Journal of Child Language (2019), **46**, 1–23
doi:10.1017/S0305000918000259

CAMBRIDGE
UNIVERSITY PRESS

ARTICLE

Children's gradient sensitivity to phonological mismatch: considering the dynamics of looking behavior and pupil dilation

Katalin TAMÁSI[1,2*], Cristina MCKEAN[3], Adamantios GAFOS[1], and Barbara HÖHLE[1]

[1]International Doctorate in Experimental Approaches to Language and the Brain, Universities of Potsdam (DE), Newcastle (UK), Groningen (NL), Trento (IT), and Macquarie University, Sydney (AU), [2]Humanities and Social Sciences, Singapore University of Technology and Design, and [3]School of Education, Communication & Language Sciences, Newcastle University (UK)
*Corresponding author: Singapore University of Technology and Design, 8 Somapah Road, Singapore 48732. E-mail: katalin_tamasi@sutd.edu.sg

(Received 20 September 2017; revised 14 May 2018; accepted 23 May 2018;
first published online 4 September 2018)

Abstract

In a preferential looking paradigm, we studied how children's looking behavior and pupillary response were modulated by the degree of phonological mismatch between the correct label of a target referent and its manipulated form. We manipulated degree of mismatch by introducing one or more featural changes to the target label. Both looking behavior and pupillary response were sensitive to degree of mismatch, corroborating previous studies that found differential responses in one or the other measure. Using time-course analyses, we present for the first time results demonstrating full separability among conditions (detecting difference not only between one vs. more, but also between two and three featural changes). Furthermore, the correct labels and small featural changes were associated with stable target preference, while large featural changes were associated with oscillating looking behavior, suggesting significant shifts in looking preference over time. These findings further support and extend the notion that early words are represented in great detail, containing subphonemic information.

Keywords: lexical development; featural distance; mispronunciation detection; eye-tracking; pupillometry

Introduction

During language acquisition, infants face the challenge of identifying the building blocks of the ambient language by parsing the auditory input into discrete units and developing categories for those units. Depending on the stage of the child's development and/or processing demands, such categories can be words, syllables, phonemes, and subphonemic[1] features. The present study is concerned with the last

[1] By *subphonemic*, we mean features such as place of articulation, manner of articulation, and voicing which characterize dimensions of contrast below the level of the phoneme. *Subphonemic* and *subsegmental* have been used in this sense in the developmental literature (e.g., Mani & Plunkett, 2011a;

category, and specifically with whether early lexical representations contain information corresponding to subphonemic features. The ability to detect small yet contrastive changes in word forms is critical as it is a prerequisite to building an adult-like lexicon that contains words differing by a single feature (e.g., *cod* vs. *god*, where the voicing feature of the initial consonant changes from voiceless to voiced).

One approach to investigating the specificity of lexical representations is to present children with correctly pronounced vs. manipulated labels (e.g., *cod* vs. *fod*). Studies using a variety of online processing paradigms have demonstrated that, by the second year, children have the ability to differentiate a correctly pronounced label from a featurally manipulated one. Differential response to correctly pronounced vs. manipulated labels has been found with a variety of methods appropriate for testing young children. In intermodal preferential looking and headturn preference paradigms, stronger target preference has been observed when children are presented with correctly pronounced target labels than with manipulated target labels (Arias-Trejo & Plunkett, 2010; Bailey & Plunkett, 2002; Ballem & Plunkett, 2005; Durrant, Luche, Cattani, & Floccia, 2015; Fikkert, 2010; Höhle, van de Vijver, & Weissenborn, 2006; Mani, Coleman, & Plunkett, 2008; Mani & Plunkett, 2010a, 2010b, 2011a, 2011b; Ramon-Casas, Swingley, Sebastián-Gallés, & Bosch, 2009; Ren & Morgan, 2011; Swingley, 2003, 2005, 2016; Swingley & Aslin, 2000, 2002; Vihman & Croft, 2007; White & Morgan, 2008). In the switch paradigm, mispronounced labels lead to longer looking at a picture showing the referent of a trained word than the correctly pronounced label (Fennell & Werker, 2003; Werker, Fennell, Corcoran, & Stager, 2002; Yoshida, Fennell, Swingley, & Werker, 2009). In single-picture paradigms, differences in event-related brain potential (ERP) signatures are found between correct vs. manipulated labels (Mani, Mills, & Plunkett, 2012); and greater pupil dilation has been documented in response to manipulated than to correctly pronounced labels (Fritzsche & Höhle, 2015; Tamási, 2017; Tamási, McKean, Gafos, Fritzsche, & Höhle, 2017). These findings suggest that early lexical representations are sufficiently specified such that the infant is sensitive to the difference between the correctly pronounced and manipulated target form.

However, the precise degree of detail in early lexical representations requires further investigation. Although the findings reviewed above have securely established that infants can detect featural manipulations, the mere detection of a change does not demonstrate that they encode information corresponding to subphonemic features. It is possible that infants respond differentially to correct versus featurally manipulated labels only because they are able to detect a mismatch between the correct form and any kind of manipulation, regardless of the number of featural changes between correct and manipulated label.

For a stricter test of the hypothesis that early lexical representations contain information corresponding to features,[2] it is necessary to test whether infants can detect the degree of phonological mismatch (in terms of the number of features)

White & Morgan, 2008). We remain agnostic on the nature of those features, i.e., whether they are acoustic/phonetic, gestural, or abstract/phonological.

[2] We say ‘information corresponding to features’ rather than the more straightforward ‘contain features’ because, as we have noted elsewhere, sounds that differ in terms of phonological features also differ acoustically (Tamási *et al.*, 2017). We do not address whether features are innate or inferred from acoustics. Moreover, it is unclear to what extent the degree of acoustic difference between the correct and incorrect forms correlates with the degree of featural distance or whether acoustic difference and featural distance independently modulate experimental results (see Tamási *et al.*, 2017, for relevant discussion). These are important issues we do not take up in this work. Nevertheless, our working

https://doi.org/10.1017/S0305000918000259 Published online by Cambridge University Press

between a correctly and an incorrectly pronounced label. If infants are sensitive to the degree of phonological mismatch, they would respond differentially not only to correct vs. manipulated labels, but also to small vs. large degrees of mismatch. This sensitivity would suggest that lexical processing makes use of subphonemic information, which in turn would indicate that early lexical representations contain subphonemic detail. Whether infants are indeed sensitive to differing degrees of mismatch between the correct label and the auditory input is not yet well established, as some findings point to the presence of gradient sensitivity (Mani & Plunkett, 2011a; Ren & Morgan, 2011; White & Morgan, 2008), while some suggest the lack thereof (Bailey & Plunkett, 2002; Swingley & Aslin, 2002). In one key study, White and Morgan (2008) introduced one, two, and three featural changes to consonants in the word onset involving place, voicing, and manner features (e.g., *keys*: correctly pronounced label; *teys*: one-feature manipulation involving a change in place; *deys*: two-feature manipulation involving changes in place and voicing; *zeys*: three-feature manipulation involving changes in place, voicing, and manner). Crucially, a condition with unfamiliar labels bearing no resemblance to the correctly pronounced label and unknown to the children was also included. White and Morgan presented children with a familiar target picture (e.g., that of keys) along with an unfamiliar distractor picture (e.g., that of a trophy) in each trial. In this study, infants' overall looks towards the target picture declined in response to the increase of phonological mismatch in the target label. The overall proportion of target looking time was highest in the correct target label condition, followed by the one-feature change condition, which in turn was followed by the two- and the three-feature change conditions. The unfamiliar label condition exhibited the lowest proportion of target looking time. A reliable difference between one- vs. two- and three-feature change conditions, but not between two- and three-feature change conditions, indicated partial sensitivity to the degree of phonological mismatch. Comparable findings have been obtained by manipulating consonants in the word coda (Ren & Morgan, 2011) and vowels (Mani & Plunkett, 2011a).

However, some studies do not corroborate such findings (Bailey & Plunkett, 2002; Swingley & Aslin, 2002). In these studies, looking behavior was not linked to the degree of phonological mismatch as no difference was obtained in the proportion of infants' target looking time when presented with one vs. more featural changes in the target word onset. White and Morgan (2008) attributed these null results to the use of familiar distractor pictures with labels known to children. They argue that studies which employ familiar target pictures and unfamiliar distractor pictures engage a word-learning mechanism called 'mutual exclusivity'. This mechanism allows infants to map labels to unfamiliar referents quickly and efficiently. When an infant encounters a familiar and an unfamiliar referent (e.g., keys and a trophy) and hears a label that does not correspond to the familiar referent label (e.g., the label *trophy* does not correspond the label *keys*), the infant will likely infer that the unfamiliar label may be considered as the label of the unfamiliar referent (Markman & Wachtel, 1988). Thus, in the studies of White and Morgan (2008), Mani and Plunkett (2011a), and Ren and Morgan (2011), infants' unfamiliarity with the distractor picture (e.g., trophy) might induce them to associate the featurally manipulated labels to the distractor. In contrast, studies employing familiar targets and distractors do not elicit

hypothesis is clear. If sensitivity to degree of featural mismatch can be demonstrated, then this is converging evidence that representations contain information that is functionally equivalent to features.

https://doi.org/10.1017/S0305000918000259 Published online by Cambridge University Press

this association between the heard label and the distractor picture (Bailey & Plunkett, 2002; Swingley & Aslin, 2002). In sum, employing an unfamiliar distractor picture seems to contribute to observing gradient sensitivity (Mani & Plunkett, 2011a; Ren & Morgan, 2011; White & Morgan, 2008) and we adopt this design aspect in our study.

We now turn to additional aspects of our design that distinguish our work from all previous work on the topic. As mentioned above, the most popular paradigm to assess children's lexical knowledge is intermodal preferential looking, typically conducted with an eye-tracker (for a recent overview, see Golinkoff, Ma, Song, & Hirsh-Pasek, 2013). Even though paradigms involving eye-tracking yield a valuable body of data, only a fraction thereof is routinely analyzed in developmental studies. The most widely reported measure is the overall proportion of target looking time (Golinkoff *et al.*, 2013).

Our present study extends the intermodal preferential looking paradigm in two main respects. First, we test whether exploration of the dynamic measures offered by eye-tracking experiments provide additional insights into early lexical processing. Although some studies do provide time-course graphs and/or offer descriptive analyses (e.g., Arias-Trejo & Plunkett, 2010; Fritzsche & Höhle, 2015; Swingley & Aslin, 2000), systematic analysis of time-course data in the developmental literature is rare (as reviewed by Luche, Durrant, Poltrock, & Floccia, 2015). To remedy this, here children's looking preferences were observed and analyzed over time in order to monitor any systematic changes due to the experimental manipulation and to identify reliable shifts in preference. Second, the looking time measure is complemented with a measure automatically collected via the eye-tracker: pupil dilation. As an early psycho-sensory reflex, greater degree of pupil dilation in children has been linked to increased cognitive effort, violation of expectation, novelty, and arousal (Beatty & Lucero-Wagoner, 2000; Karatekin, 2007), making it an appealing measure for probing infant knowledge and processing. Previous studies of social cognition that simultaneously employed looking time and pupil dilation measures found the pupillary response stable and unaffected by practice, fatigue, and test-order effects, speaking to its robustness as a measure (Jackson & Sirois, 2009; Sirois & Jackson, 2007). For this reason, pupil dilation data supplemented looking time data when the latter were uninformative due to test-order effects (Jackson & Sirois, 2009) or showing no difference across the experimental conditions (Geangu, Hauf, Bhardwaj, & Bentz, 2011; Hepach & Westermann, 2013). In the domain of language, recent work suggests that pupillometry may be a promising method in infant research. Using pupillometry, studies have reported children's sensitivity to acoustic (dis-)similarity (Hochmann & Papeo, 2014), semantic incongruity (Kuipers & Thierry, 2013), and – most crucially for the current study – featural manipulations resulting in mispronunciations (Fritzsche & Höhle, 2015; Tamási, 2017; Tamási *et al.*, 2017; Tamási, Wewalaarachchi, Höhle, & Singh, 2016).

Most relevant to our work, recent studies using single-picture pupillometry – presenting a single visual stimulus per trial – have shown that 30-month-old children respond differently to correctly pronounced labels vs. their mispronunciations. Manipulated labels were associated with larger degrees of pupil dilation than correct labels (Fritzsche & Höhle, 2015; Tamási, 2017; Tamási *et al.*, 2017). In a similar vein, a recent study that employed the intermodal preferential looking paradigm found that bilingual children exhibited an elevated pupillary response to manipulated vs. correct labels (Tamási *et al.*, 2016). The increased pupil dilation in these studies was interpreted to indicate that greater cognitive effort was

https://doi.org/10.1017/S0305000918000259 Published online by Cambridge University Press

needed to establish a link between the referent and the manipulated label than doing so with the correct label. This finding and interpretation is consistent with the studies that investigated the specificity of lexical representations with other methodologies described above.

Using the single-picture pupillometry paradigm, 30-month-old children demonstrated gradient sensitivity to the degree of mispronunciation based on their pupil dilation patterns (Tamási *et al.*, 2017). The degree of mismatch between the correct and manipulated form defined by the number of featural changes in the onset consonant was positively correlated with the degree of pupil dilation (i.e., the more featural changes were introduced to the label, the greater the resulting pupil dilation). Specifically, the one-feature change condition was associated with larger degrees of pupil dilation than the correct condition, and the two- and three-feature change conditions were in turn associated with larger degrees of pupil dilation than the one-feature change condition. These findings are again in line with intermodal preferential looking studies that demonstrated partial gradient sensitivity to the degree of mismatch, thus indicating early lexical representations to be fine-grained (Mani & Plunkett, 2011a; Ren & Morgan, 2011; White & Morgan, 2008). Note, however, that complete gradient sensitivity to the degree of mismatch, which would have been indicated by significant differences between each degree of featural manipulation, was again not demonstrated.

In the current study, one of our objectives was to test how sensitive the looking time and pupil dilation measures are to the degree of mismatch in a classical intermodal preferential looking paradigm with two pictures. Following past intermodal preferential looking studies (Ren & Morgan, 2011; White & Morgan, 2008), the degree of mispronunciation was manipulated by featural distance (0–3 featural changes to the correct label and a semantically and phonologically unrelated unfamiliar label, e.g., [b]*aby*, correct / [d]*aby*, Δ1F / [f]*aby*, Δ2F / [S]*aby*, Δ3F / *sushi*, unfamiliar label). While children were presented with familiar target and unfamiliar distractor referents and the auditory label, both their looks and pupillary responses were monitored. We expect that, as the degree of mismatch increases, children's looking preference will shift towards the distractor picture, indicating a growing tendency to associate the label with the unfamiliar distractor instead of the familiar target (Mani & Plunkett, 2011a; Ren & Morgan, 2011; White & Morgan, 2008). Extrapolating from the findings of past studies using single-picture pupillometry paradigms, mispronunciation was expected to increase the effort of recognizing the heard label and integrating it with the target picture and the corresponding lexical entry, resulting in larger degrees of pupil dilation (Fritzsche & Höhle, 2015; Tamási, 2017; Tamási *et al.*, 2017). We expect degree of phonological mismatch to be a predictor of both looking behavior (given the findings from intermodal preferential looking paradigms: Mani & Plunkett, 2011a; Ren & Morgan, 2011; White & Morgan, 2008) as well as of pupillary response (based on the findings of a single-picture pupillometry study: Tamási *et al.*, 2017).

We furthermore asked whether time-course analyses will increase the chances of uncovering complete gradient sensitivity to the degree of phonological mismatch. Complete gradient sensitivity would imply that infants are capable of differentiating not just between one vs. more (Mani & Plunkett, 2011a; Ren & Morgan, 2011; Tamási *et al.*, 2017; White & Morgan, 2008), but also between two vs. three featural changes. If early words are represented in units that correspond to features, then it follows that infants would be able to differentiate one- and two-feature changes and

https://doi.org/10.1017/S0305000918000259 Published online by Cambridge University Press

also two- and three-feature changes. Upon encountering a label, it has been posited, infants may apply a 'gradient criterion' in which the likelihood of identifying the word as novel is a function of featural distance between the heard and the stored label (Swingley, 2016). However, past research suggests that there may be a point after which increasing featural distance does not increase the probability of the novel word interpretation anymore, i.e., more than two feature changes (Mani & Plunkett, 2011a; Ren & Morgan, 2011; Tamási *et al.*, 2017; White & Morgan, 2008). It is conceivable that employing time-course analyses in an intermodal preferential looking paradigm, as we do here, better enables one to characterize the process by which infants settle on their interpretation of the label as familiar or novel. We anticipate that some conditions will exhibit stable preference for one or the other picture, and that other conditions will exhibit preference shifts over time, potentially indicating the probabilistic decision making regarding novel word status posited by Swingley (2016). With this potential for increased sensitivity to more probabilistic and gradient differences in processing, we therefore hypothesize that time-course analyses may help to uncover gradient sensitivity across the range of featural changes if they exist. Finding evidence for the additive effect of featural changes would further strengthen the claim that early word processing is affected by featural manipulations and therefore that lexical representations encode information corresponding to features.

Method

Participants

Fifty-nine 30-month-old children (*M* = 30 months 7 days, *SD* = 16 days, 32 boys), all monolingual speakers of German, were recruited from the BabyLAB Participant Pool at the University of Potsdam. Caregivers reported no sensory and developmental disorders. Children's vocabulary knowledge and familiarity with the experimental items was assessed using a parental vocabulary checklist FRAKIS (i.e., the German adaptation of the MacArthur-Bates Communicative Development Inventory: Szagun, Schramm, & Stumper, 2009) and a further vocabulary checklist including the unfamiliar referents' labels (listed in Table 1). The children's reported average vocabulary (*M* = 451.1; *SD* = 91.9) aligned closely with FRAKIS norms of German-speaking children of the same age (*M* = 439; Szagun *et al.*, 2009).

Stimuli

A total of 20 mono- and disyllabic experimental items were selected from the parental checklist (Szagun *et al.*, 2009) and were recorded by a German native speaker who produced them in an enthusiastic, child-directed manner (see Table 1). Fifteen of the experimental labels were assumed to be familiar to 30-month-old children (i.e. taken from the parental checklist) and five were assumed to be unfamiliar (not part of the parental checklist).

Degree of mispronunciation in the onsets of the familiar words was manipulated so as to create four conditions: correct (unchanged) (e.g., *Bett*, [bɛt], 'bed'); one-feature change (e.g., [pɛt], voicing change); two-feature change (e.g., [kɛt], voicing and place of articulation change); and three-feature change (e.g., [ʃɛt], voicing, place of articulation, and manner of articulation change). Each version of the task contained

Table 1. Stimulus List, Organized by Familiar–Unfamiliar Word Pairs and Condition, Noted with IPA (Labeled = Words Labeled during Trials, Corr = Correctly Pronounced Label, Δ1F = One-Feature Change, Δ2F = Two-Feature Change, Δ3F = Three-Feature Change Introduced to the Onset, Not Labeled = Words Not Labeled during Trials, Given only in English)

Labeled	Corr	Δ1F	Δ2F	Δ3F	Not labeled
Familiar					Unfamiliar
Bett (*bed*)	b	p	k	ʃ	tapir
Boot (*boat*)	b	d	z	ʃ	American pancake
Decke (*blanket*)	d	t	v	f	magenta
Dusche (*shower*)	d	t	p	f	microscope
Fahne (*flag*)	f	v	t	d	magnet
Fisch (*fish*)	f	p	z	g	ruler
Fuß (*foot*)	f	p	b	g	tarsier
Kaffee (*coffee*)	k	t	ʃ	v	coati
Pony (*pony*)	p	t	v	z	avocado
Schaf (*sheep*)	ʃ	t	d	g	static eliminator
Teddy (*teddy*)	t	p	b	v	eyelash curler
Tisch (*table*)	t	d	b	v	sun dial
Sofa (*sofa*)	z	v	b	p	butter curler
Sonne (*sun*)	z	d	f	p	caviar
Suppe (*soup*)	z	d	t	k	weasel
Unfamiliar					**Familiar**
Dodo (*dodo*)	d	–	–	–	cheese
oliv (*olive*)	o	–	–	–	scissors
Säge (*saw*)	z	–	–	–	comb
Sushi (*sushi*)	z	–	–	–	baby
Yak (*yak*)	j	–	–	–	book

six correct labels, three one-feature change labels, three two-feature change labels, and three three-feature change labels. Each manipulation changed the voicing, place of articulation, or manner of articulation features of the label onset. The type of change was counterbalanced in the one- and the two-feature change conditions. Direction of featural change (voiceless/voiced; labial/coronal/dorsal; stop/fricative) was also counterbalanced. Mispronunciations resulted in nonwords for the children.[3] Unfamiliar words were always presented with their correct pronunciation.

Easily recognizable color drawings depicting the referents of the experimental items were selected and converted to a similar size (approximately 200 × 200 pixels displayed in a 300 × 300 pixel area). The areas of interest included the 400 × 400 pixel range

[3] Two real words produced by the manipulation (*Kuppe*, 'knoll', and *Wisch*, 'note') are unlikely to be known by 30-month-olds. Reanalyses excluding those two items did not change the overall results.

https://doi.org/10.1017/S0305000918000259 Published online by Cambridge University Press

around each picture. Additional pictures, 15 unfamiliar and 5 familiar pictured referents, were chosen as distractors. These pictures were paired with labeled pictures and thus they were never labeled. This resulted in altogether 20 familiar–unfamiliar picture pairings (shown in Table 1). The side on which familiar and unfamiliar pictures appeared was counterbalanced.

Four versions of the task were created, each picture pair occurring once in each version with the mispronunciation types counterbalanced across the four versions; children never saw the same picture or heard the same label more than once. Each participant was randomly assigned to one of the versions. Participants were presented with six correctly pronounced familiar labels, five correctly pronounced unfamiliar labels, and nine incorrectly pronounced familiar labels in each version of the experiment. The proportion of correctly vs. incorrectly pronounced labels was similar to that of Experiment 1 of White and Morgan (2008), which employed the same conditions as the present study.

Procedure

Children were told that they were to watch a short movie, during which they should sit still and as a reward they could choose a booklet afterwards. After obtaining assent from the children and written informed consent from the caregiver, children were seated in their caregiver's lap and positioned such that their eyes were approximately 60 cm away from the computer screen. Their gaze direction and pupil size were monitored by a Tobii 1750 corneal reflection eye-tracker (temporal resolution: 50 Hz, spatial accuracy: $.5'$ to $1'$, recovery time after track loss: 100 ms). All visual stimuli were shown on a 17″ (1280 × 1024) TFT screen with a size of 850 × 300 pixels (the two 300 × 300 pixel experimental pictures were separated by a 250 × 300 pixel gray strip and were positioned centrally) forming a horizontal viewing angle of 10.5° and a vertical viewing angle of 7.4°. The experiment started following the calibration period (five screen positions, ≈30 seconds).

The experiment encompassed four blocks, each containing five trials (altogether 20 trials). The order of the experimental items was furthermore pseudo-randomized such that onsets were not repeated (e.g., *Bett* and *Boot* did not follow each other), target onsets were not repeated (e.g., *Bett* and *Doot* did not follow each other as *Boot*, the correct form of *Doot*, shares an onset with *Bett*), and correctness status was not repeated more than four times (e.g., *Bett*, *Decke*, *Pony*, and *Fisch* in a row was not a possible ordering). With the aim of keeping the children engaged and conveying a sense of progress throughout the experiment, a 'progression marker' was presented before each block and after the last one (i.e., silent movie clips, featuring snails that initially line up on the left and one by one crawl to the right side of the screen). The clips were played in a loop until the experimenter pressed a key to start the next block. On average, the experiment lasted 7 minutes.

Each trial consisted of a salience phase, a centering, and a naming phase (illustrated in Figure 1). In the salience phase, a pair of target and distractor pictures were simultaneously presented on a gray background for 3000 ms. In order to reorient the children towards the center of the screen, a flashing red star was presented thereon for 1000 ms during the centering phase. In the naming phase, the same pair of pictures as in the salience phase was presented again for 3000 ms and was accompanied by an auditory label.

After the experiment, caregivers were asked to complete a questionnaire in order to estimate the child's vocabulary size and their familiarity with the experimental words.

https://doi.org/10.1017/S0305000918000259 Published online by Cambridge University Press

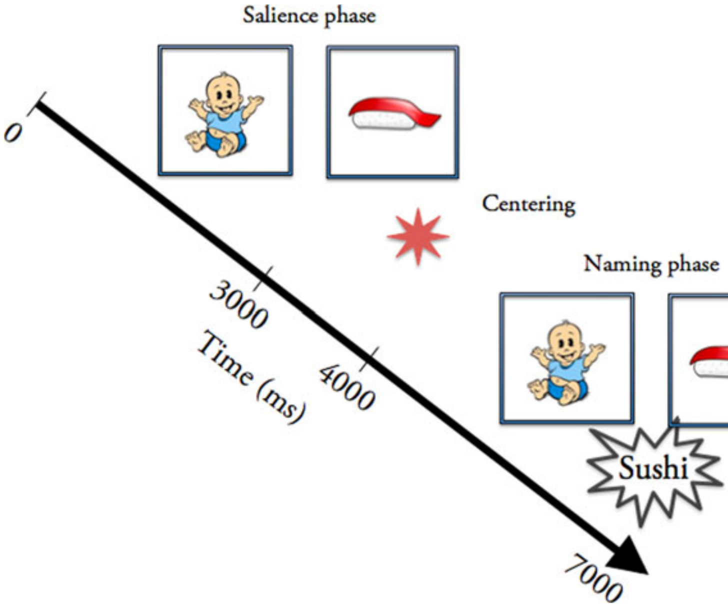

Figure 1. Trial structure. The 0–3000 ms time interval is the salience phase, whereby a pair of familiar target picture and unfamiliar distractor picture is shown. It is followed by the 3000–4000 ms centering phase, whereby a flashing star is shown. This is in turn followed by the 4000–7000 ms naming phase, which presents the same pair of pictures accompanied by a label.

Apart from the parental checklist (Szagun *et al.*, 2009), the questionnaire comprised the labels of the purportedly unfamiliar referents. On average, the questionnaire took 20 minutes to complete.

Results

To help ensure that the words intended to be familiar were part of the participants' vocabulary, only words reported to be known on the parental checklist for each child (Szagun *et al.*, 2009) were included in the analyses ($M = 74\%$, $SD = 16.9$). Conversely, among the experimental labels that were intended to be unfamiliar (i.e., the distractor labels), only those reported as such were included (of the remaining trials: $M = 93.2\%$, $SD = 10.6$). Those participants who did not reach a threshold of 50% of successful trials (trials containing pupil measures from at least half the length of the trial, following Fritzsche & Höhle, 2015) were excluded from further analyses (8 participants). Two additional children were excluded due to providing large negative difference scores (proportion of target looks during naming phase – salience phase < -0.15) in the correct condition (following White & Morgan, 2008). On average, 88% of trials per participant were retained (35.14/40 trials).

Looking behavior

The prediction that featural distance is negatively correlated with target looking time was supported by observations, as shown in the bar plot in Figure 2, and was

https://doi.org/10.1017/S0305000918000259 Published online by Cambridge University Press

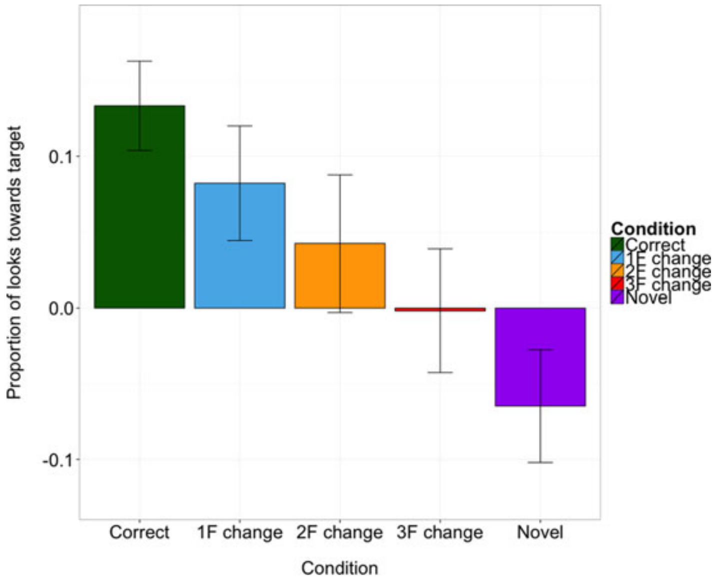

Figure 2. Mean proportion of looking time towards target in response to differing degrees of mispronunciation (error = *95% CI*). Values are baseline-corrected by subtracting the mean proportion of looking preference in the salience phase from that of the naming phase.

confirmed by statistical analyses. Linear mixed effects models were employed with random intercepts and slopes using the `lmer` function (estimates were chosen to optimize the log-likelihood criterion) in the **`lme4 R`** package **`lme4`**. **`Degree of mispronunciation`** (Correct / Δ1F / Δ2F / Δ3F / Unfamiliar) was assigned a polynomial contrast. Specifically, the first level of the contrast tested for a linear trend, the second for a quadratic trend, the third for a cubic trend, and the fourth for a quartic trend across the five conditions. **`Degree of mispronunciation`** was entered into the model as a fixed effect. Potentially confounding factors such as word frequency, and neighborhood density and phonotactic probability, all calculated from the Clearpond database (Marian, Bartolotti, Chabal, & Shook, 2012), were included as fixed effects in the model, along with children's vocabulary size that was estimated from the parental checklist. Participants ($N = 49$) and items ($N = 20$) were entered as random effects into the model. Overall proportion of looks towards the target in the naming phase – corrected for the proportion of looks in the salience phase – was used as the outcome measure. Since the salience phase establishes baseline looking preference, subtracting that from the looking preference of the naming phase indicates how preferences change in response to the experimental label. In this way, each trial can be used as its own baseline (White & Morgan, 2008). The first 200 ms of the naming phase, i.e., the period immediately after the centering phase, was excluded from analyses due to insufficient data (the minimum time required to launch fixations is longer; cf. Luche *et al.*, 2015). The linear mixed effects models were built with maximally specified random structure as justified by the design (Barr, Levy, Scheepers, & Tily, 2013; Jaeger, Graff, Croft, & Pontillo, 2011). Each intercept and slope fitted by the model was adjusted by the effect of condition nested in participants. The most parsimonious model was chosen through comparisons using Likelihood Ratio Tests (Pinheiro, Bates, DebRoy, & Sarkar, 2007) via the anova function from the **stats** package (R Core Team, 2014) and contained **degree of mispronunciation** as fixed effect. In this model, a significant negative linear trend was obtained ($\beta = -0.15$, $SE = 0.04$, $t = -4.19$) in response to the degree of mispronunciation. All other trends (quadratic, cubic, quartic) were found non-significant. Phonotactic probability was found to be a marginally significant positive predictor of target looking time ($\beta = 0.02$ $SE = 0.01$, $t = 1.81$).

Time-course analyses

To investigate the latency and the duration of contrasts between conditions, post-hoc cluster-based permutation tests (Maris & Oostenveld, 2007) were employed. Time-course analyses were used to explore when significant looking preferences emerged in response to differing degrees of mispronunciation (cf. Figure 3) using the **eyetrackingR** package (Dink & Ferguson, 2016). First, individual paired sample t-tests at each time sample were used to locate the significant t-values ($p < .05$) across the whole time-window. Second, clusters (e.g., contiguous significant t-values) were identified, for which a cluster-level t-value was given as the sum of all single sample t-values within the cluster. Third, the significance of cluster-level t-values were assessed by generating Monte Carlo distributions ($N = 2000$) thereof and determining the probability of their occurrence given the distribution. Those clusters whose t-statistic exceeded the threshold ($t = 2.8$, Bonferroni-corrected for multiple comparisons) were then tabulated for each contrast between conditions. The magnitude of contrasts in the identified clusters was then estimated by least square means (using the **lsmeans** function from the **lmerTest** package: Kuznetsova, Brockhoff, & Christensen, 2015). With this method, the following clusters were identified (using the **time_cluster_data** function in the **eyetrackingR** package: Dink & Ferguson, 2016): steady target preference was observed in the correct condition almost across the whole naming phase (300–2400 ms: $\beta = 0.13$, $SE = 0.03$, $t = 4.52$) and in a slightly shorter time-window for the Δ1 F condition (300–2200 ms: $\beta = 0.08$, $SE = 0.04$, $t = 2.17$). Preferences flipped from target to distractor and back to target for the Δ2 F condition (target preference – 300–600 ms: $\beta = 0.05$, $SE = 0.03$, $t = 1.81$, and 1500–2800 ms: $\beta = 0.06$, $SE = 0.04$, $t = 1.74$, distractor preference – 900–110 ms: $\beta = -0.05$, $SE = 0.03$, $t = 1.78$). Preferences shifted from distractor to target for the Δ3F condition (distractor preference – 500–800 ms: $\beta = -0.05$, $SE = 0.03$, $t = -1.76$, target preference: 1500–2300 ms: $\beta = 0.06$, $SE = 0.03$, $t = 2.04$). For unfamiliar items, a distractor preference was observed across almost the whole time-window (200–2300 ms: $\beta = -0.06$, $SE = 0.03$, $t = -1.82$) Positive t-values are linked to target preference and negative ones to distractor preference.

In the time-course plot in Figure 3, differences were evident across all conditions. These observations were confirmed by time-course analyses that tested the significance of contrasts between each condition pair summarized in Table 2. Each pairwise comparison (i.e., comparisons between the correct and one-feature-change conditions, between the correct and two-feature-change conditions, etc.) was found to be significant. Some comparisons identified multiple significant time intervals (e.g., the comparisons between the featural change conditions vs. the unfamiliar condition).

https://doi.org/10.1017/S0305000918000259 Published online by Cambridge University Press

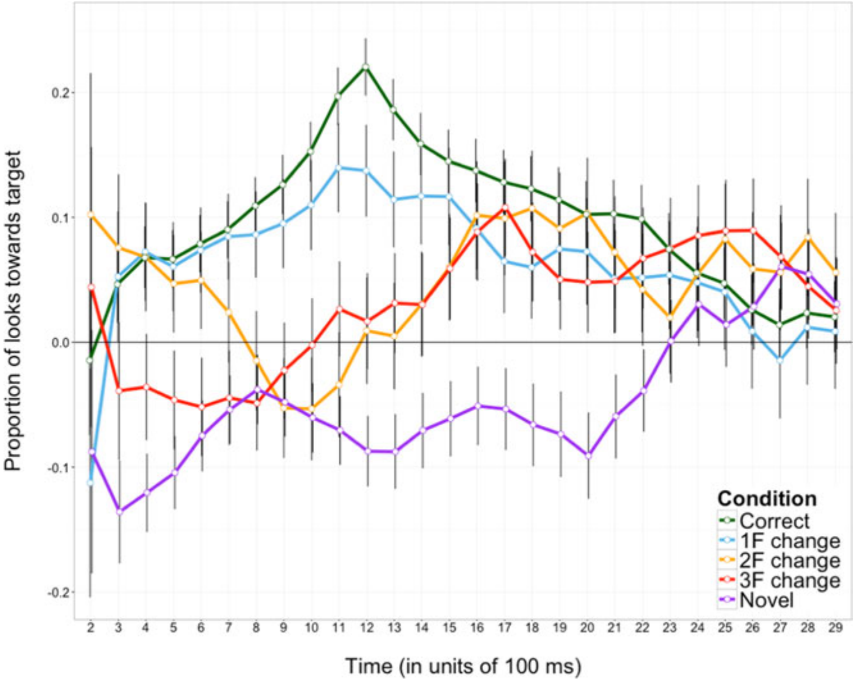

Figure 3. Proportion of looking time towards target over time in response to differing degrees of mispronunciation (error = 95% *CI*). Values are baseline-corrected by subtracting the mean proportion of looking preference in the salience phase from that of the naming phase. Time is provided in units of 100 ms.

Pupillary response

In the pupil dilation measure, a positive linear trend in pupil dilation in response to the degree of mispronunciation was obtained ($\beta = 0.05$, $SE = 0.02$, $t = 1.85$) in an analysis parallel to that of the linear mixed effects models of the looking time measure (cf. Figure 4). The only modification to the model involved the outcome measure: overall mean pupil dilation (mm) in the naming phase, baseline-corrected in each trial by the trial-wise minimum value. No other trends (quadratic, cubic, quartic) were found to be significant. In addition to **`degree of mispronunciation`**, **`vocabulary size`** was also a significant positive predictor ($\beta = 0.06$, $SE = 0.02$, $t = 3.63$).

Time-course analyses

Visual inspection of the time-course plot in Figure 5 indicates differences between the correct and three-feature change conditions, and between the unfamiliar and all the other conditions. Across-condition time-course analyses, identical to those performed on the looking time data, supported these observations. The summary of the time-course analyses is provided in Table 2. All contrasts between the unfamiliar and each of the other conditions (cf. rows 4, 7, 9, and 10 in Table 2) reached significance at the $p = .05$ criterion. The contrast between the correct and three-feature change conditions was found to be not significant at the $p = .05$ criterion, but the p value was <.1 (cf. row 3 in Table 2).

https://doi.org/10.1017/S0305000918000259 Published online by Cambridge University Press

Table 2. Significant Contrasts across Conditions in Time-course Analyses (Interval = Time Interval in the Naming Phase, $\sum t$ = Cluster-Level t-value, p = p-value Associated with Cluster-Level t, Corr = Correctly Pronounced Familiar Label, Δ1 F = One-Feature Change, Δ2 F = Two-Feature Change, Δ3 F = Three-Feature Change Introduced to the Onset, Unfamiliar = Unfamiliar Label)

	Looking time			Pupil dilation		
Contrasts	Interval (ms)	$\sum t$	p	Interval (ms)	$\sum t$	p
Corr vs. Δ1F	1200–1400	−3.31	*	–	–	–
Corr vs. Δ2F	800–1500	−19.30	**	–	–	–
Corr vs. Δ3F	400–1600	−28.69	***	1500–2900	2.33	†
Corr vs. Unfamiliar	300–2300	−78.36	***	1300–2900	41.92	*
Δ1F vs. Δ2F	900–1200	−5.33	*	–	–	–
Δ1F vs. Δ3F	400–900	−11.83	**	–	–	–
Δ1F vs. Unfamiliar	300–1800	−43.77	***	1300–2700	27.10	*
	1900–2200	−4.96	*			
Δ2F vs. Δ3F	300–800	−10.10	*	–	–	–
Δ2F vs. Unfamiliar	300–800	−16.07	**	1300–2900	35.08	**
	1500–1900	−6.90	*			
	400–600	−3.18	†	1300–1500	3.31	†
Δ3F vs. Unfamiliar	1100–1400	−5.27	*	1600–2400	16.97	*
	1600–1800	−4.48	†			

Notes. †: $p < .1$, *: $p < .05$, **: $p < .01$, ***: $p < .001$.

Discussion

Looking behavior

Our analysis using the linear mixed effects modeling indicated that children's looking behavior was modulated by degree of mismatch such that increases in the degree of mismatch between the heard label and the correct target label resulted in fewer target looks. In the case of complete mismatch (i.e., the unfamiliar condition), the looking preference flipped to the distractor picture. Following previous work (e.g., Swingley & Aslin, 2000), we interpreted looking preference to indicate association between the heard label and the picture; the earlier and the more prolonged the looking preference towards a picture in response to a given auditory label, the stronger the established association between the picture and the label. Therefore, these findings indicate gradient sensitivity to featural distance and, as such, the present study corroborates previous work conducted in intermodal preferential looking paradigms (Mani & Plunkett, 2011a; Ren & Morgan, 2011; White & Morgan, 2008). As mentioned previously, some studies did not find sensitivity to the degree of mispronunciation (Bailey & Plunkett, 2002; Swingley & Aslin, 2002). In our study, just as in the studies of Ren and Morgan (2011), Mani and Plunkett (2011a), and White and Morgan (2008), the demonstrated sensitivity was possibly uncovered by using unfamiliar distractor pictures that could serve as plausible referents for the unfamiliar and manipulated labels and also by making use of the dynamics of the looking behavior in the analysis.

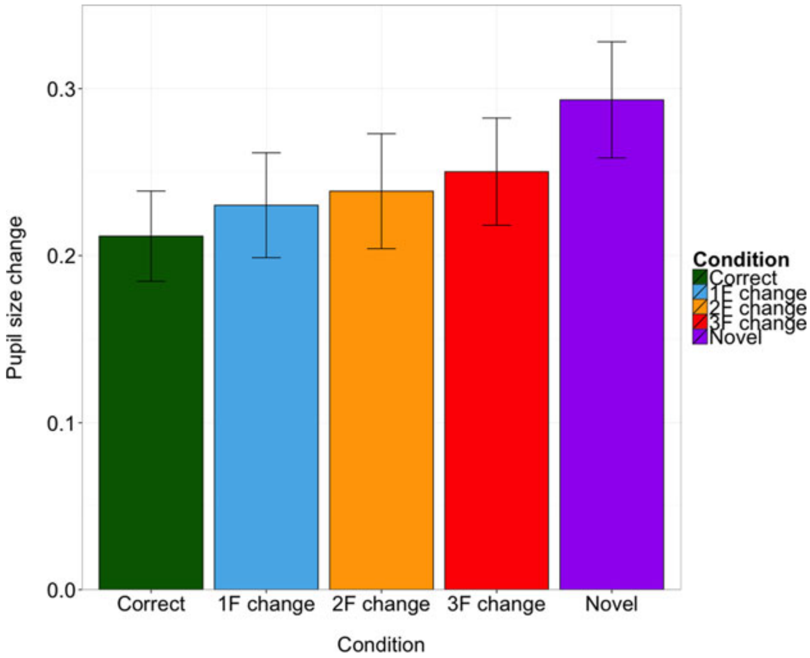

Figure 4. Mean pupil size change (mm) in response to differing degrees of mispronunciation (error = *95% CI*).

Time-course analyses of looking preference further revealed how stable those preferences were across the conditions. Steady target preference was recorded in response to the correct and Δ1F conditions and distractor preference in response to the unfamiliar condition, which is in line with earlier results that averaged over the naming phase (Mani & Plunkett, 2011a; Ren & Morgan, 2011; White & Morgan, 2008). Note that in these conditions, looking preferences did not significantly shift; that is, children were more inclined to look at one picture over another for the duration of almost the whole naming phase.

Most importantly, the results from time-course analyses extended previous work in two respects. First, they enabled the detection of significant looking preferences in featurally manipulated conditions wherein children showed different patterns of oscillation between distractor and target preferences across the conditions. Detecting shifts in looking preference would not have been possible with averaging techniques. The dynamic shifts in looking preference when presented with Δ2F and Δ3F labels suggest that children attempted to form links between those – largely mispronounced – labels and both pictures, yet stable link formation was disrupted by the mispronunciation manipulation of the current study: changed onset coupled with unchanged rhyme. In particular, Δ2F labels initially mirrored the pattern of correct and Δ1F conditions by exhibiting target preference (as a possible sign of attempting to associate the label with the target picture). This suggests that the Δ2F condition resembled the correctly pronounced target label so as to partially activate the target label. Preference then switched to the distractor (a potential attempt to map the label with the distractor picture), and finally shifted, reverting back to the target. Since even in the featurally manipulated conditions, the rhyme of the word was always

https://doi.org/10.1017/S0305000918000259 Published online by Cambridge University Press

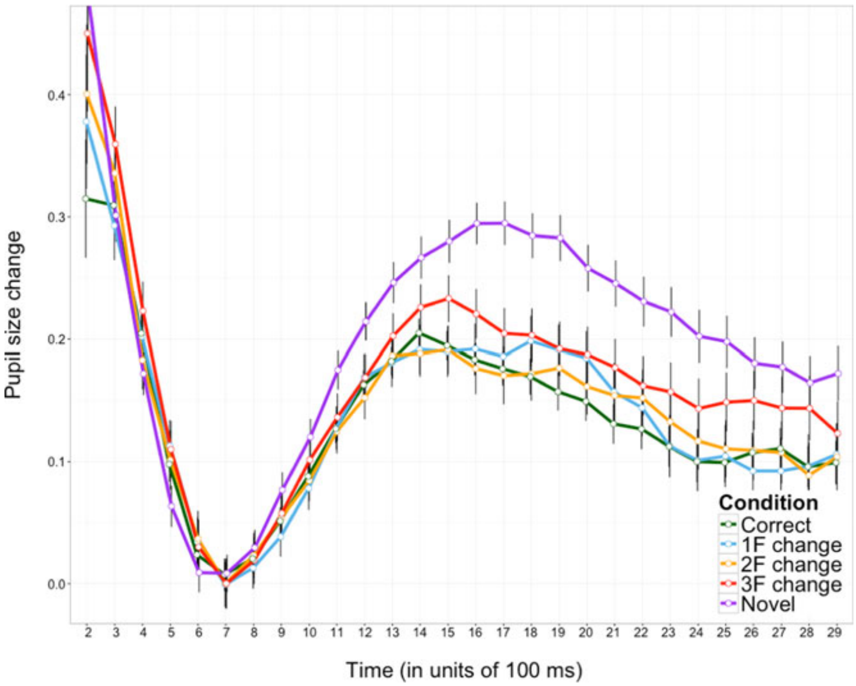

Figure 5. Pupil size change (mm) over time in response to differing degrees of mispronunciation (error = 95% *CI*). Time is provided in units of 100 ms.

produced correctly, this may have facilitated the – albeit interrupted – retrieval of the correct word form and its mapping to the target picture (i.e., rhyme effect). Note, however, that the rhyme effect alone would not be able to capture the gradedness obtained in response to the degree of featural manipulation. The Δ3F condition, unlike the Δ2F condition, followed a trajectory similar to the unfamiliar condition by exhibiting an initial phase of distractor preference, which signals an attempt to link the manipulated label with the distractor picture. Apparently, the combination of three featural changes to the onset resulted in a featural distance that failed to activate the target label. Eventually, however, looking preference shifted to the target picture, which, similarly to the shift in the Δ2F condition, could have been caused by rhyme identity with the correct label. Thus, the distractor-to-target shift that occurred around 1000–1500 ms in response to the Δ2F and Δ3F conditions was probably due to delayed consolidation of the largely mispronounced label with the correct lexical entry, and in turn delayed association with the target picture. We call this final preference shift observed in the large featural change conditions CONVERGENCE TO THE TARGET. Moving beyond the findings of previous work that have described overall target looking preference, the present findings expose how lexical activation is modulated differently over time, a result which remains opaque to averaging looking preference measures over the entire window of analysis. Specifically, degree of featural mismatch strongly drives target activation early on, as the locus of featural mismatch is localized to the word onset, shifting to the later

https://doi.org/10.1017/S0305000918000259 Published online by Cambridge University Press

convergence to the target as the effect of featural mismatch diminishes (and information about the rhyme becomes available).

The second set of novel findings involving time-course analyses of the looking time data revealed significant differences ACROSS ALL CONDITIONS (cf. Table 2). This is the first time that complete gradient sensitivity to the degree of mispronunciation was observed, as past research did not report differentiation between conditions containing large degrees of mispronunciation, Δ2F and Δ3F (Mani & Plunkett, 2011a; Ren & Morgan, 2011; Tamási *et al.*, 2017; White & Morgan, 2008). Time-course analyses in the present study found that differentiation between the looking patterns of correct pronunciation and small degree of mispronunciation (Δ1F) emerged relatively late at 1200 ms and only lasted for a short time (200 ms) in the naming phase (see Table 2). Featural distance between correct and manipulated forms positively predicted the latency and duration of differentiation from the correct pronunciation. That is, the larger the featural distance, the earlier and longer the differentiation: Δ2F – at 800 ms for 700 ms; Δ3F – at 400 ms for 1200 ms; unfamiliar label – at 300 ms for 3000 ms. In fact, this finding could be generalized to differences between any given condition pair: the more featural mismatch across conditions, the longer the differential response (one-step distance: 200–400 ms duration, two-step distance: 400–700 ms duration, three-step distance: 1200–1500 ms duration, four-step distance: 2000 ms duration; see Table 2).

It is worth noting that this finding of a gradient response across the range of featural distance does not necessarily imply a linear relationship between featural distance and word recognition. Although it is in principle possible that the difference between the one- and two- featural changes and that between the two- and three-featural changes are comparable in degree, it is more probable that featural changes interact, i.e., features sometimes produce a smaller and sometimes a larger effect relative to the sum of their separate effects. To address this issue, a systematic investigation of the unique contribution of each featural manipulation is required.

Even though our study was not designed to assess the role of lexical factors by employing them as control variables only in this task, some considerations are offered below on those that were retained in the most parsimonious models. Phonotactic probability was found to be a marginal positive predictor of looking preference. The relationship was such that the more frequent the target label's sublexical sequences were, the stronger the target looking preference. This may be expected, as phonotactic probability has been shown to have a facilitatory effect on adult word recognition (Vitevitch & Luce, 1998, 1999) and infant word learning (McKean, Letts, & Howard, 2013). In addition, based on previous findings with children, words with frequently occurring phonotactic sequences can be accessed more quickly (Edwards, Beckman, & Munson, 2004; Munson, Kurtz, & Windsor, 2005; Zamuner, Gerken, & Hammond, 2004).

Pupillary response

The prediction that degree of mispronunciation affects magnitude of pupil dilation was borne out given the positive trend obtained by the linear mixed effects models, in line with the findings of Tamási *et al.* (2017). Considering pupil dilation to be a direct measure of cognitive effort, this finding can be interpreted such that the more featural manipulations were introduced to the target label, the more cognitive resources were recruited to link the label and the target picture (Tamási *et al.*, 2017). In the absence of a phonological relationship between the unfamiliar label and the

https://doi.org/10.1017/S0305000918000259 Published online by Cambridge University Press

label of the target picture, children attempted to associate the unfamiliar label with the also unfamiliar distractor picture instead. The positive trend found by linear mixed effects models in pupil dilation appeared to be driven by the differences between the correct and Δ3F conditions and those between the unfamiliar and all other conditions, as no other contrasts were found significant in the time-course analyses. Namely, no significant differences were observed across correct and small-feature-change (Δ1F and Δ2F) conditions and across the manipulated label conditions (Δ1F vs. Δ2F, Δ1F vs. Δ3F, Δ2F vs. Δ3F). To account for the relatively suppressed gradient response to degree of mispronunciation in contrast to previous findings (Tamási *et al.*, 2017), we consider potential reasons that can stem from differences in design. To this end, it is important to revisit the findings that were obtained by analyzing children's pupillary response.

Recall that studies that employed the single-picture pupillometry paradigm found that the pupillary response of monolingual children was higher in single-feature mispronunciations relative to correctly pronounced labels (Fritzsche & Höhle, 2015; Tamási *et al.*, 2017). On the other hand, the one study that employed the intermodal preferential looking paradigm detected no significant difference between the pupillary responses of monolingual children given to correctly pronounced labels vs. single-feature and tonal mispronunciations (Tamási *et al.*, 2016). These results, plus our finding of no difference between the correct vs. Δ1F and Δ2F condition in the pupillary response measure, can be due to methodological differences between the single-picture pupillometry paradigm and the intermodal preferential looking paradigm.

We now turn to discuss potential methodological differences between the studies that may have suppressed the emergence of a differential response to small featural manipulations in the pupillary response measure. The first obvious difference between the intermodal preferential looking and the single-picture set-up is the inclusion of a distractor picture. To accommodate this change, the display size was enlarged, which in turn lead to a slightly larger viewing angle (from 7.4° × 7.4° to 10.5° × 7.4°). Furthermore, the inclusion of the distractor picture also increased complexity in the visual modality. In order to process the increased visual information presented to them, children presumably needed to launch relatively more fixations, which may have disrupted the emergence of the pupillary response. For this reason, we quantified children's fixation patterns (number of fixation changes between target and distractor pictures, latency of first fixation towards target and distractor pictures, and longest duration of fixation towards the target and distractor pictures) and assessed whether they had any bearing on the degree of pupil dilation. We found that, in the naming phase, children launched their first fixation towards the target after 798 ms ($SD = 711$) and towards the distractor after 960 ms ($SD = 897$), fixated the longest at the target for 1538 ms ($SD = 849$) and at the distractor for 1122 ms ($SD = 749$), and changed fixations between the target and the distractor pictures 1.72 ($SD = 1.12$) times. Considering that the naming phase lasted for 3000 ms, the longest fixation durations towards the target and the distractor pictures together account for most of the duration of the trial. However, since the latency to reach peak pupil dilation, 1169 ms (682 ms), is relatively long compared to some of the observed fixation durations, it is conceivable that the emergence of pupil dilation was interrupted by changes in fixation behavior. To test this, we entered the five above-mentioned variables characterizing fixation as predictors and degree of pupil dilation as the outcome measure in linear mixed

https://doi.org/10.1017/S0305000918000259 Published online by Cambridge University Press

effects models (Baayen, Davidson, & Bates, 2008) and selected the most parsimonious model by likelihood ratio tests (Pinheiro *et al.*, 2007) (the exact procedure is described in the 'Results' section in more detail). We found that none of the fixation variables predicted the degree of pupil dilation (models containing fixed effects did not significantly improve on models containing random intercepts and slopes only: all χ^2s < 1.99, *p*s > .15). This result suggests that the degree of pupil size change cannot be ascribed to fixation patterns.[4] Besides increasing the visual angle and complexity, both leading to changes in fixation behavior, the addition of the distractor picture also introduced changes to the implicit requirements of the task. In the single-picture pupillometry paradigm, a single referent is to be associated with the heard label, while in the intermodal preferential looking paradigm, the target and the distractor compete on matching the heard label. Consequently, the two-picture paradigm may have created weaker expectations as to the status of the upcoming auditory label relative to the single-picture set-up, where the picture was always semantically related to the label. Therefore, the two-picture set-up may have made the establishment of a semantic link between picture and heard label less straightforward.

The second difference between the studies is the relative timing of visual and auditory stimuli. In single-picture pupillometry paradigms (Fritzsche & Höhle, 2015; Tamási, 2017; Tamási *et al.*, 2017), the onset of visual stimuli preceded that of the auditory stimuli by 1000 ms, while, in the current study, only the visual stimuli were presented in the salience phase and the two types of stimuli were presented simultaneously in the naming phase. In the single-picture pupillometry studies, the silent presentation of visual stimuli for 1000 ms was introduced to provide time to process seeing the picture as well as to adjust to the luminance of the picture. In the present study, processing of the visual and auditory stimuli in the naming phase was simultaneous (similarly to other intermodal preferential looking studies), making the disambiguation of the effect of visual and auditory stimuli on the pupillary response difficult.[5]

Finally, children's vocabulary size was found to be a positive predictor in the pupil dilation measure. In the linear mixed effects models built on pupil dilation (discussed in the 'Results' section), the main effects model containing the experimental manipulation degree of mispronunciation and vocabulary size was more parsimonious than the interaction model, as determined by likelihood ratio tests (Pinheiro *et al.*, 2007). This indicates that the effect of vocabulary size on pupil dilation remained unaffected by the degree of mispronunciation, thus the general pattern is such that the larger the

[4] Using a method parallel to the one described, no significant relationships emerged between fixation patterns and the experimental manipulation, that is, degree of mispronunciation, either (all χ^2s < 1.91, *p*s > .21). This indicates that fixation behavior remains unaffected by the experimental manipulation. Thus, the only significant relationship was between the experimental manipulation and the degree of pupil dilation (as described in the 'Results' section), and neither of them was associated with fixation patterns.

[5] In fact, by looking at Figure 5, a contracting pupillary response can be observed for the duration of around 700 ms prior to dilation, a pattern that has also been noted in response to the visual stimuli in single-picture pupillometry paradigms in the first second of the trial (Study 3 in Tamási, 2017; Tamási *et al.*, 2017). Analyses in the present study included identical time-windows for the looking time and pupil dilation measures (200–3000 ms). In order to confirm that the initial constriction of the pupil in the naming phase did not influence the results, the mixed effects modeling was replicated with a 700–3000 ms time-window. The results were comparable to the analysis conducted in the whole time-window (linear trend: $\beta = 0.05$, $SE = 0.02$, $t = 2.02$).

https://doi.org/10.1017/S0305000918000259 Published online by Cambridge University Press

vocabulary size, the larger the degree of pupil dilation. Comparable results were obtained when the models were re-run with degree of mispronunciation collapsed into a binary variable (correctly vs. incorrectly pronounced target labels). The independent effect of vocabulary size on pupil dilation may be due to differences in cognitive functioning. Individual cognitive characteristics such as fluid intelligence and verbal and arithmetic skills are known to affect both baseline pupil size and task-evoked pupillary response in adults (Ahern & Beatty, 1979; Tsukahara, Harrison, & Engle, 2016; van der Meer *et al.*, 2010). Although such research is lacking with children, the findings of the present study and those gained from single-picture pupillometry (Tamási *et al.*, 2017) may be indicative of an association between individual differences in cognitive abilities as measured by vocabulary size and pupil dilation. Future research is needed to test whether children with greater vocabulary size do invest more cognitive effort in word recognition, and, if so, whether this reflects increased competition of denser lexical networks or larger overall recruitment of cognitive capacity. Note that this association between vocabulary size and performance was only obtained in the pupil dilation, not in the looking time measure in our study, an observation that mostly holds in the wider literature (no effect of vocabulary size found using looking time;[6] Bailey & Plunkett, 2002; Ballem & Plunkett, 2005; Swingley, 2005; Swingley & Aslin, 2000; Zesiger, Lozeron, Lévy, & Frauenfelder, 2011). Future studies should assess the extent to which the pupil dilation measure is sensitive to individual differences in cognitive and language skills. In particular, research using online methodologies that studies these factors remains scarce. Further investigation is warranted to assess the degree to which pupillometry may be sensitive to lexical and sublexical factors such as vocabulary size, neighborhood density, and phonotactic probability in children's word recognition.

Conclusions

The current study analyzed looking behavior in conjunction with pupil dilation data collected from a standard intermodal preferential looking paradigm to explore children's gradient sensitivity to the degree of mismatch. In our looking preference results, each level of increase in featural distance inhibited the association of the label with the target picture further. This novel finding regarding children's looking behavior was made possible by employing time-course analyses. Changing one feature in the target label onset weakened target preference and thus the overall association between label and target picture, indicating a weakening in the retrieval of the appropriate lexical entry. Changing two features introduced oscillation between target and distractor preference (target → distractor → target), suggesting entertaining the identification of the mispronounced lexical entry first as the target label, then as the label of the novel referent, but eventually switching back to target. Changing three features induced initial distractor preference that flipped to target preference, suggesting an attempt to link the label first with the distractor and then with the target picture and, as such, a delay in recovering the correct lexical representation. Therefore, despite perturbations introduced by the large featural changes, the looking behavior in both of those conditions eventually converged to

[6] A counter-example is Werker *et al.* (2002), which reported vocabulary size to be correlated with the head turn preference of 14-month-olds, but not that of 20-month-olds.

https://doi.org/10.1017/S0305000918000259 Published online by Cambridge University Press

the target, but with the within-trial shifts in preference clearly indicating differential effects of degree of featural mismatch as discussed above. Finally, in the case of complete mismatch between the target label and the heard label, children preferred the distractor picture, indicating their attempt at establishing a link between the two (due to mutual exclusivity). Taken together, these findings provide for the first time evidence for complete gradient sensitivity to phonological mismatch. As such, these findings add further support to the thesis that early lexical representations are fine-grained enough to encode subphonemic detail.

Furthermore, the pupil dilation measure has shown promise in complementing the intermodal preferential looking paradigm. The present findings were consistent with past research that found degree of phonological mismatch to have influenced cognitive effort as measured by pupillary response, although this effect was attenuated compared to previous work using other paradigms (Tamási *et al.*, 2017). Possible reasons for such attenuation involve changes in visual angle, complexity, and task requirements due to the addition of a distractor picture, the addition of the unfamiliar label condition, and simultaneous presentation of the visual and auditory stimuli in the naming phase, all of which can be addressed in future research.

Acknowledgements. This research was supported by grants EMJD 520101–1–2011–1–DE–ERA to B. Höhle and K. Tamási and ERC AdG 249440 to A. Gafos. The authors would like to thank the participants and their families; the BabyLAB personnel (Tom Fritzsche, Katja Schneller, Carolin Jaekel, Steffi Meister, Vanessa Löffler, Elisabeth Metz and Lennard Gottmann) for participant recruitment and testing; participants of ATTLIS, CLEAR, EWCN2016, and the IDEALAB Winter School 2016; and members of the Phonetics Lab (University of Potsdam), the Language Acquisition Colloquium (University of Potsdam), the Phonology and Phonetics Group (Newcastle University), and the Child Language Research Group (Newcastle University) for discussions and feedback.

References

Ahern, S., & Beatty, J. (1979). Pupillary responses during information processing vary with scholastic aptitude test scores. *Science*, *205*(4412), 1289–92.

Arias-Trejo, N., & Plunkett, K. (2010). The effects of perceptual similarity and category membership on early word–referent identification. *Journal of Experimental Child Psychology*, *105*(1), 63–80.

Baayen, R. H., Davidson, D. J., & Bates, D. M. (2008). Mixed-effects modeling with crossed random effects for subjects and items. *Journal of Memory and Language*, *59* (4), 390–412.

Bailey, T. M., & Plunkett, K. (2002). Phonological specificity in early words. *Cognitive Development*, *17*(2), 1265–82.

Ballem, K. D., & Plunkett, K. (2005). Phonological specificity in children at 1;2. *Journal of Child Language*, *32*(1), 159–73.

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: keep it maximal. *Journal of Memory and Language*, *68*(3), 255–78.

Bates, D., Maechler, M., Bolker, B., & Walker, S. (2014). lme4: linear mixed-effects models using Eigen and S4 [Computer software manual]. Retrieved from <http://CRAN.R-project.org/package=lme4> (R-package-version 1.1-6).

Beatty, J., & Lucero-Wagoner, B. (2000). The pupillary system. In J. T. Cacioppo, L. G. Tassinary, & G. G. Berntson (Eds.), *Handbook of psychophysiology*, 2nd ed., Vol. 2 (pp. 142–62). Cambridge University Press.

Dink, J., & Ferguson, B. (2016). eyetrackingR [Computer software manual]. Retrieved from <http://www.eyetracking-R.com> (R package version 0.1.6).

Durrant, S., Luche, C. D., Cattani, A., & Floccia, C. (2015). Monodialectal and multidialectal infants' representation of familiar words. *Journal of Child Language*, *42*(2), 447–65.

https://doi.org/10.1017/S0305000918000259 Published online by Cambridge University Press

Edwards, J., Beckman, M. E., & Munson, B. (2004). The interaction between vocabulary size and phonotactic probability effects on children's production accuracy and fluency in nonword repetition. *Journal of Speech, Language, and Hearing Research*, *47*(2), 421–36.

Fennell, C. T., & Werker, J. F. (2003). Early word learners' ability to access phonetic detail in well-known words. *Language and Speech*, *46*(2/3), 245–64.

Fikkert, P. (2010). Developing representations and the emergence of phonology: evidence from perception and production. In C. Fougeron, B. Kuehnert, M. Imperio, & N. Vallee (Eds.), *Laboratory phonology*, Vol. 10 (pp. 227–60). Berlin: de Gruyter.

Fritzsche, T., & Höhle, B. (2015). Phonological and lexical mismatch detection in 30- month-olds and adults measured by pupillometry. *Proceedings of ICPhS XVIII*. Retrieved from <http://www.ling.uni-potsdam.de/~fritzsche/assets/files/Fritzsche&Hoehle(2015)ICPhS.pdf>.

Geangu, E., Hauf, P., Bhardwaj, R., & Bentz, W. (2011). Infant pupil diameter changes in response to others' positive and negative emotions. *PLoS One*, *6*(11), e27132. doi: 10.1371/journal.pone.0027132

Golinkoff, R., Ma, W., Song, L., & Hirsh-Pasek, K. (2013). Twenty-five years using the intermodal preferential looking paradigm to study language acquisition: What have we learned? *Perspectives on Psychological Science*, *8*(3), 316–39.

Hepach, R., & Westermann, G. (2013). Infants' sensitivity to the congruence of others' emotions and actions. *Journal of Experimental Child Psychology*, *115*(1), 16–29.

Hochmann, J.-R., & Papeo, L. (2014). The invariance problem in infancy: a pupillometry study. *Psychological Science*, *25*(11), 2038–46.

Höhle, B., van de Vijver, R., & Weissenborn, J. (2006). Word processing at 19 months and its relation to language performance at 30 months: a retrospective analysis of data from German-learning children. *International Journal of Speech-Language Pathology*, *8*(4), 356–63.

Jackson, I., & Sirois, S. (2009). Infant cognition: going full factorial with pupil dilation. *Developmental Science*, *12*(4), 670–9.

Jaeger, T. F., Graff, P., Croft, W., & Pontillo, D. (2011). Mixed effect models for genetic and areal dependencies in linguistic typology. *Linguistic Typology*, *15*(2), 281–320.

Karatekin, C. (2007). Eye-tracking studies of normative and atypical development. *Developmental Review*, *27*(3), 283–348.

Kuipers, J.-R., & Thierry, G. (2013). ERP-pupil size correlations reveal how bilingualism enhances cognitive flexibility. *Cortex*, *49*(10), 2853–60.

Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2015). Package 'lmertest' [Computer software manual]. Retrieved from <http://CRAN.R-project.org/package=lmerTest> (R package version 2.0-29).

Luche, C. D., Durrant, S., Poltrock, S., & Floccia, C. (2015). A methodological investigation of the intermodal preferential looking paradigm: methods of analyses, picture selection and data rejection criteria. *Infant Behavior and Development*, *40*, 151–72.

Mani, N., Coleman, J., & Plunkett, K. (2008). Phonological specificity of vowel contrasts at 18-months. *Language and Speech*, *51*(1/2), 3–21.

Mani, N., Mills, D. L., & Plunkett, K. (2012). Vowels in early words: an event-related potential study. *Developmental Science*, *15*(1), 2–11.

Mani, N., & Plunkett, K. (2010a). In the infant's mind's ear: evidence for implicit naming in 18-month-olds. *Psychological Science*, *21*(7), 908–13.

Mani, N., & Plunkett, K. (2010b). Twelve-month-olds know their 'cups' from their 'keps' and 'tups'. *Infancy*, *15*(5), 445–70.

Mani, N., & Plunkett, K. (2011a). Does size matter? Subsegmental cues to vowel mispronunciation detection. *Journal of Child Language*, *38*(3), 606–27.

Mani, N., & Plunkett, K. (2011b). Phonological priming and cohort effects in toddlers. *Cognition*, *121*(2), 196–206.

Marian, V., Bartolotti, J., Chabal, S., & Shook, A. (2012). CLEARPOND: cross-linguistic easy-access resource for phonological and orthographic neighborhood densities. *PLoS One*, *7*(8), e43230. doi: 10.1371/journal.pone.0043230.

Maris, E., & Oostenveld, R. (2007). Nonparametric statistical testing of EEG-and MEG-data. *Journal of Neuroscience Methods*, *164* (1), 177–90.

Markman, E. M., & Wachtel, G. F. (1988). Children's use of mutual exclusivity to constrain the meanings of words. *Cognitive Psychology*, *20*(2), 121–57.

https://doi.org/10.1017/S0305000918000259 Published online by Cambridge University Press

McKean, C., Letts, C., & Howard, D. (2013). Functional reorganization in the developing lexicon: separable and changing influences of lexical and phonological variables on children's fast-mapping. *Journal of Child Language*, *40*(2), 307–35.

Munson, B., Kurtz, B. A., & Windsor, J. (2005). The influence of vocabulary size, phonotactic probability, and wordlikeness on nonword repetitions of children with and without specific language impairment. *Journal of Speech, Language, and Hearing Research*, *48*(5), 1033–47.

Pinheiro, J. C., Bates, D., DebRoy, S., & Sarkar, D. (2007). Linear and nonlinear mixed effects models. R-Package-version, 3.1-86. Retrieved from: <http://CRAN.R-project.org/package=nlme>.

R Core Team (2014). R: a language and environment for statistical computing [Computer software manual]. Vienna. Retrieved from <http://www.R-project.org/>.

Ramon-Casas, M., Swingley, D., Sebastián-Gallés, N., & Bosch, L. (2009). Vowel categorization during word recognition in bilingual toddlers. *Cognitive Psychology*, *59*(1), 96–121.

Ren, J., & Morgan, J. L. (2011). Sub-segmental details in early lexical representation of consonants. In *Proceedings of the 17th International Congress of Phonetic Sciences*. Retrieved from: <http://titan.cog.brown.edu/research/infant_lab/Ren_sub%20segmental%20details%20early%20lexical.pdf>.

Sirois, S., & Jackson, I. (2007). Pupil dilation and infant cognition. In *2007 IEEE 6th International Conference on Development and Learning*. IEEE. doi: 10.1109/devlrn.2007.4354056.

Swingley, D. (2003). Phonetic detail in the developing lexicon. *Language and Speech*, *46*(2/3), 265–94.

Swingley, D. (2005). 11-month-olds' knowledge of how familiar words sound. *Developmental Science*, *8*(5), 432–43.

Swingley, D. (2016). Two-year-olds interpret novel phonological neighbors as familiar words. *Developmental Psychology*, *52*(7), 1011–23.

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, *76*(2), 147–66.

Swingley, D., & Aslin, R. N. (2002). Lexical neighborhoods and the word-form representations of 14-month-olds. *Psychological Science*, *13*(5), 480–4.

Szagun, G., Schramm, S. A., & Stumper, B. (2009). *Fragebogen zur frühkindlichen Sprachentwicklung (FRAKIS) und FRAKIS-K (Kurzform)*. Frankfurt: Pearson Assessment.

Tamási, K. (2017). *Measuring children's sensitivity to phonological detail using eye tracking and pupillometry* (Doctoral dissertation, Universities of Potsdam, Newcastle, Groningen, Trento, and Macquarie University, Sydney). Retrieved from <http://nbn-resolving.de/urn:nbn:de:kobv:517-opus4-395954>.

Tamási, K., McKean, C., Gafos, A., Fritzsche, T., & Höhle, B. (2017). Pupillometry registers toddler's sensitivity to degrees of mispronunciation. *Journal of Experimental Child Psychology* (*153*), 140–8.

Tamási, K., Wewalaarachchi, T. D., Höhle, B., & Singh, L. (2016). Measuring sensitivity to phonological detail in monolingual and bilingual infants using pupillometry. In *Proceedings of the 16th Speech Science and Technology Conference*. Retrieved from <http://www.assta.org/sst/2016/SST2016_Proceedings.pdf#page=108>.

Tsukahara, J. S., Harrison, T. L., & Engle, R. W. (2016). The relationship between baseline pupil size and intelligence. *Cognitive Psychology*, *91*, 109–23.

van der Meer, E., Beyer, R., Horn, J., Foth, M., Bornemann, B., Ries, J., … Wartenburger, I. (2010). Resource allocation and fluid intelligence: insights from pupillometry. *Psychophysiology*, *47*(1), 158–69.

Vihman, M., & Croft, W. (2007). Phonological development: toward a 'radical' templatic phonology. *Linguistics*, *45*(4), 683–725.

Vitevitch, M. S., & Luce, P. A. (1998). When words compete: levels of processing in perception of spoken words. *Psychological Science*, *9*(4), 325–9.

Vitevitch, M. S., & Luce, P. A. (1999). Probabilistic phonotactics and neighborhood activation in spoken word recognition. *Journal of Memory and Language*, *40*(3), 374–408.

Werker, J. F., Fennell, C. T., Corcoran, K. M., & Stager, C. L. (2002). Infants' ability to learn phonetically similar words: effects of age and vocabulary size. *Infancy*, *3*(1), 1–30.

White, K. S., & Morgan, J. L. (2008). Sub-segmental detail in early lexical representations. *Journal of Memory and Language*, *59*(1), 114–32.

Yoshida, K. A., Fennell, C. T., Swingley, D., & Werker, J. F. (2009). Fourteen-month-old infants learn similar-sounding words. *Developmental Science*, *12*(3), 412–18.

https://doi.org/10.1017/S0305000918000259 Published online by Cambridge University Press

Zamuner, T. S., Gerken, L., & Hammond, M. (2004). Phonotactic probabilities in young children's speech production. *Journal of Child Language*, *31*(3), 515–36.

Zesiger, P., Lozeron, E. D., Lévy, A., & Frauenfelder, U. H. (2011). Phonological specificity in 12- and 17-month-old French-speaking infants. *Infancy*, *17*(6), 591–609.

Cite this article: Tamási K, Mckean C, Gafos A, Höhle B (2019). Children's gradient sensitivity to phonological mismatch: considering the dynamics of looking behavior and pupil dilation. *Journal of Child Language* **46**, 1–23. https://doi.org/10.1017/S0305000918000259

https://doi.org/10.1017/S0305000918000259 Published online by Cambridge University Press