Received: 29 March 2021 | Revised: 2 July 2023 | Accepted: 13 August 2023

DOI: 10.1111/desc.13444

Developmental Science WILEY

REGISTERED REPORT

# The role of systematicity in early referent selection

Ming Yean Sia[1,2] | Emily Mather[3] | Matthew W. Crocker[4] | Nivedita Mani[1,2]

[1]Department for Psychology of Language, University of Göttingen, Göttingen, Germany

[2]Leibniz Science Campus Primate Cognition, Göttingen, Germany

[3]School of Psychology and Social Work, University of Hull, Hull, UK

[4]Department of Language Science & Technology, Saarland University, Saarbrücken, Germany

**Correspondence**
Ming Yean Sia, Psychology of Language Research Group, University of Göttingen, Göttingen, Germany. Email: sia.mingyean@uni-goettingen.de

**Funding information**
RTG 2070 Understanding Social Relationships; SFB 1528 Cognition of Interaction; Deutsche Forschungsgemeinschaft (DFG, German Research Foundation), Grant/Award Number: Project-ID454648639

## Abstract

Previous studies showed that word learning is affected by children's existing knowledge. For instance, knowledge of semantic category aids word learning, whereas a dense phonological neighbourhood impedes learning of similar-sounding words. Here, we examined to what extent children associate similar-sounding words (e.g., rat and cat) with objects of the same semantic category (e.g., both are animals), that is, to what extent children assume meaning overlap given form overlap between two words. We tested this by first presenting children ($N = 93$, Mage = 22.4 months) with novel word-object associations. Then, we examined the extent to which children assume that a similar sounding novel label, that is, a phonological neighbour, refers to a similar looking object, that is, a likely semantic neighbour, as opposed to a dissimilar looking object. Were children to preferentially fixate the similar-looking novel object, it would suggest that systematic word form-meaning relations aid referent selection in young children. While we did not find any evidence for such word form-meaning systematicity, we demonstrated that children showed robust learning for the trained novel word-object associations, and were able to discriminate between similar-sounding labels and also similar-looking objects. Thus, we argue that unlike iconicity which appears early in vocabulary development, we find no evidence for systematicity in early referent selection.

KEYWORDS
leveraged learning, referent selection, semantic networks, word meaning arbitrariness

## 1 | INTRODUCTION

Despite considerable overlap in more basic cognitive abilities across species, our ability to learn and use language in communication with one another distinguishes us from even our nearest primate cousins. While animals solve even linguistically challenging tasks such as rejecting a familiar object as the referent for a novel label (akin to fast-mapping; Kaminski et al., 2004) or executing different actions depending on the syntax of the instructions provided (Herman et al., 1984), none of these feats scale up to the complexity of the human language. Outlining the defining features of human language, Hockett and Hockett (1960) highlighted the arbitrariness of the symbols used in linguistic communication (see also Greenberg, 1957). As they put it, "man is the only animal that can communicate by means of abstract symbols". Indeed, the arbitrariness of word-referent mappings is immediately apparent in comparing the labels for the same object in various languages (e.g., dog; *chien*—French; *inu*—Japanese; *anjing*—Malay). More often than not, there appears to be little reason why an object is labelled the way it is—why, otherwise, would the word "boot" refer to footwear in English but a watercraft in German?

Arbitrary mappings between word forms and meanings populate natural languages, potentially because they allow us near-unlimited possibilities of expressing ourselves beyond iconic representations (i.e., one can (theoretically) attach any string of sound (e.g., *gibo*) to any object (e.g., a starfish)). However, as we discuss below, there may be advantages to non-arbitrary mappings in language, particularly when it comes to language learning. Such advantages may also underlie recent findings that language structures may be more systematic than initially assumed them to be (Dingemanse et al., 2015) and that artificial neural networks prefer simpler naming systems (akin to non-arbitrariness) despite the possibility of miscommunications (Chaabouni et al., 2021). For instance, using texts from Wikipedia, Dautriche et al. (2017) examined the degree of semantic and phonological distance between pairs of words from 100 languages from diverse language families. They found a positive correlation between semantic and phonological distance in most languages, suggesting that phonological minimal pairs are often likely to also be semantically related. In other words, there is a high chance that words that sound or look alike belong to the same or related semantic category. Given the consistent finding of such systematicity across numerous languages, this study suggests that some degree of language systematicity may possibly be the norm rather than the exception it was previously thought to be.

The potential for non-arbitrariness in language is captured by the concepts of iconicity and systematicity. Iconicity refers to the overlap between word form and word meaning, most observable in onomatopoeia, where the word phonologically overlaps with a sound associated with a particular object, for example, *woof* (Haiman, 2015; Imai & Kita, 2014; Winter et al., 2017). For instance, Gasser (2004) suggested that iconicity may be useful in constraining form-meaning relations, thus facilitating acquisition of word meanings. In other words, reduced effort is required to learn a non-arbitrary word-meaning mapping because the learner can leverage their already existing knowledge in acquiring a novel word-object mapping (Monaghan et al., 2011). This may especially be the case with young infants who have few other cues that they can rely on with regard to the labels referring to objects in their environment (Gasser, 2004; Monaghan et al., 2014). In keeping with this suggestion, Asano et al. (2015) found that even young 11-month-old infants are sensitive to the sound-symbolic correspondences between words and the objects they are presented with, showing increased processing effort when the sound-symbol mapping presented was incongruent (e.g., *kipi* mapped to a round shape instead of to a spiky shape).

Systematicity, on the other hand, refers to regularities in the relationship between particular word forms and their meanings within a language, which may or may not be iconic (e.g., *gl-* is often associated with words related to the concept of light in English, as in *glitter* and *glow*; Abramova & Fernandez, 2016; Perry et al., 2015). Systematicity also appears to aid learning, with children learning artificial categories more easily when there is greater phonological overlap between words referring to the categories (Brooks et al., 1993; Monaghan et al., 2005). Taking this argument further, Monaghan et al. (2014) hypothesised that were non-arbitrariness to help constrain word meanings, there should be increased prevalence of systematic sound-meaning relations in words learned by younger children relative to older children. Given that Perry et al. (2015) found adjectives to be more iconic than nouns, Monaghan et al. (2014) found, after controlling for such a word-type effect, that words learned earlier have more systematic sound-meaning relations than words learned later in development. Taken together, the studies reviewed above present considerable evidence for early sensitivity to systematicity in languages in even young children as well as potential benefits for such systematicity in learning.

**RESEARCH HIGHLIGHTS**

- We test how systematicity in word form-meaning mappings impacts referent selection in young children.
- We expect that toddlers are able to exploit the systematic relations between word forms and their meanings during referent selection.
- While children neither favoured nor avoided form-meaning systematicity during referent selection, they were sensitive to differences between phonological neighbours and also between perceptually similar objects.
- We found that vocabulary size is negatively associated with systematicity, which may indicate that systematicity arises earlier than the ages tested in the current study.

On the other hand, while children may be sensitive to systematicity in word-meaning relations, combined phonological and semantic overlap between words may make it difficult for children to discriminate between overlapping word-pairs (Monaghan et al., 2011). In other words, arbitrariness may help to distinguish words (and their meanings) from one another—such that words that sound similar but mean different things may be discriminated along the semantic dimension while words that mean similar things but sound dissimilar may be discriminated along the phonological dimension. Overlap across both dimensions may lead to words being easily confused with one another, especially in immature language learners (see also Gasser, 2004). There are also suggestions that this may be important from an evolutionary perspective, with arbitrariness in word-meaning mappings leading to words and concepts that are essential in potentially life-endangering situations being less confusable for one another (Corballis, 2002). To a certain extent, this reasoning may also explain why languages have evolved to favour arbitrariness over systematicity, despite potential benefits of systematicity during acquisition. In other words, systematicity may aid learning novel words due to learners being able to leverage their already existing knowledge, but arbitrariness may be more important in lexical processing by making words that potentially co-occur in the same contexts (Roy et al., 2013) more discriminable from one another.

Such distinction may be particularly important in early development, given the fact that contextual distinctiveness (distinctiveness with regard to where and when particular words are uttered and what other words co-occur in conversations) has been shown to be a strong predictor of children's vocabulary development. Specifically, words that are used more broadly across different contexts, that is, are more distinctive from other words, are acquired earlier than words that co-occur in the same context (Roy et al., 2013). Against this background, the current study will examine the extent to which children associate similar-sounding words with objects that overlap on the semantic dimension. If a child learns that *maacke* refers to a novel animal, to what extent does the child then associate a similar-sounding word (*maasche*) with another novel animal or an object from a different semantic category?

Importantly, in examining systematicity in early vocabulary development, it is necessary to ensure that children perceive overlapping word-object mappings as distinct mappings. In other words, are children sensitive to the fact that *maacke* and *maasche* are distinct lexical entries and that the objects they refer to in the context of the study are distinct? Therefore, we next examine the literature on children's sensitivity to semantic and phonological categories in early word learning.

## 1.1 | The role of overlap in early vocabulary development

The early vocabulary is highly connected, with links between words that overlap along semantic and phonological dimensions (Fourtassi et al., 2019; Steyvers & Tenenbaum, 2005). It comes as no surprise, then, that children are sensitive to the semantic and phonological relationships between words that belong to the same category. To illustrate, at 18-months of age, toddlers show robust sensitivity to phonological overlap between words, by looking more at labelled images when primed by phonologically-related words relative to unrelated words (Mani & Plunkett, 2010a; Mani et al., 2012). Similarly, at an early stage of development, children show sensitivity to the semantic (Arias-Trejo & Plunkett, 2009; Avila-Varela et al., 2021; Bergelson & Aslin, 2017; Delle Luche et al., 2014; Willits et al., 2013) and/or perceptual relationships between objects that words refer to (Johnson & Huettig, 2011; Mani et al., 2013), again by looking more at labelled images when primed by related object-label pairs than unrelated pairs. Such findings have been taken to suggest that there is interconnectivity in the early lexicon, such that even at a young age, children are sensitive to overlap between words and the objects they refer to and that this overlap impacts processing of these familiar words (see Mani & Borovsky, 2017 for a review).

Equally, studies suggest that children are also sensitive to small changes to the phonological characteristics of early words from a young age, with even 12-month-olds being able to discriminate a familiar word from a small mispronunciation of this word (Mani & Plunkett, 2010b; see Mani & Plunkett, 2011 for evidence of children's fine-grained graded sensitivity to such mispronunciations). By 14-months of age, children also show sensitivity to small mispronunciations of novel word-object mappings suggesting that sensitivity to the phonological representations of words is not limited to words children are robustly familiar with (Ballem & Plunkett, 2005; Mani & Plunkett, 2008). Hence, previous studies show that children from early on are able to discriminate between minimally different lexical entries and perceive phonologically overlapping entries as distinct.

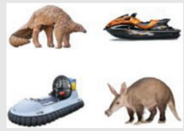
However, the studies reviewed above capture the influence of phonological or semantic overlap on children's *processing* of lexical entries. To what extent, then, is there evidence in the literature for a similar influence of overlap on children's learning novel word-object mappings? Does the influence of overlap in processing align with the influence of overlap on learning, or, as suggested above, do arbitrariness and systematicity play different roles in processing and learning?

Indeed, there is considerable evidence for an influence of connectivity on vocabulary development (see Mani & Ackermann, 2018 for a review). With regard to outcomes, typically developing children have a more connected semantic network than late talkers (Beckage et al., 2011). In addition, the structure of children's early vocabularies appears to be related to children's biases in language learning. In particular, children whose vocabularies show increased lexical connectivity exhibited increased novelty biases, suggesting that they may be better at learning words (Yurovsky et al., 2012). Furthermore, Borovsky et al. (2016) demonstrated that 2-year-olds may leverage their knowledge of a semantic category in learning new words, showing increased learning of novel words in large semantic categories than in smaller semantic categories (see also Ackermann et al., 2020) and that coherence in category structures improves word learning (Borovsky & Elman, 2006). Thus, children may be able to leverage their semantic knowledge in learning word-object pairings that are semantically related to words they already know (Barabási & Albert, 1999).

However, evidence for phonological leveraging is not as straightforward. Phonological neighbours are word pairs that differ by one phoneme, either through addition (e.g., *cat*—*scat*), deletion (e.g., *cat*—*at*) or substitution (e.g., *cat*—*mat*; Luce & Pisoni, 1998). Several studies suggest that toddlers younger than 18 months appear to have difficulties learning novel word-object associations that sound similar to familiar words children already know, suggesting that phonological neighbours may impede early word learning (Mather & Plunkett, 2011; Swingley & Aslin, 2007). On the other hand, other studies find that word-form overlap may boost children's segmentation of words from fluent speech (Altvater-Mackensen & Mani, 2013) and that children learn words more easily when these words belong to larger phonological neighbourhoods (Newman et al., 2008) such that there is a *facilitative* neighbourhood effect on learning novel word-object associations in 20- and 24-month-old toddlers. Using corpus analyses, Storkel (2004) also found support for a facilitative neighbourhood effect on vocabulary growth, but only for low frequency words.

Some of the differences in the studies reviewed here may be explained by considering the strength of children's knowledge. For instance, brief exposure to novel phonological neighbours has been suggested to promote novel word learning while prolonged exposure to these neighbours jeopardises word learning (Hollich et al., 2002). This study is in line with Kucker et al. (2020) finding that children retain novel word-object associations better if the distractor

**TABLE 1** Overall Study Structure and Examples of Stimuli for Each Phase.

| Phase | Total trials | Total novel labels | Total novel objects | Example of a trial | | |
|---|---|---|---|---|---|---|
| | | | | Visual stimuli | Label | Target |
| Training | Eight (four per block: two per category) | Four (two per block: one per category) | Four (two per block: one per category) |  | *maacke* | - |
| Retention | Four (two per block: one per category) | Four (two per block: one per category) | Four (two per category: one trained, one semantic neighbour) |  | *maacke* | Pangolin |
| Leveraging | Four (two per block: one per category) | Four phonological neighbours (two per block: one per category) | Four (two objects per category: one trained, one semantic neighbour) |  | *maasche* | Aardvark |

*Note*. There are two blocks in the present study. Categories used are ANIMAL and VEHICLE. Trained objects are novel objects presented in the training phase, whereas semantic neighbours are super-novel objects that share perceptual features with the trained novel object.

objects are weakly known (i.e., just learned) as opposed to when the distractors are well-known. In other words, when highly familiar phonological neighbours are activated (as in Swingley & Aslin, 2007), they may compete with novel labels, resulting in unsuccessful word learning. Thus, there may be an interaction between familiarity and systematicity with regard to word learning, such that systematicity may boost learning when children are not familiar with the content they are presented with.

In sum, there is robust evidence for separate influences of semantic and phonological overlap on early lexical acquisition and processing, suggesting that systematicity in novel word-object mappings may boost word learning. At the same time, no study to-date has examined systematicity per se regarding how similarity in both form and meaning impacts novel word learning.

## 1.2 | The current study

Against this background, the current study trained children on novel word-object mappings from two familiar categories (i.e., animal and vehicle, see Table 1). In other words, children were taught a distinct novel label for a novel object from each category (e.g., *maacke* for a pangolin, see first row of Table 1). Following training, children were presented with an array of objects in the visual world paradigm, where pairs of objects belong to the categories that children were exposed to earlier. That is, children saw two novel objects they were recently told the names of (e.g., a pangolin and a hovercraft) as well as two super-novel objects from the same category as the previously presented objects (e.g., an aardvark and a jet ski). Across trials, toddlers were then asked to locate the referent of either a previously presented novel label (i.e., retention test phase, see second row of Table 1) or a novel label which phonologically overlaps with one of the previously presented novel labels (e.g., *maasche*; i.e., leveraging test phase, see last row of Table 1).

With regard to retention trials, we hypothesise that exposure to the novel word-object mappings (i.e., during training phase) will lead to robust recognition of the pairings at test. Using the example in Table 1, children will fixate the pangolin when presented with the trained label for this object, *maacke* (see second row of Table 1). We hypothesise that systematicity in word-form mappings will leverage referent selection, such that children associate the phonologically similar label (i.e., *maasche*) with the super-novel object from the same category (aardvark in the example above) as opposed to the super-novel object from the other category (here, jet ski; see last row of Table 1).

The contrast between retention trials and leveraging trials will allow us to test whether children discriminate between the two labels presented, since we expect that children fixate the previously trained objects in retention trials (where they are presented with the labels they were previously trained on) and the super-novel semantic neighbour objects in leveraging trials (where they are presented with labels that sound similar to the original labels they were trained on). In contrast, similar looking behaviour to the trained object (or indeed, the similar-looking object) across both retention and leveraging trials would be interpreted as a failure to discriminate either the two similar-sounding labels from one another or the two similar-looking objects from one another. Similarly, in leveraging trials, we can rule out the possibility that children treat similar-sounding labels as mispronunciations of the trained novel labels: If the similar-sounding labels are treated as mispronunciations, children ought to fixate the trained object when presented with the similar label, albeit to a lesser extent than in retention test trials (Mani & Plunkett, 2011; Swingley & Aslin, 2000; von Holzen & Bergmann, 2018). If, on the other hand, similar-sounding labels leverage referent selection, as under examination here, children should look preferentially at the novel semantic neighbour object when presented with the similar-sounding label.

Finally, our inclusion of four objects at test controls for the possibility that children merely fixate any super-novel object in response to a novel label. Thus, children have the option to select either of the super-novel objects as referents for the similar-sounding label. Of interest is whether they systematically select the similar-looking target, that is, the object that looks similar to the trained object whose label also sounds similar to the test label, or the different-looking object, that is, the object that looks similar to the object whose label does not sound similar to the test label. The former pattern of results would suggest that systematicity leverages referent selection while the latter would suggest that systematicity is not preferred in referent selection.

## 2 | METHODS

### 2.1 | Participants

We recruited 108 children through our database. Six of these recruited children did not provide any eye-tracking data: fussiness (3) or technical problems (3). One child was excluded from the analysis because she was identified as a bilingual. Of the remaining 101 children, one was excluded from the analysis because the eye-tracker could track less than 20% of the child's eye-gaze, another seven children were excluded for not providing data for at least two retention trials and two leveraging trials (see *Data analysis* for details). The final sample size was 93 monolingual German children (48 boys, 45 girls), whose mean age was 22.4 months old (range: 20–28 months). This includes five children more than pre-registered, based on simulation of data from a study examining children's responding in a task similar to that planned (Schmid et al., 2019; see *sample size calculation* section for details). We chose to test 24-month-olds based on previous findings of semantic leveraging effects in word learning at this age (Borovsky et al., 2016) and suggestions that systematic form-meaning relations facilitate lexical acquisition in younger children (Monaghan et al., 2014).

As we presented visual and auditory stimuli to children, we only recruited typically-developing children who are born full-term and do not report having any vision or hearing problem. In order to control for vocabulary size, we measured children's linguistic experience using an adapted, computerized version (Mayor & Mani, 2019) of the German vocabulary checklist (FRAKIS; Szagun et al., 2009).

As the novel objects used in the present study were real-world objects, we checked whether children had any previous knowledge about these objects by asking caregivers to fill in a short questionnaire where they indicate whether their child is familiar with these novel objects. To control for the effect of vocabulary size and semantic neighbourhood size on children's performance, we also requested caregivers to complete an adapted version of the German vocabulary checklist (Mayor & Mani, 2019) to indicate which words their child produces. Following Ackermann et al. (2020) and Borovsky et al. (2016), neighbourhood size was measured by calculating the absolute number of known members within each semantic category. We have obtained ethics approval from the ethics committee of University of Göttingen.

### 2.2 | Stimuli

Four rare members from each of the semantic categories ANIMAL and VEHICLE were used as novel objects in the current study. These categories were selected because they were reported to be common and familiar to 2-year-old children (Borovsky et al., 2016). Of the four members from each category, we chose pairs of objects that share visual features, such that children saw four pairs of similar-looking novel objects. In particular, *aardvark—pangolin* and *flying fox—flying squirrel* were used for the category ANIMAL and *hovercraft—jet ski* and *rickshaw—tuk-tuk* were used for the category VEHICLE. Each object image was placed on a white background and had a resolution of 1024 × 768 pixel. We specifically ensured that the visual features of the novel object pairs are highly similar to emphasise the semantic relation between the object pairs, due to the findings that perceptually dissimilar objects activate taxonomic relations more slowly in children (Chow et al., 2017; although toddlers do understand that objects of the same semantic category need not necessarily share similar features, Arias-Trejo & Plunkett, 2010).

Eight novel labels were used in the present study, all of which are bi-syllabic and in keeping with German phonotactic rules. These novel labels formed four pairs of phonological neighbours (*schufi—schuri*, *maacke—maasche*, *gissel—gibbel*, and *peto—pewo*). We examined the number of phonological neighbours known to children at this age using Wordbank (Frank et al., 2016; Szagun et al., 2009), an online open repository for vocabulary data of children between 18 and 30 months of age. We did not find any words known to be familiar to children at this age that were phonologically related to the words presented in the study.

All novel labels were referred to using the German neutral article, *das*. As the novel objects were labelled six times in each trial of the training phase, six different carrier sentences were used in the training phase: "That is an X! Wow, an X! Do you see the X? Look, an X! Oh, what a fantastic X! I see an X!". In the retention trials, novel labels were embedded in the sentence "You know that now! Now, where is the X? The X!", whereas in the leveraging trials, novel labels were embedded in the sentence "And now something new! Where is probably the X? The X!" (see Appendix A). Carrier phrases in the leveraging trials were specifically chosen to ensure that children know that they are being asked something different across retention and leveraging trials. The primary purpose of the study was to examine whether systematicity leverages referent selection—where we expect them to look at the super-novel semantic neighbour—or whether systematicity is avoided in referent selection—where we expect them to look at the super-novel semantic non-neighbour. While the biasing sentences ensured that children are made aware of the phonological difference between the trained labels and the similar-sounding labels, they do not bias children with regard to the role of systematicity in referent selection.

All sentences were recorded by a female native German speaker in an enthusiastic, infant-directed manner.

## 2.3 | Design

Stimuli presentation were split into two blocks, where each block trained children on two distinct novel word-object mappings (one mapping from each category) as well as presented them with a set of related retention test and leveraging test (see Table 1). After the first block, children were presented with the second block following the same phase order but with different stimuli. Hence, at the end of the experiment, children would have received a total of four novel word-object mappings over two blocks of training phase. The fixed phase order within each block allowed us to first train children on novel word-object mappings (see first row of Table 1) before testing whether they have retained these novel word-object associations (see second row of Table 1) and how they use this newly-acquired knowledge to leverage referent selection for similar-sounding novel labels (see final row of Table 1).

In every trial of the training phase, a novel object appeared at the centre of the screen for 20 seconds (see Appendix A). The auditory stimulus began around 200 ms after the image onset. The novel object was labelled six times in different carrier phrases, with an interval of about 1000 ms between carrier phrases. There were four training trials in every block, such that each novel object from each semantic category will be presented in two trials, labelled for a total of 12 times. Based on previous studies (e.g., Ackermann et al., 2020; Borovsky et al., 2016; Dautriche et al., 2015) showing successful word learning in 18- and 30-months-old using a similar design (where each novel object was labelled 10 times across two trials), we anticipated the number of training trials to be sufficient for the toddlers to learn the novel word-object associations. The order of object presentation (i.e., whether the novel object is presented in the training phase of the first or second block), phase of object presentation (i.e., whether the novel object was presented as a trained object or a super-novel object) and the pairing between the novel labels and the novel objects were counter-balanced across children.

In each trial of retention and leveraging test phases, four object images appeared on the screen (one object in each corner of the screen) for 8000 ms before the next trial began (see Appendix A). The four objects were the two previously-presented novel objects and two perceptually-similar super-novel objects (see Table 1). The position of these images were counter-balanced across trials. One thousand ms after the onset of the images, toddlers were asked to locate the referent of a particular label, such that the toddlers heard the target word at approximately 4000 ms after the onset of images. The target word was repeated 1000 ms after the offset of the previous sentence to boost the naming effect as toddlers' attention is reportedly captured by super-novel objects (Horst et al., 2011; Mather & Plunkett, 2012).

Following Chow et al. (2017) and Mather and Plunkett (2011), the target word onset split each test trial into pre-naming and post-naming phases. Importantly, children were presented with the previously-trained novel label in the retention test phase, and the novel similar-sounding labels in the leveraging test phase (see Table 1). Each novel label was presented once, thereby providing children with two retention and two leveraging test trials in each block (i.e., four retention and leveraging trials each across blocks). As the data were analysed by collapsing blocks and aggregating all trials within each test phase, each child contributed four trials per test phase. The number of test trials presented was in accordance with previous studies (e.g., Ackermann et al., 2020; Arias-Trejo & Plunkett, 2009; Kucker et al., 2020; Pomper & Saffran, 2019).

## 2.4 | Procedure

Caregivers and toddlers were invited into the waiting room where the researcher explained the aim and procedure of the study to the caregivers. Once the caregivers agreed to participate in the study, they were given two questionnaires to complete, one that asked them if their child knew the novel objects used and another that measured children's vocabulary knowledge. After that, both the caregiver and the child were led to the eye-tracking room.

The child was seated comfortably in a high chair in front of a TV screen with the Tobii X 120 eye-tracker (sampling rate of 40 Hz) positioned below the screen. Auditory stimuli were played through two loud-speakers placed above the TV screen. Caregivers were asked to be seated silently behind the child while keeping their eyes closed[1] so that their eye movements are not captured by the eye-tracker.

We used a 5-point calibration, where a red dot moved in random directions across the screen while the researcher encouraged the child to follow the movement of the red dot. When calibration was successful, the experimenter started the study by presenting the first training trial of the first block. As we only measured toddlers' looking behaviour and did not require them to make behavioural responses, each trial was presented for a fixed duration (see *design* section for details) before the next trial began. Every child received the same order of trials shown in Table 1, that is, first the training phase followed by the retention test phase and finally the leveraging test phase. The study ended with the last leveraging test trial of the second set of stimuli. The experiment took about 5 min.

At the end of the experiment, the child and caregiver were thanked for their participation and given a book as compensation.

## 2.5 | Data analyses

We first pre-processed data by removing trials where children looked at visual stimuli for less than 20% of the trial or where children's eye-gaze could be tracked for less than 20% of the trial. We then excluded children who fail to provide data for at least two trials for each type of test trial. For the remaining data, we computed the proportion of target looking (PTL) separately for pre- and post-naming phases for each participant in each test phase. This was done by dividing toddlers' total looking time at only the target by their total looking time at the target and three other distractor objects (see Table 1 for predicted targets in each test trial).

In order to examine whether systematicity in word form-meaning mappings leveraged referent selection of similar-sounding labels, primary analyses used generalized linear mixed models (GLMM, Baayen et al., 2008, see model specification below). As the blocks followed the same pattern with merely a different set of stimuli, looking behaviour during retention and leveraging trials from both blocks were aggregated prior to data analyses. *Test* (retention, leveraging test) and *Phase* (pre-, post-naming) were included as predictors and vocabulary size, neighbourhood size and sex (child) as control predictors. The full model was compared to a null model excluding *Phase* to examine whether adding this factor improved model fit. The model that was fitted comprised random intercept effects for child ID, the target object and all theoretically identifiable random slopes (Barr et al., 2013; Schielzeth & Forstmeier, 2009). We added one further random intercept when fitting the model, namely, *trial*, which is crucial because children's looking behaviour in pre- and post-naming phases differs from trial to trial. The random slopes of *test*, *phase* and *sex* were manually dummy-coded and then centred so that the model output is not dependent on the reference level of the factors. As slight changes to our models' specification (e.g., whether we centre the random slopes of factors) may cause convergence problems, we always fitted two models, one with all random slopes of factors centred and one without. Provided that both models converged, we compared their log-likelihood and used the model with a higher log-likelihood as an index of better data fit. We also z-transformed the covariates, that is, vocabulary size and neighbourhood size, to help alleviate the model convergence issue.

Hence, the model was (in lme4 notation for brevity):

$$
\begin{aligned}
\text{Model}_{\text{beta}} = \ & \text{glmmTMB(PTL} \sim \text{Test} * \text{Phase} + \text{vocabulary\_size} \\
& + \text{neighbourhood\_size} + \text{sex} + (1 + \text{Test} * \text{Phase}||\text{id}) \\
& + (1 + \text{Phase}||\text{trial\_id}) + (1 + \text{Test} * \text{Phase} \\
& + \text{vocabulary\_size} + \text{neighbourhood\_size} + \text{sex}||\text{Object}), \\
\text{Data} = \ & \text{data, family} = \text{beta\_family}, \\
\text{weights} = \ & \text{total.looking.time})
\end{aligned}
$$

We did not include correlations among random intercepts and random slopes because the beta model did not converge otherwise. We also weighted the contribution of the individual data points by their respective total looking time for a more accurate measure of children's looking behaviour.

A significant interaction between *Test* and *Phase* would be further broken down across test trials to examine whether there is an effect of phase in different test trials. In particular, we would split the data by trial (i.e., retention or leveraging) and fit the same model as above but excluding *Test* from the fixed and random effects part, separately for the two subsets. Exploratory analyses included the PTL in retention test trials as a covariate into a model examining whether children's performance in retention test trials influences their performance in leveraging test trials. Details regarding the models are provided in Appendix B. All study materials, anonymised data (https://doi.org/10.17605/OSF.IO/DWGPT), analyses scripts as well as registered protocol (https://osf.io/u8vzh) are publicly available at https://osf.io/7mbsh/

## 2.6 | Determination of sample size

We determined sample size using a power analysis which was based on simulated data. We simulated data sets comprising 40–96 individuals (increment: 8), and simulated 1000 data sets per sample size. We simulated a beta distributed response (i.e., one that is bound between zero and one) and originally analysed it with a corresponding beta regression mixed model, as we will be measuring PTL, where our response measure ranges from zero to one.

Power analysis revealed that the probability of our model converging was less than 0.05 and not much affected by sample size. Therefore, we conducted a simulation in which we fitted a Gaussian model as specified in Appendix B instead (with PTL being *arcsine* transformed), where we obtained 100% model convergence. Based on the results of the power analysis for this model, we decided to recruit 88 children to achieve a power of 0.90. Further details are provided in Appendix B.

# 3 | RESULTS

For the main beta model, we retained the full model with all random slopes of factors centred because the log-likelihood value was higher ($2LL_{centred} = 810.32$; $2LL_{not\ centred} = 801.10$). We evaluated model stability for the retained full model by dropping levels of random effects one at a time, fitting the full model to each of the subsets, and comparing the estimates of these models with those obtained for the full data set. The full model was found to be stable, that is, we did not find any influential level of any of the random effects in the model which may affect the model output.

The full-null model comparison was significant, suggesting that adding *Phase* to the model improved model fit ($\chi^2$ (2) = 10.76, $p = 0.005$). *Drop1* tests revealed a significant interaction between *Test* and *Phase*, and a significant main effect of neighbourhood size (see Table 2). As shown in Table 2 and Figure 1, there was a significant difference in PTL between retention and leveraging trials in the post-naming phase (reference level). Furthermore, the significant effect of *Phase* suggests increased looking to the distractor in the post-naming phase relative to the pre-naming phase in leveraging trials (reference level: post-naming phase, leveraging trials).

To further analyse the interaction between *Test* and *Phase*, we fitted the full model separately for retention and leveraging trials, such that *Test* was removed from the fixed and random effects part of these models. For the retention trials, the full model with all random slopes of factors centred has a higher log-likelihood value ($2LL_{centred} = 345.37$; $2LL_{not\ centred} = 338.06$) and was retained. The null model was exactly

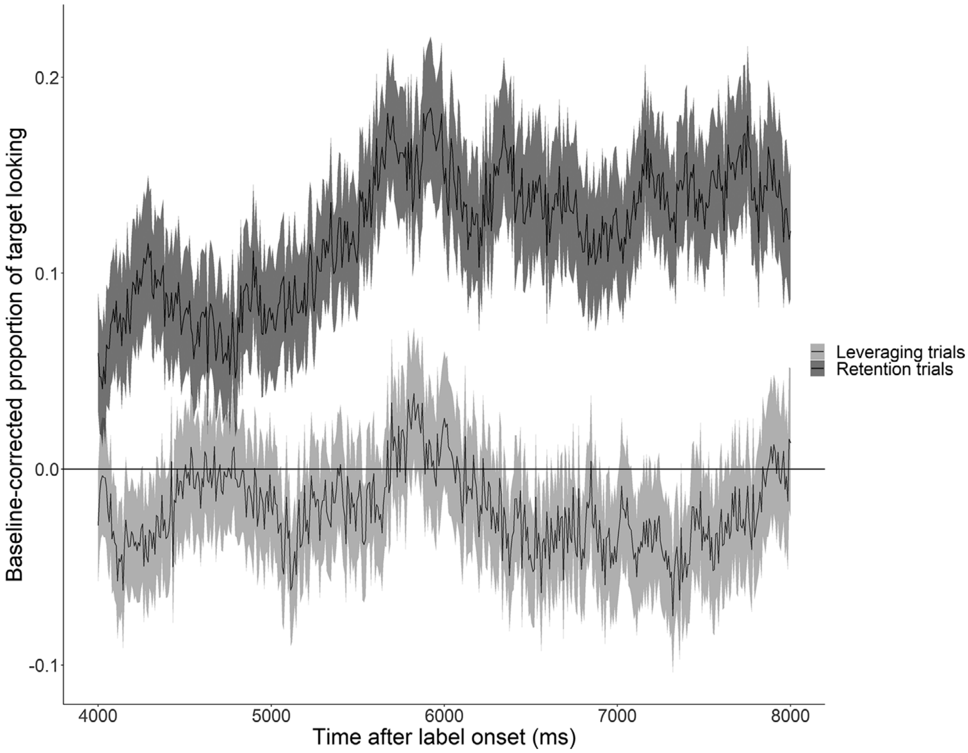



**FIGURE 1** Baseline-corrected PTL in retention and leveraging trials. Note. Horizontal line at 0 indicates no preference in looking at either the target or any of the three distractors. A positive baseline-corrected proportion indicates a preference for the target whereas a negative baseline-corrected proportion indicates a preference for the distractors.

**TABLE 2** Estimates, standard errors and *P*-values of predictor variables included in the full main registered model.

| Terms | Estimates | Standard error | *P*-value |
|---|---|---|---|
| Intercept | −1.335 | 0.094 | <0.001 |
| Test (reference level: leveraging) | 0.468 | 0.120 | <0.001 |
| Phase (reference level: post-naming) | 0.279 | 0.103 | 0.007 |
| Vocabulary size | −0.023 | 0.050 | 0.650 |
| Neighbourhood size | 0.122 | 0.050 | 0.014 |
| Sex (reference level: female) | −0.069 | 0.081 | 0.397 |
| Interaction (test*phase) | −0.520 | 0.136 | <0.001 |

*Note*. All covariates have been z-transformed. All random slopes of factors have been centred. The reference level of factors was determined alphabetically by default in R.

the same as the full model except that *Phase* was removed. The full-null model comparison was significant, suggesting that adding *Phase* improved model fit ($\chi^2$ (1) = 4.45, $p$ = 0.035, see Table 3). Specifically, children were more target-oriented in the post-naming phase than the pre-naming phase (see Table 3). There were no other significant effects.

For the leveraging trials, we fitted a slightly different model in that we included the PTL of retention trials as a covariate. Here, we matched leveraging and retention trials based on the perceptual

**TABLE 3** Estimates, standard errors and *P*-values of predictor variables included in the full retention-only model.

| Items | Estimates | Standard error | *P*-value |
|---|---|---|---|
| Intercept | −0.829 | 0.109 | <0.001 |
| Phase (reference level: post-naming) | −0.271 | 0.112 | 0.016 |
| Vocabulary size | 0.028 | 0.077 | 0.718 |
| Neighbourhood size | 0.107 | 0.076 | 0.163 |
| Sex (reference level: female) | −0.107 | 0.147 | 0.246 |

*Note*. All covariates have been z-transformed. All random slopes of factors have been centred. The reference level of factors was determined alphabetically by default in R.

similarity of the target objects (see *Stimuli* for details). The inclusion of PTL of the associated retention trials allows us to examine whether children's retention of the trained novel word-object associations (e.g., *maacke*-pangolin) affects their reliance on the systematicity information (e.g., *maasche*-aardvark) during referent selection in the leveraging trials. The full model of the leveraging trials is thus specified as:

$$\text{Model}_{\text{leveraging}} = \text{glmmTMB(PTL} \sim \text{Phase} + \text{vocabulary\_size} + \text{neighbourhood\_size} + \text{sex} + \text{retention\_PTL}$$

**TABLE 4** Estimates, standard errors and *P*-values of predictor variables included in the full leveraging-only model.

| Items | Estimates | Standard error | *P*-value |
|---|---|---|---|
| Intercept | −1.345 | 0.116 | <0.001 |
| Phase (reference level: post-naming) | 0.279 | 0.103 | 0.007 |
| Vocabulary size | −0.100 | 0.070 | 0.154 |
| Neighbourhood size | 0.174 | 0.069 | 0.012 |
| Sex (reference level: female) | 0.055 | 0.104 | 0.595 |
| PTL in retention trials | 0.056 | 0.057 | 0.327 |

*Note.* All covariates have been z-transformed. All random slopes of factors have been centred. The reference level of factors was determined alphabetically by default in R.

**TABLE 5** Estimates, standard errors and *P*-values of predictor variables included in the full mispronunciation model.

| Items | Estimates | Standard error | *P*-value |
|---|---|---|---|
| Intercept | 0.127 | 0.088 | 0.149 |
| Test (reference level: leveraging) | 0.279 | 0.106 | 0.009 |
| Phase (reference level: post-naming) | −0.150 | 0.104 | 0.149 |
| Vocabulary size | 0.039 | 0.051 | 0.445 |
| Neighbourhood size | 0.015 | 0.051 | 0.774 |
| Sex (reference level: female) | −0.120 | 0.085 | 0.159 |
| Interaction (test*phase) | −0.258 | 0.145 | 0.076 |

*Note.* All covariates have been z-transformed. The random slopes of factors were not centred. The reference level of factors was determined alphabetically by default in R.

$$+ (1 + \text{Phase} + \text{retention\_PTL}||\text{id}) + (1 + \text{Phase}$$
$$+ \text{retention\_PTL}||\text{trial\_id}) + (1 + \text{Phase}$$
$$+ \text{vocabulary\_size} + \text{neighbourhood\_size} + \text{sex}$$
$$+ \text{retention\_PTL}||\text{Object}), \text{Data} = \text{leveraging\_trials},$$
$$\text{family} = \text{beta\_family}, \text{weights} = \text{total.looking.time})$$

The full model without centring the random slopes of factors did not converge, hence, we retained the model with all random slopes of factors centred. The null model was the same as the full model but with *Phase* removed. The full-null model comparison was significant, indicating that adding *Phase* to the model improved model fit ($\chi^2$ (1) = 6.80, $p = 0.009$). A breakdown of the full model revealed that children looked more at one of the distractors in the post-naming phase relative to the pre-naming phase (see Table 4, Figures 1 and 2).

## 3.1 | Exploratory analyses

Post-hoc exploratory analyses examined which of the three distractors children were fixating in leveraging trials. In particular, we fitted two more models on a subset of data to examine whether children fixated the trained semantic neighbour and whether they fixated the super-novel dissimilar-looking object.

First, we compared looking behaviour towards the trained objects and their respective super-novel semantic neighbour in the retention and leveraging trials. As explained earlier, if children discriminated between two phonologically similar labels and two perceptually similar objects, they would fixate the corresponding trained objects to a lesser extent in leveraging trials relative to retention trials. Otherwise, there should be no difference in their looking behaviour. To test this hypothesis, we computed a new PTL which only compares the semantic neighbour object pairs, that is, we divided the amount of time children spent looking at the trained object by the amount of time children spent looking at the trained object and its super-novel semantic neighbour. We then fitted the same full model as the main beta model. The full model with all random slopes of factors centred had a higher log-likelihood. However, as the null model did not converge, we retained the model without centring random slopes of factors. Since we were predominantly interested in whether there was a difference between retention and leveraging trials in this analysis, we dropped *Test* in the null model. The full-null model comparison was significant, suggesting that adding *Test* significantly improved model fit ($\chi^2$ (2) = 6.82, $p = 0.033$). Specifically, children fixated the trained object more in retention trials relative to leveraging trials, suggesting that they were able to discriminate between two similar-sounding labels and two similar-looking objects (see Table 5 and Figure 3).

Next, we examined whether children preferred to look at the similar-looking super-novel object (i.e., the intended target) or the dissimilar-looking super-novel object (i.e., the intended distractor) in leveraging trials. This examines whether children employ or avoid systematicity between word form and meaning during referent selection. To achieve this comparison, we computed a new value of PTL, where we divided the amount of time children spent looking at the target by the amount of time they spent looking at the target and the distractor. *Test* was not included in the model because we analysed only leveraging trials. The full model with all random slopes of factors centred was retained because the full model without centring random slopes of factors did not converge. Neither could we compare the full model to a null model excluding *Phase*, since the null model did not converge. Nonetheless, we obtained a summary of estimates of the predictor variables. As shown in Table 6, children were not biased to either the target or the distractor during the post-naming phase (see also Figure 4). However, children's looking behaviour was significantly influenced by their vocabulary size and neighbourhood size, such that reliance on systematicity, that is, increased looking towards the similar-looking super-novel object upon hearing the similar-sounding label was associated with a lower vocabulary size but a higher neighbourhood size.

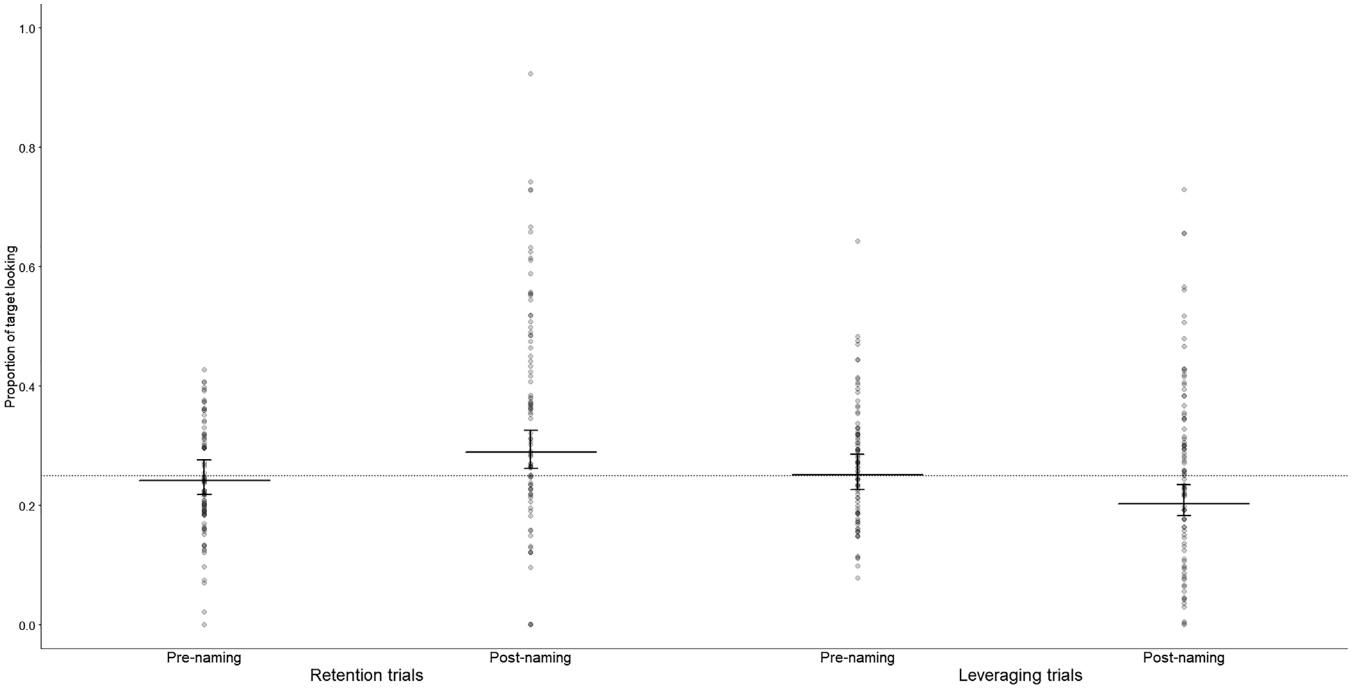



**FIGURE 2** PTL in pre-naming and post-naming phases in retention and leveraging trials of the fitted registered main model. Note. Each translucent point represents one data point. The horizontal solid lines represent the fitted model with sex dummy coded and centred. The error bars represent 95% confidence interval of the fitted model. The horizontal dotted line indicates chance level at 0.25.

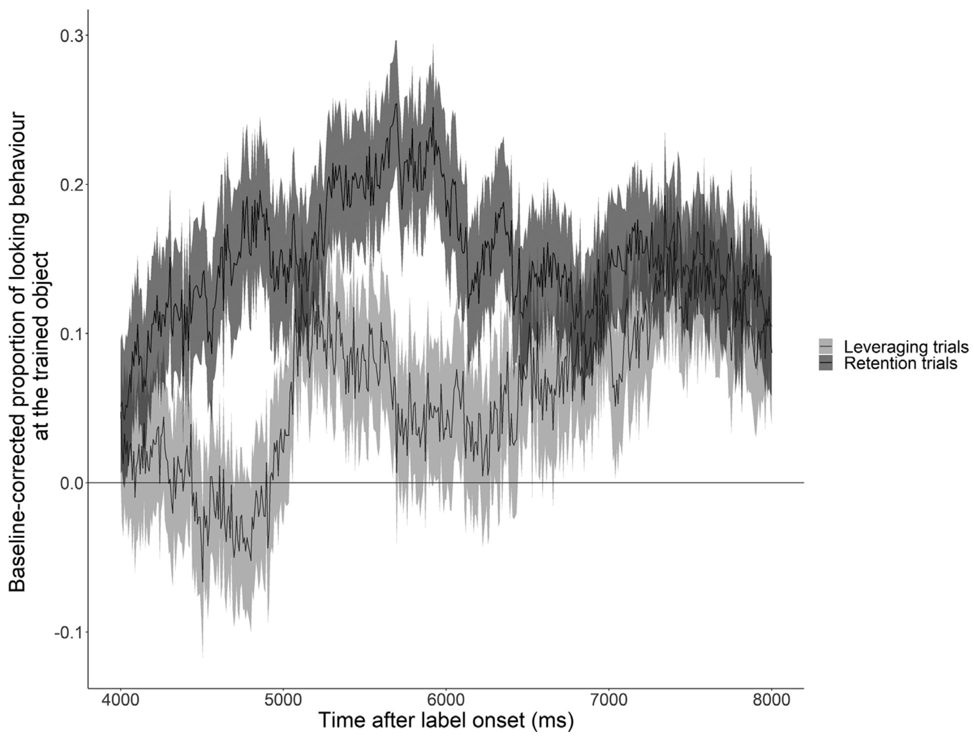



**FIGURE 3** Baseline-corrected proportion of looking at the trained objects as opposed to their respective super-novel semantic neighbours in retention and leveraging trials. Note. The trained objects are identified as targets in the retention trials but as semantic neighbours in the leveraging trials. Horizontal line at 0 indicates no preference in looking at either the trained object or the super-novel semantic neighbour. A positive baseline-corrected proportion indicates a preference for the trained object whereas a negative baseline-corrected proportion indicates a preference for the super-novel semantic neighbour.

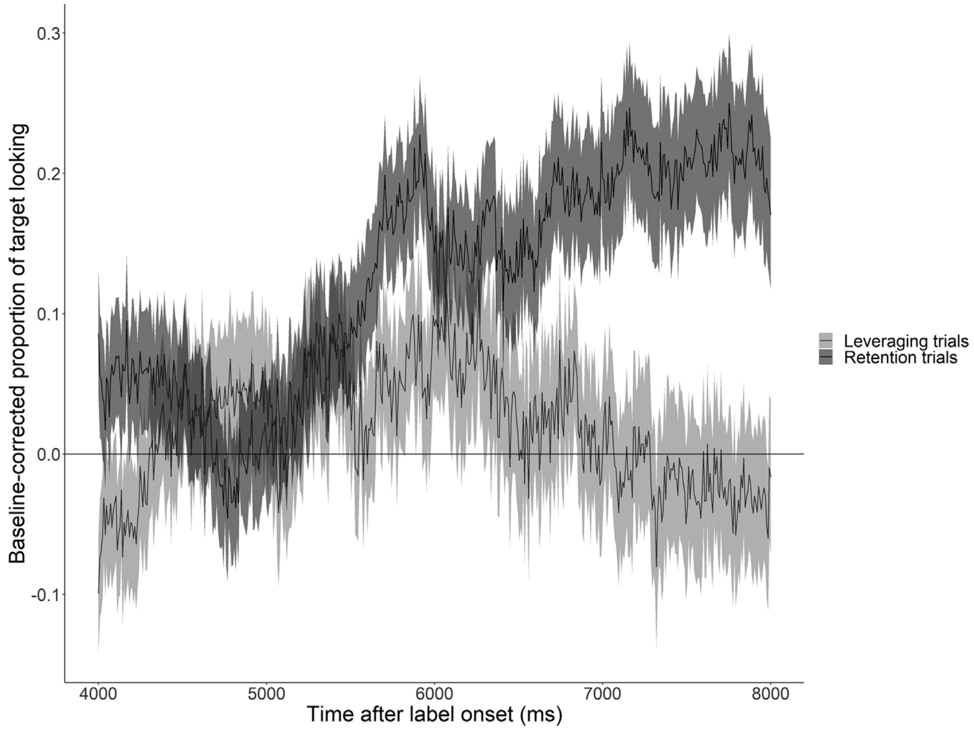



**FIGURE 4** Baseline-corrected proportion of looking at the target as opposed to its distractor in retention and leveraging trials. *Note*. The target-distractor pairs are the two trained objects in the retention trials and the two super-novel objects in the leveraging trials. Horizontal line at 0 indicates no preference towards either of the two objects. A positive baseline-corrected proportion indicates a preference for the target whereas a negative baseline-corrected proportion indicates a preference for the distractor.

**TABLE 6** Estimates, standard errors and *P*-values of predictor variables included in the full systematicity model.

| Items | Estimates | Standard error | *P*-value |
|---|---|---|---|
| Intercept | 0.053 | 0.111 | 0.630 |
| Phase (reference level: post-naming) | −0.131 | 0.112 | 0.242 |
| Vocabulary size | −0.158 | 0.074 | 0.033 |
| Neighbourhood size | 0.149 | 0.073 | 0.042 |
| Sex (reference level: female) | 0.104 | 0.110 | 0.347 |

*Note*. All covariates have been z-transformed. All random slopes of factors have been centred. The reference level of factors was determined alphabetically by default in R.

## 4 | DISCUSSION

In the current study, we examined the extent to which systematicity guided referent selection in early development. In particular, we examined whether children were more likely to assume that the referents of two words which sound similar to one another also overlap in meaning, here, look similar to one another. To examine this, we first trained children on two novel word-object associations (e.g., *maacke*-pangolin). We then presented children with four images at test, two of which they had seen before during training, and two of which were super-novel semantic neighbours of these trained objects. In retention trials, we examined whether children fixated the labelled trained object when presented with the label. In leveraging trials, we examined whether children relied on word form-meaning systematicity during referent selection, that is, when presented with a novel phonological neighbour of one of the trained labels (e.g., *maasche*), whether children fixated the super-novel object that overlapped perceptually with the trained object with the similar-sounding label (e.g., aardvark). Children showed robust evidence for learning of the trained word-object associations. However, they neither favoured nor avoided form-meaning systematicity in leveraging trials. Exploratory analyses suggested that children were sensitive to the phonological differences between similar-sounding labels and to the perceptual differences between similar-looking objects.

With regard to the retention trials, we note that children successfully identified the target upon hearing the trained label. This is despite the fact that our design, presenting children with two newly-learned objects and two perceptually overlapping super-novel objects, is arguably more difficult than other word-learning tasks (e.g., Bion et al., 2013; Diesendruck & Markson, 2001; Pomiechowska & Gliga et al., 2019). While adding to the bulk of evidence that young children show robust evidence of word-object association learning (e.g., Wojcik & Saffran, 2013), this finding allows us to examine our critical hypothesis, namely, that children leverage this newly learned mapping in future referent selection.

In leveraging trials, children saw two trained objects and two super-novel objects while hearing a novel phonological neighbour of one of the trained labels. We assumed that were children to rely on form-meaning systematicity, they would fixate the super-novel object which looks similar to the trained object whose label is similar to the test label (e.g., fixate the aardvark for the test label *maasche* when the trained word-object association was *maacke*-pangolin). However, we found no evidence that systematicity influenced referent selection in leveraging trials. In particular, children tended to fixate one or more of the distractors, rather than the target in leveraging trials (see Table 4).

Exploratory analyses further examined which of the distractors children fixated in leveraging trials. On the one hand, children may fixate the trained similar-looking object in leveraging trials, as an index of their difficulty in discriminating the similar-sounding labels and/or similar-looking objects. Alternatively, children may fixate the super-novel semantic neighbour of the other trained objects, thereby showing a preference for arbitrariness in early referent selection.

With regard to the first option, we found no statistical evidence that children looked at the trained object upon hearing the similar-sounding label in the leveraging trials. Furthermore, we obtained a significant difference in children's fixations to the trained object in the post-naming phase of retention and leveraging trials. Thus, children were sensitive to differences between two similar-sounding labels and two similar-looking objects. However, we note that the time course plot in Figure 3 suggests that there were brief windows where children did gaze at the trained object in leveraging trials, suggesting that children may overlook such sensitivity in order to find the nearest possible match in more ambiguous contexts. Such a finding is in keeping with previous findings suggesting that very young toddlers have difficulties learning similar-sounding labels (Mather & Plunkett, 2011; Swingley & Aslin, 2007). Nevertheless, our findings suggest that children were sensitive to the distinction between the similar-sounding labels and similar-looking objects—there was a difference between fixations to the trained object across retention and leveraging trials—but that they may have a non-significant tendency towards considering the trained object as a referent of the similar-sounding label. This does, however, raise doubt with regard to the role that systematicity may play in early referent selection inasmuch as systematic word pairs, that is, word pairs that sound similar and mean similar things, may be mistaken for one another in early development. At the very least, this result suggests that there was no clear preference for the super-novel semantic neighbour, that is, there was no evidence for systematicity in children's referent selection in leveraging trials.

Regarding the second option, that is, whether children showed a tendency to *avoid* word form-meaning systematicity, we found no evidence to support this claim either. In other words, children in the present study did not actively fixate the dissimilar-looking super-novel object in leveraging trials. However, we did find that children's looking behaviour was modulated by both vocabulary size and neighbourhood size. We found that children with larger vocabularies looked more towards the dissimilar-looking super novel object, that is, there was a bias towards systematicity in children with smaller vocabularies. We speculate that children with larger vocabularies may have greater exposure to arbitrary word-form meaning relations, given that systematicity does not increase proportionally with vocabulary size. Thus, with greater lexical maturity, children may be more biased to expect arbitrariness in new mappings. At the same time, larger vocabulary sizes are likely to also increase neighbourhood sizes, which we found to be positively associated with systematicity. The positive association between neighbourhood size and the systematicity effect reported may be explained by the suggestion that children who knew more semantic neighbours were able to leverage their greater semantic knowledge about the objects in the service of learning (Barabasi & Albert, 1999; Mani & Ackermann, 2018 for a review), while also being able to better discriminate the semantically similar objects. Thus, there appears to be nuances with regard to how systematicity and arbitrariness are impacted by and impact lexical development.

Importantly, unlike iconicity, which appears early in vocabulary development, that is, around 10–12 months (Laing, 2014), we did not find supporting or contradicting evidence for form-meaning systematicity during referent selection. Our finding appears to contrast with most previous studies that words which are acquired early (i.e., around 18–24 months) tend to be more iconic (Monaghan et al., 2014; Perry et al., 2015, 2018), giving children a basis to guide early referent selection (Cassani & Limacher, 2022; Cassani et al., 2020). Currently, we cannot ascertain when—and if—children will demonstrate a systematicity bias in early referent selection. On the one hand, our findings above suggest that greater lexical maturity—in terms of vocabulary size—may be associated with a bias towards arbitrariness. This would suggest that systematicity may indeed arise early in development, earlier than the ages tested in the current study. On the other hand, across different analyses, more dense semantic neighbourhoods were robustly associated with a bias towards systematicity, suggesting that the structure of the vocabulary may be linked to greater systematicity. Indeed, similar changes to vocabulary structure appear to be implicated in the word recognition literature. Thus, older children with more dense networks and larger vocabularies show greater interference effects in semantic and phonological priming studies (Arias-Trejo et al., 2022; Avila-Varela et al., 2021). In line with this developmental trajectory, the reliance on form-meaning systematicity may emerge later, with greater lexical maturity and more complex vocabulary structure. In other words, once a particular word-object association is well consolidated in the child's lexicon, information such as semantic category and phonology similarity may be extracted fairly automatically (as in adults; Gatti et al., 2023), allowing children to start exploiting systematicity in the service of learning.

Crucially, we note that the absence of a systematicity effect in the current study is unlikely to be due to a lack of power since our final sample size was 93 children and the power analysis showed that we need 88 children in order to achieve a power of 90%. It is also unlikely that the null effect is due to a poor fitting model, given the stability of our main registered beta model, that is, there were no outliers which significantly influenced model output. Instead, we argue that the absence of a systematicity effect may arise from the complex mechanisms behind word form-meaning systematicity. For instance, we may be more likely to find an effect of systematicity had the trained similar-looking object not been simultaneously presented, or were the super-novel labels to sound similar to words that children are already familiar with. Such manipulations may enable us to tap into potentially weaker effects of systematicity in early referent selection. Ongoing studies in our lab are currently exploring these possibilities. In conclusion, therefore, while the current study finds that children were reliably able to learn the trained novel word-object associations and to discriminate between two similar-sounding labels and two similar-looking objects, children did not rely on form-meaning systematicity in early referent selection. We found no evidence that children were more likely to assume that the referents of two words which sound similar to one another also overlap in meaning, here, look similar to one another.

## ORCID

*Ming Yean Sia* https://orcid.org/0000-0001-8180-8164
*Emily Mather* https://orcid.org/0000-0001-9924-1637
*Matthew W. Crocker* https://orcid.org/0000-0003-3452-3064
*Nivedita Mani* https://orcid.org/0000-0002-9843-4629

## ENDNOTE

[1] We found that this was virtually impossible as children sometimes turned back to look back at their parents. Thus, parents were asked to wear a pair of sunglasses instead.

## ACKNOWLEDGMENTS

We are very thankful to Roger Mundry for his input regarding statistical analyses and sample size calculation. We are also very thankful to the families who participated in the study and to the research assistants who helped in data collection. This publication was funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) - Project-ID 454648639 - SFB 1528 Cognition of Interaction and RTG 2070 Understanding Social Relationships.

Open access funding enabled and organized by Projekt DEAL.

## ETHICS STATEMENT

We have obtained ethics approval from the ethics committee of University of Göttingen for this project.

## CONFLICT OF INTEREST STATEMENT

We declare no potential conflict of interest.

## DATA AVAILABILITY STATEMENT

All study materials, anonymised raw data (https://doi.org/10.17605/OSF.IO/DWGPT), analyses scripts as well as the registered protocol are publicly available at https://osf.io/7mbsh/

## REFERENCES

Abramova, E., & Fernandez, R. (2016). Questioning arbitrariness in language: A data-driven study of conventional iconicity. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, 343–352. Retrieved from https://www.aclweb.org/anthology/N16-1038.pdf

Ackermann, L., Hepach, R., & Mani, N. (2020). Children learn words easier when they are interested in the category to which the word belongs. *Developmental Science*, *23*(3), e12915. https://doi.org/10.1111/desc.12915

Altvater-Mackensen, N., & Mani, N. (2013). The impact of mispronunciations on toddler word recognition: Evidence for cascaded activation of semantically related words from mispronunciations of familiar words. *Infancy*, *18*(6), 1030–1052. https://doi.org/10.1111/infa.12022

Arias-Trejo, N., Angulo-Chavira, A. Q., Avila-Varela, D. S., Chua-Rodriguez, F., & Mani, N. (2022). Developmental changes in phonological and semantic priming effects in Spanish-speaking toddlers. *Developmental Psychology*, *58*(2), 236–251. https://doi.org/10.1037/dev0001290

Arias-Trejo, N., & Plunkett, K. (2009). Lexical-semantic priming effects during infancy. *Philosophical Transactions of The Royal Society B*, *364*, 3633–3647. https://doi.org/10.1098/rstb.2009.0146

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*, 63–80. https://doi.org/10.1016/j.jecp.2009.10.002

Asano, M., Imai, M., Kita, S., Kitajo, K., Okada, H., & Thierry, G. (2015). Sound symbolism scaffolds language development in preverbal infants. *Cortex*, *63*, 196–205. https://doi.org/10.1016/j.cortex.2014.08.025

Avila-Varela, D. S., Arias-Trejo, N., & Mani, N. (2021). A longitudinal study of the role of vocabulary size in priming effects in early childhood. *Journal of Experimental Child Psychology*, *205*, 105071. https://doi.org/10.1016/j.jecp.2020.105071

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*, 390–412. https://doi.org/10.1016/j.jml.2007.12.005

Ballem, K. D., & Plunkett, K. (2005). Phonological specificity in children at 1;2. *Journal of Child Language*, *32*, 159–173. https://doi.org/10.1017/S0305000904006567

Barabási, A., & Albert, R. (1999). Emergence of scaling in random networks. *Science*, *286*, 509–512. https://doi.org/10.1126/science.286.5439.509

Barr, D. J., Levy, R., Scheepers, C., & Tily, H. J. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory Language*, *68*, 255–278. https://doi.org/10.1016/j.jml.2012.11.001

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software*, *67*, 1–48. https://doi.org/10.18637/jss.v067.i01

Beckage, N., Smith, L., & Hills, T. (2011). Small worlds and semantic network growth in typical and late talkers. *PLoS ONE*, *6*(5), e19348. https://doi.org/10.1371/journal.pone.0019348

Bergelson, E., & Aslin, R. N. (2017). Nature and origins of the lexicon in 6-mo-olds. *Proceedings of the National Academy of Sciences*, *114*(49), 2916–12921. https://doi.org/10.1073/pnas.1712966114

Bion, R. A. H., Borovsky, A., & Fernald, A. (2013). Fast mapping, slow learning: Disambiguation of novel word-object mappings in relation to vocabulary learning at 18, 24, and 30 months. *Cognition*, *126*(1), 39–53. https://doi.org/10.1016/j.cognition.2012.08.008

Borovsky, A., & Elman, J. L. (2006). Language input and semantic categories: A relation between cognition and early word learning. *Journal of Child Language*, *33*, 759–790. https://doi.org/10.1017/S0305000906007574

Borovsky, A., Ellis, E. M., Evans, J. L., & Elman, J. L. (2016). Lexical leverage: Category knowledge boosts real-time novel word recognition in 2-year-olds. *Developmental Science*, *19*(6), 918–932. https://doi.org/10.1111/desc.12343

Brooks, P. J., Braine, M. D. S., Catalano, L., Brody, R. E., & Sudhalter, V. (1993). Acquisition of gender-like noun subclasses in an artificial language: The contribution of phonological markers to learning. *Journal of Memory and Language*, *32*, 76–95. https://doi.org/10.1006/jmla.1993.1005

Brooks, M. E., Kristensen, K., van Benthem, K. J., Magnusson, A., Berg, C. W., Nielsen, A., Skaug, H. J., Mächler, M., & Bolker, B. M. (2017). glmmTMB balances speed and flexibility among packages for zero-inflated generalized linear mixed modeling. *The R Journal*, *9*, 378–400. https://doi.org/10.3929/ethz-b-000240890

Cassani, G., Chuang, Y. Y., & Baayen, R. H. (2020). On the semantics of nonwords and their lexical category. *Journal of Experimental Psychology—*

*Learning, Memory and Cognition*, *46*(4), 621–637. https://doi.org/10.1037/xlm0000747
Cassani, G., & Limacher, N. (2022). Not just form, not just meaning: Words with consistent form-meaning mappings are learned earlier. *Quarterly Journal of Experimental Psychology*, *75*(8), 1464–1482. https://doi.org/10.1177/17470218211053472
Chaabouni, R., Kharitonov, E., Dupoux, E., & Baroni, M. (2021). Communicating artificial neural networks develop efficient colour-naming systems. *Proceedings of the National Academy of Sciences*, *118*(12), e2016569118. https://doi.org/10.1073/pnas.2016569118
Chow, J., Davies, A. A., & Plunkett, K. (2017). Spoken-word recognition in 2-year-olds: The tug of war between phonological and semantic activation. *Journal of Memory and Language*, *93*, 104–134. https://doi.org/10.1016/j.jml.2016.08.004
Corballis, M. C. (2002). *From hand to mouth: The origins of language*. Princeton University Press.
Dautriche, I., Swingley, D., & Christophe, A. (2015). Learning novel phonological neighbours: Syntactic category matters. *Cognition*, *143*, 77–86. https://doi.org/10.1016/j.cognition.2015.06.003
Dautriche, I., Mahowald, K., Gibson, E., & Piantadosi, S. T. (2017). Word-form similarity increases with semantic similarity: An analysis of 100 languages. *Cognitive Science*, *41*, 2149–2169. https://doi.org/10.1111/cogs.12453
Delle Luche, C., Durrant, S., Floccia, C., & Plunkett, K. (2014). Implicit meaning in 18-month-old toddlers. *Developmental Science*, *17*(6), 948–955. https://doi.org/10.1111/desc.12164
Diesendruck, G., & Markson, L. (2001). Children's avoidance of lexical overlap: A pragmatic account. *Developmental Psychology*, *37*(5), 630–641. https://doi.org/10.1037//0012-1649.37.5.630
Dingemanse, M., Blasi, D. E., Lupyan, G., Christiansen, M. H., & Monaghan, P. (2015). Arbitrariness, iconicity, and systematicity in language. *Trends in Cognitive Sciences*, *19*(10), 603–615. https://doi.org/10.1016/j.tics.2015.07.013
Fourtassi, A., Regan, S., & Frank, M. C. (2021). Continuous developmental change explains discontinuities in word learning. *Developmental Science*, *24*(2), e13018. https://doi.org/10.1111/desc.13018
Fourtassi, A., Scheinfeld, I. L., & Frank, M. C. (2019). The development of abstract concepts in children's early lexical networks. Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics, pp. 129–133. Retrieved from https://www.aclweb.org/anthology/W19-2914.pdf
Frank, M. C., Braginsky, M., Yurovsky, D., & Marchman, V. A. (2016). Wordbank: An open repository for developmental vocabulary data. *Journal of Child Language*, *44*(3), 677–694. https://doi.org/10.1017/S0305000916000209
Gasser, M. (2004). The origins of arbitrariness in language. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *26*, 434–439. Retrieved from https://escholarship.org/uc/item/34g8355v
Gatti, D., Marelli, M., & Rinaldi, L. (2023). Out-of-vocabulary but not meaningless: Evidence for semantic-priming effects in pseudoword processing. *Journal of Experimental Psychology: General*, *152*(3), 851–863. https://doi.org/10.1037/xge0001304
Greenberg, J. H. (1957). *Essays in linguistics*. University of Chicago Press.
Haiman, J. (2015). Iconicity in linguistics. *International Encyclopaedia of the Social & Behavioural Sciences*, *11*, 512–515. https://doi.org/10.1016/B978-0-08-097086-8.52006-6
Herman, L. M., Richards, D. G., & Wolz, J. P. (1984). Comprehension of sentences by bottlenosed dolphins. *Cognition*, *16*, 129–219. https://doi.org/10.1016/0010-0277(84)90003-9
Hockett, C. F., & Hockett, C. D. (1960). The origin of speech. *Scientific American*, *203*(3), 88–97. Retrieved from https://www.jstor.org/stable/24940617 https://doi.org/10.1038/scientificamerican0960-88
Hollich, G., Jusczyk, P., & Luce, P. (2002). Lexical neighbourhood effects in 17-month-old word learning. Proceedings of the 26th Annual Boston University Conference on Language Development (pp. 314–323) Cascadilla Press. Retrieved from http://hincapie.psych.purdue.edu/NH2002.pdf
Horst, J. S., Samuelson, L. K., Kucker, S. C., & McMurray, B. (2011). What's new? Children prefer novelty in referent selection. *Cognition*, *118*, 234–244. https://doi.org/10.1016/j.cognition.2010.10.015
Imai, M., & Kita, S. (2014). The sound symbolism bootstrapping hypothesis for language acquisition and language evolution. *Philosophical Transactions of The Royal Society B*, *369*(1651), 2013–0298. https://doi.org/10.1098/rstb.2013.0298
Johnson, E. K., & Huettig, F. (2011). Eye movements during language-mediated visual search reveal a strong link between overt visual attention and lexical processing in 36-month-olds. *Psychological Research*, *75*, 35–42. https://doi.org/10.1007/s00426-010-0285-4
Kaminski, J., Call, J., & Fischer, J. (2004). Word learning in a domestic dog: Evidence for "fast mapping". *Science*, *304*, 1682–1683. https://doi.org/10.1126/science.1097859
Kucker, S. C., McMurray, B., & Samuelson, L. K. (2020). Sometimes it is better to know less: How known words influence referent selection and retention in 18- to 24-month-old children. *Journal of Experimental Child Psychology*, *189*, 104705. https://doi.org/10.1016/j.jecp.2019.104705
Kuznetsova, A., Brockhoff, P. B., & Christensen, R. H. B. (2017). lmerTest Package: Tests in linear mixed effects models. *Journal of Statistical Software*, *82*(13), 1–26. https://doi.org/10.18637/jss.v082.i13
Laing, C. E. (2014). A phonological analysis of onomatopoeia in early word production. *First Language*, *34*(5), 387–405. https://doi.org/10.1177/0142723714550110
Luce, P. A., & Pisoni, D. B. (1998). Recognising spoken words: The neighbourhood activation model. *Ear Hear*, *19*(1), 1–36. https://doi.org/10.1097/00003446-199802000-00001
Luke, S. G. (2017). Evaluating significance in linear mixed-effects models in R. *Behaviour Research Methods*, *49*, 1494–1502. https://doi.org/10.3758/s13428-016-0809-y
Mani, N., & Ackermann, L. (2018). Why do children learn the words they do? *Child Development Perspectives*, *12*(4), 253–257. https://doi.org/10.1111/cdep.12295
Mani, N., & Borovsky, A. (2017). Building a lexical network. In Westermann, G., & Mani, N. (Eds.), *Early word learning (Current issues in developmental psychology)* (pp. 57–69). Routledge.
Mani, N., Durrant, S., & Floccia, C. (2012). Activation of phonological and semantic codes in toddlers. *Journal of Memory and Language*, *66*, 612–622. https://doi.org/10.1016/j.jml.2012.03.003
Mani, N., Johnson, E., McQueen, J. M., & Huettig, F. (2013). How yellow is your banana? Toddlers' language-mediated visual search in referent-present tasks. *Developmental Psychology*, *49*(6), 1036–1044. https://doi.org/10.1037/a0029382
Mani, N., & Plunkett, K. (2008). Fourteen-month-olds pay attention to vowels in novel words. *Developmental Science*, *11*(1), 53–59. https://doi.org/10.1111/j.1467-7687.2007.00645.x
Mani, N., & Plunkett, K. (2010a). In the infant's mind's ear: Evidence for implicit naming in 18-month-olds. *Psychological Science*, *21*(7), 908–913. https://doi.org/10.1177/0956797610373371
Mani, N., & Plunkett, K. (2010b). Twelve-month-olds know their cups from their keps and tups. *Infancy*, *15*(5), 445–470. https://doi.org/10.1111/j.1532-7078.2009.00027.x
Mani, N., & Plunkett, K. (2011). Phonological priming and cohort effects in toddlers. *Cognition*, *121*, 196–206. https://doi.org/10.1016/j.cognition.2011.06.013
Mather, E., & Plunkett, K. (2011). Mutual exclusivity and phonological novelty constrain word learning at 16 months. *Journal of Child Language*, *38*, 933–950. https://doi.org/10.1017/S0305000910000401
Mather, E., & Plunkett, K. (2012). The role of novelty in early word learning. *Cognitive Science*, *36*, 1157–1177. https://doi.org/10.1111/j.1551-6709.2012.01239.x
Mayor, J., & Mani, N. (2019). A short version of the MacArthur–Bates communicative development inventories with high validity. *Behaviour*

*Research Methods*, *51*, 2248–2255. https://doi.org/10.3758/s13428-018-1146-0

Monaghan, P., Chater, N., & Christiansen, M. H. (2005). The differential role of phonological and distributional cues in grammatical categorisation. *Cognition*, *96*, 143–182. https://doi.org/10.1016/j.cognition.2004.09.001

Monaghan, P., Christiansen, M. H., & Fitneva, S. A. (2011). The arbitrariness of the sign: Learning advantages from the structure of the vocabulary. *Journal of Experimental Psychology: General*, *140*(3), 325–347. https://doi.org/10.1037/a0022924

Monaghan, P., Shillcock, R. C., Christiansen, M. H., & Kirby, S. (2014). How arbitrary is language? *Philosophical Transactions of The Royal Society B*, *369*(1651), 2013–0299. https://doi.org/10.1098/rstb.2013.0299

Newman, R., Samuelson, L., & Gupta, P. (2008). Learning novel neighbours: Distributed mappings help children and connectionist models. *Proceedings of the Annual Meeting of the Cognitive Science Society*, *30*, 29–34. Retrieved from https://escholarship.org/uc/item/80p7d1h2

Perry, L. K., Perlman, M., & Lupyan, G. (2015). Iconicity in English and Spanish and its relation to lexical category and age of acquisition. *PLoS ONE*, *10*(9), e0137147. https://doi.org/10.1371/journal.pone.0137147

Perry, L. K., Perlman, M., Winter, B., Massaro, D. W., & Lupyan, G. (2018). Iconicity in the speech of children and adults. *Developmental Science*, *21*(3), e12572. https://doi.org/10.1111/desc.12572

Pomper, R., & Saffran, J. R. (2019). Familiar object salience affects novel word learning. *Child Development*, *90*(2), e246–262. https://doi.org/10.1111/cdev.13053

Pomiechowska, B., & Gliga, T. (2019). Lexical acquisition through category matching: 12-month-old infants associate words to visual categories. *Psychological Science*, *30*(2), 288–299. https://doi.org/10.1177/0956797618817506

R Core Team (2021). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing.

Roy, B. C., Frank, M. C., DeCamp, P., Miller, M., & Roy, D. (2013). Predicting the birth of a spoken word. *Proceedings of the National Academy of Sciences*, *112*(41), 12663–12668. https://doi.org/10.1073/pnas.1419773112

Schielzeth, H., & Forstmeier, W. (2009). Conclusions beyond support: Overconfident estimates in mixed models. *Behavioural Ecology*, *20*, 416–420. https://doi.org/10.1093/beheco/arn145

Schmid, B., Mani, N., & Behne, T. (2019). The cognitive underpinnings and early development of children's selective trust. Retrieved from https://ediss.uni-goettingen.de/handle/21.11130/00-1735-0000-0005-12B5-A

Smithson, M., & Verkuilen, J. (2006). A better lemon squeezer? Maximum-likelihood regression with beta-distributed dependent variables. *Psychological Methods*, *11*, 54–71. https://doi.org/10.1037/1082-989X.11.1.54

Steyvers, M., & Tenenbaum, J. B. (2005). The large-scale structure of semantic networks: Statistical analyses and a model of semantic growth. *Cognitive Science*, *29*, 41–78. https://doi.org/10.1207/s15516709cog2901_3

Storkel, H. L. (2004). Do children acquire dense neighbourhoods? An investigation of similarity neighbourhoods in lexical acquisition. *Applied Psycholinguistics*, *25*, 201–221. https://doi.org/10.1017/S0142716404001109

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, *76*, 147–166. https://doi.org/10.1016/S0010-0277(00)00081-0

Swingley, D., & Aslin, R. N. (2007). Lexical competition in young children's word learning. *Cognitive Psychology*, *54*(2), 99–132. https://doi.org/10.1016/j.cogpsych.2006.05.001

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

Von Holzen, K., & Bergmann, C. (2018). A meta-analysis of infants' mispronunciation sensitivity development. In CogSci: Annual conference of the cognitive science society (pp. 1157–1162). Retrieved from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6476320/

Willits, J. A., Wojcik, E. H., Seidenberg, M. S., & Saffran, J. R. (2013). Toddlers activate lexical semantic knowledge in the absence of visual referents: Evidence from auditory priming. *Infancy*, *18*(6), 1053–1075. https://doi.org/10.1111/infa.12026

Winter, B., Perlman, M., Perry, L. K., & Lupyan, G. (2017). Which words are most iconic? Iconicity in English sensory words. *Interaction Studies*, *18*(3), 443–464. https://doi.org/10.1075/is.18.3.07win

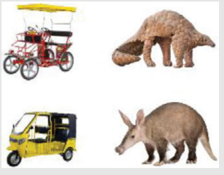
Wojcik, E. H., & Saffran, J. R. (2013). The ontogeny of lexical networks: Toddlers encode the relationships among referents when learning novel words. *Psychological Science*, *24*(10), 1898–1905. https://doi.org/10.1177/0956797613478198

Yurovsky, D., Bion, R., Smith, L. B., & Fernald, A. (2012). Mutual exclusivity and vocabulary structure. In N. Miyake, D. Peebles, & R. P. Cooper (Eds.), Proceedings of the 34th annual meeting of the cognitive science society (pp. 1197–1202) Cognitive Science Society. Retrieved from http://callab.uchicago.edu/papers/ybsf-cogsci-2012.pdf

**How to cite this article:** Sia, M. Y., Mather, E., Crocker, M. W., & Mani, N. (2023). The Role of Systematicity in Early Referent Selection. *Developmental Science*, 19, e13444. https://doi.org/10.1111/desc.13444

## APPENDIX A

List of stimuli in every trial (for one counterbalanced condition)

| Block | Phase | Trial | Visual stimuli | Auditory stimuli |
|---|---|---|---|---|
| 1 | Training | 1 |  | That is a *shufi*! Wow, a *shufi*! Do you see the *shufi*? Look, a *shufi*! Oh, what a fantastic *shufi*! I see a *shufi*!<br>(*Das ist ein X! Wow, ein X! Siehst du das X? Schau mal, ein X! Oh, was für ein tolles X! Ich sehe ein X!*) |
| | | 2 |  | That is a *peto*! Wow, a *peto*! Do you see the *peto*? Look, a *peto*! Oh, what a fantastic *peto*! I see a *peto*! |
| | | 3 |  | That is a *shufi*! Wow, a *shufi*! Do you see the *shufi*? Look, a *shufi*! Oh, what a fantastic *shufi*! I see a *shufi*! |
| | | 4 |  | That is a *peto*! Wow, a *peto*! Do you see the *peto*? Look, a *peto*! Oh, what a fantastic *peto*! I see a *peto*! |
| | Retention | 1 |  | You know that now! Now, where is the *shufi*? The *shufi*!<br>(*Das weißt du jetzt! Wo ist nun das X? Das X!*) |
| | | 2 |  | You know that too! Now, where is the *peto*? The *peto*! |
| | Leveraging | 1 |  | And now something new! Where is probably the *shuri*? The *shuri*!<br>(*Und jetzt etwas Neues! Wo ist wohl das X? Das X!*) |
| | | 2 |  | And now another new thing! Where is probably the *pewo*? The *pewo*! |
| 2 | Training | 1 |  | That is a *gibbel*! Wow, a *gibbel*! Do you see the *gibbel*? Look, a *gibbel*! Oh, what a fantastic *gibbel*! I see a *gibbel*! |
| | | 2 |  | That is a *maacke*! Wow, a *maacke*! Do you see the *maacke*? Look, a *maacke*! Oh, what a fantastic *maacke*! I see a *maacke*! |

| Block | Phase | Trial | Visual stimuli | Auditory stimuli |
|---|---|---|---|---|
| | | 3 |  | That is a *maacke*! Wow, a *maacke*! Do you see the *maacke*? Look, a *maacke*! Oh, what a fantastic *maacke*! I see a *maacke*! |
| | | 4 |  | That is a *gibbel*! Wow, a *gibbel*! Do you see the *gibbel*? Look, a *gibbel*! Oh, what a fantastic *gibbel*! I see a *gibbel*! |
| | Retention | 1 |  | You know that now! Now, where is the *gibbel*? The *gibbel*! |
| | | 2 |  | You know that too! Now, where is the *maacke*? The *maacke*! |
| | Leveraging | 1 |  | And now something new! Where is probably the *maasche*? The *maasche*! |
| | | 2 |  | And now another new thing! Where is probably the *gissel*? The *gissel*! |

**APPENDIX B: Details on model specification and data simulation to determine the sample size**

The models used to analyse the results will be fitted in R (v4.1.1 or higher; R Core Team, 2021) using the function glmmTMB of the identically-named package (v1.1.1 or higher; Brooks et al., 2017) for beta models or lmer of the package lme4 (v1.1-27.1 or higher; Bates et al., 2015) for Gaussian models. For beta models, we shall determine the significance of each fixed effect using likelihood ratio tests, in that we compare the full model with a reduced model lacking the effects in question (R function drop1), whereas for Gaussian models, we shall determine the significance of each fixed effect using the Satterthwaite approximation (Luke, 2017), as implemented in the function lmer of the package lmerTest (v3.1-3 or higher; Kuznetsova et al., 2017). Confidence intervals of fixed effects' estimates and fitted values will be determined by means of a parametric bootstrap (functions simulate or bootMer of the packages glmmTMB and lme4, respectively). In cases where the response comprises values being exactly 0 or 1 when a beta model is used, we shall transform the response as recommended in Smithson and Verkuilen (2006).

Power analysis was conducted by simulating the dataset from Schmid et al. (2019) based on the model just described. We determined the random effects (intercepts and slopes) associated with child ID and also the precision parameter phi to be simulated based on a model fitted to the dataset of Experiment 1b in Schmid (2019) where 5-year-olds were presented with an array of four objects during test (see https://ediss.uni-goettingen.de/handle/21.11130/00-1735-0000-0005-12B5-A for details and data of the study). Data used for the power simulation were from 2-year-olds taking part in the same study. As the dataset from Schmid et al. (2019) comprises only two targets, the standard deviation of the target's random intercepts (in link space) was set in a way that the expected average of absolute difference between mean responses for each target is equal to the fixed effect of the target in the model. For all random slopes within the target, we simulated the effects to be zero.

We determined the coefficients to be simulated for the fixed effects based on the following; we assumed average PTL to be 0.25 (for pre-naming phase in retention and leveraging tests), 0.35 (for post-naming phase in retention test), and 0.30 (for post-naming phase in leveraging test) in our simulation. For all other fixed effects, we simulated an effect of zero.

To simulate the vocabulary size and the neighbourhood size for categories ANIMAL and VEHICLE, we first sampled children's age from a uniform distribution with a minimum of 21 months and a maximum of 27 months (rounded to the nearest integer). Using this sampling method, a child is randomly selected from the full child-by-word data downloaded from www.wordbank.com to get a proportion of known words for the full vocabulary list and for the semantic categories ANIMAL and VEHICLE. These proportions were then used to assess the vocabulary size and neighbourhood size of our participants by sampling from a binomial distribution with a sampling size of 25, 45 and 14 for the general vocabulary and the semantic categories ANIMAL and VEHICLE, respectively. The resulting proportions were added as predictors to the simulated data set.

We then simulated the response with regard to the fixed and random effects in link space and then logit-transformed it to proportions. Numbers from a beta distribution were then randomly sampled with a respective mean and phi to generate the response (i.e., one that is bound between zero and one). Following this, we fitted a corresponding beta regression mixed model with logit link function (as we will be measuring PTL), such that our response measure ranged from zero to one. However, if this does not converge, we shall *arcsine* transform the response (i.e., PTL of children) and use a Gaussian model instead:

$\text{Model}_{\text{Gaussian}}$ = lmer (PTL ~ Test*Phase + vocabulary_size + neighbourhood_size + sex + (Test*Phase || id) + (Test*Phase + vocabulary_size + neighbourhood_size + sex | Object), Data = data, REML = F, control = contr, weights = total.looking.time)

As convergence is unlikely to be an issue with Gaussian models, in this latter model we could include parameters for the correlations among random intercepts and slopes within *Object*. We determined power as the proportion of simulated data sets which revealed a significant ($p < 0.05$) full-null model comparison.