DOI: 10.1111/cdev.14010

SPECIAL SECTION

CHILD DEVELOPMENT

A network model of referent identification by toddlers in a visual world task

Mihaela Duta[1] | **Kim Plunkett**[2]

[1]Doctoral Training Centre, University of Oxford, Oxford, UK

[2]Department of Experimental Psychology, University of Oxford, Oxford, UK

Correspondence
Kim Plunkett, Department of Experimental Psychology, University of Oxford, Oxford, UK.
Email: kim.plunkett@psy.ox.ac.uk

Mihaela Duta, Doctoral Training Centre, University of Oxford, 1-4 Keble Road, Oxford OX1 3NP, UK.
Email: mihaela.duta@dtc.ox.ac.uk

Funding information
Leverhulme Trust

Abstract

We present a neural network model of referent identification in a visual world task. Inputs are visual representations of item pairs unfolding with sequences of phonemes identifying the target item. The model is trained to output the semantic representation of the target and to suppress the distractor. The training set uses a 200-word lexicon typically known by toddlers. The phonological, visual, and semantic representations are derived from real corpora. Successful performance requires correct association between labels and visual and semantic representations, as well as correct location identification. The model reproduces experimental evidence that phonological, perceptual, and categorical relationships modulate item preferences. The model provides an account of how language can drive visual attention in the inter-modal preferential looking task.

INTRODUCTION

The Intermodal Preferential Looking task (IPL, e.g., Spelke, 1976; Thomas et al., 1981; Golinkoff et al., 1987) has been widely used to study linguistic and cognitive development during infancy and early childhood. Also, commonly referred to as the looking-while-listening task (e.g., Fernald et al., 2001), it has been used to investigate infant sensitivity to temporal invariances in audio-visual relationships (Spelke, 1979), vocabulary development (Thomas et al., 1981), phonological development (Swingley & Aslin, 2000), semantic development (Meints et al., 1999), and conceptual development (Sučević et al., 2022), to mention just a few topics.

In a typical IPL task, pictures of objects are displayed on a computer monitor and one of them is named over a loudspeaker while the eye-movements of the listener are monitored either automatically or manually using static camera(s). The logic behind the task is that selective visual attention to the objects can be used to index the listener's interpretation of the auditory and visual stimuli used, thereby gaining insights into the corresponding representations and processing of these stimuli. For example, selective attention to a picture of a dog in an array of other animals upon hearing the word "dog" might be used to infer that the listener understands the word. Failure to pay selective visual attention to the picture of a dog upon hearing the sound sequence "tog" might be used to infer that the listener has a detailed representation of the word "dog," or conversely if the listener still selectively attends to the dog upon hearing "tog," that the listener has a flexible or underspecified representation of this lexical item. Similarly, if the listener only fixates an image of a Labrador but not a Poodle upon hearing the word "dog," the investigator might be tempted to conclude that the listener's understanding of the word is deviant, or at least unadult-like. Provided appropriate control conditions are incorporated into the design of such experiments, the IPL task can be used to investigate development across a broad range of linguistic and cognitive domains and ages. A critical advantage of this task is that it requires minimal activity (eye movements) from the infant to demonstrate their sensitivity to correspondences between auditory and visual stimuli.

Abbreviations: IPL, Intermodal Preferential Looking task; OCDI, Oxford Communicative Development Inventory; VWP, Visual World Paradigm.

Furthermore, infant response latencies in this task are sufficiently rapid that automatic processing can often be assumed.

The IPL task can be considered a special case of the Visual World Paradigm (VWP) which has been used to investigate psycholinguistic representations and processes in adults (Cooper, 1974; Tanenhaus et al., 1995). The linking hypothesis of the VWP is the same as IPL: selective visual attention to objects in a display are used to index the mental processes involved in interpreting a linguistic (usually auditory) stimulus. The rapid nature of eye movements allows us to monitor these mental processes on a millisecond time scale in relatively natural contexts. However, in contrast to IPL which typically uses just 2 object stimuli, the VWP may use many objects in the visual display. In order to launch a rapid saccade toward an appropriate referent in the display, the listener can take advantage of their memory for the identity of objects at the various locations in the display. However, since the opportunity to preview the display is quite brief (typically a few seconds) and the displays vary from trial-to-trial, the listener needs to continually update their binding of object identities to their locations in working memory (Huettig et al., 2011) in order to achieve rapid referent identification. The simpler visual displays used in IPL tend to make them more suited to younger participants, although more complex displays have also been used with toddlers (see Chow et al., 2017). Nonetheless, even with the simpler displays used in IPL, the response latencies of infant and toddler listeners benefit from maintaining a working memory of the identity of the displayed objects and their locations.

The cognitive and linguistic demands of both the VWP and IPL are multi-faceted. Magnuson (2019) provides an excellent summary of these requirements. Fundamental skills for successful referent identification in IPL are:

1. The ability to recognize the word(s) uttered in the auditory stimulus.
2. The ability to identify the objects represented in the visual stimulus.
3. An appreciation of the relationships between the perceived auditory and visual stimuli.
4. Memory for the unfolding auditory stimulus.
5. Memory for the visual stimuli *and* their locations.
6. The ability to launch a saccade to a spatial location based on the preceding information.

Each of these skills involves a complex set of constituent parts, which are worthy objects of investigation in their own right. A proper understanding of how individuals behave during the IPL requires us to make explicit our assumptions about the mechanisms, representations, and process underlying these constituent skills and how they operate together to enable meaningful behavior. One way to achieve such an understanding is to build a formal model able to mimic aspects of human behavior on the task. This model should clarify the researcher's theoretical stance on these skills, whether they be radical or neutral. To be sure, no model will provide a complete account of the fundamental skills (1–6) listed above: any model of the IPL will focus of some subset of these skills. The model presented in this article focuses on components 3–5 in the above listing.

A common theme raised by many researchers using IPL or VWP concerns the impact of *context* on the identification of a referent in the visual array in response to an auditory signal. Here, *context* usually refers to both the auditory context and visual context and how they interact with each other. Contexts can be *facilitatory* or *inhibitory* depending on the nature of the stimuli used in the task. For example, stimuli in the visual array that look alike can interfere with, or inhibit, referent identification. Likewise, visual stimuli that have similar sounding names can also inhibit referent identification. In contrast, extended preview of the visual stimuli can accelerate, or facilitate, referent identification upon hearing a subsequent name. Likewise, pre-exposure to the auditory stimuli can also be facilitatory, by helping listeners anticipate the likely target in a visual array, as well as predict the name of the referent: "He is eating a …." engenders faster saccades to a visible edible item than a neutral phrase such as "Look at the ….". Identification of these facilitatory and inhibitory effects provide important clues as to nature of the processes and representations involved in these situations.

In this paper, we model the impact of visual and phonological similarity on referent identification in an IPL task and identify the processes and representations that are needed to reproduce some established findings in the development literature regarding context effects in the IPL.

Overview

We describe a neural network model of the Intermodal Preferential Looking (IPL) task. The model is trained on stimuli derived from real-world corpora and is shown to mimic toddler behavior in IPL tasks using bottom-up processes alone. The model captures an important characteristic of toddler behavior in this task, namely that preferences in later development are dominated by semantic properties of the referent, whereas visual properties play a greater role in earlier development (Mandler, 2000). In contrast to earlier network models of semantic development (see, e.g., Mayor & Plunkett, 2010; McMurray, Horst & Samuelson, 2012;

Plunkett et al., 1992; Rogers & McClelland, 2004), we use realistic and dynamically unfolding representations of phonological and visual stimuli. The use of real-world corpora also distinguishes this model from other attempts to simulate adult behavior in the visual world task (Smith et al., 2017). This is made possible by the recent availability of more powerful architectures and learning procedures, as well as the availability of pre-trained models for image recognition and co-occurrence statistics.

Our goal in this work is to demonstrate that the pattern of eye fixations toward a named target item in a visual scene can be explained by the visual and semantic similarity of the items, and the phonological relationships between their labels. Furthermore, we demonstrate that it is not necessary to postulate top-down, feedback processes to account for the increasing reliance of the system on semantic information over visual information during the course of learning. In particular, the learning process itself leads to an increasing reliance on the semantic properties of the items in the visual scene, despite the system exploiting phonological and visual information in an entirely bottom-up fashion. In using the term *bottom-up*, we mean that there are no top-down connections in the network whereby activation can flow from higher levels of the system to lower levels.

We review some experimental investigations of toddler behavior in an IPL task in which the auditory and visual stimuli, and the visual and semantic relationships between them, are carefully manipulated in order to identify some of the factors that influence the regularities in the eye movements observed in the task (Angulo-Chavira et al., 2023; Arias-Trejo & Plunkett, 2010; Bergelson & Aslin, 2017). We focus on modeling the development of toddler behavior on the assumption that studying the manner in which the building blocks required to perform the task are assembled during development will provide insights into the functioning of the fully mature system. By constructing a computer model in which the visual, phonological, and semantic relationships between objects and their labels are systematically manipulated, we show how these relationships influence referent selection in a manner that mimics toddler behavior. The assumptions embodied in the model constitute a theory about linguistic and cognitive processes involved in the VWP. In the empirical part of the paper, we report on previously published and unpublished experimental work that systematically manipulates the categorical (semantic) and visual similarity of the items in the visual scene. These experimental data show that both semantic and visual factors have an impact on the toddler's ability to identify the target referent of the auditory stimulus. In particular, semantic factors come to dominate referent choices as development proceeds. We also show how phonological similarity affects performance in the task.

The centerpiece of the manuscript is a neural network model of toddler behavior in the same IPL task used by Arias-Trejo and Plunkett (2010). The model receives inputs from real-word corpora reflecting the vocabularies of young toddlers and images of typical exemplars of the referents of the words in these vocabularies. Each vocabulary item is assigned a dynamic phonological representation of the item label corresponding to the unfolding speech pattern. The visual inputs are unfolding representations (corresponding to sequences of fixations) of two items. The vocabulary item always refers to just one of the image inputs. The task of the network is to activate a semantic representation of the item that is named at the target output location, and to suppress the semantic representation of the visual input item at the distractor output location. This training regime forces the network to learn the associations between phonological, visual, and semantic representations of an item and to use these associations to identify the location of the named item by activating its semantic representation at the appropriate output location. Anthropomorphizing, we might say that the training regime encourages the network to pay *selective attention* to the target output location and to suppress attention to the distractor. In this regard, the model provides an account of how language can drive visual attention in the inter-modal preferential looking task.

The visual inputs to the model are vector representations derived from a deep neural network pre-trained on ImageNet (Deng et al., 2009). The semantic outputs reflect real-world corpora insofar as they are vector representations taken from the GloVe model pre-trained on aggregated global word–word cooccurrence statistics from a 6 billion token corpus (Pennington et al., 2014). Phonological inputs consist of an unfolding sequence of standard phonological features used to uniquely identify each word in the toddler's vocabulary (Karaminis, 2018). These training stimuli have been previously used in a neural network model of spoken word recognition (Duta & Plunkett, 2021). That model captured the phonological preference effect (PPE) whereby both toddlers and adults demonstrate an early preference to fixate a phonological distractor over a semantic distractor in a visual world task. For example, on hearing the word "trousers" both adults and toddlers will fixate a picture of a "train" before they fixate a picture of a "hat" in a target-absent trial (Chow et al., 2017; Huettig & McQueen, 2007). The model presented here builds on this work but introduces the additional component of *object location* which is absent in earlier models of referent selection in a visual world task (including Plunkett et al., 1992; Allopenna et al., 1998; Rogers & McClelland, 2004; Mayor &

Plunkett, 2010; McMurray et al., 2012; Mayor & Plunkett, 2014; Duta & Plunkett, 2021).[1] The inclusion of location is an important elaboration of the earlier model as it requires an account of how the identity of an object is *bound* to a particular location. As we shall see, the model must learn these identity-location bindings in order to generalize them to object pairings to which it has never previously been exposed. An understanding of how infants and toddlers learn object-location bindings and generalize this knowledge to new situations is critical for any theory of child development. The neural network model mimics the IPL task in that it is presented with audio-visual inputs and is directed to select the appropriate semantic representation and ignore the alternative.

The model is trained on a broad range of audio-visual pairings (and their corresponding semantic representations) using standard machine learning techniques from the PyTorch library (Paszke et al., 2019). It is then tested on a carefully sculpted set of audio-visual input combinations, not used during training, that mimic the experimental conditions used in Arias-Trejo and Plunkett's (2010) IPL task with young toddlers. In particular, we systematically manipulate the phonological (P), visual (V), and semantic (S) similarity of the test stimuli, where each factor (P, V, and S) is either related (R) or unrelated (U) in level of similarity. We compare the model's performance with that of young toddlers.

The model is quite successful in mimicking toddler behavior. Visual and semantic similarity lead to interference effects as observed in toddlers: both visually and semantically similar items reduce preference for the target compared to unrelated items (though target preference remains above chance in all cases). Likewise, phonologically similar items, here defined as items having labels with identical onsets, suffer interference whereby phonological unrelated items are easier to identify than related ones. However, the model demonstrates an unpredicted interaction between semantic and phonological relatedness: semantic similarity overpowers phonological similarity as training progresses.

In addition to providing a formal model of observed behavior, the model provides a framework for conceptualizing the processes involved in this toddler version of the IPL task. The model instantiates an entirely bottom-up, feedforward account of the task. There are no top-down, feedback connections from the semantic outputs to the auditory and visual inputs. Nevertheless, the model exhibits strong semantic category effects which increasingly drown out the impact of visual similarity and phonological similarity as training progresses. We show how these effects can be attributed to the architectural and representational features embedded in the model, and outline how these outcomes would change if the assumptions underlying these features were changed. Nevertheless, the model paints a picture of the developing toddler, attributing increasing weight to semantic information as learning proceeds, even to the extent that visual and phonological ambiguity is suppressed. We speculate about the impact of architectural and representational modifications of the model and their implications for explaining toddler behavior in the IPL task.

[1]A notable exception being Smith, Monaghan, and Huettig (2017). However, these authors did not use real-world stimuli in their model which consequently did not exhibit any visual or semantic similarity effects which is one of the main goals of the current work.

Experimental findings

Upon hearing a word describing a visual scene, both adult and child listeners selectively attend to the item matching the word's referent (Allopenna et al., 1998; Chow et al., 2017). For example, on seeing a display with a train and a bear, listeners hearing the word *train* selectively attend to the image of the train. Previous studies with infants and young children show that the proportion of time spent looking at the named target image is influenced by the relationships between the target and distractor items (Arias-Trejo & Plunkett, 2010). Listeners spend proportionally more time looking at the target if the distractor is visually dissimilar and this effect is amplified if the two items are also from different semantic categories. This suggests that both perceptual similarity and category membership of the items lead to competition effects in word recognition and referent identification, in the context of the visual world task.

Arias-Trejo and Plunkett (2010) compared toddler (Figure 1b) and adult (Figure 1c) performance in this task and found that although adult preference for the target is diminished when the target is visually similar to the distractor, the impact of visual similarity is much greater in toddlers, particularly when target and distractor come from the same semantic category. Figure 1b shows that for toddlers, target recognition is abolished when item pairs are taken from the same semantic category and are perceptually similar, for example, for *cat* and *dog*.

The outcome of the Arias-Trejo and Plunkett (2010) study converges with results from previous investigations of early knowledge of category membership (Mandler et al., 1991; Mandler & Bauer, 1988; Mcdonough, 2002) and extends these findings to show that infants exploit global category information when identifying the referent of a basic level term. Mandler and Bauer (1988) showed that differences between members of the same category are minimized, whereas differences between members of different categories are maximized, in a categorization task with infants under 2 years of age. Arias-Trejo and Plunkett (2010) found that these differences work in a very similar way when lexical categorization is involved. In other words, when infants see two similar items from the same category (e.g., *cat* and *dog*) and are asked to look at one of them, disambiguation is

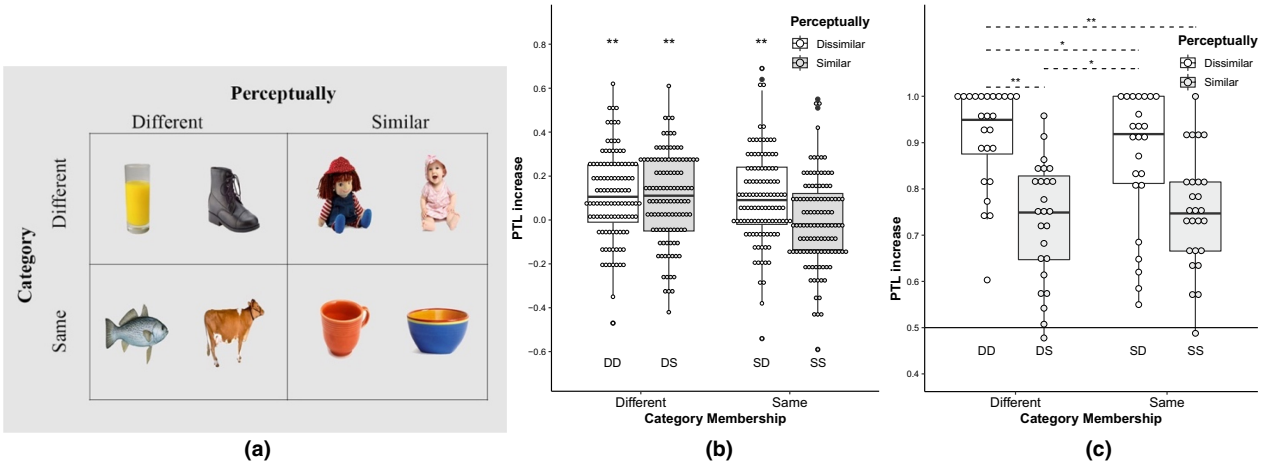

FIGURE 1 (a) Examples of object pairings in the Arias-Trejo and Plunkett (2010) study (b) Mean increase in the proportion of target looking for 18–24 month olds (replotted using the original dataset). Item pairs can be from the same semantic or different semantic categories. Furthermore, each item pair is selected to be perceptually similar or dissimilar to yield a 2×2 design. (c) Mean proportion of target looking for adults using the same 2×2 design as the toddler study (replotted using the original dataset).

difficult because their commonalities, at both the perceptual and conceptual levels, are maximized. However, when infants are presented with two items from different categories (e.g., *ball* and *apple*), their similarities are minimized. Similar results have been reported by Bergelson and Aslin (2017) with much younger 6-month olds: the infants found it easier to identify the named target referent when the competitor images were semantically unrelated than when they semantically related. The nature of the mechanism(s) underpinning these effects on categorization and referent identification in infants and young toddlers is not well-understand. One of the goals of the current work is to demonstrate how these effects could emerge mechanistically in a system trained to construct semantic representations from visual and phonological stimuli.

In a recent study, Angulo-Chavira et al. (2023) presented 18–24-month olds with pairs of objects taken from the same or different semantic categories. Furthermore, the labels of the objects in a pair could be related or unrelated. Related labels share the same phonological onset, whereas unrelated labels had no overlap (see Figure 2).

As with the Arias-Trejo and Plunkett (2010) study, the toddlers readily identified the target referent when the objects belonged to different semantic categories, irrespective of phonological overlap in the labels of the two objects. However, when the objects belonged to the same semantic category, phonological overlap in the objects' labels had a profound effect on target recognition: Phonological overlap abolished target recognition with object pairs taken from the same semantic category. This finding highlights the dominance of phonological information over semantic information in identifying the referent of a target word in the earlier stages of lexical development.

The general pattern of results emerging from these studies suggests that target recognition is easily disrupted in infants and young toddlers when additional sources of similarity between target and distractor (either visual or auditory) are introduced. Such disruption occurs despite the fact that the same participants can readily identify the same target object when the viewing conditions are more distinctive, visually or auditorily. In other words, *context* can have a dramatic impact on performance in such tasks. In contrast, adults are much less prone to the deleterious effects of similarity (see right-hand panel of Figure 1). We suggest that overcoming this *tyranny of similarity* in obstructing target referent identification in IPL tasks is achieved by a process of exposure and learning. To demonstrate the plausibility of the claim that learning itself can lead to the decontextualization of responding in the IPL task, we present a neural network model of the IPL task in which we evaluate network performance at different points in learning. Performance is evaluated on a training set to demonstrate mastery of the task and carefully constructed test sets of stimuli to demonstrate generalization of performance to previously unseen combinations of objects.

METHODS

Lexicon

The vocabulary to which the model is exposed consists of 200 imageable nouns from the toddler lexicon, as documented by the Oxford Communicative Development Inventory (OCDI) data (Hamilton et al., 2000). The OCDI consists of a checklist of 416 words that are commonly known (understood or used) by infants and toddlers.

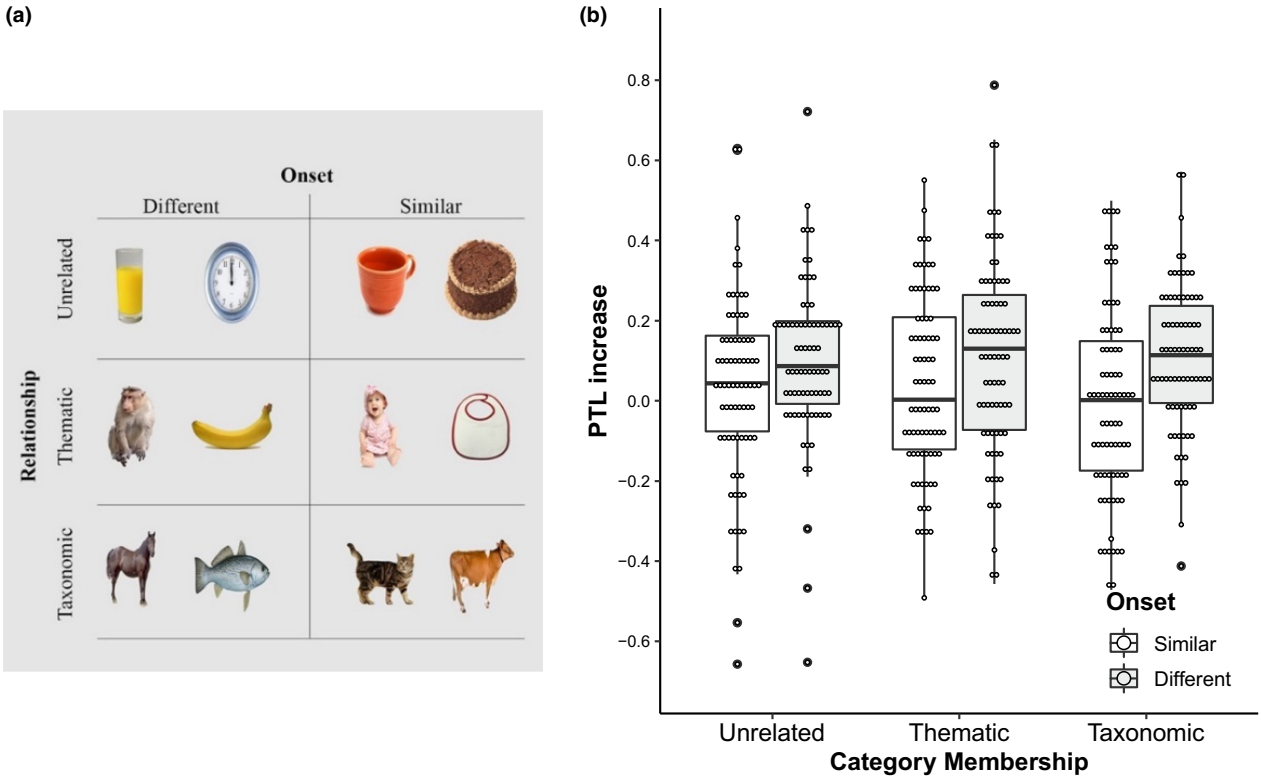

FIGURE 2 (a) Examples of object pairs used in Angulo-Chavira et al. (2023) study. Item pairs can be from the same semantic (thematic or taxonomic) or different semantic categories. Furthermore, each item pair is selected to have the same or different phonological onset in 3×2×2 design. (b) Mean increase in proportion of target looking for toddlers from the pre- to post-naming phases of a trial.

Parents are asked to indicate whether their child understands or says each word. Hamilton et al. (2000) provide norming data for these words, indicating which words are likely to be known at different ages from 10- to 30 months of age. These norms were used by Arias-Trejo and Plunkett (2010) and Angulo-Chavira et al. (2023) to select stimuli for their experiments based on the imageability of the words and the likelihood that the participants know the words. Likewise, in this modeling study, we chose 200 imageable words from the OCDI that were likely to be known by at least 50% of 24-month olds, as a representative vocabulary for training our model. The words were taken from a range of 12 distinct semantic categories, with a majority (64%) belonging to the categories of animals, food/drink, or household objects. Distribution plots for category membership, label length and onset phone identity across the entire vocabulary are provided in the Appendix (Figures A1 and A2). Items were chosen to facilitate the creation of realistic training scenarios whereby items could be combined in such a way that category membership and visual similarity could be manipulated independently of each other.

Each vocabulary item was assigned a dynamic phonological and visual representation and a static semantic representation. The dynamic phonological representation of an item was an unfolding sequence of phonemes corresponding to the label. The dynamic visual representation of an item consisted of an unfolding sequence of visual features that gradually revealed all the visual features of the image. The dynamic nature of these representations is intended to mirror the toddler's real-world experience: the toddler hears a word unfold over time and fixates different locations of the visual scene over time. In contrast, the static semantic representation is intended capture the internal representation of the item's meaning, including its relationship to the meanings of other items.

The model is trained to take these phonological and visual representations as inputs and output the semantic representation of the target item. The target output corresponds to the semantic representation of the visual item which is labeled (by the phonological representation) on a given trial. The model must learn to activate the target in its correct output location. The model also learns to suppress the semantic representation of the unlabeled visual item (the distractor) in its correct output location.

Representations

Dynamic phonological representations

Dynamic phonological representations are constructed from encodings of the phones making up the item's label. Each phone is assigned a distributed binary

encoding based on 20 articulatory and phonological features (Karaminis, 2018). The dynamic phonological representation for each vocabulary item is a matrix composed of the phonological feature representations of its phones in the order in which they appear as the spoken word unfolds, in which each row corresponds to a time step in the unfolding word. To account for phone co-articulation, the transition between two consecutive phone representations is achieved via two intermediate rows between the phonological representations of the two phones, so that the transition between the feature values 1 and 0 consists of two intermediate values of 0.95 and 0.05, and vice versa. A segmentation character for which all 20 phonological features are set to 1 was introduced to mark the offset of all labels. Ten phone slots are assigned to accommodate the longest vocabulary item including the segmentation character. Therefore, each dynamic phonological representation is a 20×32 feature matrix (10 rows for the phone representations and 22 rows for intermediate co-articulations steps between consecutive phones including ramping up to the first phone and ramping down from the segmentation character). For short labels the inputs are padded with 0.5 values after the segmentation character so that all labels are represented by a 20×32 feature matrix. These dynamic phonological representations are intended to capture the inherent time-dependent nature of the auditory input, in a fashion similar to that adopted by the TRACE model of speech perception (McClelland & Elman, 1986).

Dynamic visual representations

The visual representation for each vocabulary item is derived from the response to an illustration of the item of a *resnet18* deep neural network pre-trained on ImageNet, using the 512-dimensional activation vector for the *avgpool* layer (Deng et al., 2009; He et al., 2016; Paszke et al., 2019). The raw visual vectors are pre-processed to replace outliers (vector values with a z-score >2) with the median value for the corresponding dimension. They are further processed using principal component analysis to reduce their dimensionality to 150 (cumulative variance explained: 95%), then digitized using two bins (one below and one above the median value).

The 150-dimensional visual vectors are split into five 30-dimensional chunks stacked into a dynamic visual representation matrix in which each 30-dimensional chunk lasts for 4 timesteps. The dynamic phonological and visual matrices both have 32 rows corresponding to the 32 timesteps needed to unfold the longest word in the lexicon. After the 20*th* timestep at which the visual representation has completely unfolded, the remainder of the dynamic visual representation is padded with 0.5 values to accommodate the remaining timesteps for the unfolding of the phonological representations. These dynamic visual representations are intended to capture the inherent time-dependent nature of the processing of the visual input.

Static semantic representations

The semantic representations are 100-dimensional word vectors taken from the GloVe model, pre-trained on aggregated global word-word co-occurrence statistics from a 6 billion token corpus composed of the Gigaword5 and Wikipedia 2014 dump (Pennington et al., 2014).

The raw semantic vectors are pre-processed to replace outliers (vector values with a z score > 2) with the median value for the corresponding dimension and then digitized using two bins (one below and one above the median value).

Item similarity

The between-item similarity of the semantic and visual representations is evaluated with the Jaccard index. The Jaccard index between two representations is given by the number of features concomitantly turned on (value of 1) in the two vectors as a percentage of all the features that have the value of 1—the intersection of the active representation values as a percentage of their union. The Jaccard index provides a percentage measure of similarity between two representation vectors: the larger the index, the higher the similarity between the vectors' Jaccard indices for pairs of semantic representations are narrowly and normally distributed in the range .12:.69 with a mean of .35 and a SD of .06. The Jaccard indices for pairs of visual representations are similarly distributed (range = .16:.50, mean = .33, SD = .03). The Jaccard indices for pairs of phonological representations are more broadly distributed with a positive skew (range = 0:.93, mean = .27, SD = .11), reflecting the variable length of words used in the training environment. Semantic and visual vectors are defined to have the same lengths, that is, 100 and 150, respectively. See the Appendix for a graphic of the distribution of Jaccard indices. Two items are considered to be semantically or visually related (SR or VR) if the Jaccard index of their corresponding representations is in the top 15*th* percentile; they are semantically or visually unrelated (SU or VU) if their Jaccard index is in the bottom 15*th* percentile.

These definitions of item similarity are important because they parallel the manner in which Arias-Trejo and Plunkett (2010) designed their IPL experiment: item pairs could be taken from the same or different semantic categories and could be visually similar or distinct yielding a classic 2×2 experimental design (see Figures 1 and 2). Judgments about the similarity of experimental materials were obtained from adult raters by Arias-Trejo and Plunkett (2010). In the current

model, we define similarity mathematically in terms of the Jaccard index between the 2 vectors representing two items, semantic or visual, and then calculate the distribution of Jaccard indices to determine which items are most closely related semantically and least closely related semantically, and then choose, within these semantically related and unrelated groups, the items that are most visually similar and least visually similar. This yields a modeling analogue of the experimental 2×2 design adopted by Arias-Trejo and Plunkett (2010). In the model, we also had the freedom to choose items which were phonologically related or unrelated for each of the cells of this 2×2 design. Items are considered to be phonologically related (PR) if they share the onset phone, and phonologically unrelated (PU) otherwise. This manipulation mimics the design of the experiment conducted by Angulo-Chavira et al. (2023).

Model architecture

The model is designed to associate the unfolding of the dynamic phonological and visual representation of an item with its static semantic representation when the visual representation of the target item is accompanied by the visual representation of an unnamed distractor. The real world equivalent is the identification of a target item in a two-item visual display (2AFC). The functional architecture of the model is displayed in Figure 3.

The core of the architecture consists of a layer of Gated Recurrent Units (GRU) (Cho et al., 2014) whose inputs are:

- A 20-dimensional vector with the encoding of the current phone of the target label
- A 60-dimensional vector in which the first and the second half of the vector contain the current unfolding of the visual features of the item on the left and right of the visual display, respectively.

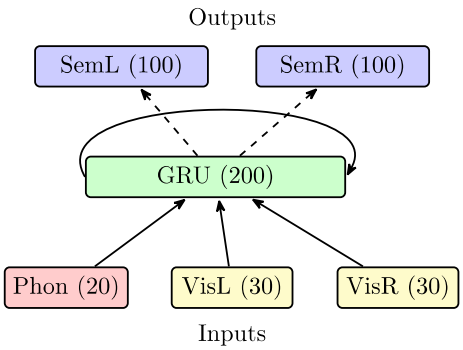

FIGURE 3 The functional architecture is a feedforward recurrent network starting with a layer of phonological and visual (left and right) inputs (Phon, VisL, and VisR), respectively. These inputs feed directly into a Gated Recurrent Layer (GRU) which generates the semantic representations (SemL and SemR) associated with the current input stimuli. The phonological and visual input stimuli unfold over time, providing dynamic input to the GRU layer at each time step. See text for further explanation. The number of units in each layer is indicated in parentheses.

The model output consists of a 200-dimensional semantic representation vector, the first and second half of which are associated with the left and right visual items, respectively (see Figure 3).

A GRU is a recurrent neural network particularly well-suited for processing sequential information, such as the unfolding over time of the phonological and visual representations of an item. A more detailed explanation of the architecture of a GRU is given in the Appendix.

Model training

A training set with all possible item pairs (each item as a target on the left and right for each pair) was generated. A subset of the target and distractor pairs were set aside for the test (6940 trials), while the rest of the pairs were used for training (72,052 trials). Test trials were designed to test the generality of the model, that is, the model was never tested with an item pair on which it had previously been trained. However, model training ensured that all items were viewed as targets by the model in both locations. We used a form of supervised learning—back propagation through time—to train the model to output the correct static semantic representations for the two unfolding inputs. The model was trained for 500,000 epochs using batch update and stochastic gradient descent.[2] Weights were dumped every 50,000 epochs, providing a snapshot of the state of the network during the course of training. Performance on the training set was evaluated at all snapshot points to provide an estimation of how well the network had learnt the IPL task for the words and picture pairs to which it had been exposed. Performance on the test set was also evaluated at all snapshot points to estimate how well the network could generalize what it had learnt to item pairs that it had never seen before.

The processing cycle for an individual visual display consists of the number of timesteps required to fully unfold the phones in the target item's label (including the intermediate steps accounting for phone co-articulation and the segmentation character) and the visual features of the display items.

In a training trial, the desired model outputs are set for the entire duration of the trial as follows: (a) the

[2]One epoch of training consists of the single presentation of all unique training pairs. Batch update refers to the updating of connection weights after a single epoch of training. It is useful to contrast batch update with pattern update when connection weights are updated after the presentation of every training pair. Batch update is more efficient that pattern update and generally leads to similar learning outcomes.

desired model outputs on the side of the target image are set to the semantic representation of the target item, while (b) the desired semantic representation of the side of the distractor image are set to 0.5, corresponding to distractor suppression. Setting the desired values of the semantic features on the suppressed (distractor) side to 0.5 ensures that the desired output on the distractor side is equidistant from all potential semantic representations.

In order for the model to activate the correct semantic representation, it must be successful in two respects:

1. It must select the visual input vector that matches the target word. In other words, it must learn the association between the label and its image. The matrix of connections feeding into the entire semantic output vector is responsible for encoding this set of associations for all 200 label-object pairs.
2. It must select the correct location of the semantic representation. In other words, it must activate the semantic representation of the target-image association only for the appropriate output location. Since targets can appear on the left or the right, the network must encode all label-item associations for both semantic output slots, and use the location of the target visual input vector to suppress the inappropriate semantic output location.

The model achieves this using the error signal generated at the output (on the semantic representations) and back-propagating this error through time to adjust the weights in the network until, over repeated training trials, the error is reduced to a very small value. Training was performed on the entire training set using batch update and stochastic gradient descent (learning rate: 0.4, momentum: 0.4 and Nesterov momentum enabled, Sutskever et al. (2013)). A training epoch consisted of the presentation to the model of all the trials in the training set.

Output analysis

To assess the model's likelihood of fixating the named target, we evaluated the level of activation of the target and distractor semantic vectors in the *target output* position. The activations of the target and distractor semantic representations are calculated using the proximity of the output vector in the target slot to the target and distractor semantic representations as defined in the training set. To calculate the activation of the semantic representation of a visual item a^I (where the item I can either be the target T or distractor D), first the Euclidean distance d^I from model output to the target or distractor semantic representation was calculated. The item activation was then calculated as one minus the ratio of d^I from the maximum possible Euclidean distance between two semantic representations $d^{\max}$ (which for the 100-dimensional semantic space is equal to $\sqrt{100} = 10$) is given by

$$a^I = 1 - \frac{d^I}{d^{\max}} \quad (1)$$

At every timestep t of the model's processing of the phonological and visual input, the activations of target and distractor items on the target output side were calculated ($a^T(t)$, $a^D(t)$). The semantic activation levels of the displayed items are then transformed into the probability of fixation following Luce (1959), an approach adopted in other models of looking behavior (e.g., Allopenna et al., 1998; Mayor & Plunkett, 2014). We assume that total looking time is split entirely between the target and distractor objects, enabling us to convert the activation strengths into the probability of fixation using the Luce choice rule. The probability of looking to the target $\text{prob}^T(t)$ at time t is given by

$$\text{prob}^T(t) = \frac{e^{2\times a^T(t)}}{e^{2\times a^T(t)} + e^{2\times a^D(t)}} \quad (2)$$

We also calculate how successfully the network has learnt to suppress semantic activations in the distractor slot. Recall that the teacher signal in the distractor output slot was set to a neutral value of all 0.5 s, a value that is equidistant from all valid semantic representations. Recall also that the location of the suppressed slot depends on the location of the named target. Consequently, it is appropriate to ask how well during training the network is able reduce the distance of the vector in the suppressed output slot to the neutral value. We, therefore, report on the normalized distance of the output vector in the suppressed slot to the neutral value (all 0.5 s) as

$$d^I_{\text{norm}} = \frac{d^I}{d^{\max}} \quad (3)$$

UNDERLYING ASSUMPTIONS

The Methods section provides a fairly detailed account of the model's architecture and training environment, as well as how model performance is evaluated. The definition of the architecture imposes constraints on the type of computations that the model can use to analyze the training set, and the latter further constrains the type of structures that the model can discover. Essentially, the model is a statistical inference machine: It needs the right kind of machinery to discover the underlying regularities in the environment to which it is exposed. If there are no systematic regularities, no amount of mechanical sophistry can discover them. Likewise, regularities in the training environment will pass unnoticed if the computational power embodied in the model architecture is inadequate to discover them. It is therefore important to lay bare the critical features of the architecture and the

training data that underpin model performance. After all, these features characterize the theory underlying this model's explanation of performance in the IPL task.

Assumptions underlying the model's architecture

Each of the representations (semantic, visual, and phonological) associated with an object are distributed: every item activates multiple units in each of the stimulus vectors on every training trial. In order to learn about items in the training set and react to items in the test set, the network must process the inputs and their constituent units in parallel. This architectural assumption is referred to as Parallel Distributed Processing and is often used in connectionist modeling. The learning regime employed in the model involves supervised learning. The target or teaching signal in any training trial is the semantic representation. The network must learn to produce the correct configuration of semantic outputs for any given configuration of phonological and visual inputs. This is achieved using a learning algorithm called back propagation through time whereby errors in the output from the network on any given trial are used to change the connections in the network so that a repetition of the same input configuration will tend to reduce the error at the output. Repeated presentations of training trials eventually reduce the output error to a very small value. The generality of the solution discovered by the network during training is evaluated using a test set of object pairings that the network has never seen before.

Recurrent architecture

The heart of the network architecture is a layer of gated recurrent units (GRUs). A GRU is a recurrent layer where units are connected to each other as well as themselves. This means that the activity of a unit depends upon its activity at previous time steps and the activity of the other units in the same layer at previous time steps. This connectivity is particularly well suited to predicting the next stimulus in a sequence of stimuli. In the current model, the input stimuli consist of an ordered sequence of phonological representations and visual representations. The GRU is particularly good at solving prediction problems as compared to early recurrent networks such as simple recurrent networks (SRNs, Elman, 1993) which required manipulation of the training set or memory parameters to achieve learning of long distance dependencies in any sequence. The GRU is also computationally more efficient than recent antecedents such as long short-term memory (LSTM) architectures.

We assume separate input pathways for the auditory and visual stimuli. The auditory input accepts a sequence of phonological representations at the same time as the visual input processes a sequence of visual representations. The visual input itself is split into 2 pathways, each pathway corresponding to the left and right locations of the visual representations of the 2 distinct objects used in a given trial. The visual representations of each object unfold over time in a fashion intended to simulate a sequence of fixations of the visual image. Unlike the real world, these fixations are chosen at random and both images are fixated in parallel. This artificial sequencing of visual fixations simplifies model training. We have yet to determine the impact of a more lifelike fixation regime on model performance.

Activity from the auditory and visual inputs feeds to the GRUs in parallel over a period of 32 time steps. Thirty-two time steps are needed to accommodate the longest words in the training vocabulary, including the ramping up and down of phonological feature activities involved in the transition (coarticulation) between phonological segments within a word. An end-of-word signal (all 1s) provides a segmentation cue to the network that the word has ended. This is particularly important for identifying words which have lexical embeddings, such as (bee in beat). For shorter words, all phonological inputs are thereafter set to 0.5 which provides a neutral stimulus regarding phonological identity. As can be seen in the Appendix, the majority of words were 3–5 phonemes in length so that the unfolding of the visual stimuli continued beyond the end of the word in most trials. The unfolding of the visual stimuli ceased in all cases after 20 timesteps, beyond which the visual inputs were padded with neutral inputs (0.5s) so that there were no biases toward one object over another.

The relative timing of the auditory and visual stimuli is a parameter that can be manipulated in the model. In the current simulations (training and test), we have chosen to set the onset of both modalities to be synchronized so as not to bias attention to one source of information over another. This means that for shorter words, the informative part of the auditory stimulus will complete before the offset of the visual stimulus. For longer words, the opposite will hold. In real-life experiments, synchronous as well as asynchronous stimulus onsets have been used to show that the relative timing of auditory and visual stimuli have an impact on gaze patterns in the visual world task. Further work will be needed to determine the impact of asynchronous onsets on model performance and how this relates to human performance. For the moment, it is sufficient to point out that neither type of asynchrony (visual before auditory or vice versa) is typically used as a cue to referent identity or location in experimental research, or in the current modeling work.

Assumptions underlying the training environment

The visual and semantic representations used in this modeling endeavor are derived from previous statistical

analyses of real-world stimuli. The visual representations are taken from a deep learning model of image identification and classification, that include thousands of pictures of everyday objects. The algorithm used in the deep learning model that generated the visual representations guarantees that images that have a similar appearance will have similar visual representations. Semantic representations are taken from co-occurrence statistics of words in a very large corpus of utterances. It is assumed that words with similar meanings have similar patterns of co-occurrence. Note that the referents of words with similar meanings need not be similar visually. Hence, there is an assumption of independence (or orthogonality) of semantic representations from visual representations inherent in the training set used by the model. At the same time, there are many objects that appear similar that also belong to the same semantic category. The complex (and often opaque) relationship between semantic and visual representations is thereby honored by the training environment used in the model. The phonological representations use a standard set of linguistic features to define a phonological segment, with a few tweaks to accommodate co-articulation effects. An important characteristic of the model is that these segments are fed to the network one segment at a time, mimicking the unfolding of speech in time. In the visual world task, speech also unfolds in time and the time course of speech recognition is known to have an impact on the pattern of fixations that a listener directs to the objects in a visual display from 1 s to the next.

Our model attempts to capture the patterns of fixations when a dynamic label (an unfolding phonological representation) is presented contemporaneously with the visual representations of two objects which occupy left and right positions in the visual input to the network. The network must learn to activate the correct semantic representation in the appropriate output location and suppress the semantic representation of the unnamed distractor image. To achieve this output, the network must be able to activate the target semantics in either location. If the target only ever occurred on the left, then learning would be relatively trivial: simply activate the semantic representation in the left output location corresponding to the left phonological signal or left visual input signal (assuming that one of the objects is named in every trial), and suppress activity in the right output location. In our model (and virtually all visual world experiments), the location of the named target varies from one trial to the next.

RESULTS

Performance on the training set

We begin by assessing whether the network can perform the task successfully, that is whether at the end of the training process it activates the relevant semantic vectors in the training set at their appropriate output locations. The probability of fixating the named target at each time step as the target label and visual inputs unfold, is calculated using the Luce choice rule (see Equation 2). We report on the two parts of the training set, phonologically related pairs (PR) and phonologically unrelated pairs (PU), and obtain an average for each of these constituent parts of the training set. Figure 5 depicts the probability of fixating the target at each time step in a trial after 500,000 epochs of training. Note that chance performance corresponds to 0.5.

Figure 4 shows a rapidly increasing likelihood of fixating the named target, well above chance, as the input stimuli unfold for both constituents of the training set (PR and PU). However, the target items paired with phonologically unrelated distractors produce a more robust target preference than those paired with the phonologically related distractors. By the 20*th* timestep of a trial, the probability of fixations has reached an asymptotic level of 0.7 for PU pairs and 0.65 for PR pairs, respectively. These results show that the network has succeeded in learning the task of fixating the named target by the end of training (500,000 epochs) but that it is more robust in achieving this for phonologically unrelated pairs than for phonologically related ones.

Figure 5 (left panel) provides a longitudinal picture of network performance at different stages in the training process. During training, the weights are dumped every 50,000 epochs providing a snapshot of the state of network which can then be used to evaluate performance as training proceeds. At each snapshot of the training process, we plot the asymptotic level of fixation probabilities achieved after 20 timesteps of the unfolding of the input stimuli (c.f. Figure 4). Again, the training set is divided into phonologically related pairs (PR) and phonologically unrelated pairs (PU). The probability of fixating the named target increases monotonically throughout training, until reaching the same asymptotic levels reported in Figure 4. By

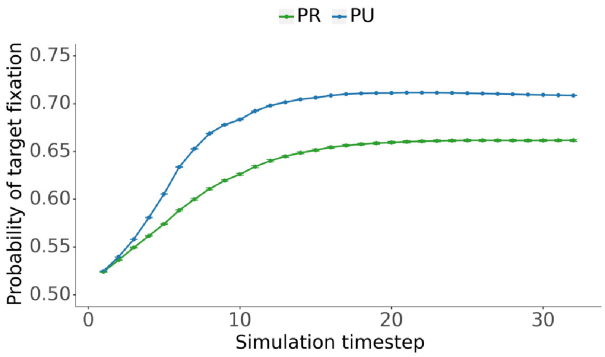

FIGURE 4 Probability of fixating the named item after 500,000 epochs of training as the target label and image unfold. Grand averages over the training set. Each combination of phonological relatedness is plotted separately. PR indicates performance on phonologically related items and PU indicates performance on phonologically unrelated items. Bars: 95% confidence intervals.

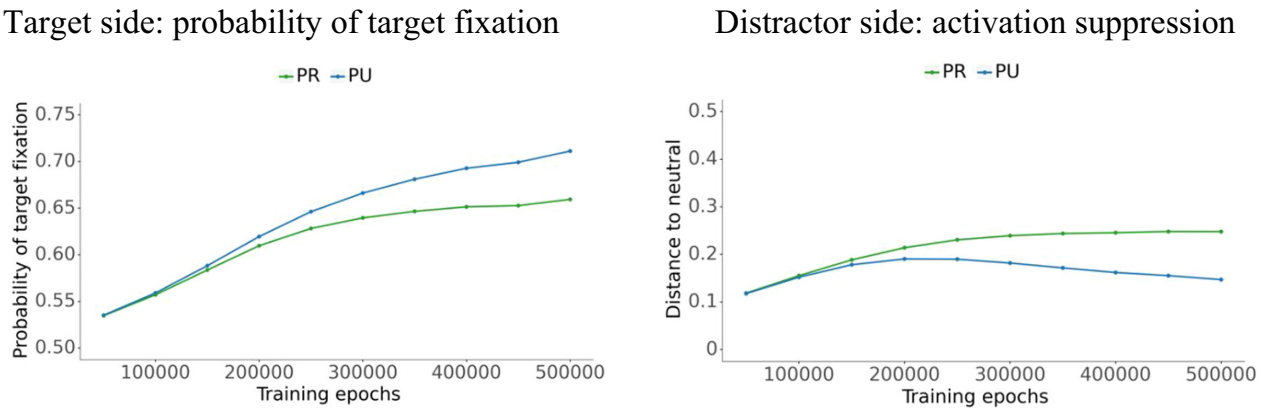

FIGURE 5 Probability of fixating the named target (left) and distractor side activation suppression (right) as training proceeds, at regular intervals of 50,000 epochs of training. Grand averages over the training set. Each combination of phonological relatedness is plotted separately. Bars: 95% confidence intervals.

150,000 epochs of training, target fixation probability for phonologically unrelated pairs (PU) begins to diverge and become stronger than that for phonologically related pairs (PR). Thereafter, the difference between PU and PR increases, indicating that the network produces higher levels of confidence in selecting the named target for phonologically unrelated than related pairs throughout most of the training process.

Figure 5 (right panel) plots the normalized Euclidean distance (d^I_{norm}, see Equation 3) between the output vector at the suppressed location and the neutral vector (all 0.5 s) on that side at different points during training. Again, phonologically related and unrelated pairs are plotted separately. At the start of training, the average distance to the neutral vector increases but then levels off for the phonologically related pairs and decreases again for phonologically unrelated pairs. This pattern of results shows that the network is successful in suppressing semantic activity in the distractor slot but finds this suppression more difficult for phonologically related pairs. The d^I_{norm} distances are lowest at the beginning of training because the network starts in a random state. Neutral outputs (our definition of suppression) are therefore the natural state at the output at the beginning of training. However, as each output slot is equally likely to be a target location as be suppressed, the network must learn to suppress the appropriate location for each trial. This is particularly difficult (though achieved) for the phonological related pairs.

Overall, we can conclude that the network has learnt to perform the task it has been assigned on the training set: it fixates the named target and suppresses the unnamed location.

Performance on the test set

Given the powerful learning algorithms and architecture employed in this neural network, it is perhaps unsurprising that the model succeeds in mastering the complex arbitrary mappings defined by the input–output training stimuli. An important test of *any* model of learning is how well it is able to generalize to stimuli that it has not been trained on. In the current context, this implies asking how well the model performs on stimulus pairs that it has not seen before. Note that this is *not* a test of word learning: the analysis of the performance on the training set has shown that the network has succeeded already in learning the label-item associations and using these associations to identify the location of a named item. The question we ask now is whether the network can identify the semantic representation of a named item in an *item pair* it has never seen before, and activate that representation in the appropriate location. We evaluate the generalization characteristics of the model by comparing its performance on two test sets involving previously unseen item pairs. These test sets are intended to parallel experimental sets used in real IPL studies with infants and toddlers that were reviewed in the Introduction section. Infant performance in these studies can also be considered evaluations of generalization ability as it is unlikely that the infants would have previously seen the particular combinations of objects used in the experiments.

The effect of semantic and visual relatedness

Recall that the picture pairs in the Arias-Trejo and Plunkett (2010) study were semantically related (SR) or unrelated (SU), and visually related (VR) or unrelated (VU), as depicted in Figure 2c. The same division was created for the picture pairs used as the input stimuli in a test set for the model (as defined in the section on Item similarity), to yield a 2 × 2 design analogous to that used by Arias-Trejo and Plunkett (2010).

Figure 6 shows model performance as the target word and visual images unfold on these previously unseen

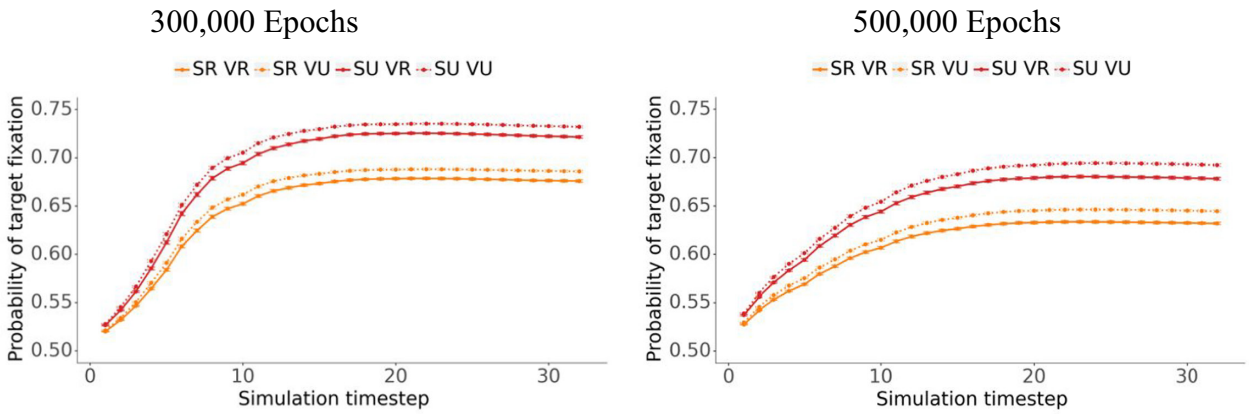

FIGURE 6 Probability of fixating the named target as the target label and image unfold: after 300,000 epochs of training (left) and after 500,000 epochs (right). Grand averages over the test set. Each combination of visually relatedness and semantic relatedness is plotted separately. Bars: 95% confidence intervals.

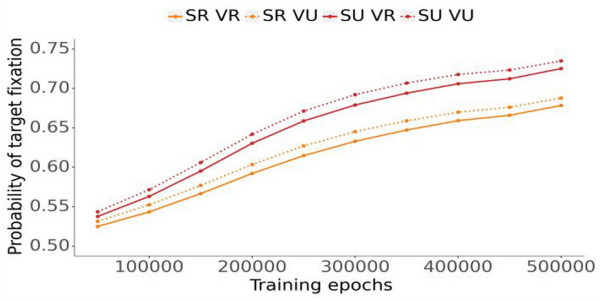

FIGURE 7 Probability of fixating the named target as training proceeds, at regular intervals of 50,000 epochs. Each combination of visual relatedness and semantic relatedness is plotted separately. Bars: 95% confidence intervals.

combinations of items at 2 snapshots in the training process, after 300,000 epochs and after 500,000 epochs. Each condition (SU VU, SR VU, SU VR, SR VR) is plotted separately for these 4 different types of combinations of targets and distractors. The probability of fixating the named target increases monotonically as the target label and image unfold, reaching asymptotic levels of probability within 20 time steps. The probability of fixating the target item is well above chance shortly after target label onset. Performance is most robust for stimuli pairs that are visually and semantically unrelated, and least robust for pairs that are visually and semantically related. Picture pairs that are visually related and semantically unrelated, or visually unrelated and semantically related, result in intermediate levels of target fixation probability, although semantic relatedness suppresses fixation probability more than visual relatedness, throughout the unfolding of the target label and image. The overall levels of target fixation probability increase from 300,00 to 500,000 training epochs. The impact of visual relatedness diminishes with greater levels of training.

In Figure 7, we plot the level of network performance at successive steps in the training process. The asymptotic levels of fixation probability on the named target, for each condition, are shown every 50,000 epochs. Figure 7 shows that performance increases monotonically in all conditions, with the relative robustness of fixating the named target consistent throughout training, that is, SU VU>SU VR>SR VU>SR VR. During the earliest stages of training the impact of visual similarity and semantic similarity are equivalent. As training proceeds, the impact of semantic similarity becomes stronger than that of visual similarity.

These findings demonstrate that semantic and visual context have a robust impact on target identification in the model, just as we see in infant, toddler, and adult performance on this task: The greater the visual and semantic similarity between the objects used in these visual world tasks, the more difficulty participants have in identifying the target referent. In the model, these effects are entirely due to competition between semantically and/or visually similar representations of the objects involved. Similar competition effects may be at work in the infant, toddler, and adult participants.

The effect of phonological and semantic relatedness

In a related experiment (see Figure 2), Angulo-Chavira et al. (2023) presented toddlers with pairs of objects taken from the same semantic category or different semantic categories but instead of systematically varying the visual similarity between the objects, they systematically varied the phonological similarity of the object labels: the object labels could be related or unrelated. Related labels shared the same phonological onset, whereas unrelated labels had no overlap (see Introduction). We created an analogous test set of previously unseen item pairs using the same definition of semantically related and unrelated pairings as given in the section on item

similarity. In addition, for both levels of semantic similarity, the labels of the item pairs were either related (PR) or unrelated (PU), again yielding a 2×2 design. No attempt was made to control for visual similarity, reflecting the design of the Angulo-Chavira et al. (2023) study. Preference for the target item at the target output location was calculated in the same manner as the previous test set for each of the four conditions: PU SU, PR SU, PU SR, PR SR.

Figure 8 shows model performance on these previously unseen pairs of items, each plotted separately for the 4 different types of combinations of targets and distractors, as the target label and images unfold. Again, the probability of fixating the named target increases monotonically as the target label and images unfold, at both levels of training depicted (300,000 epochs of training and 500,000 epochs of training). As before, the most unrelated pairs (PU SU) produce the highest probability of fixating the named target, while the most related pairs (PV SV) produce the least robust fixation probabilities. During the earlier level of training (300,000 epochs), semantic similarity has a stronger impact on the probability of target fixation than phonological similarity (PU SU > PR SU > PU SR > PR SR). However, by 500,000 epochs, the impact of phonological and semantic similarity is roughly equivalent (PU SU > PR SU = PU SR > PR SR). It is noteworthy, however, that there is a transitory period within a trial (between time steps 5 and 15) when phonological overlap is a greater impediment to target identification than is semantic similarity, even after 500,000 epochs of training.

Finally, in Figure 9 we plot the level of network performance at successive steps in the training process. As before, the asymptotic levels of fixation probability on the named target, for each condition, are shown every 50,000 epochs. Figure 9 shows that performance increases monotonically in all conditions, with the relative robustness of fixating the named target quite consistent across conditions throughout training, that is, PU SU > PR SU >= PU SR > PR SR.

In contrast to other *longitudinal* plots, there is a period of training between 250,000 and 400,000 epochs when PU SR pairs make significant gains on PR SU pairs. This indicates the phonologically unrelated pairs are able to partially overcome the impact of semantic relatedness as training proceeds, mimicking the impact of phonological overlap early in a trial, depicted in Figure 8.

Just as we saw in the previous set of simulations, stimulus similarity has a robust impact on the probability of target identification in the model, as has been observed in infants, toddlers, and adults. In this case, we observe the impact of phonological and semantic similarity: The more similar the test objects are phonologically and semantically, the greater the difficulty the model and human participants have in identifying the named target. Once again, in the model, these effects are entirely due to competition between semantically and phonologically similar representations of the objects involved. Similar competition effects may be at work in the infant, toddler and adult participants. Moreover, the general principle that similarities between objects makes referent identification difficult seems to hold both in the model and human participants, even when multiple sources of similarity from different modalities (phonological, visual, and semantic) are involved.

DISCUSSION

The primary goal in building the neural network model of the Intermodal Preferential Looking task presented here was to evaluate the adequacy of the assumptions underlying the processes, representation and architecture implemented in the model as a theory of aspects of human behavior in this task. Insofar as the model is able

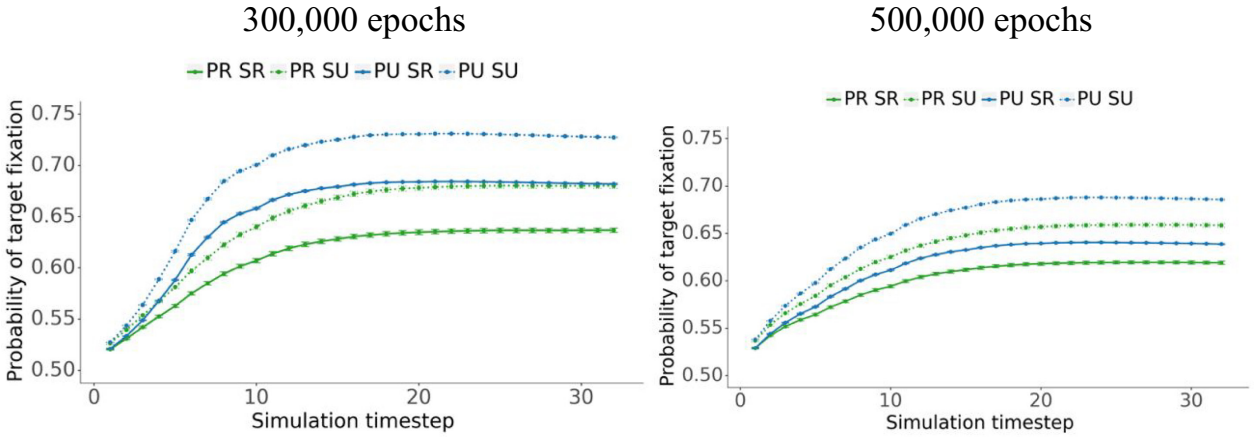

FIGURE 8 Probability of fixating the named target as the target label and image unfold: after 300,000 epochs of training (left) and after 500,000 epochs (right). Each combination of phonological similarity and semantic similarity is plotted separately. Bars: 95% confidence intervals.

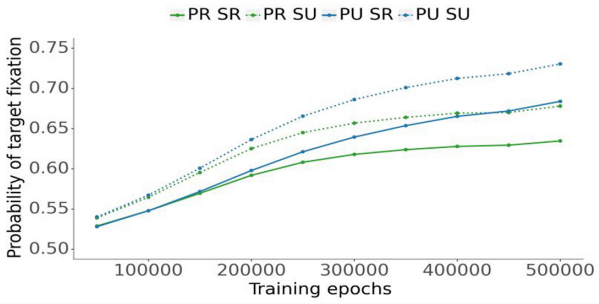

FIGURE 9 Probability of fixating the named target as training proceeds, at regular intervals of 50,000 epochs. Each combination of phonological relatedness and semantic relatedness is plotted separately. Bars: 95% confidence intervals.

to mimic successfully aspects of behavior in several experimental tasks, we may conclude that the model constitutes a plausible theory of the constituent mechanisms underlying the task. The model permits the presentation of dynamic phonological and visual inputs using a powerful recurrent architecture (GRU) that maps these inputs to static semantic representations. The training regime forces the model to learn the associations between labels and images (word leaning), and use these associations to bind the location of the named image to a specific location by activating the semantic representation of the named image in the target output location and suppressing the activation of the unnamed image in the distractor output location (identity-location binding). The learning process itself consists of a standard supervised training regime whereby the error at the output is used to adjust the weights in the network so that error at the output is gradually reduced as training proceeds. The training stimuli themselves, visual, phonological, and semantic, are all realistic representations of the words and objects that young toddlers are likely to know during the second half of the second year of life. An important aspect of using these realistic representations of the training stimuli is that they afford the possibility of making precise mathematical evaluations of their similarity to each other—an aspect that is far more difficult when assessing the perceived similarity between stimuli for human observers.

The target characteristics of toddler behavior concern the manner in which the phonological, visual, and semantic similarities between items in an IPL task impact their ability to identify a target referent. Experimental studies with toddlers have shown that similarity between items on all three dimensions impairs this ability. The simulations reported here demonstrate exactly this characteristic: whether evaluated on a trial-by-trial basis (e.g., as the word unfolds over time) or longitudinally (at different snapshots in the training process), semantically, visually and phonologically unrelated items all produced stronger target fixation probabilities than semantically, visually and phonologically related items. This finding holds true irrespective of whether the model is evaluated on the training set or on previously unseen pairings of items (the test set). Target fixation probabilities are consistently higher for SU VU pairs than SR VR pairs (see Figures 6 and 7) and PU SU pairs consistently higher than PR SR pairs (see Figures 8 and 9). This behavior of the model mimics the behavior of toddlers in an IPL task where it is explained in terms of the competition between stimuli. The model provides a formal instantiation of this competition. For example, similar semantic vectors have more overlap than dissimilar semantic vectors (see the discussion of the Jaccard index in the section on Item similarity). Likewise, similar image vectors have more overlap than dissimilar image vectors. Hence, when activity propagates through the network from similar pairs, there will be a tendency to produce similar outputs (the tyranny of similarity). The Luce Choice Rule (see Equation 2) is sensitive to these similarities so that the choice of output location is affected both by the similarity of inputs and the output representations.

These well-understood properties of neural network models make them well-suited to understanding the impact of similarity on behavior. However, they also enable us to understand more nuanced effects of similarity, particularly when different levels of similarity are combined in the same trial, such as semantically unrelated items with visually related items. Consider first the effects of semantic and visual relatedness, modeled as a simulation of the Arias-Trejo and Plunkett (2010) study. Figures 6 and 7 show that SR VU and SU VR pairs result in intermediate levels of target fixation as compared to fully related SR VR or unrelated SU VU pairs. This is unsurprising given the sensitivity of these network models to similarity relations. However, it is apparent from Figure 7 that the impact of semantic similarity increases as training proceeds. After 100,000 epochs the effects of semantic and visual similarity are additive and approximately equally impactful. However, by 300,000 epochs, semantic similarity is clearly the most disruptive factor affecting target fixation, with visual similarity disrupting target fixations to the same albeit minor extent, irrespective of semantic similarity. This pattern of responding remains in the later stages of training while overall levels of target fixation probability continue to increase across the board until 500,000 epochs. The increasing impact of semantic similarity relative to visual similarity is also apparent in the timestep (trial) analyses presented in Figure 6.

The experimental data reported by Arias-Trejo and Plunkett (2010) and depicted in Figure 1 demonstrates the clear effects of semantic and visual similarity on the probability of fixating a target referent in an IPL tasks for both adults and toddlers. Figure 1c (for adults) shows an overall suppression of target looking for semantically related pairs, with visual similarity suppressing target looking for both semantically related and unrelated pairs. For the toddlers' target looking, depicted in Figure 1b, the interaction of semantic (category)

similarity with visual similarity is quite dramatic: If the object pairs are taken from different semantic categories, then these toddlers are relatively impervious to perceptual differences. They readily identify the target object irrespective of visual similarity. In contrast, if the objects are taken from the same semantic category, visual similarity essentially abolishes target looking whereas visually dissimilar items merely modulate target looking relative to the other conditions. Mandler and Bauer (1988) suggested that this type of result can be explained by assuming that differences between members of the same category are minimized, whereas differences between members of different categories are maximized. Interestingly, our model does not behave quite like the toddlers in the Arias-Trejo and Plunkett (2010) study since it *never* entirely abolishes target looking for the named item. In fact, the model behaves more like the adults in that study, where target looking is maintained across all conditions but is increasingly attenuated by increasing levels of similarity in the object pairs.

The central demonstration of the model using pairs of items that are semantically and visually related and unrelated is the graded impact of overall similarity on the probability of target fixations. Perhaps the most important aspect of this demonstration is the increasingly dominant impact of semantic similarity as training proceeds: the main effect of semantic similarity increases whereas the main effect of visual similarity becomes proportionally smaller. In other words, the network gets better at ignoring visual differences and pays increasing attention to semantic differences. The reason for this asymmetry between semantic and visual similarity is that the network is *trained* on semantic targets rather visual targets. Hence, semantic differences are more important to the network—semantic targets provide the error signal that drives learning in the network, including learning to ignore visual differences if they are irrelevant to producing the correct output. If the network had been designed with visual targets in mind, then visual differences would remain a substantial factor in determining the probability of target fixations.

The simulations involving item pairs where semantic relatedness and phonological relatedness is systematically manipulated, revealed a convergent story regarding the impact of similarity effects. Item pairs that were unrelated (PU SU) produced higher levels of target fixations than item pairs that were related (PR SR)—see Figure 8 for trial-based similarity effects and Figure 9 for longitudinal similarity effects, respectively. Intermediate levels of similarity (PR SU and PU SR) produced intermediate levels of target fixation probabilities. During the earlier stages of training (see left-hand panel of Figure 8), the impact of semantic and phonological similarity are approximately additive with phonologically unrelated pairs producing consistently higher levels of target fixations when semantic similarity is held constant. This finding is consistent with the experimental study by Angulo-Chavira et al. (2023), reviewed in the Introduction, in which phonological overlap had a deleterious effect on target recognition. Phonological overlap even eliminated target recognition when the visual stimuli were drawn from the same semantic category. Again, in the model, the probability of fixating the target above chance is never abolished in any condition, suggesting that other factors are influencing the toddlers' behavior which are not included in the model. Nevertheless, the convergent impact of phonological similarity in the model and in the toddlers remains clear.

It is noteworthy that there is a transient effect of phonological overlap as the target word unfolds in the more highly trained network (see righthand panel of Figure 8). Between timesteps 7 and 14, phonologically related pairs (PR SU) suppress the probability of target fixation more than phonologically unrelated pairs (PU SR). This difference disappears as the target word and images unfold. This transient advantage of phonologically unrelated pairs is consistent with the experimental literature demonstrating that toddlers show faster target recognition when the picture pairs do not have labels beginning with the same sound as compared to picture pairs with onset overlap (Swingley et al., 1999). It is also consistent with experimental findings reporting that toddlers will fixate objects that share phonological properties with the referent of a target word before they fixate objects that share only semantic properties with the target word (Chow et al., 2017).

The behavior of the model described is achieved with an architecture that permits activation to flow in one direction only—from input to output. Although the model has recurrent connections within a layer of the network (the GRU), there are no connections that feedback to lower levels in the network. We consider this architecture to be a bottom-up processor, lacking in any top-down components. Consequently, all the effects observed in the model, including the increasing impact of semantic similarity on target recognition, are attributable entirely to bottom-up processes. This is *not* to claim that top-down connectivity is absent in the toddler's phonolexical-semantic system. Rather we mean to claim that much can be explained without appealing to such top-down processing.

Some limitations of the model

To be sure, the model is far from a perfect implementation of toddler behavior in an IPL task. Witness the failure of the model to abolish target recognition under certain combinations of phonological, visual, and semantic relatedness. It is possible that the inclusion of top-down connectivity in the architecture could achieve this outcome. Likewise, other modifications of the training set are likely to impact behavior. For example, most 24-month olds (and probably even 18-month olds) have a vocabulary far exceeding 200 words (Mayor & Plunkett, 2011). The composition of the training regime inevitably has a

major impact on model performance. Hence, future models should attempt to scale up to larger vocabularies and image sets to evaluate the impact of such changes. We contend, however, that the impact of phonological, semantic and visual similarity will follow the same general principles observed to be at work in the current model.

The model has no output lexicon (target output representations of the entire input label that consolidate as phonological input unfolds). It is not implausible that including an output lexicon will accentuate phonological effects in the model. For example, embedded words or cohort or rhyme competitors might well influence lexical retrieval, as in TRACE (McClelland and Elman (1986) or Shortlist (Norris, 1994). Including an output lexicon would also enable the evaluation of comprehension and production in the model (c.f. Plunkett et al., 1992), thereby enabling a broader range of comparisons with empirical data and evaluation of the adequacy of the model.

Perhaps one of the most striking limitations of the current model is its inability to escape the binding of the identify of objects to their locations during the training process: If the training environment included trials where a particular object, say a dog, only ever occurred in a given location, say the left location, then it would be unable to make reliable predictions about the output location of the named target if that object, say dog, always occurred in, say, the right location. One might speculate whether infants and toddlers also demonstrate similar limitations in identifying target objects that are restricted to a limited range of locations within an experimental session. However, visual world experiments with infants (and adults) strive to avoid these contingencies in the service of good experimental design. It is possible that exposure to multiple similar objects which varied their locations during training would enable the model to generalize to those objects with no such variability in location. The model could piggyback off their similarity relations to overcome location-binding limitations. Likewise, if an object occurred in many locations during training, it might be able to generalize to absent locations. However, in the current model, only 2 locations are available. This constraint will mitigate against any such generalization. It remains to be seen whether the similarities in identity and location distributions are sufficient to overcome the identify-binding constraints inherent in the current model. It is also possible that architectural modifications of the model are needed (Foldiak, 1991; Mareschal et al., 1999) for the model to abstract away from restricted identity-location binding in target recognition.

Finally, we have avoided investigating the impact of the relative timing of auditory and visual signals for the time course of target recognition in visual world experiments: Does the auditory signal occur before the onset of the visual signal or vice versa. There is an extensive literature with adults and toddlers demonstrating that timing matters in this regard (e.g., Apfelbaum et al., 2021; Chow et al., 2022; Huettig & McQueen, 2007). In the current model, the timing of the onset of the auditory and visual signals is synchronous. However, the use of asynchronous timings in the model is readily achievable without any architectural modifications and is a worthwhile avenue for future investigations of the adequacy of the model.

CONCLUSION

We conclude that phonological and visual representations mapped dynamically in a *bottom-up* fashion to semantic representations are *sufficient* to capture important aspects of phonological, semantic, and visual preference effects often reported in visual world tasks. The growing impact of semantics on target preference as learning proceeds, as well as the early effects of phonological overlap within a trial do not require *top-down* feedback from a semantic or visual system. The use of semantic representations as targets in the learning process is adequate to achieve these differences between the impact of semantic similarity on the one hand and visual and phonological similarity on the other.

FUNDING INFORMATION

This research was funded by a Leverhulme Trust research project grant to Kim Plunkett.

DATA AVAILABILITY STATEMENT

The code needed to replicate the simulation is available at: https://github.com/mihaeladuta/child-development-ipl-model.

ORCID

Mihaela Duta https://orcid.org/0000-0002-0435-571X

REFERENCES

Allopenna, P. D., Magnuson, J. S., & Tanenhaus, M. K. (1998). Tracking the time course of spoken word recognition using eye movements: Evidence for continuous mapping models. *Journal of Memory and Language*, *38*(4), 419–439.

Angulo-Chavira, A., Arias-Trejo, N., & Plunkett, K. (2023). The effects of phonological and semantic similarity on early word recognition. In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society*, Sydney, Australia.

Apfelbaum, K. S., Klein-Packard, J., & McMurray, B. (2021). The pictures who shall not be named: Empirical support for benefits of preview in the visual world paradigm. *Journal of Memory and Language*, *121*, 104279.

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*(1-2), 63–80.

Bergelson, E., & Aslin, R. N. (2017). Nature and origins of the lexicon in 6-month-olds. *Proceedings of the National Academy of Sciences*, *114*(49), 12916–12921. https://doi.org/10.1073/pnas.1712966114

Cho, K., Van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning phrase

representations using rnn encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

Chow, J., Angulo-Chavira, A. Q., Spangenberg, M., Hentrup, L., & Plunkett, K. (2022). Bottom-up processes dominate early word recognition in toddlers. *Cognition*, *228*, 105214.

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

Cooper, R. M. (1974). The control of eye fixation by the meaning of spoken language: A new methodology for the real-time investigation of speech perception, memory, and language processing. *Cognitive Psychology*, *6*, 84–107.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *In 2009 IEEE conference on computer vision and pattern recognition* (pp. 248–255).

Duta, M., & Plunkett, K. (2021). A neural network model of lexical-semantic competition during spoken word recognition. *Frontiers in Human Neuroscience*, *15*, 700281. https://doi.org/10.3389/fnhum.2021.700281

Elman, J. L. (1993). Learning and development in neural networks: The importance of starting small. *Cognition*, *48*, 71–99.

Fernald, A., Swingley, D., & Pinto, J. P. (2001). When half a word is enough: Infants can recognize spoken words using partial phonetic information. *Child Development*, *72*(4), 1003–1015.

Foldiak, P. (1991). Learning invariance from transformation sequences. *Neural Computation*, *3*, 194–200.

Golinkoff, R., Hirsh-Pasek, K., Cauley, K., & Gordon, L. (1987). The eyes have it: Lexical and syntactic comprehension in a new paradigm. *Journal of Child Language*, *14*, 23–46.

Hamilton, A., Plunkett, K., & Schafer, G. (2000). Infant vocabulary development assessed with a british communicative development inventory. *Journal of Child Language*, *27*(3), 689–705. https://doi.org/10.1017/S0305000900004414

He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770–778).

Huettig, F., & McQueen, J. M. (2007). The tug of war between phonological, semantic and shape information in language-mediated visual search. *J. Mem. Lang.*, *57*, 460–482. https://doi.org/10.1016/j.jml.2007.02.001

Huettig, F., Olivers, C. N. L., & Hartsuiker, R. J. (2011). Looking, language, and memory: Bridging research from the visual world and visual search paradigms. *Acta Psy- Chologica*, *137*, 138–150. https://doi.org/10.1016/j.actpsy.2010.07.013

Karaminis, T. (2018). The effects of background noise on native and non-native spoken-word recognition: A computational modelling approach. In *40th annual conference of the cognitive science society: Changing minds, CogSci 2018* (pp. 1902–1907). The Cognitive Science Society.

Luce, R. (1959). *Individual choice behavior*. Wiley.

Magnuson, J. S. (2019). Fixations in the visual world paradigm: Where, when, why? *Journal of Cultural Cognitive Science*, *3*(2), 113–139.

Mandler, J. M. (2000). Perceptual and conceptual processes in infancy. *Journal of Cognition and Development*, *1*(1), 3–36.

Mandler, J. M., & Bauer, P. J. (1988). The cradle of categorization: Is the basic level basic? *Cognitive Development*, *3*, 247–264.

Mandler, J. M., Bauer, P. J., & Mcdonough, L. (1991). Separating the sheep from the goats: Differentiating global categories. *Cognitive Psychology*, *23*, 263–298.

Mareschal, D., Plunkett, K., & Harris, P. (1999). A computational and neuropsychological account of object-oriented behaviours in infancy. *Developmental Science*, *2*(3), 306–317.

Mayor, J., & Plunkett, K. (2010). A neuro-computational account of taxonomic responding and fast mapping in early word learning. *Psychological Review*, *117*(1), 1–31.

Mayor, J., & Plunkett, K. (2011). A statistical estimate of infant and toddler vocabulary size from cdi analysis. *Developmental Science*, *14*(4), 769–785.

Mayor, J., & Plunkett, K. (2014). Infant word recognition: Insights from TRACE simulations. *Journal of Memory and Language*, *71*(1), 89–123.

McClelland, J. L., & Elman, J. L. (1986). The TRACE model of speech perception. *Cognitive Psychology*, *18*, 1–86.

Mcdonough, L. (2002). Basic-level nouns: First learned but misunderstood. *Journal of Child Language*, *29*, 357–377. https://doi.org/10.1017/S030500090200507X

McMurray, B., Horst, J. S., & Samuelson, L. K. (2012). Word learning emerges from the interaction of online referent selection and slow associative learning. *Psychological Review*, *119*(4), 831.

Meints, K., Plunkett, K., & Harris, P. L. (1999). When does and ostrich become a bird? The role of typicality in early word comprehension. *Developmental Psychology*, *35*(4), 1072–1078.

Norris, D. (1994). Shortlist: A connectionist model of continuous speech recognition. *Cognition*, *52*(3), 189–234.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., Desmaison, A., Kopf, A., Yang, E., DeVito, Z., Raison, M., Tejani, A., Chilamkurthy, S., Steiner, B., Fang, L., … Chintala, S. (2019). Pytorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché Buc, E. Fox, & R. Garnett (Eds.), *Advances in neural information processing systems 32* (pp. 8024–8035). Curran Associates, Inc.

Pennington, J., Socher, R., & Manning, C. D. (2014). Glove: Global vectors for word representation. In *Empirical methods in natural language processing (EMNLP)* (pp. 1532–1543), October 25–29, 2014, Doha, Qatar.

Plunkett, K., Sinha, C., Møller, M. F., & Strandsby, O. (1992). Symbol grounding or the emergence of symbols? Vocabulary growth in children and a connectionist net. *Connection Science*, *4*, 293–312.

Rogers, T., & McClelland, J. (2004). *Semantic cognition: A parallel distributed processing approach*. Bradford Books.

Smith, A. C., Monaghan, P., & Huettig, F. (2017). The multimodal nature of spoken word processing in the visual world: Testing the predictions of alternative models of multimodal integration. *Journal of Memory and Language*, *93*, 276–303.

Spelke, E. (1976). Infants' intermodal perception of events. *Cognitive Psychology*, *8*, 553–556.

Spelke, E. S. (1979). Perceiving bimodally specified events in infancy. *Developmental Psychology*, *15*(6), 626–636. https://doi.org/10.1037/0012-1649.15.6.626

Sučević, J., Althaus, N., & Plunkett, K. (2022). Discovering category boundaries: The role of comparison in infants' novel category learning. *Infancy*, *27*(3), 533–554.

Sutskever, I., Martens, J., Dahl, G., & Hinton, G. (2013). On the importance of initialization and momentum in deep learning. In *International conference on machine learning* (pp. 1139–1147). PMLR.

Swingley, D., & Aslin, R. N. (2000). Spoken word recognition and lexical representation in very young children. *Cognition*, *76*(2), 147–166.

Swingley, D., Pinto, J. P., & Fernald, A. (1999). Continuous processing in word recognition at 24 months. *Cognition*, *71*, 73–108.

Tanenhaus, M. K., Spivey-Knowlton, M. J., Eberhard, K. M., & Sedivy, J. C. (1995). Integration of visual and linguistic information in spoken language comprehension. *Science*, *268*(5217), 1632–1634.

Thomas, D. C., Campos, J. J., Shucard, W., Ransay, D. S., & Shucard, J. (1981). Semantic comprehension in infancy: A signal detection analysis. *Child Development*, *52*, 798–903.

How to cite this article: Duta, M., & Plunkett, K. (2023). A network model of referent identification by toddlers in a visual world task. *Child Development*, *00*, 1–20. https://doi.org/10.1111/cdev.14010

APPENDIX A

GRU functionality is achieved via a *reset gate* and an *update gate*, each a trainable vector used in conjunction with the GRU's current input and output from the previous timestep, to filter out irrelevant information and retain pertinent information (Figure A3). The role of the update gate vector is to select the information from the previous processing timestep to be kept for the processing of subsequent timesteps. The role of the *reset gate* is to determine which information from the previous timesteps is irrelevant and therefore does not need to be kept for the processing of subsequent timesteps.

To obtain the update gate vector u_t and the reset gate vector r_t at each time step t, the current input x_t and the output from the previous timestep $h_t - 1$ are each multiplied with their respective gate weights (W^u and H^u for the update gate and W^r, and H^r for the reset gate) and added together before applying a sigmoidal function σ to constrain the vector values between 0 and 1:

$$u_t = \sigma\left(W^u \times x_t + H^u \times h_t - 1\right)$$

$$r_t = \sigma\left(\mathrm{W}^{\mathrm{r}} \times \mathrm{x}_t + H^r \times h_t - 1\right)$$

The reset gate is used in conjunction with the current input and the GRU's output at the previous timestep to select the relevant information from the current time step in the intermediate memory h'_t. First, the current input x_t and the output at the previous timestep $h_t - 1$ are both weighted with the W and U weight vectors, respectively. The element-wise product between the reset vector and the weighted output from the previous timestep are added to the weighted current input before applying a *tanh* function:

$$h'_t = tanh\left(W \times x_t + r_t \odot U \times h_{t-1}\right)$$

The relevant information from the current timestep is taken as the element-wise product between one minus the update gate vector and the intermediate memory h'_t, while the relevant information from the previous timestep is selected as the element-wise product between the update gate vector and the output at the previous timestep h_t. The output of the GRU at the current timestep h_t is then the sum of the relevant information from the current timestep and the relevant information from the previous timestep:

$$h_t = u_t \odot h_{t-1} + \left(1 - u_t\right) \odot h'_t$$

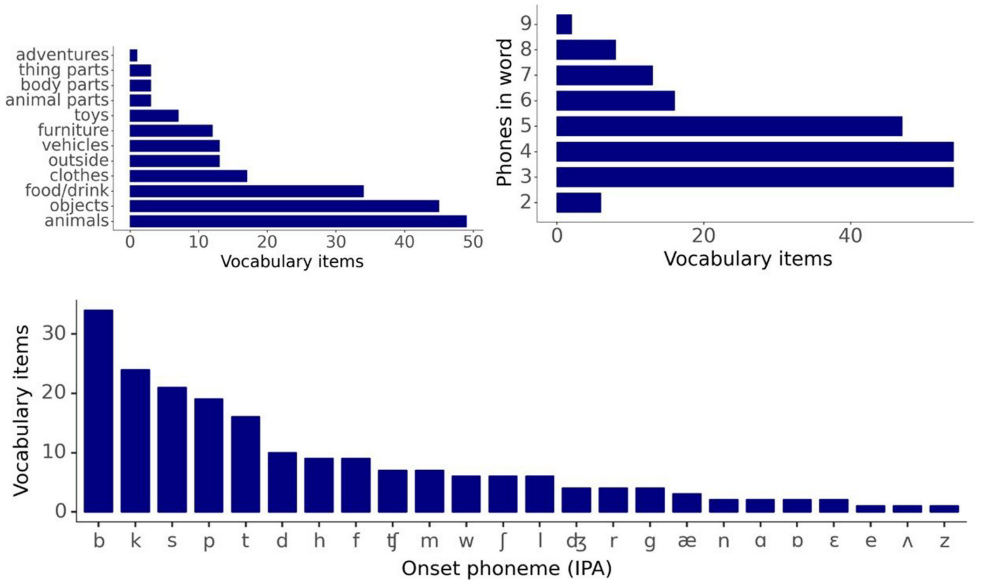

FIGURE A1 Descriptive statistics for vocabulary items: word category membership, word length distribution, and cohort size distribution.

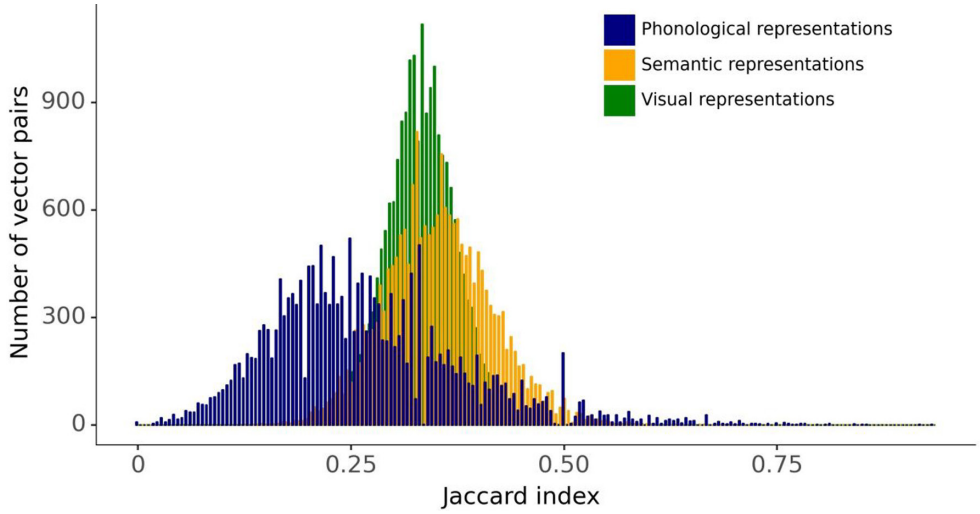

FIGURE A2 Distributions for Jaccard index between pairs of phonological, semantic, and visual vectors.

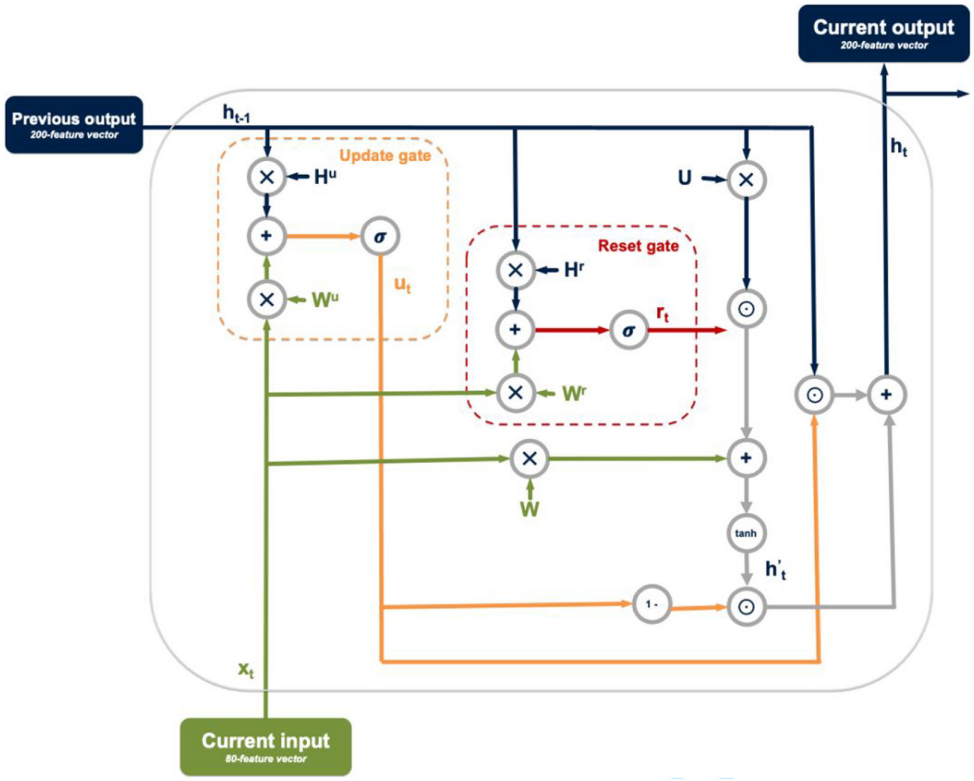

FIGURE A3 Detailed illustration of the GRU functionality.