff*); and (3) *Ambiguous* anchors, to be read using the subordinate pronunciation out of the two available (*bive* rhyming with *give*, not with *drive*).

Training was spread over two alternate days (Days 1 and 2), so as to maximize learning by spaced practice (Ebbinghaus, 1885/1964; see Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006, for a review). As generalization has been shown to be promoted by sleep (or sometimes just time alone), and comes for free as part of memory consolidation (Earle & Myers, 2015; Fenn, Nusbaum, & Margoliash, 2003; Gaskell et al., 2014; Gómez, Bootzin, & Nadel, 2006; Tamminen, Davis, & Rastle, 2015; see Stickgold & Walker, 2013, for a review), it was tested only after another 48 hours, by probing participants with new (nonword) orthographic rhyme neighbors of the learnt anchors (e.g., *brax* for *grax*). Warping was expected to reduce regularization of probes for newly learnt exception anchors (in favor of the new pronunciation), compared to probes for regular anchors. Such a reduction in regularizations was expected to be even greater for probes of ambiguous anchors, due to the subordinate pronunciation already being coded in the network (see above).

To test these predictions, we used tempo naming (Kello & Plaut, 2000). Participants heard a train of five tones. Their task was to read aloud the letter string which appeared on the fifth tone and synchronize their pronunciation with when a sixth tone would have occurred. To push participants to their limits, the tempo was derived from the participant’s own baseline speed estimated in a pretest, from which 150, 100, 50 (or 0) ms were subtracted.

Tempo naming is minimally affected by speed-accuracy trade-offs. In addition, a fast tempo seems to prevent the emergence of the strategic biases seen in standard naming, such as lexicality effects caused by lists containing nonwords (e.g., Andrews and Scaratt, 1998). Further, the fast nature of tempo naming response reduces the possibility that it would be influenced by slower processes, such as “reading by analogy” to either lexical representations (Glushko, 1979; for discussion, see Zevin & Seidenberg, 2006) or episodic traces (Düzel et al., 1999), including those of the recently learnt anchors. On another front, because it encourages fast and constant latencies, tempo naming shows effects mostly on errors. Their frequency and their types should give us a window into the nature of the learnt representations and how these generalize.

Method

Participants. Eighty-three American-English native speakers were tested, 36 in the main experiment and 47 in a control condition with no training. All were Ohio State undergraduates and had no visual, hearing or language impairment. They received course credits or were compensated monetarily.

Materials. The key materials were three sets of ten to-be-learned *regular, exception* or *ambiguous* anchors, as described above (All stimuli retained after data clean up, described in the results section, are reported in Appendix A). The regular and exception anchors were derived from the "regular consistent" nonwords from Glusko (1979). The choice to use this stimulus set as the foundation for our stimuli was based on the popularity of this work and its associated stimuli in both the empirical and computational literatures, including Kello and Plaut's (2000) tempo naming study. We assumed that elaborating from these stimuli would enhance the comparability of our findings with prior work.

The regular anchors that we sampled were not changed from the original study and they shared the standard pronunciation common to many English words (e.g., *blit* in *wit, mit, hit, fit*). For the exception anchors, we created a new pronunciation for each vowel, so that these items precisely did not rhyme with their neighbors[1] (e.g., *suff* rhymed with *roof*, not *cuff*). Note, however, that the new imposed mapping may have already been present in non-neighboring existing words (e.g., *flu*). In the no training condition, probes for exception anchors were

[1] For the purpose of creating the regular, exception, and ambiguous anchors, a neighbor of an item was any word that can be created by substituting the first grapheme of the target word with another grapheme.

therefore expected to produce identical performance to the regulars, as all of these items were regular consistent items taken from Glusko (1979).

The ambiguous anchors were identified via a corpus search for words whose rhyme neighbors broke down into at least two sets of multiple words sharing one or the other pronunciation (e.g., *bive* rhymed with *drive/hive*, not *give/live*). We also checked that the Plaut et al. (1996) model produces two pronunciations for each of item, when asked to read them multiple times (in some cases, a third pronunciation was produced extremely infrequently). As will be seen below, participants in our no training condition also produced multiple pronunciations for the ambiguous anchors, but not for the regular and exception anchors (frequency of most frequent pronunciation for exceptions as a function of all responses: 95%, SE = 1%; regulars: 93%, SE = 2%; ambiguous: 55%, SE = 15%). The most frequent pronunciation of each anchor in the model was considered to be the regularized pronunciation of that word.

Each anchor was associated with four rhyming nonword probes (e.g., *chuff*, *druff*, and *vuff* for the anchor *suff*), identical to the anchor except for the onset consonant (or consonant cluster). As can be seen in Appendix B, the three sets of anchors and probes were reasonably well matched on a number of lexical and sublexical variables, though some differences persisted given inevitable language constraints (As reported in the Results section, covariance analyses demonstrated that none of these differences could account for our findings). As was the case for the anchors, the no-training participants produced the regularized pronunciation for all probes (93%, SE = 1%), except, of course, the ambiguous ones. For these, they produced the regularized pronunciation from the anchor less frequently (46%, SE = 7%) and relied instead on other alternate mappings (predominantly, the training-consistent pronunciation). Descriptive statistics for the probes and anchors related to several psycholinguistic properties are presented in Table B1.

Procedure.

Training. Participants in the training condition completed two days of training before test. There was a 48 hour break between the training sessions on Day 1 and day 2, and between the last training session and test on Day 3 (e.g., training sessions could occur on Monday and on Wednesday, with the test following on Friday). The training session consisted of three tasks that were intended to promote learning. The tasks were cycled through twice. An overview of the training and testing procedure is presented in Figure 3.

A cycle began with *mere exposure* to the stimuli, in which participants saw and heard all 30 anchors in random order and had to learn how each was pronounced. A trial began with the simultaneous presentation of the written and spoken form of an anchor. The written form then remained on the screen and after 2500 ms, the spoken form was repeated once. After a 1750 ms inter-stimulus interval, the next trial began.

Participants then moved onto a *sound-to-spelling association* task. This required them to indicate by button press which of two similar ways of spelling the auditory form that appeared on the screen (e.g., *blit* vs. *blitt*) was the spelling of the anchor played through headphones. There were 8 possible foils per anchor (see Appendix A). Each trial began with a 750 ms fixation cross, followed by the simultaneous presentation of the auditory stimulus and the two strings on the left and right sides of the screen. Participants indicated their response by pressing the corresponding left and right control keys. Trials timed out if a response was not made within 4000 ms. After a response or time out, the next trial began after an inter-stimulus interval of 250 ms.

The cycle ended with a *spelling-to-sound association* task, analogous to the previous task. Participants had to indicate which of two spoken forms corresponded to the anchor printed on the screen. Each foil diverged from its anchor only minimally, with equal numbers of foils that differed either at onset or offset (e.g., *blit* vs. *blick*). Each trial began with a 750 ms fixation cross, followed by the presentation of the spelling of an anchor for 1000 ms. Participants then heard the two spoken forms one after the other, and then indicated by button press whether the first (left control key) or second (right control key) auditory form was the correct pronunciation. As a reminder, and to increase the parallels between the two association tasks, the numbers “1” and “2” appeared on the left and right sides of the screen as a reminder of which key to use to make a response. As in the prior task, after a response or the trial timed out (4000 ms), the next trial began after 250 ms. In both association tasks, participants were instructed to respond as quickly as possible without making errors.

Each task was run twice during the first cycle, and once during the second cycle, except for mere exposure, which was always run twice. During the first cycle, auditory and visual feedback (accuracy only) was provided in both association tasks. Auditory feedback consisted of a bell or a buzzer sound for correct and incorrect responses, respectively. Visual feedback was also presented for 2000 ms and consisted of changing the color of the written forms (sound-to-spelling task) or number indicating whether the correct pronunciation appeared first or second (spelling-to-sound task) to red if the response was incorrect or to green if correct. The font size of the indicated response was also increased, whereas the font size of the non-indicated response was decreased. During the second cycle, feedback was turned off in both association tasks in order to assess learning. To focus attention on the spelling-to-sound mappings, no other type information (e.g., visual, semantic, emotional, etc.) was provided during learning (cf. Dumay, Feng & Gaskell, 2004).

Test. The testing session occurred 48 hours after the last training session on Day 3. It involved participants from both the no training (control) condition and the training condition. First, participants performed a standard naming task, which required them to name as quickly as possible 19 words and 19 nonwords presented in random order (see Kello & Plaut, 2000). The stimuli used in this task were sampled from the original study, after removing a couple of items duplicated in our set. There was no systematic relationship between the specific items used in standard naming and those used later during test, and the main reason for including the standard naming task was to derive automatically each participant's naming speed (using Checkvocal; Protopapas, 2007). This served as a baseline for tempo naming, which immediately followed. In this second task, each trial started with a 5-to-1 countdown, at one of four possible tempos. The tempo was conveyed by the successive deletions of one of initially five pairs of flankers on the screen, accompanied by five repetitions of a 50 ms/1000 Hz tone (see Figure 3, right). The stimulus appeared between the last two flankers (e.g., > brax <). Participants had to name it in synchrony with when the next tone would have occurred, even at the cost of a mispronunciation. All responses were recorded and, at a later point, sorted in various categories by two raters. To make sure participants followed the tempo, feedback was provided in the form of a scale that indicated the deviation of their response latency (in ms) from the tempo. The four tempos were established by subtracting 150, 100, 50, or 0 ms from the baseline, and were randomly assigned to one of four blocks. Each block contained 30 probes (one probe for each anchor) and 36 filler words and nonwords, rotated across participants so that each participant saw each probe only once. The session ended with one last block with only the anchors, at the fastest tempo (-150). The experiment was controlled using Psychopy (Peirce, 2007).

Results

Training. Prior to analysis of the sound-spelling and spelling-sound matching tasks, we eliminated all trials with latencies below 200 ms or above 3000 ms, or that were more than 2.5 standard deviations above or below the mean latency for that item type in a given session for a given participant. This eliminated 3% of the data. In both association tasks, accuracy was near ceiling (> 98%), while latencies decreased by an average of 316 ms across sessions. Thus, participants rapidly learned the 30 new orthographic anchor strings and their pronunciations. Variability across item types was small relative to the training effect, as shown in Figure 4 and Figure 5, respectively. Additionally, Figures 4 and 5 show that there was little constancy in the rank ordering of performance for the three item types, or in how these rank orderings line up with the amount of regularization observed for each item type in tempo naming (Figure 6, described below). Collectively, these observations preclude differential performance across item types during training as an alternative explanation of the tempo naming results.

Test (tempo naming).

Data Screening. Prior to analysis, we dropped three ambiguous anchors, along with their associated probes, as well as three ambiguous probes and three exception probes. These items were eliminated because of the similarity between many of participants' responses and a high frequency word, which made it difficult to delineate between the two and which was likely to have biased pronunciation (e.g., the pronunciation of ambiguous anchor *brear*, trained to rhyme with *rare*, and the actual word *rare*). In the naming and tempo naming tasks, participants who had poor automatic onset detection rates (> 50%; 1 participant from the training condition, 5 from the no training condition) or atypically high numbers of responses that were neither correct pronunciations based on the learned items or regularizations thereof (> 50%; 1 participant from the training condition, 2 from the no training condition; or < 2.5 SD below the performance of the other participants, 1 from the no training condition) were removed. This eliminated 2 participants in the main experiment and eight controls. All empty recording files were then removed (0.5% of trials) before screening individual trials and dropping those outside of 2.5 standard deviations of the mean for a given participant, block, and item type (0.4% of trials).

Our main results relate to the proportion of regularized responses (*suff* rhyming with *cuff*) as compared to responses that were training-consistent and reflected a newly learned pronunciation (*suff*, trained to rhyme with *roof*, not *cuff*), or a spillover from it (e.g., to neighboring nonword *druff*). Regularized pronunciations (see Appendix A) were defined as the most frequent pronunciation of each anchor and probe by an implementation of the Plaut et al. (1996) model, which has been shown to yield human-like pronunciations of words and nonwords (Plaut et al, 1996, pp. 69-70). In addition, use of this definition facilitated comparison of the behavioral data with model performance. These two types of responses (regularized and training-consistent) accounted for 94% of the data and therefore provide a relatively clean and transparent index of representational warping. The vast majority of errors were pronunciations that did not fit into one of these categories (e.g., *kek* for *kest;* 5% of all trials), with most of the remaining errors being stutters or silence (< 1% of all trials). Critically, error rates for each type of probe and target were essentially identical across the two conditions regardless of training (all error rate differences < 2%, except for ambiguous anchors, at 4%). Thus, including these trials in the subsequent analyses only adds a small amount of noise without changing the patterning of the main effects of interest, which involve modulations of regularization rates of at least 25%.

As can be seen in the left graphs (top and bottom) in Figure 6, training had the intended effect of inhibiting regularization of ambiguous and exception anchors in the training condition (where a new pronunciation was learned), but not that of regular anchors.[2] These data, in conjunction with the data from the training tasks, further confirm that both exceptions and ambiguous anchors were processed with a similar degree of proficiency. In contrast, the no training condition showed near ceiling regularization rates for regulars and exceptions, and considerably higher regularized responses for ambiguous anchors than the training condition. These patterns of responding in the training condition are consistent with a vast body of prior computational and empirical research indicating that both the dominant and alternative pronunciations should be produced in relatively high numbers for ambiguous items. However, only a single pronunciation should be produced for regular items (which in the no training condition also should include the exception items because they correspond to regular consistent nonwords from Glusko, 1979; see also Plaut et al., 1996; Seidenberg & McClelland, 1989 for related computational simulations).

To provide quantitative support for the overall patterns observed in Figure 6, we conducted a series of mixed-effect model analyses that compared the effects of training on performance on each item type. We begin by describing the omnibus model, and later elaborate on how we pared down the model to run additional targeted comparisons. The dependent measure was whether a pronunciation was pronounced in the training consistent manner (coded as 0) or was regularized (coded as 1), which we modeled using a binomial distribution. In terms

[2] Because of the very low error rates, the overall difference in regularization rates between the training and no training conditions is due to participants in the training condition producing training-consistent responses (e.g., reading *suff* as a training-consistent rhyme *roof*, not the regularized rhyme *cuff*).

of independent measures, training was included as a fixed factor with the no training condition acting as the baseline level. Thus, the sign of the training regression slope indicates the change in performance relative to baseline (positive = increase relative to baseline, negative = decrease relative to baseline). Similarly, item type was also included as a fixed effect, with the regular items serving as a baseline for two separate contrasts with the other item types (regular vs. ambiguous; regular vs. exceptions). Both training and item type were allowed to interact.

Additionally, to rule out possible confounds, we also included fixed effects of orthographic neighborhood size, as measured via orthographic Levenshtein distance[3], length in letters, and positional bigram frequency, each of which was allowed to interact with training. The inclusion of these covariates in the same model that tests for the critical interaction between item type and training model rules them out as an alternative explanation of the effects. Additional details regarding the covariate analyses are presented in Appendix C.

Model convergence issues prevented the omnibus analysis from being run on the anchor and probe data simultaneously, even when the fixed effects related to potential confounds were removed, so we conducted separate analyses of the anchor and probe data. Pilot analyses revealed that models which included random slopes often failed to converge (particularly in the case of the smaller set of anchor data). Therefore, random slopes were not included in any of the models to facilitate comparisons across models (for discussion of this analytical approach, see Barr, Levy, Scheepers, & Tily, 2013; Bates, Kliegl, Vasishth, & Baayen, 2015). In the cases where slightly more complex random slope structure could be added (e.g., in the larger set of

[3] Orthographic Levenshtein distance is an alternative measure of how dense the lexical neighborhood is and counts words that can be created by adding, removing, or substituting letters in a given word. This measure has been shown to be a sensitive predictor of performance (Yap et al., 2009) relative to the classic measure of orthographic neighborhood size (Coltheart's N) which only counts neighbors created by substitution (Coltheart et al., 1977).

probe data), the same patterns of significance emerged, so the choice to perform separate analyses does not distort the critical comparisons to a substantial degree.

Anchors. Analyzing performance on the anchors allowed us to examine the strength of the mappings acquired (or reinforced) during training, on the new words themselves (i.e., independently of generalization). The contrast between regular anchors and exception anchors was not significant ($t < 1$), whereas the contrast between regular anchors and ambiguous anchors indicated that regularizations were significantly less frequent for ambiguous anchors overall (b = -2.81, SE = 0.54, n = 1904, z = -5.2, p < .001). Corroborating the visual impressions that regularizations were vastly less frequent for exceptions and ambiguous words in the training condition, the interactions between the item type contrasts (regulars vs. ambiguous; regulars vs. exceptions) and training indicated that regularizations were significantly less frequent for ambiguous anchors (b = -2.99, SE = 0.50, n = 1904, z = 6.0, p < .001) and for exception anchors (b = -4.56, SE = 0.54, n = 1904, z = - 5.8, p < .001) in the training condition. To evaluate whether this reduction in regularization rates was significantly greater for the exception items relative to the ambiguous items, we re-ran the same model after resetting the baseline (intercept) for the item type contrasts to the ambiguous anchors. This allowed us to test the interaction between ambiguous and exception anchors and training directly, which confirmed that differentially less regularizations were observed for exception items in the training condition (b = -2.23, SE = 0.36, n = 1904, z = -6.18, p < .001). In other words, training reduced regularization responses for the exception anchors to a greater degree than for the ambiguous anchors. To determine whether the greater decrease in regularizations for the exceptions led to equal regularization rates for ambiguous and exception items in the training condition, a follow-up analysis was conducted. The results indicated that there was not a significant difference in regularization rates between the ambiguous and exception anchors in the training condition, although the exceptions were regularized marginally more frequently (*b* = 1.20, *SE* = 0.66, *n* = 547, *z* = 1.82, *p* = 0.07). The similar regularization rates for exception anchors and ambiguous anchors rules out the possibility that any differences in regularization rates in the probes, which test of the effects of warping on generalization, are due to different regularization rates in the anchors.

In sum, the analyses corroborated the initial visual impressions concerning performance in the no training and training conditions. In the no training condition, regular and exception anchors were essentially always regularized and ambiguous items were regularized the majority of the time. In the training condition, participants learned to produce training-consistent pronunciations for exception and ambiguous anchors to a high and similar degree. Regularization rates for regular anchors, in contrast, remained near ceiling.

Probes. The key predictions from connectionist models regarding generalization fromnewly learned word forms were tested in the probe data (Figure 6, right). To this end, the same mixed effect models used in the analyses of the anchors were used in the analogous analyses of the probes. The central prediction that falls out of the warping mechanism is that learning a new representation that violates the regularities of the domain, in this case spelling-sound mappings, should impact neighboring regions of the representational space, as indexed by nonword probes that are orthographic neighbors of the anchors. Additionally, the warping should be highly restricted to a particular local area around the newly learned representation if the violation of the regularity is relatively unique (as in the case of exception words) whereas the effects of warping will extend to a larger portion of the representational space if multiple items are inconsistent with the overall regularity of the domain (ambiguous items).

Critical to the predictions of spillover of a newly learned inconsistent spelling, the interaction between training and item type was significant, with reduced regularization rates for both item types in the training condition (interaction with ambiguous probes: *b* = -0.74, *SE* = 0.22, *n* = 7199, *z* = -3.41, *p* < .001; with exception probes: *b* = -1.72, *SE* = 0.21, *n* = 7199, *z* = -8.35, *p* < .001). To test whether there was less regularization for the exception items relative to the ambiguous items in the training condition, we ran a follow-up comparison including only the ambiguous and exception data from that condition, with the ambiguous items serving as the baseline for item type (the regulars were removed because of a convergence warning). This model showed a greater relative decrease in regularizations for the exception probes compared to the ambiguous probes (*b* = -0.74, *SE* = 0.19, *n* = 4366, *z* = -3.86, *p* < .001). Finally, we tested the critical predictions of warping on the probes after training by comparing the regularization rates for regular, ambiguous, and exception probes in the training condition only. This model showed that relative to regular probes, there were fewer regularizations for exception probes (*b* = -1.97, *SE* = 0.29, *n* = 3382, *z* = -6.89, *p* < .001) and for ambiguous probes (*b* = - 4.04, *SE* = 0.34, *n* = 3382, *z* = -11.97, *p* < .001). To determine whether there was significantly less regularization in the training condition for the exception probes relative to ambiguous probes, we re-ran the same model after re-leveling item type to use ambiguous words as a baseline. This model showed that there were significantly more regularizations of exception probes relative to ambiguous probes (*b* = 2.07, *SE* = 0.30, *z* = 6.97, p < .001).

Thus, despite equal regularization rates for the ambiguous and exception anchors, the frequency of regularization was substantially greater for exception than ambiguous probes, which is precisely what warping predicts: anchor learning should result in the formation of subordinate representations more easily in the ambiguous than in the exception condition.

According to the warping account, this is because the subordinate mapping (*bive* rhyming with *live*, not *drive*) has already been established by other words (e.g., *give*). Consequently, the representational space needs to be warped less to accommodate the new anchor. This, in turn, results in greater generalization to the training-consistent pronunciation rather than to the dominant (regularized) pronunciation. That the pattern across item types is virtually identical across the four tempos (Appendix C) demonstrates the reliability and stability of the findings.[4]

Latency. As was our aim in using the tempo naming task, participants were sensitive to the tempo and responded more rapidly when the tempo increased, from 652 ms (*SE* = 4 ms) at the slowest, to 511 ms (*SE* = 3 ms) at the fastest tempo. Tracking the tempo in this manner also biased performance so that the effects of training manifested themselves primarily in the regularization rate data, described below, rather than in the latency data, which were similar across item types for both regularized responses and training-consistent responses at every tempo. For completeness, however, we have included the full analyses of the latency data in Appendix C.

In summary, we observed a number of strong effects of regularity as a function of training, which held up when potential confounds were included as covariates. Representational

[4] The regularization rates were found to not vary as a function of tempo in additional analyses of the probe data including item type, tempo and training, and all interactions (all *zs* < 1.63, *p*s > .10). These models generated convergence warnings when orthographic Levenshtein distance, length in letters, and bigram frequency were included and allowed to interact. Given these issues and their reduced comparability with the analyses of the anchors, which only occurred at one tempo, we focus on the simpler models. The absence of an effect of tempo on regularization rate in our data was likely caused by our participants responding slightly more slowly in the baseline naming task. This is consistent with the Kello and Plaut (2000) data, which had faster overall naming latencies and which only showed significant effects as a function of tempo for the fastest tempo conditions. At any rate, the absence of effects of tempo on regularization rates bears no relevance on our core findings and claims.

warping predicts these patterns of generalization as a function of word regularity, and therefore appears to be the most plausible and parsimonious explanation of the data, as is further supported by the simulation reported next.

Simulation

The strength and consistency of the generalization effects observed in the behavioural experiment provide strong initial support for warping as neurocomputational mechanism. To seek converging evidence that warping underlies these large effects, we assessed whether the Plaut et al. (1996) model can simulate the behavioral data (i.e., those reported in Figure 6). As mentioned in the Introduction, the notion of warping originated from a detailed analysis of this model (Kim et al., 2013). To the extent that it is a viable approximation of representation formation in readers, and in quasiregular domains more generally, the model should reproduce the qualitative patterns found in the data using the same anchor and probe stimuli. We did not engage in any quantitative comparison, fitting the model to the empirical data or performing inferential statistics on the model-generated data. To do either in a meaningful way requires also modeling participant variability, which has yet to be introduced into the current model. In addition, in model evaluation, it is most important to first establish a good qualitative fit.

The Plaut et al (1996) model is a three-layer, feed forward neural network that contains 105 grapheme input units, 100 hidden units, and 61 phoneme output units. The hidden units make it possible for the model to learn complex mappings between input and output that are necessary in a deep orthography like English, and warping occurs in the connections between layers. We used the implementation of the model developed by Kim et al. (2013), which was shown to perform equivalently to the Plaut et al. version. To ensure the stability of our results, the simulation results that we report are averages over 50 different instances of the model, each initialized using different random initial weights. Because of the very small variability across the different instances of the model, error bars were imperceptible and have been omitted from the simulation plots.

Our simulations began by training the model on the spelling-sound correspondences of 2998 English words for 400 epochs. Default parameter settings were used throughout the simulation, except for some minor adjustments when the anchors were introduced, which are noted below. Thus, as in prior work, the model learnt by receiving feedback (i.e., cross-entropy error) on pronunciation errors. Error was scaled by each word's log-transformed word frequency (ln(2+frequency), obtained from Kucera and Francis,1967). Word frequencies ranged from 1 to 69,971 words per million, although the bulk of words had frequencies between 1 and 42 (frequency for 25th percentile: 2; 75th percentile: 42). Throughout the simulation the global learning rate was .0008. To speed overall learning, momentum was enabled and set to .9 as of epoch 10, and connection-specific learning rates were tuned using the delta-bar-delta method (initial multiplier = 1.0, rate increment = .1, rate decrement = .9; Jacobs, 1988). All connections were subject to a small amount of weight decay (.00001). We treated the model’s representations at epoch 400 as corresponding to those of our no-training control participants and measured the network’s regularization rates for all three types of anchors and probes (similar performance was obtained if we trained the model for an additional 50 epochs without altering the initial vocabulary). The most active vowel was taken as the network’s pronunciation of that vowel (as in Plaut et al., 1996 and Kello & Plaut, 2003). The model was considered to have made a regularized response if the most active vowel after training was the same as before training. The model was considered to have made a training-consistent response if the most active vowel was that of the trained pronunciation. Collectively, these two response categories accounted for the vast majority of the model's responses, and relative changes in these response categories represent the critical test of warping.

At epoch 401, the 27 anchors with novel target pronunciations were introduced to the training corpus and the model was run for additional 50 epochs to learn to pronounce the new words. To encourage the rapid learning of the anchors and the uniform preservation of all previously learned word knowledge, the error for all of the words in the initial corpus were hereafter scaled by ln(2), whereas the error terms for the anchors were scaled by ln(10+2). The word frequency for the anchors was selected from pilot simulations to strike a balance between learning the new regularities rapidly (lower values required more training) and avoiding catastrophic interference when anchor word frequencies were substantially higher than for the all other words in the corpus. The exact value was not critical and a range of values could be used to generate similar results. The only other change was to reset the connection-specific learning rates using the delta-bar-delta method to their initial values, because those tuned values were not necessarily valid with the addition of the anchors to the corpus.[5]

To simulate performance in the training condition, regularization rates of all three types of anchors and probes were compared at epoch 450, which was chosen because (1) it represents a point at which the new pronunciations of the ambiguous and exception anchors were learned reasonably well in the training condition and (2) regularization rates stabilize around this epoch.

[5] The delta-bar-delta method speeds learning by making larger changes to a connection that changes in a consistent direction across epochs (i.e., a connection that is either consistently increasing in strength or decreasing in strength), and smaller changes otherwise. The addition of the anchors to the training corpus meant that the evidence accumulated over the initial 400 epochs indicating that a very small or very large change to a given connection was justified was no longer valid, however. Thus, resetting the learning rates avoided impairing learning of the expanded corpus due to making very large changes to a connection when very small changes were needed, or vice versa.

As is the case for our other parameter choices, the exact epochs selected, including whether the anchors are introduced earlier or later in network training (-50 or +50 epochs) are not critical; other time points could be sampled to generate similar results, as we elaborate on below. Also, we note that introduction of anchors caused minimal (1.2%) and only short-term forgetting of training words in the model. By epoch 450, the model correctly pronounced all words again.

The mean proportion of regularized pronunciations produced by the model for the anchors and the probes are presented in Figure 7. Because no substantial changes in regularization rates were observed in our behavioral data, likely because our participants responded slightly more slowly in the regular naming baseline task than in Kello & Plaut (2000), we did not attempt to simulate changes in pronunciation as a function of speed by varying input gain (as in Kello & Plaut, 2003). Overall, the model's performance was qualitatively similar to that of the participants, as can be observed by comparing the behavioral regularization rates in Figure 6 with the simulation regularization rates in Figure 7. Starting with the anchors, in the no-training condition the model showed perfect regularization for regulars and exceptions, whereas ambiguous words were regularized 79% of the time. This data pattern paralleled the behavioral data (Figure 6), although the ambiguous items were regularized at a quantitatively higher rate in the model compared to the behavioral data. In the training condition, the model approximated the frequency of regularization across all three items types. Regularization was at ceiling in the simulation and behavioral data for the regular anchors, and is low for both the ambiguous anchors and exception anchors. A minor discrepancy is that the model regularizes exceptions slightly less frequently than ambiguous words (8% vs. 29%, respectively), although this discrepancy was small relative to the substantial changes in regularization rates due to training.

The model's propensity to generalize to novel instances of the three word types also paralleled that of the participants. As expected, in the no-training condition regularization of regulars and exception words was near ceiling in both the simulation and the behavioral data. The ambiguity inherent in pronouncing ambiguous words persisted and dropped slightly when pronouncing ambiguous probes; participants exhibited a similar drop as well. In the training condition, regularization of the regular words remained virtually unchanged in both the simulation and behavioral data. For the exception probes, the crucial drop in regularization that was seen with participants was also produced by the model, albeit visibly smaller. For ambiguous probes, regularization rates dropped in the simulation and behavioral data, and by a similar amount in both.

Taken together, the standard Plaut et al. (1996) model simulated the behavioral results relatively well and captured all of the qualitative patterns in the behavioral data. Importantly, the model simulated the key finding that exception learning impedes regularization of probes as a function of anchor regularity using the stimuli of the behavioural experiment. The learning mechanism in which the representations of exceptions are warped is responsible for the reduced ability to generalize. Warping is highly local to neighbors of the anchor (Kim et al, 2013), which is why generalization of the newly learned pronunciation is low and regularization itself is high. Inevitably, there are some differences at a quantitative level either in terms of the initial no-training regularization rates or in the relative change in regularization rates after training. But, as noted previously, given the vast differences between the simulation and the human participants on many fronts (e.g., total vocabulary size, amount of consolidation of existing vocabulary vs. the anchors, age of acquisition effects, speed of response, etc.), the similarity between the behavioral data and simulation data provides important general validation that a representational mechanism akin to warping underlies readers' abilities to simultaneously generalize regularities and learn exceptions.

Before turning to the general implications of the simulation and behavioral work, however, we provide a broader picture of how the model learned the anchors and how this knowledge was generalized to the probes. Figure 8 graphs the rates of regularized and training-consistent responses from epoch 350 to 500. The red vertical bar denotes when the anchors were added in the training condition. Values immediately before this point designate performance in the no-training condition, and values after it show how training impacted regularization as a function of training. The behavioral data (from Figure 6) are plotted for reference at Epoch 400 (before the introduction of anchors) and at 450 (after 50 epochs of anchor learning). These two time points correspond to the no training and after training snapshot of model performance presented in Figure 7.

Several insights about model behavior can be gained by examining this plot. First, for both the probes and the anchors, there is a trade-off between regularizations and training-consistent responses, such that reduction in regularized responses is being replaced by training-consistent responses. This is to be expected for the anchors given that the mapping between spelling and sound for those items was trained specifically. It is more telling in the context of the probes, for which the trade-off between the two types of responses indicates a very specific re-shaping of the representational space consistent with the predictions derived from the warping mechanism. If learning the new anchors had reshaped the representational space in some other way, this trade-off between the two response types would not have occurred, and other types of responses (e.g., production of a third alternative pronunciation) would have been found, instead.

A detailed inspection of the relative changes in trajectories for the anchors and probes provides insights into the structure of the representations that are being learned and how they underlie generalization. To begin, consider the anchor data on the left side of Figure 8, which depicts the effects of learning new words through explicit training. In the case of regular anchors, which are consistent with the statistics of English, no training is actually required to pronounce those items correctly—that is, the network can already generalize from existing word knowledge to pronounce those items correctly. In the case of the ambiguous and exception anchors, the network produces regularized responses prior to training and the network must now learn to produce training consistent responses. This is particularly the case for the exception anchors, which before training were regularized 100% of the time, and less the case for the ambiguous anchors, for which the training-consistent response is already familiar to the network and produced 29% of the time. Why the difference? As the functions show, the exception anchors are effectively regular items up until training on the new anchors begins, and therefore exhibit ceiling levels of regularization. In contrast, the presence of a competing pronunciation of ambiguous words causes regularization to be below ceiling and training-consistent pronunciations to be above floor. The difficulties of violating established regularities in the case of the exceptions anchors is reflected in the additional training needed to produce training-consistent responses instead of regularized responses. Both of these changes require altering the network's pre-training behavior to a larger extent for the exception anchors relative to the ambiguous anchors. By the time the network has received extensive training on all items at epoch 450, however, all item types are associated with near ceiling rates of training-consistent responses, and this pattern persists even with additional training.

Next, we turn to the probe data on the right side of Figure 8, which shows how the effect of learning the new anchors is intertwined with how the newly learned pronunciation generalizes. Prior to the introduction of the anchors on epoch 400, the probes behave similarly to the (as yet untrained) anchors (compare regularization rates for both at this point in training). After the anchors are introduced at epoch 401, for regular probes, the network continues to produce responses that are both training-consistent and regularized because these items remain consistent with the overall regularities of English. The patterns are quite different for the ambiguous and exception probes, however. For both of these item types, the changes in the training-consistent and regularized responses associated with learning the anchors carrying over to the probes. This *echo* is particularly faint for exception probes, which show only a small increase in training-consistent responses and a small decrease in regularized responses, but is stronger for ambiguous probes (again, compare functions across adjacent graphs). As was the case for the anchors, the response rates remain relatively stable from Epoch 450 onward. The stability of the results after this point demonstrates that the effects of warping are not transient. Warping has a permanent effect on generalization, which as the simulations suggest, is intimately tied to learning.

In addition to demonstrating the relatively rapid change in response patterns for our anchors and probes and the temporal stability of our overall pattern of results, the learning trajectories also reveal an intriguing transient dynamic in the learning time-course between epochs 400 and 450. Although the same qualitative patterns of less regularization for ambiguous and exception anchors and probes are present throughout the learning trajectory following the introduction of the anchors, the probes show the highest rates of training-consistent generalization responses around epoch 430. This is well before the anchors have reached their asymptotic levels of training-consistent responses, although learning of the training-consistent pronunciation of the anchors appears to slow substantially in that time window as well.

Given the complementary and symmetrical relationship between regularized and training-consistent responses, these dynamics appear to reflect the gradual emergence of a warped representation as a result of learning the new inconsistent pronunciation amid other words in the local neighborhood. Initially, no explicit representation is encoded in the specific location in the representational space for an ambiguous or exception word, so distorting the representation of knowledge at this specific location of the representational space can proceed unimpeded. Because of the graded and continuous nature of the internal representations of the model, however, the warping involved in representing the inconsistent representation gradually spreads out to impact surrounding words that do not share the inconsistent pronunciation. Given that these words are still part of the model's vocabulary and training regime (consistent with complementary learning systems theory to avoid catastrophic interference; McClelland et al., 1995), the existing word representations push back to undo this overgeneralization and preserve their pronunciations.[6] Consequently, fully learning the ambiguous and exception anchors proceeds more slowly because a very locally constrained nonlinearity in the representational space is needed to correctly encode the inconsistencies while minimizing broader contamination.

As a final observation, it is worth noting that during the transient period when the new anchor representations are stabilizing, the network passes through a period of time (Epochs 430-440) in which the ambiguous and exception anchors both produce similar, higher rates of

[6] The residual effects of this push-back remain apparent in the simulation if, instead of examining the model's "response" in terms of whether the training-consistent or regularized pronunciation was activated most strongly, the activity of the training-consistent response is examined instead. This reveals that even when the network's is always activating the training-consistent response to a higher level than the regularized response, the absolute level of activity in the training-consistent response plateaus at a lower level for the exceptions than for the ambiguous items. This is consistent with the predictions derived from warping concerning greater push-back from neighbors of exceptions and reduced spill-over of the exceptional pronunciation.

regularized responses (~25-30%). This post hoc observation opens the possibility that our participants also transiently generated comparable rates of regularized responses when the ambiguous anchor regularization rate is beginning to stabilize but the exception anchor regularization rates is still descending toward a lower final level. Of course, having only measured the behavioral effects at one time-point after training, and having no independent method for equating training epochs in the model with the effects of training elicited with our behavioral tasks, drawing strong conclusions in this vein would be premature. However, this alignment suggests that additional value can be derived from sampling multiple time-points during learning to understand better how it leads to a lasting effect of warping.

General Discussion

How humans encode quasiregularity is a central question in the cognitive sciences. The ability for single-route connectionist models to accommodate both regular and exception items via a warping mechanism is the means by which connectionist (PDP) networks learn to structure knowledge. We here identified a causally induced behavioral signature of warping, by means of a word learning paradigm testing for generalization of a newly learnt (or re-learnt) pronunciation in reading aloud nonword orthographic neighbours. We found different amounts of generalization (spillover) for exception and ambiguous pronunciations, despite equally strong learning of the pronunciation of the novel anchors.

The broad patterns of empirical effects were also reproduced in simulations of new word learning, thereby providing explicit evidence that representational warping is a viable explanation of our core findings. The simulations also highlighted the value of building models that integrate theories of learning and representation, suggesting how these facets of cognition can interact to produce novel predictions about how representations gradually take shape. Such insights can provide valuable guidance for targeted empirical research that is not available from “static” models simulating only the end-state of learning (for related discussion, see Lerner, Armstrong, & Frost, 2014).

Despite a relatively high level of agreement between the simulated and behavioral data (the qualitative patterns are the same), not all of the simulation results line up exactly at the quantitative level. Such discrepancies are not unexpected given our coordinated computational and empirical research strategy. We explored representational warping using a modification of an established connectionist model that maps between spelling and sound. This test bed allowed us to hone in on warping in sublexical representations but necessarily does not do full justice to the many differences between human participants and the model that could further color performance (e.g., vocabulary size, age of acquisition effects, relative amounts of training, contributions from episodic and semantic memory, response strategies, word frequency, etc.). By couching our work within the domain-general connectionist framework, our approach can be naturally extended to examine how warping interacts with these and other aspects of word learning and representation. As a first example, the presence of warping in the Kim et al. (2013) simulation, in which all words were included from the onset of training, and in the present simulation, in which the anchors were introduced later, points to a role for warping regardless of age of acquisition. Future work, however, could target how warping could interact with age of acquisition in more detail. For example, how does warping differ depending on when exception words are learned relative to the regularities in the language? Such investigations should have implications for many issues of long-standing and broad theoretical import.

Alternative Models of Reading. Of course, a connectionist account is not the only way in which to think about our findings. Learning, representation, and generalization are fundamental issues relevant to any mechanistic account of reading and other similar quasiregular domains, and we expect that our paradigm and results can be informative in shaping understanding in other frameworks as well. In focusing on how a connectionist model maps between orthography and phonology, our aim has not been to show how such a model is necessarily the only account of the data. Rather, it has been to understand how such a model could overcome the competing pressures of generalizing regulars and representing exceptions within a single representational space. Nevertheless, it is useful to consider how other models could explain the current findings.

The most long-standing and popular alternative to a PDP account is the DRC model (Coltheart et al., 2001), which relies on two qualitatively different mechanisms to read, direct lexical access to whole word forms and grapheme-phoneme correspondences (GPCs) linking individual letters or letter clusters to specific phonemes. Of interest to the current study is how the model pronounces newly learned words and neighboring nonwords. To date, most computational investigations using the DRC have focused on proficient reading without considering how learning shapes performance (but see Coltheart, Curtis, Atkins, & Haller, 1993; Seidenberg, Petersen, MacDonald, & Plaut, 1996). This is because principles of learning and representation have yet to be developed in that framework in the way that they are in the PDP framework. Clearly then, it would be unfair to evaluate the DRC in the context of our findings. Instead, we ask how the DRC might be updated.

Several means of achieving such integration appear possible, a couple of which are laid out below, although each would need to overcome some challenges as well. Most critically, however, and overarching these different possible modifications, is the requirement that learning must be modulated in a way that allows for a graded decrease in generalization as a function of the spelling-sound consistency of a newly learned word—which is our core finding. That is, the model would need to move towards treating learning and representation as two integral facets of the proficient reading system.

One possibility that is available to the DRC, but not to the PDP model that we have used, would be to allow for new lexical representations to strongly shape the pronunciation of nonword probes. Such an influence from the lexical pathway is, in principle, possible already, although analysis of the implemented model has shown that, in effect, the DRC reads nonwords strictly on the basis of GPCs (Zevin & Seidenberg, 2006). A new parameterization of the model would need to be developed that simultaneously satisfies two criteria: First, the model's ability to read nonwords in general using the sublexical route cannot be impaired. Second, the model cannot leverage the lexical route to pronounce words and their neighbors in such a way that performance is shaped by lexicality list effects, because tempo naming is insensitive to such effects (Kello & Plaut, 2000). Alternatively, the GPCs could be adjusted via learning so that nonword neighbors can be read via that pathway. At first glance, this would appear to be a very similar solution to that adopted by the PDP network. However, in the details, GPC rules are quite different, in that they reflect the core statistical structure of the language, only at a very low grain size (usually 1-2 letters per phoneme). Consequently, teaching participants a single new exception anchor should not be sufficient to override the overall GPCs of the entire language. A larger, more flexible GPC grain size could allow for new learning to occur in select GPCs without altering the representation of the language as a whole. Indeed, effects consistent with a larger grain size of representation are reported for both spelling-sound and sound-spelling correspondences, where context sensitivity increases with exposure to the language (Treiman & Kessler, 2006; Treiman et al., 2006).

A more recent dual-route model, CDP++ (Perry, Ziegler, & Zorzi, 2010), has already adopted an architecture more in line with the preceding proposal, and so is likely to also be able to accommodate our finding. In particular, CDP++ employs an architecture for representation formation and generalization in the sublexical pathway that is very similar to that in the Plaut et al. (1996) model. Consequently, in principle, it likely can produce similar performance to our simulation by relying only on the learning and representation principles in that pathway, just as we have done. However, the more complex dual-mechanism architecture of CDP++, which also includes lexical representations and processing dynamics, could potentially offer an account of our data that leverages both pathways. Whether the model would actually do so while still simulating other phenomena such as an absence of list effects—and more importantly, whether such an account offers an advantage over a single-mechanism account—is an open question for future work.

Our generalization test could be leveraged to differentiate these alternative theoretical accounts. For example, at one extreme, an even stronger demonstration that our results are necessarily sublexical in nature could involve training participants on two anchors for each pronunciation half as often. This would lead to the same amount of practice of the sublexical component of the representation, but would weaken possible lexical contributions (cf. Dumay & Gaskell, 2007, 2012; Qiao & Forster, 2013).[7] In contrast, increased practice with the existing set of anchors and the use of standard naming, in which latencies are slower and in which lexicality

[7] This manipulation would also make the exception items slightly less exceptional, although we would still expect the overall contrasts between the different item types to hold.

list effects have been observed (Andrews & Scaratt, 1998), should increase the contributions from the lexical pathway according to dual route accounts. Pursuing such a line of research would therefore help understand data focused on subtle differences in distinct but related factors on warping (e.g., proportion vs. frequency of neighbors) that have proven challenging for PDP and dual-route accounts. The success or failure of each account in this context could be especially useful for advancing all accounts, in addition to evaluating the necessity of lexical representations in proficient reading.

Language Learning. The common learning, representation, and processing mechanisms of the connectionist framework also imply that warping is relevant to other aspects of language and cognition beyond proficient reading in English. For example, the current results should have broader implications for domains such as first and second language instruction and studies of cross-linguistic differences. In the case of first language instruction, children learning to read must discover the regularities underlying the mapping between print and sound to facilitate new word learning and to generalize this knowledge to new unfamiliar words. This process has been shown to be enhanced when a regularity such as a particular pronunciation of a vowel is embedded is a training list with greater variation in spelling (e.g., the “a” in *fan*, *pat*, *pal*, *lap*, *ram*, *cab*), as compared to words with more similar spellings (e.g., *bat*, *hat*, *pat*, *cat*, *pal*, *bad;* Apfelbaum, Hazeltine, & McMurray, 2013). Considered in the context of the present work, these findings are consistent with the notion that increased variability reinforces a broad generalization of the vowel, and a commensurate reduction of warping, just as we observed for regular words.

Similarly, second language learners strive to extend regularities to new words without sacrificing accuracy for exceptions. How efficiently they are able to do so depends on the interaction between several factors, including whether the initial training items are regulars or exceptions, and on the consistency (opacity) of the native and second languages (Ijalba & Obler, 2015). To improve the efficiency of L2 acquisition, we must understand the structure of the different languages, and how these structures will be internalized via learning, as well as the constraints of the native language. A detailed understanding of quasiregular learning and generalization can serve as a basis for identifying training vocabularies with optimized proportions of regular, ambiguous, and exception items. This could help learners carve their internal representations of the language to maximize accuracy and generalization as efficiently as possible, while simultaneously providing new insight into how two different quasiregular domains can coexist in a single representational store. These predictions could be readily tested.

Similarly, the potential power and flexibility inherent in the warping mechanism is well illustrated by considering the role of warping as a function of the regularity of the domain in question. At one end of the regularity continuum, some mappings, such as those mediating between orthography and phonology in transparent languages (e.g., Serbo-Croatian), are unambiguous (i.e., regular) and single graphemes always map to single phonemes. In this case, no warping is necessary and generalizations can be extremely broad without being constrained to specific warped local neighborhoods around context-sensitive representations of vowels, as in our case (Treiman, Mullennix, Bijeljac-Babic, & Richmond-Welty, 1995; see also Peereman & Content, 1997; Treiman, Kessler, & Bick, 2003; Treiman et al., 2006, for evidence that in English, spelling-sound correspondences are more dependent on local context than as predicted by GPCs).

At the other end, the mappings between surface forms and meaning illustrate a case where extreme warping is needed to ensure that similar surface forms do not activate the meanings of their neighbors to a substantial degree (a similar argument applies to mappings between surface forms and episodic memory). This is because the mappings between the structure of the surface form and these other representations is arbitrary in nature, as opposed to quasiregular (for discussion, see Plaut, 1997; Frost et al., 2015; McClelland et al., 1995). Extreme warping is necessary to insulate representations from this arbitrariness and avoid incorrect generalizations. In other words, the representational system gains an advantage rather than pays a price by using extreme warping to reduce generalization. For example, without warping to constrain the spread of a surface form to meaning mapping, the word CAT would strongly activate an extremely broad set of features related to many living (e.g., RAT) and nonliving things (e.g., CAR, HAT). Thus, the relationship between newly learned representations and the overall regularity (or lack thereof) in a domain can interact to produce a rich set of possible outcomes despite relying on the same underlying computational principle. The degree of warping implicitly determines how existing representations are generalized, and flexibility would seem to be advantageous to capture the extent of generalization across domains.

These different degrees of overall warping needed depending on the regularity, quasiregularity, or arbitrariness of the domain undoubtedly have important roles in shaping processing dynamics within a “full” triangle model that includes semantic representations. In such models, a standard conclusion is that semantic processing can contribute to processes such as naming, particularly in the case of exception words (e.g., Balota, Cortese, Sergent-Marshall, Spieler, & Yap, 2004; Harm & Seidenberg, 2004; Strain, Patterson, & Seidenberg, 1995). This outcome suggests that despite the general arbitrariness of surface form to meaning mappings, there are likely some detailed distinctions in how exception words map with semantics, possibly in terms of the detailed characteristics of the warped representation (e.g., in terms of exactly how far the warping extends and how extreme it is). Coordinated computational and empirical work targeting these issues may therefore reveal important subtleties beyond the initial characterization of warping that we offer here, for instance, in terms of how semantics can differentially shape the pronunciation of exceptions. Combining other tasks such as standard naming with our own tempo naming paradigm, as well as associating specific new meanings with our newly learned words and studying warping in different languages, may be particularly relevant to this end. Here, we focused on a training task and a testing paradigm that aimed to minimize contributions from representations other than those that map directly between orthography and phonology as an initial, simple test of the warping mechanism. As a theoretical construct, warping would seem to have considerable promise, but it is important that future work evaluate its strengths and limitations comprehensively.

Theories of learning and memory. In broader strokes, the preceding discussion highlights the often underemphasized issue of learning and generalization in theories of representation in quasiregular domains, which abound in language (e.g., speech, reading, grammar, discourse; McClelland, 2015) and permeate other aspects of cognition (e.g., semantic cognition; Rakison & Butterworth, 1998). That is, it is desirable for models not only to explain the performance of a proficient "end-state," but also to explain how learning leads to the formation of more stable representations over time. In cases where the developmental trajectory has been placed at center stage (e.g., the past tense debate; Pinker & Ullmann, 2002; Seidenberg & Plaut, 2014), the field has borne witness to advances not only in the target domain, but in our understanding of the alternative theoretical approaches in general (e.g., symbolic vs. subsymbolic processing).

In this vein, the current findings can inform not only theories of representation and proficient processing in quasiregular domains, but also theories of how newly learned word forms are gradually integrated and consolidated in the lexical system. As noted previously, how regularization rates—including some non-monotonic effects—vary as a function of learning make targeted predictions regarding the more detailed structure of a warped representation and how it comes to be constrained by other words. Similarly, the step-like improvements in performance in our training tasks between training days but not within training days suggests that slower offline learning processes (and possibly sleep-specific consolidation processes) are shaping our effects in important ways. Future investigations of these issues could provide insights for theories of complementary memory systems regarding how knowledge of a quasiregular domain is initially learned and gradually integrated with existing knowledge (McClelland, McNaughton, & O'Reilly, 1995; O'Reilly, Bhattacharyya, Howard, & Ketz, 2014).

The warping account we advance also offers a potential explanation for an important paradox in the statistical learning literature, which includes—but extends well beyond—studies of different facets of language learning (Frost, Armstrong, Seigelman, & Christiansen, 2015). On the one hand, statistical learning is typically considered to be a domain-general mechanism by which cognitive systems discover the underlying regularities of a particular input (vision, sound, speech, multi-modal cuing, motor learning, etc.). For example, participants may be trained to learn that the first two elements in a sequence tend to be repeated, and that the last element is different (i.e., AAB cues), where the elements could be tones, syllables, visual symbols, etc. (e.g., three syllable nonwords such as "leleje, wiwije", etc.). On the other hand, the rates of generalization of a particular regularity like this vary considerably and are most often associated with relatively high degrees of modality (and sometimes stimulus-level) specificity rather than a much more abstract and widely-applicable regularity (e.g., participants would not generalize the latent AAB structure to "jijili"; for a review, see Frost et al.).

Warping can resolve this paradox by elucidating how and why some newly learned representations generalize whereas others do not. Insofar as at least some aspects of representing the newly learned stimuli occur in the modality-specific systems used to encode and derive an internal representation of the stimuli (e.g., visual cortex, auditory cortex), the effects of learning and generalization would be bound to that particular representational space. Moreover, whether the effects of learning generalize to other similar stimuli within the modality (e.g., whether learning one set of tones in an AAB sequence would extend to other novel tones) would depend on how regular that input appears to be. If there is only a limited amount of consistent structure across items, participants may generate warped representations that show minimal generalization. For example, infants that learn the auditory forms “leleje, wiwije, jijije” generalize what they have learned to other words ending in “je” but not to other words with the same overall AAB structure such as “jijili” because they have no evidence supporting the latter regularity in the input (Gerken, 2006). In contrast, if a consistent regularity is present across a much broader set of representations, a more generalizable representational structure may emerge overall. For example, infants that learn the word forms “leledi, wiwije, jijili, dedewe” will generalize their pronunciations to “dedeje” and other word forms with similar AAB structure because the training set emphasizes the AAB structure and minimizes evidence that this structure is restricted to one specific terminal syllable. Warping therefore offers a flexible single mechanism for accounting for these data without recourse to separate statistics-based and rule-based systems (for discussion and related proposals, see Aslin & Newport, 2012; Christiansen & Curtin, 1999). Similar principles also appear to be involved in semantic cognition in the context of label learning: children will more readily extend a novel nonword label (e.g., “glim”) to other dogs if it has been used to refer to many dogs (i.e., reflects a regularity) versus if it has been used to refer to only a single dog (and is therefore more of an exception; Xu & Tenenbaum, 2007). Collectively, these results from the statistical learning literature indicate that the structure of the input gives the learner a cue about what regularities exist (or not) in the learning domain, which in turn drives generalization to novel items. A warping mechanism, therefore, can gracefully modulate generalization behavior across diverse learning domains.

In conclusion, the warping mechanism clearly raises a number of exciting possibilities for understanding the learning, representation, and processing principles underlying quasiregular domains such as reading, and how warped representations relate to many other aspects of cognition. It is a powerful theoretical construct that bridges learning quasiregularities with their representation.

Acknowledgments

This work was supported by a *Marie Curie Fellowship* (PIIF-GA-2013-627784) to Blair C. Armstrong and by a grant from the Spanish *Ministry of Science and Innovation* (PSI2011-24048) to Nicolas Dumay. Additional support was provided by the *Ohio State University*. We thank Christine Kay for helping with data collection and scoring, as well as David Plaut, Manuel Carreiras, Ram Frost, Yevdokiya Yermolayeva, and three anonymous reviewers for commenting on earlier drafts. We presented a brief overview of the warping mechanism, the main simulation and behavioural results related to learning and generalizing from new made up words, and the implications of these results for dual route models of the lexical system and statistical learning theory at the *International Workshop on Reading and Developmental Dyslexia* (Bilbao, 5-7 May 2016), the *First vs. Second Language Learning: from Neurobiology to Cognition Workshop* (Jerusalem, 26-29 September, 2016) and at the *57th Annual Meeting of the Psychonomic Society* (Boston, 17-20 November 2016).

References

Apfelbaum, K. S., Hazeltine, E., & McMurray, B. (2013). Statistical learning in reading: variability in irrelevant letters helps children learn phonics skills. *Developmental Psychology,* 49(7), 1348-1365.

Andrews, S., & Scaratt, D. R. (1998). Reading and analogy mechanisms in reading nonwords: hough Dou Peapel Rede Gnew Wirds? *Journal of Experimental Psychology: Human Perception and Performance, 24*(4), 1052-1086.

Aslin, R. N., & newport, E. L. (2012). Statistical learning: From acquiring specific items to forming general rules. *Current Directions in Psychological Science, 21*(3), 170-176.

Balota, D. A., Cortese, M. J., Sergent-Marshall, S. D., Spieler, D. H., & Yap, M. J. (2004). Visual word recognition of single-syllable words. *Journal of Experimental Psychology: General, 133*, 283-316.

Barr, D., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keeping it maximal. *Journal of Memory and Language, 68*(3), 255-278.

Bates, D., Kliegl, R., Vasishth, S., & Baayen, H. (2015, under review). Parsimonious mixed models. *Journal of Memory and Language*.

Cepeda N. J., Pashler H., Vul E., Wixted J. T., Rohrer D. (2006). Distributed practice in verbal recall tasks: a review and quantitative synthesis. *Psychological Bulletin, 132*, 354-380.

Christiansen, M. H., & Curtin, S. (1999). Transfer of learning: rule acquisition or statistical learning? *Trends in Cognitive Sciences, 3*(8), 290-291.

Coltheart, M., Curtis, B., Atkins, P., & Haller, M. (1993). Models of reading aloud: Dual-route and parallel-distributed-processing approaches. *Psychological Review*, *100*(4), 589-608.

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, 204-256.

Düzel, E., Cabeza, R., Picton, T. W., Yonelinas, A. P., Scheich, H., Heinze, H.-J. Tulving, E. (1999). Task- and item-related processes in memory retrieval: A combined PET and ERP study. *Proceedings of the National Academy of Sciences, 96*, 1794–1799.

Dumay, N., & Gaskell, M. G. (2007). Sleep-associated changes in the mental representation of spoken words. *Psychological Science*, 18, 35-39.

Dumay, N., & Gaskell, M.G. (2012). Overnight lexical consolidation revealed by speech segmentation. *Cognition*, 123, 119-132.

Dumay, N., Gaskell, M. G. & Feng, X. (2004). A day in the life of a spoken word. In K. Forbus, D. Gentner, and T. Regier (Eds.), *Proceedings of the Twenty-Sixth Annual Conference of the Cognitive Science Society* (pp. 339-344). Mahwah, NJ: Erlbaum.

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology* (H. A. Ruger, C. E. Bussenius, & E. R. Hilgard, Trans.). New York: Dover Publications. (Original work published 1885)

Earle, F. S., & Myers, E. B. (2015). Sleep and native language interference affect non-native speech sound learning. *Journal of Experimental Psychology: Human Perception and Performance, 41*, 1680-1695.

Fenn, K. M., Nusbaum, H. C., & Margoliash, D. (2003). Consolidation during sleep of perceptual learning of spoken language. *Nature, 425*, 614-616.

Qiao, X., & Forster, K. I. (2012). Novel word lexicalization and the prime lexicality effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39, 64-74.

Frost, R., Armstrong, B. C., Seigelman, N., & Christiansen, M. H. (2015). Domain generality versus modality specificity: The paradox of statistical learning. *Trends in Cognitive Sciences, 19,* 117-125.

Gaskell, M. G., Warker, J., Lindsay, S., Frost, R., Guest, J., Snowdon, R., & Stackhouse, A. (2014). Sleep underpins the plasticity of language production. *Psychological Science, 25*, 1457-1465.

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

Glushko, R. J. (1979). The organization and activation of orthographic knowledge in reading aloud. *Journal of Experimental Psychology: Human Perception and Performance, 5,* 674-691.

Gómez, R. L., Bootzin, R., & Nadel, L. (2006). Naps promote abstraction in language learning infants. *Psychological Science*, *17*, 670-674.

Harm M. W., Seidenberg M. S. (2004). Computing the meanings of words in reading: cooperative division of labor between visual and phonological processes. *Psychological Review, 111,* 662–720.

Hudson, P. T. W., & Bergman, M. W. (1985). Lexical knowledge in word recognition: Word length and word frequency in naming and lexical decision tasks. *Journal of Memory and Language, 24*(1), 56-58.

Ijalba, E., & Obler, L. K. (2015). First language grapheme-phoneme transparency effects in adult second-language learning. *Reading in a Foreign Language 27*, 47-70.

Jacobs, R. A. (1988). Increased rates of convergence through learning rate adaptation. *Neural Networks, 1*(4), 295-307.

Kello, C. T., & Plaut, D. C. (2000). Strategic control in word reading: Evidence from speeded responding in the tempo naming task. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 26,* 719-750.

Kim, W., Pitt, M. A., & Myung, I. J. (2013). How do PDP models learn quasiregularity? *Psychological Review, 120,* 903-916.

Lerner, I., Armstrong, B. C., & Frost, R. (2014). What can we learn from learning models about sensitivity to letter-order in visual word recognition? *Journal of Memory and Language, 77,* 40-58.

McCLELLAND, J. L. (1998). Complementary learning systems in the brain: A connectionist approach to explicit and implicit cognition and memory. *Annals of the New York Academy of Sciences*, *843*(1), 153-169.

McClelland, J. L. (2015). Capturing gradience, continuous change, and quasi-regularity in sound, word, phrase, and meaning. In The Handbook of Language Emergence (1st Ed.; Eds. B. MacWhinney & W. O’Grady), pp. 53-80. Wiley & Sons.

McClelland, J. L., McNaughton, B. L., & O'Reilly, R. C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review, 102*, 419-457.

O’Reilly, R. C., Bhattacharyya, R., Howard, M. D. & Ketz, N. (2014), Complementary Learning Systems. *Cognitive Science, 38,* 1229–1248.

Peereman, R, & Content, A. (1997). Orthographic and Phonological Neighborhoods in Naming: Not All Neighbors Are Equally Influential in Orthographic Space. *Journal of Memory and Language, 37*, 382–410.

Peirce, J. W. (2007). PsychoPy - Psychophysics software in Python. *Journal of Neuroscience Methods, 162*, 8-13.

Perry, C., Ziegler, J. C., & Zorzi, M. (2010). Beyond single syllables: Large-scale modeling of reading aloud with the Connectionist Dual Process (CDP++) model. *Cognitive Psychology, 61*, 106-151.

Pinker, S., & Ullman, M. T. (2002). The past and future of the past tense. *Trends in Cognitive Sciences, 6*, 456-463.

Plaut, D. C. (1997). Structure and function in the lexical system: Insights from distributed models of word reading and lexical decision. *Language and Cognitive Processes*, 12, 767-808.

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

Plaut, D.C., and Shallice, T. (1993). Deep dyslexia: A case study of connectionist neuropsychology. Cognitive Neuropsychology, 10, 377-500.

Protopapas A. (2007). CheckVocal: A program to facilitate checking the accuracy and response time of vocal responses from DMDX. *Behavior Research Methods*, 39, 859–862.

Qiao, X., & Forster, K.I. (2013). Novel word lexicalization and the prime lexicality effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 1064-1074.

Rakison, D. H., & Butterworth, G. E. (1998). Infants' use of object parts in early categorization. *Developmental Psychology, 31*, 49-62.

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

Seidenberg, M. S., Petersen, A., MacDonald, M. C., & Plaut, D. C. (1996). Pseudohomophone effects and models of word recognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 22*, 48-62.

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

Stickgold, R. (2005). Sleep-dependent memory consolidation. *Nature, 437*(27), 1272-1278.

Stickgold, R., & Walker, M. P. (2013). Sleep-dependent memory triage: Evolving generalization through selective processing. *Nature Neuroscience, 16,* 139-145.

Strain E., Patterson K., Seidenberg M. S. (1995). Semantic effects in single-word naming. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 21*, 1140–1154.

Tamminen, J., Davis, M. H. & Rastle, K. (2015). From specific examples to general knowledge in language learning. *Cognitive Psychology, 7*9, 1-39.

Taraban, R. & McClelland, J. L. (1987). Conspiracy effects in word pronunciation. *Journal of Memory and Language, 26*, 608-631.

Treiman, R., Kessler, B., & Bick, S. (2003). Influence of consonantal context on the pronunciation of vowels: A comparison of human readers and computational models. *Cognition, 88,* 49-78.

Treiman, R. & Kessler, B. (2006). Spelling as Statistical Learning: Using Consonantal Context to Spell Vowels. *Journal of Educational Psychology, 98,* 642-652.

Treiman, R., Kessler, B., Zevin, J. D., Bick, S., & Davis, M. (2006). Influence of consonantal context on the reading of vowels: Evidence from children. *Journal of Experimental Child Psychology, 92*, 1-24.

Treiman, R., Mullennix, J., Bijeljac-Babic, R., & Richmond-Welty, E.D. (1995). The special role of rimes in the description, use, and acquisition of English orthography. *Journal of Experimental Psychology: General*, *124*, 107-136.

Xu, F., & Tenenbaum, J. B. (2007). Word learning as Bayesian inference. *Psychological Review*, *114*, 245-272.

Zevin, J. D., & Seidenberg, M. S. (2002). Age of acquisition effects in word reading and other tasks. *Journal of Memory and Language*, *47*, 1-29.

Zevin, J. D., & Seidenberg, M. S. (2006). Consistency effects and individual differences in nonword naming: A comparison of current models. *Journal of Memory and Language, 54,* 145-160.

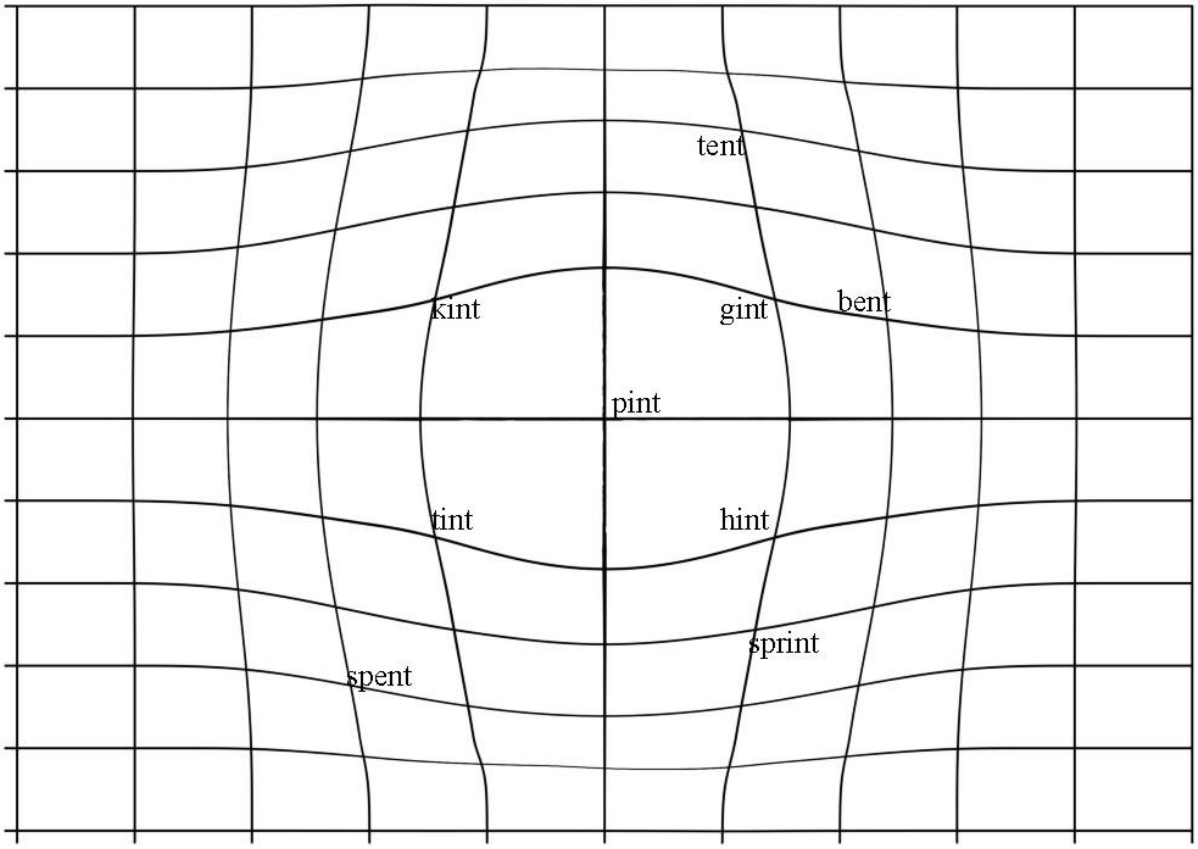

Figure 1. Simplified depiction of the “warping” in hidden-unit space that is required to represent an exception word (*pint*) along with its neighbors. Although warping is localized to the region occupied by *pint*, there is substantial spillover to neighboring nonwords (e.g., *kint, gint*), and to neighboring words (e.g., *hint tint*). However, because word pronunciations were learned during training, warping does not disrupt their pronunciation. For words with ambiguous pronunciations (e.g., *bive* rhyming with *give,* not with *drive*), similar principles apply but the amount of warping needed to accommodate those words is reduced, leading to greater spillover to their neighbors. That is, less warping allows greater generalization. See Kim et al. (2013) for details.

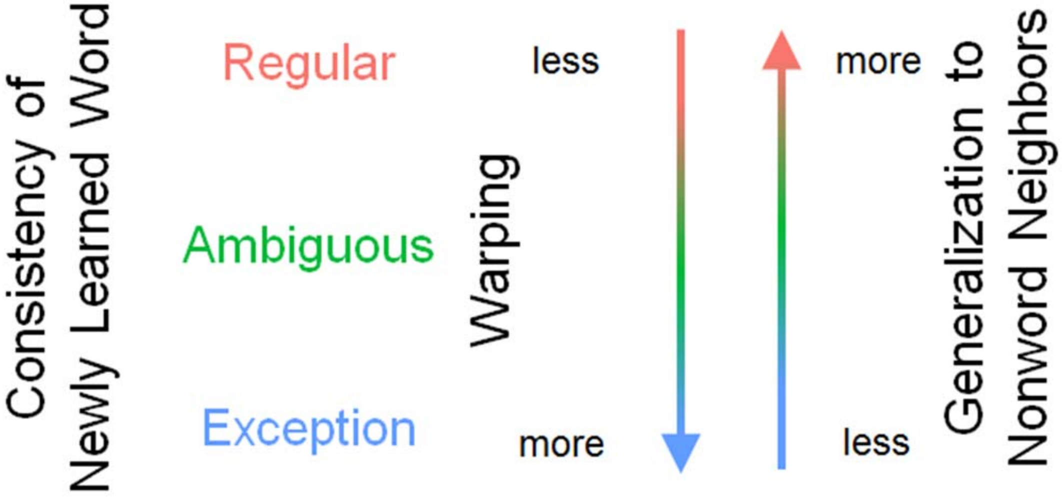

Figure 2. Schematic illustration of the relationship between regular, ambiguous, and exception words, representational warping, and generalization of a word's pronunciation to neighboring nonwords.

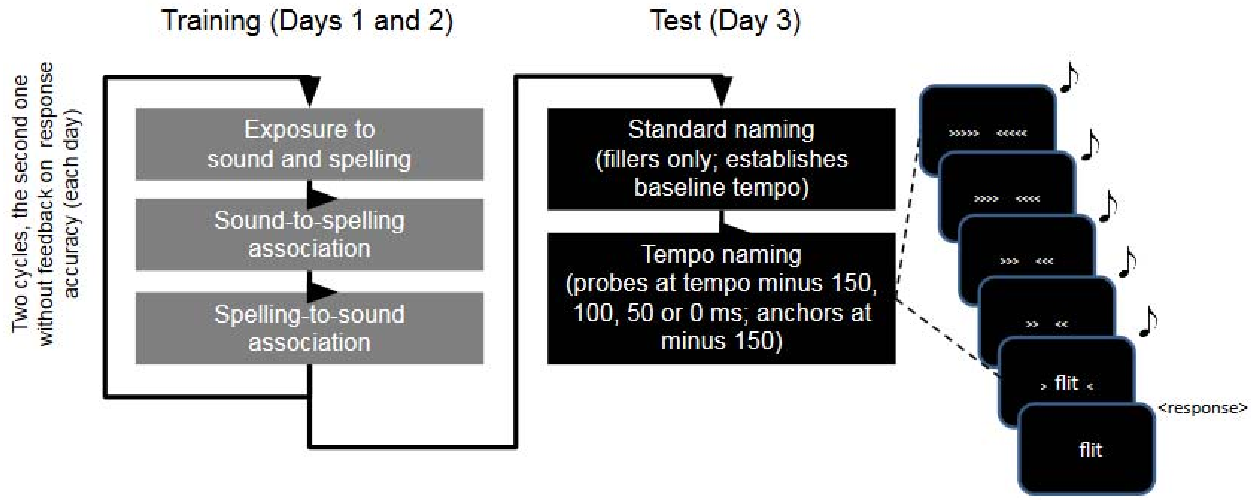
Training (Days 1 and 2)
Test (Day 3)
Two cycles, the second one without feedback on response accuracy (each day)
Exposure to sound and spelling
Sound-to-spelling association
Spelling-to-sound association
Standard naming (fillers only; establishes baseline tempo)
Tempo naming (probes at tempo minus 150, 100, 50 or 0 ms; anchors at minus 150)
>>>>> <<<<<
>>>> <<<<
>>> <<<
>> <<
> flit <
<response>
flit

Figure 3. Overview of the training and test procedure.

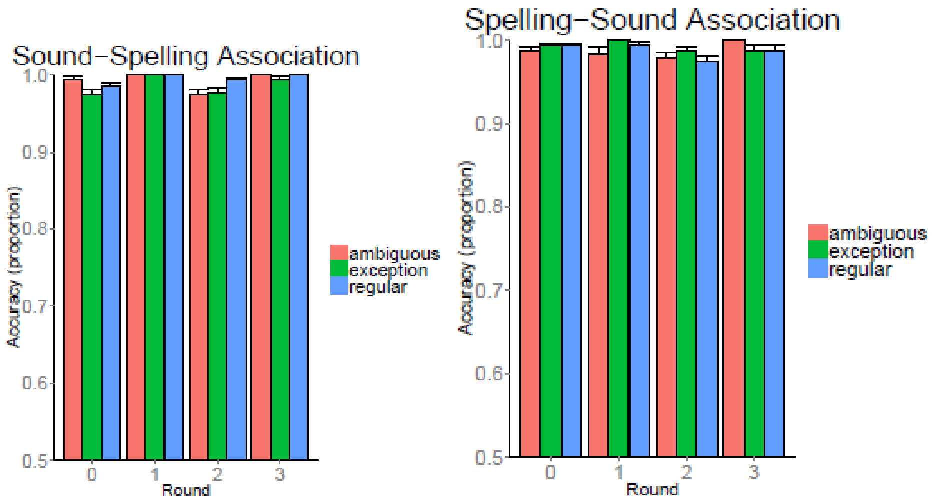

Figure 4. Accuracy for each anchor type for each round of training. Rounds 0 and 1 were conducted on training day 1, and round 2 and 3 were conducted on training day 2. Error bars represent the standard error of the mean.

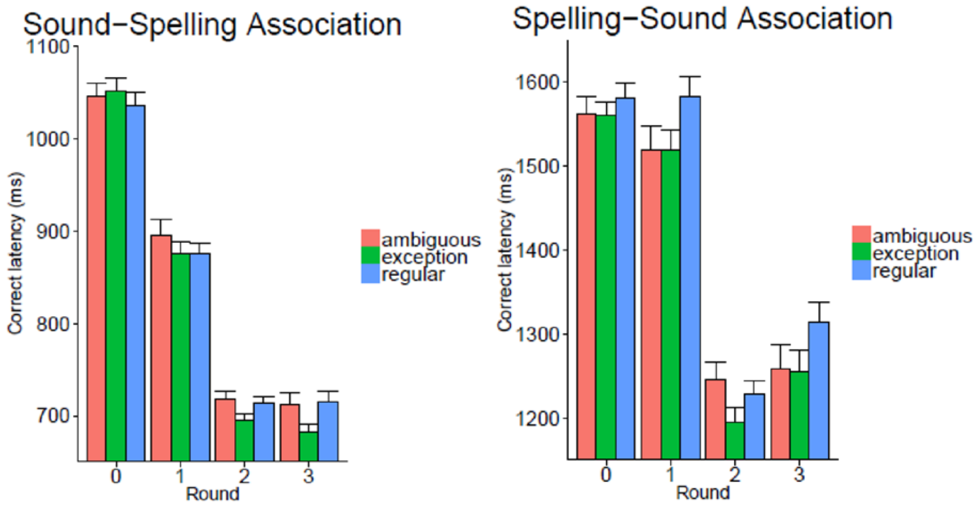

Figure 5. Correct latency for each anchor type for each round of training.

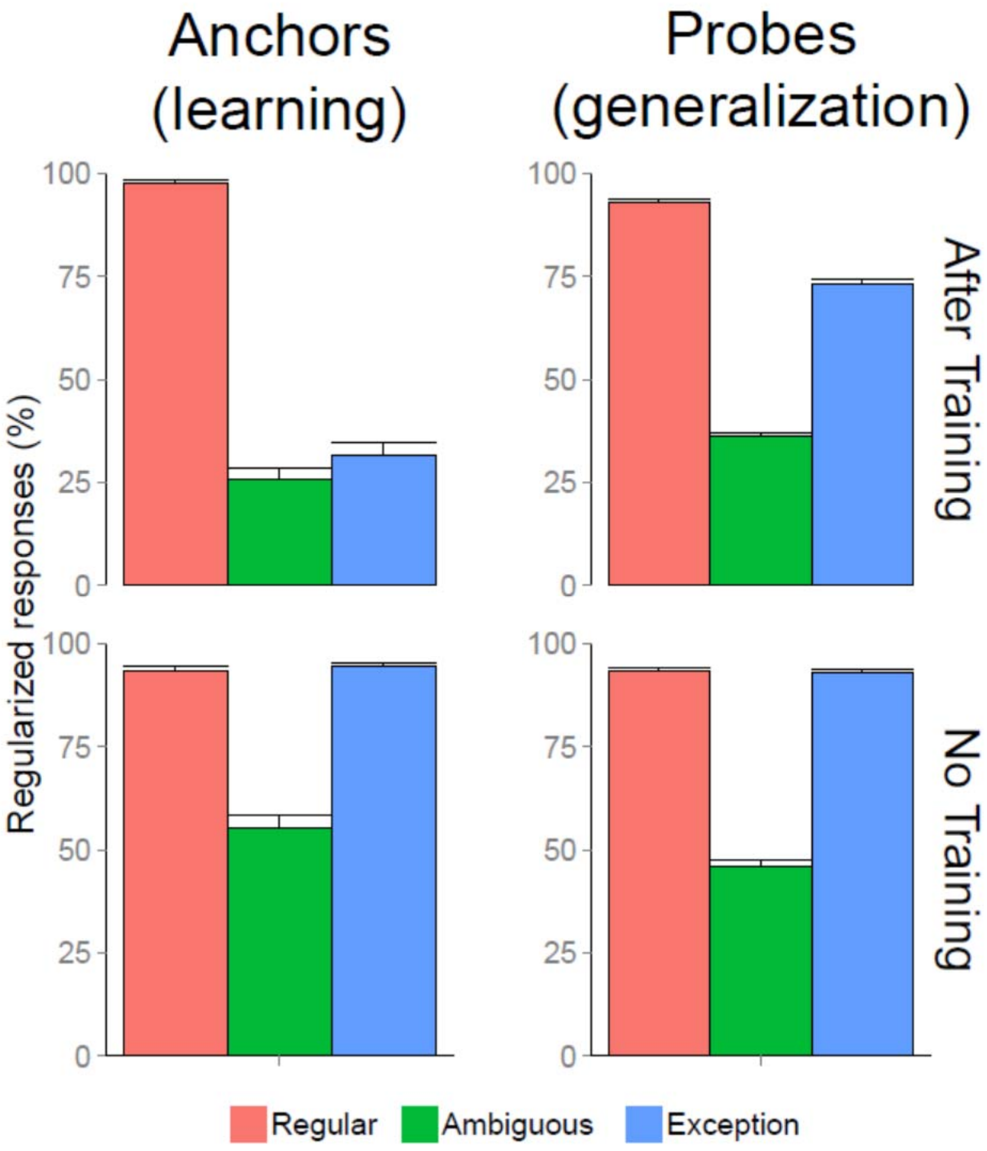

Figure 6. Percent of regularized responses as a function of training for regular, ambiguous, and exception anchors [left], and for the corresponding probes [right]. Error bars in this and all subsequent plots represent estimates of the standard error.

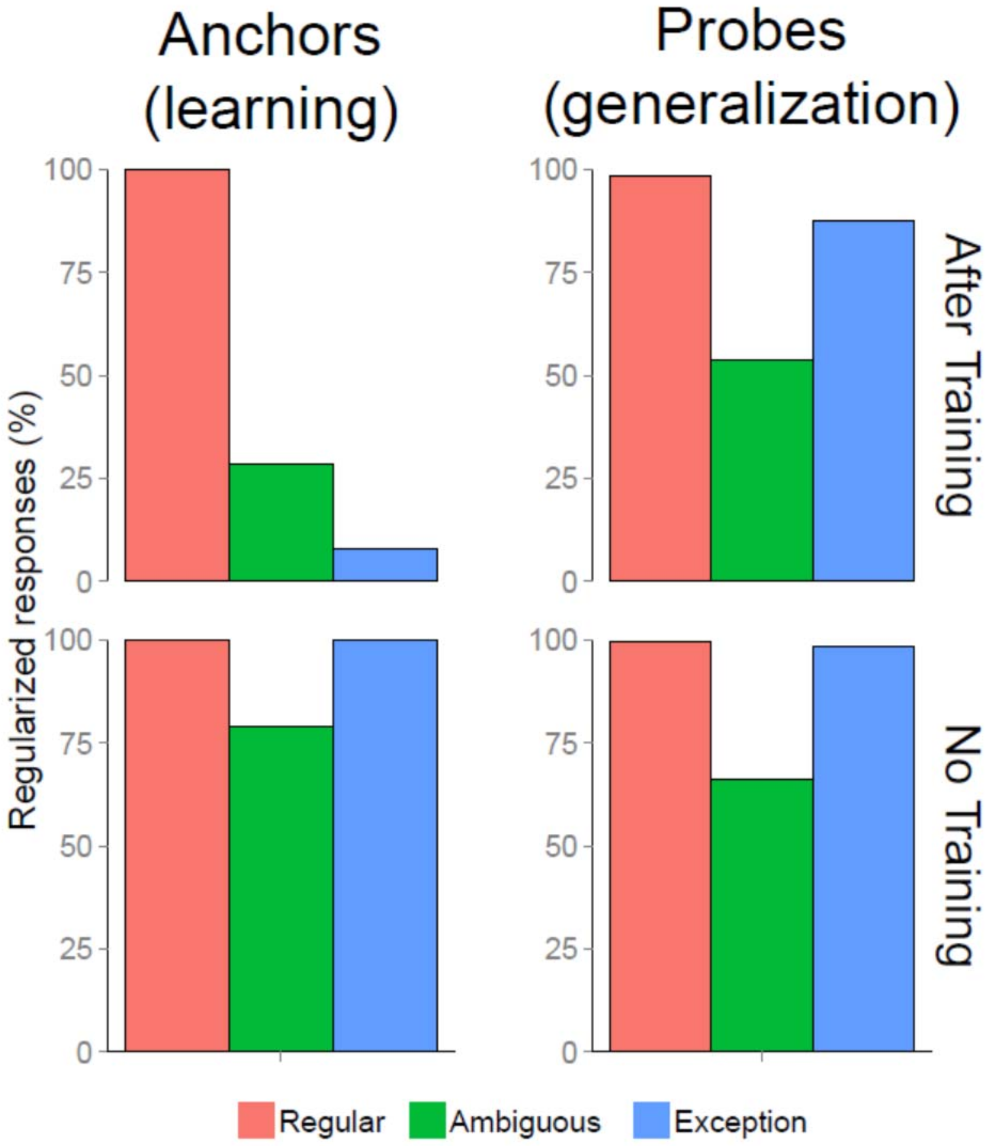

Figure 7. Simulation results. Percent of regularized responses as a function of training for regular, ambiguous, and exception anchors [left], and for the corresponding probes [right]. Standard errors were perceivably small and so were omitted from the plot.

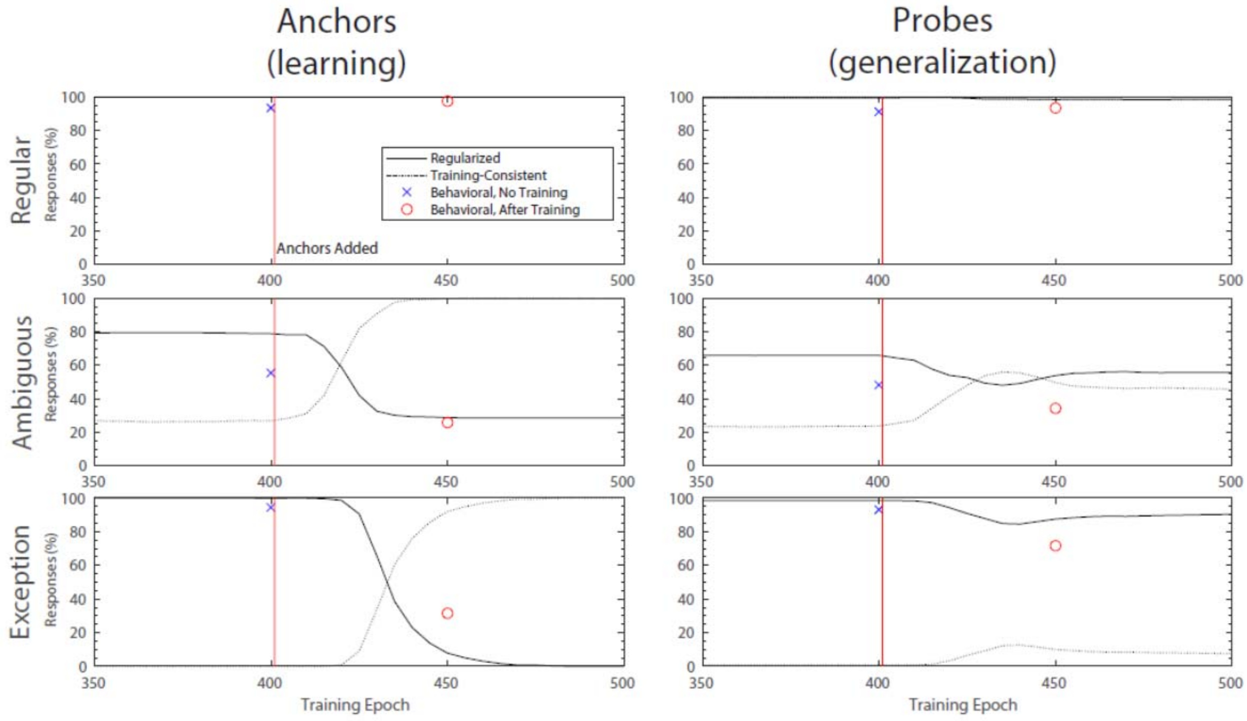

Figure 8. Model simulation results from epoch 350 to 500. Percent of regularized responses (solid lines) as a function of training epoch for regular (top), ambiguous (middle), and exception (bottom) anchors [left], and for the corresponding probes [right]. The solid red line denotes the training epoch at which the new anchors were introduced to the training set. The time point just before the new anchors were introduced is used to simulate the no training control, and the ceiling performance shows that the model regularized the pronunciation of all words perfectly. The after- training condition was taken to be performance at epoch 450, that is, after 50 epochs of training that included the new anchors. Also included is the percent of training-consistent responses across epochs (dashed lines), which shows that regularized responses are being replaced by training-consistent responses when learning new ambiguous and exception anchors, and when generalizing to probes. The means from the probe conditions in Figure 6 are plotted for reference (hollow red circles and solid blue Xs).

Appendix A: Experimental Stimuli

Table A1. Critical Experimental Stimuli from the Tempo Naming Task

Type	Anchor	Rhyme Word	Probes			
regular	blit	wit	flit	glit	plit	trit
	grax	wax	brax	drax	prax	shrax
	kleef	reef	bleef	gleef	pleef	preef
	krim	him	blim	clim	drim	frim
	nisp	lisp	bisp	kisp	risp	tisp
	plig	big	blig	clig	flig	slig
	preld	weld	breld	creld	dreld	steld
	scark	dark	blark	crark	plark	slark
	shing	wing	ging	jing	ning	ting
	slape	tape	blape	clape	glape	plape
exception	brot	wrote	crot	drot	grot	prot
	chell	peel	brell	crell	drell	prell
	crill	mile	blill	brill	clill	prill
	dest	beast	clest	glest	plest	trest
	drace	draw	frace	krace	prace	vrace
	fank	honk	lank	vank		
	geam	gem	cheam	fleam	keam	peam
	kipe	chip	bipe	fipe	gipe	nipe
	nust	roost	chust	pust	tust	vust
	suff	roof	chuff	druff	vuff	
ambiguous	bive	drive	kive	mive	pive	tive
	blome	home	clome	flome	grome	prome
	clead	led	glead	pread	smead	kread
	frow	how	clow	trow		
	grour	flower	brour	drour	prour	trour
	plone	on	blone	frone	slone	glone
	slood	mud	glood	klood	plood	

Note. A small number of targets and probes eliminated during data screening have been omitted, as described in the results section. A few probes (e.g., *flit*, *chuff*) are in fact very low frequency words (< 0.5 words per million in Brysbaert & New, 2009). However, an informal test of nine lab members showed that only the regular probe *flit* was sometimes recognized as a word, and the item-level data showed that performance for this probe was comparable to that of other probes. This suggests that the impact of such low frequency words was negligible.

Table A2: Visual Foils used during Training

Type	Anchor	Rhyme	Visual Foils							
regular	blit	wit	blitt	blitte	bliht	blihte	blytt	blytte	blyht	blyt
	grax	wax	graks	graxe	graques	gracs	gracks	gracques	grakx	grackes
	kleef	reef	kleaf	klefe	kleaph	kleeph	kleif	kleiph	cleaf	cleaph
	krim	him	krimn	chrym	chrimme	chrim	krimme	kryme	krym	krymm
	nisp	lisp	knisp	mnisp	knispe	mnispe	nispe	knysp	nysp	mnysp
	plig	big	pligh	plygg	pligue	plihg	plyg	plygue	plyhg	pligg
	preld	weld	prelled	prehled	prehlde	prelde	preald	prealde	preldt	pwreld
	scark	dark	skarck	skarc	squark	squarck	squarque	skarque	scarque	skark
	shing	wing	shingg	shingue	shyng	shyngg	shyngue	shingh	shyngh	shinghe
	slape	tape	slaype	slaeype	slaipe	slaiype	sleype	sleipe	sleip	slayp
exception	brot	wrote	browt	broht	brohte	browte	browtt	brohtte	broat	broate
	chell	peel	cheel	cheale	cheal	cheele	chiel	chiele	cheall	cheell
	crill	mile	creil	creile	cryle	cryl	krill	kreil	kryll	kryle
	dest	beast	deeste	deast	diest	deased	deesed	dieced	deaced	deaste
	drace	draw	dross	drosse	drauss	drausse	drauce	droce	drawce	draus
	fank	honk	phank	phanck	faunck	fanque	phanque	faunk	fonk	phonk
	geam	gem	guemm	guem	guemme	gheme	guemn	ghem	jeam	guelm
	kipe	chip	kip	kippe	kyp	chyp	kipp	kihp	kypp	kyppe
	nust	roost	newst	noost	neust	noowst	nuest	neuwst	niewst	nieust
	suff	roof	souf	soof	cewf	seuf	seuff	souph	sooph	seuph
ambiguous	bive	drive	triv	triyve	tryve	trighve	tryeve	treyeve	twryve	twrive
	blome	home	bloame	blowm	blaume	blawm	blawme	blomme	bloamm	blaumme
	clead	led	cled	kled	cledd	kledd	clehd	klehd	chled	chledd
	frow	how	frawe	frowe	froaw	phrow	phraw	phroaw	phrawe	phrah
	grour	flower	qrower	growir	grauwer	groawer	growre	grouwer	groweur	groaweur
	plone	on	plown	plowne	ploane	ploan	plaune	ploghne	ploahn	plonne
	slood	mud	slud	sludd	sloed	sloode	slude	sludde	sloede	sluhd

Table A3: Auditory Foils used during Training

Type	Anchor	Rhyme	AuditoryFoils							
			Onset Foils				Offset Foils			
regular	blit	wit	krit	stit	frit	prit	blish	bliff	blith	blick
	grax	wax	spax	skax	vrax	frax	grakt	grast	grasht	grav
	kleef	reef	sleef	fleef	creef	sreef	kleep	kleesh	kleege	kleeth
	krim	him	plim	flim	shrim	vrim	krin	krid	krish	kriv
	nisp	lisp	risp	sisp	visp	shisp	nist	nisk	nism	nilt
	plig	big	drig	crig	stig	grig	plid	plip	plick	plin
	preld	weld	treld	gleld	cleld	bleld	prelt	preln	prelm	prend
	scark	dark	flark	smark	snark	frark	skarp	skart	skarsh	scarm
	shing	wing	hing	ming	ving	fing	shinged	shingz	shinje	shingked
	slape	tape	brape	srape	smape	flape	slake	slabe	slafe	slame
exception	brot	wrote	frote	vrote	clote	plote	broce	brophe	broshe	brope
	chell	peel	cheeb	cheem	cheed	cheeg	treel	gueel	theel	beal
	crill	mile	thrile	shrile	plile	clile	crine	crive	cribe	crithe
	dest	beast	jeast	meast	bleast	reast	deeft	deekt	deeshed	deesp
	drace	draw	bloss	ploss	closs	sloss	drosh	drof	drock	drotch
	fank	honk	pank	slank	donk	wonk	fonch	fonse	fong	fondge
	geam	gem	nem	vem	frem	trem	gueb	guesh	guep	guell
	kipe	chip	yip	vip	smip	thip	kif	kib	kith	kish
	nust	roost	loost	bloost	foost	shoost	nooft	noosk	noosp	noosht
	suff	roof	luff	foof	moof	toof	soosh	sootch	soov	sook
ambiguous	bive	drive	vive	plive	thive	sive	bime	bine	bice	bife
	blome	home	brome	vrome	shome	trome	blol	bloth	blon	bloc
	clead	led	sred	vred	shled	vled	clet	cleth	cleb	kless
	frow	how	throuw	strow	srow	zow	fral	fraws	fram	frab
	grour	flower	clower	blower	glower	vrower	growesh	growen	groweb	growed
	plone	on	smone	trone	krown	srown	plowm	plowv	plowg	plowse
	slood	mud	frud	smood	shlud	prud	slun	slubb	sluv	slull

Note. The pronunciation of the vowels remains the same across anchors and foils.

Appendix B: Properties of Experimental Items and Additional Analyses

Table B1. Descriptive statistics for the item types

Item Type	Number of letters	Orthographic Levenshtein Distance (OLD20)	Coltheart's N (Orthographic)	Positional Bigram Frequency
Ambiguous Probe	4.8 (0.4)	1.6 (0.3)	5.0 (4.2)	1684 (1893)
Exception Probe	4.6 (0.5)	1.6 (0.3)	5.3 (4.2)	1379 (729)
Regular Probe	4.4 (0.5)	1.6 (0.3)	4.2 (3.1)	831 (725)
Ambiguous Anchor	4.7 (0.5)	1.6 (0.3)	5.3 (3.6)	1617 (1072)
Exception Anchor	4.3 (0.5)	1.3 (0.2)	8.2 (3.9)	1740 (935)
Regular Anchor	4.5 (0.5)	1.6 (0.3)	4.1 (3.1)	945 (1324)

Note. Standard deviations are in parentheses. Orthographic Levenshtein Distance - 20 was calculated based on all words in the SUBTL database (Brysbaert & New, 2009) with a frequency of at least 1.0, using the tool provided by Yarkoni, Balota, and Yap (2008). All other measures were obtained from MCWord (http://www.neuro.mcw.edu/mcword/).

Appendix C. Supplementary Analyses and Results

Analyses of regularization rates.

Effects of tempo. Figure C1 plots regularization rate as a function of item type and training presented in Figure 6, now broken down by tempo. Note the consistency across the tempos, which reflects the stability of our results.

Effects of covariates. Although we were able to match our stimuli relatively well on a number of covariates that could have confounded our results, a few significant differences between conditions remained because of the constraints of English (see Table B1). We therefore included each of these variables as a covariate in the analyses reported in the main text.

For the analyses of the probe data, these potential confounds were all added to a single model. To avoid convergence issues in the models of the smaller anchor dataset, we ran three separate variants of the omnibus model, each of which contained only one of the three covariates. The effects related to training and item type were similar in all cases. In the main text, we report the results of the analyses that contained the orthographic Levenshtein distance predictor. All of the potential confounds were scaled to have a mean of 0 and a standard deviation of 1 prior to entry in the model. The fact that the analyses still yielded significant effects of item type and regularization rules out these alternative accounts. However, some of these variables did have an independent effect on performance, which shows us how these variables are influenced by training, independently of warping. The details of these effects are reported below.

Orthographic Levenstein distance. Prior work has established that denser orthographic neighborhoods impact word naming (Zevin & Seidenberg, 2006). To rule out an account of our findings based on interactions between training and lexical neighborhood, we included orthographic Levenshtein distance and its interaction with. Despite not suppressing the critical interaction effects between item type and training that we attribute to warping, this variable did have an independent effect in its own right: (anchors: *b* = 0.578, *SE* = 0.21, *n* = 1904, *z* = 2.70, *p* = 0.006; the main effect was not significant, *t* < 1; probes: *b* = 0.44, *SE* = 0.13, *n* = 7199, *z* = -3.46, *p* < .001; the main effect was also significant: *b* = 0.38, *SE* = 0.18, *n* = 7199; *z* = 2.02, *z* = 0.04). The interaction terms indicate that regularizations rates were differentially higher in sparser neighborhoods after training, potentially because these sparser neighborhoods are less warped overall.

Positional bigram frequency. We also tested whether our effects could be due to other sublexical differences, such as bigram frequency. Insofar as this was the case, it would open the possibility that relative exposure to particular sublexical orthographic patterns could alter how readily an existing or a new pronunciation could be associated with them. However, the fact that our critical pattern of effects related to item type and training persisted despite including this predictor rules out an alternative account based on bigram frequency. Bigram frequency did, however, interact with training to a significant degree in the case of the probes and to a marginal degree in the case of the anchors (anchors; *b* = -0.33, *SE* = 0.18, *n* = 1904, *z* = -1.82, *p* = 0.06; the main effect was not significant, *t* < 1; probes: *b* = -0.22, *SE* = 0.08, *n* = 7199, *z* = -2.70, *p* = .007; the main effect was not significant, *t* < 1). These interaction terms indicate that regularization rates were lower for more frequent bigrams after training, independent of the effects of training on item type.

Length in letters. Finally, we examined whether our effects could be due to differences in word length (in letters), which previous work has shown to covary with naming latencies (Hudson & Bergman, 1985). This variable did not suppress the item type and training effects reported in the main text. For the anchors, the variance of the omnibus model that included length failed to converge, likely because of the smaller size of the anchor data set and the small variability in word length across items. For the probes, a significant interaction between length and training was observed in the omnibus analysis (*b* = 0.57, *SE* = 0.14, *n* = 1904, *z* = 3-50, *p* < .001; the main effect was not significant; *b* = -0.26, *SE* = 0.20, *n* = 7199, *z* = -1.33, *p* = 0.18). This interaction indicates that regularizations rates were higher for longer words after training, independent of the effects of training on item type.

.

Analyses of latency data.

Figure C2 plots latency of regularized responses as a function of tempo and training for ambiguous and exception anchors [left], and for the corresponding probes [right], and shows that there were no significant differences between the ambiguous and exception items on this measure. Figure C3 reveals that the same outcome is found for training-consistent responses. This is supported by a series of mixed effects models on latencies, which were run separately for probes and anchors, and for the regularized and training-consistent responses, none of which showed an effects of item type. These analyses are reported next.

We begin with the models for the regularized responses. All of these models included random intercepts for participant and item, as well as fixed effects of item type (with regular responses used as the baseline level) and training (with no training used as the baseline level), and the interaction between these two factors. For probes, the models also included tempo, introduced as a continuous variable denoting the difference between the current tempo and the participant's naming baseline. (The main effect of tempo was always significant in every latency analysis that we report, *p* < .001.) Tempo was also allowed to interact with both item type and training. No random slope terms were included due to convergence issues. All analyses assumed a Gaussian distribution of errors and degrees of freedom were determined using the Satterwaithe approximation.

No significant effects were detected involving the anchors (*t*s < 1.31, *df* >= 37.7, *p*s > .2), except for an interaction indicating that the exception anchors were regularized more slowly in the training condition relative to the regular anchors (*b* = 22.8, *SE* = 11.0, *t(1259.9)* = 2.07, *p* = .04). This supports the notion that competition from the training-consistent response is slowing responding. For the probes, no significant effects were detected, although there were a few marginal trends in the latter case described below (all other *t*s < 1.47; *df* >= 77, *p*'s > .14). These trends indicate that the exception probes were responded to marginally more slowly in the training condition relative to the regular probes, as assessed by an item type by training interaction (*b* = 17.3, *SE* = 9.8, *n* = 5551, *t*(5432) = 1.77, *p* = .08). Again, this outcome is consistent with the notion that competition between the training-consistent and regularized pronunciation of each exception probe is producing competition and slowing responding. The three way interaction, in which tempo interacted with the previous two way interaction, showed a marginal trend for the difference between regular and exception items to shrink for longer tempos (*b* = 0.20, *SE* = 0.10, *n* = 5551, *t*(5434) = 1.94; *p* = 0.05) in the training condition, potentially because tracking the slower tempo was slightly easier. The ambiguous probes were also responded to marginally more rapidly than the regular probes, as reflected by a main effect of item type when contrasting the ambiguous and regular probes (*b* = - 16.9, *SE* = 10.2, *n* = 5551, *t*(863) = 1.66, *p* = .10).

We now turn to the latency data for training-consistent responses. There were no training-consistent responses for regular items in any condition, or for exception items in the no training condition (see Figure C3). As a result, we could not use exactly the same statistical model used to analyze the latency data for regularized responses because some cells in the design are empty. Instead, we used two main variations of the model used to analyze regularized responses to avoid these empty cells, each of which was run separately for the anchors and the probes so as to allow tempo to be entered as a predictor in the analyses of the probes. All of these models included random intercepts for participant and item and no random slopes terms to avoid convergence issues. Except as noted below, the same baseline levels for the fixed effects used in the prior analysis were again used in these analyses.

The first set of analyses focused on the latencies for the training-consistent responses to ambiguous words as a function of training by including training as a fixed effect. In the analysis of the probes, which were presented at multiple tempos, we also included tempo as a fixed effect that could interact with the fixed effect of training. No significant effects involving item type were detected in any of these analyses (all *t*s < 1).

The second set of analyses tested for differences between the ambiguous and exception items in the training condition. The ambiguous items were used in place of the regular items as the baseline level for item type. In these analyses, item type was included as a fixed effect. As in the prior analyses, tempo was also included as a fixed effect that could interact with item type in the analyses of the probes. No significant effects involving item type were detected in any of these analyses (all *t*s < 1), except for a marginal trend suggesting latencies were slower for exception anchors than ambiguous anchors in the training condition (*b* = 21.1, *SE* = 11.0, *n* = 485, *t*(21.2) = 1.91, *p* = 0.07; ambiguous words served as the baseline). This is consistent with the idea that there is more competition between the regularized and training-consistent responses in the case of the exception anchors.

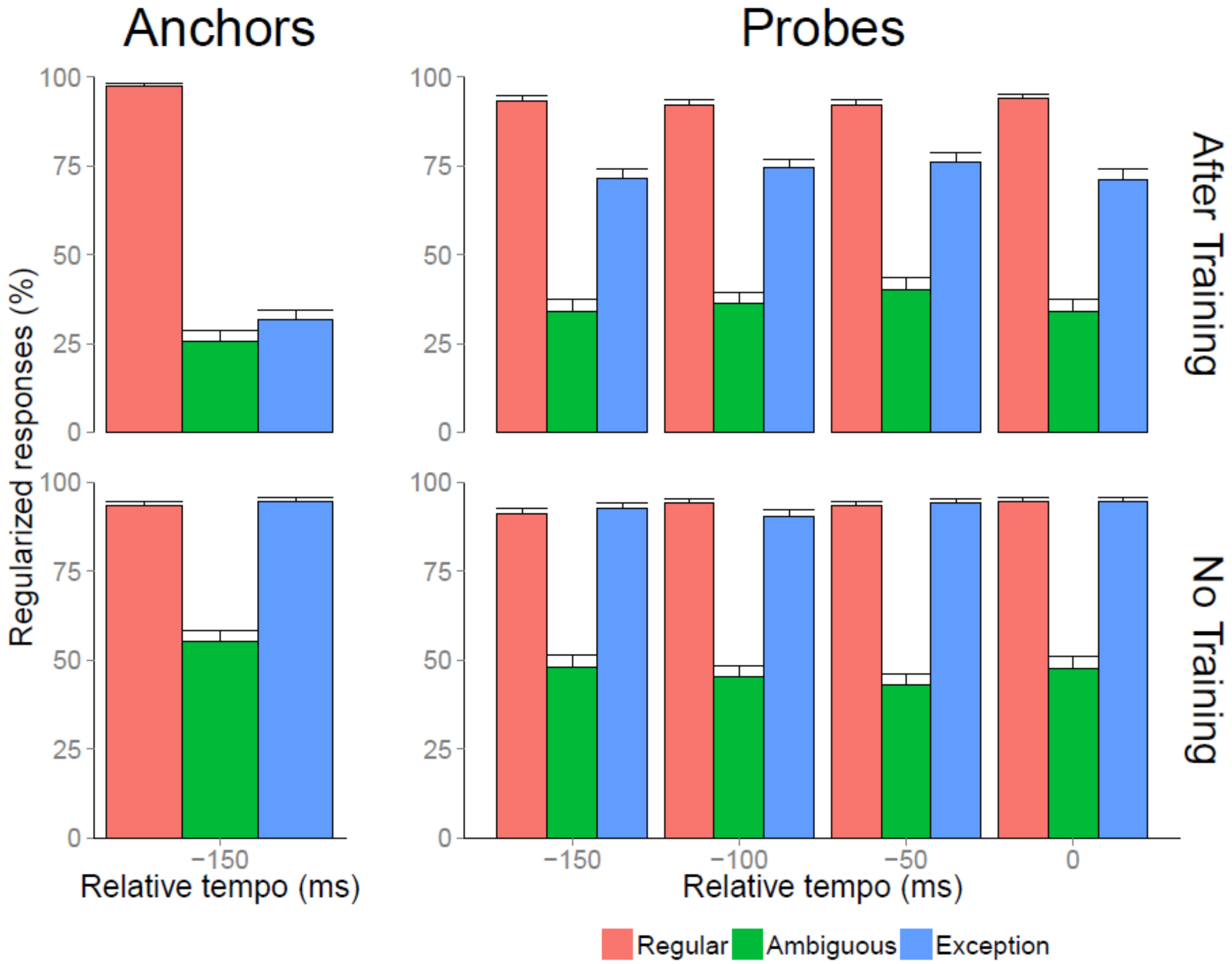

Figure C1. Percent of regularized responses as a function of relative tempo and training for regular, ambiguous, and exception anchors [left], and for the corresponding probes [right].

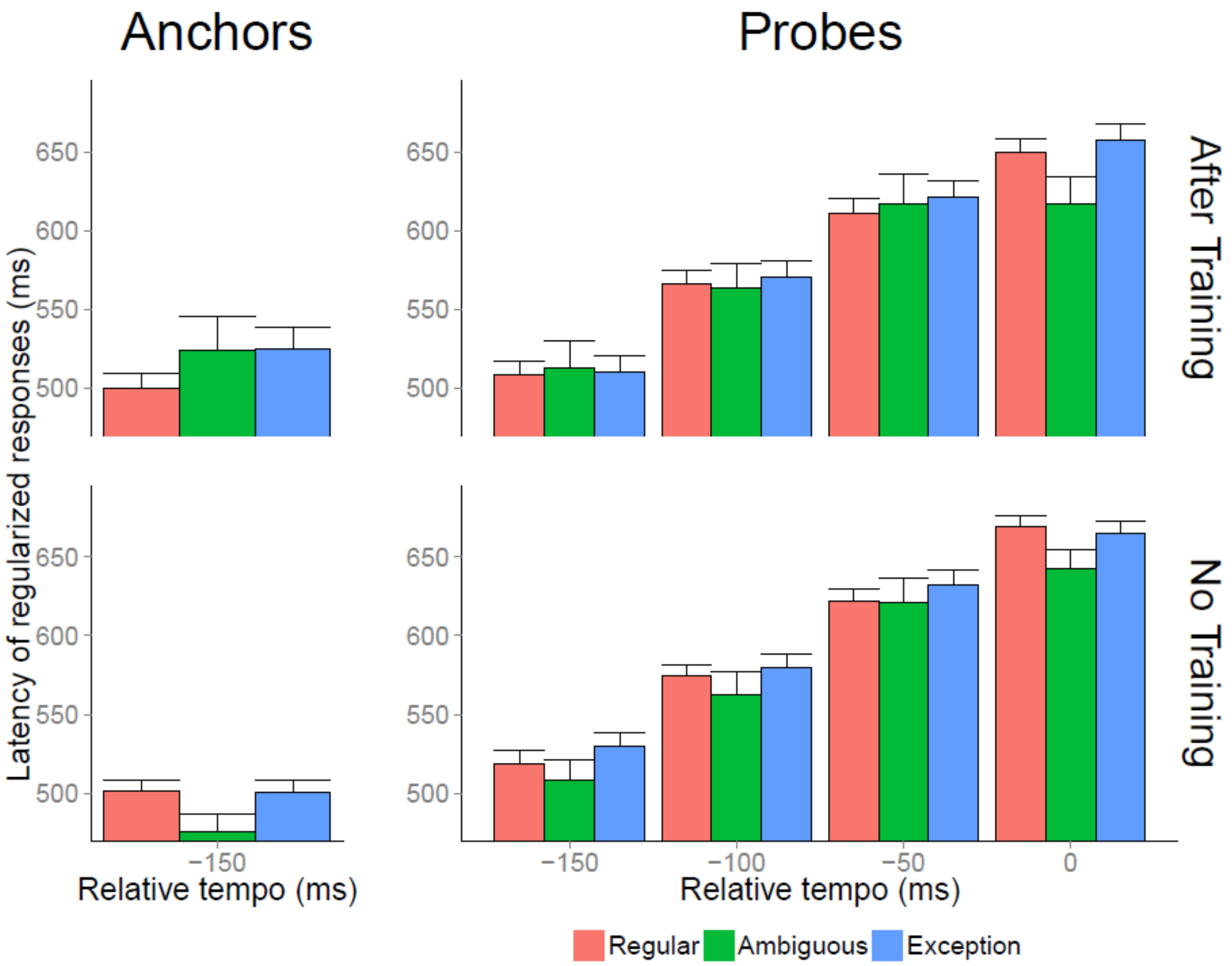

Figure C2. Latency of regularized responses as a function of relative tempo and training for regular, ambiguous, and exception probes [left], and for the corresponding anchors [right].

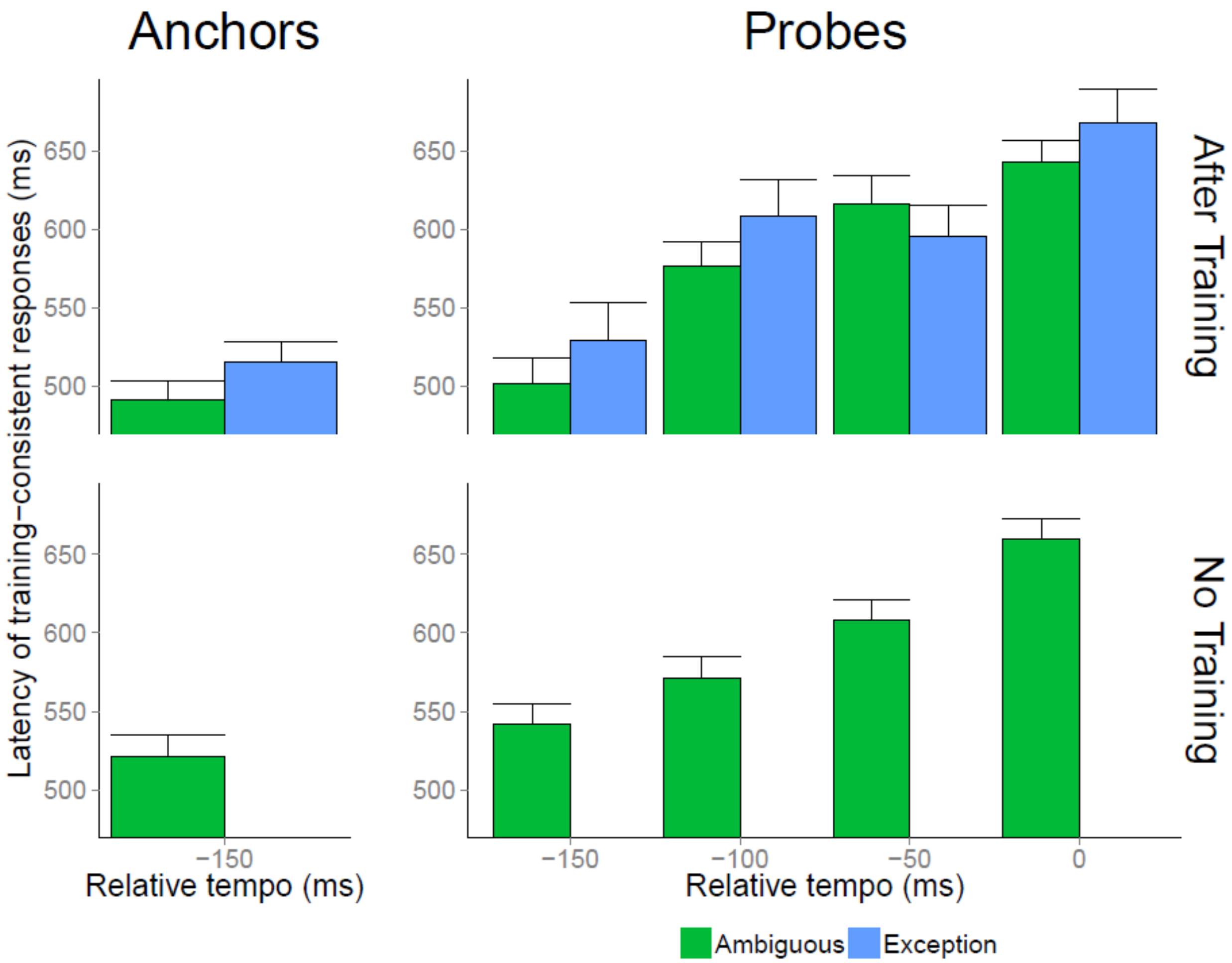

Figure C3. Latency of training-consistent responses as a function of relative tempo and training for ambiguous, and exception anchors [left], and for the corresponding anchors [right].