

Continuous Development in Learning Similar-sounding Words

Abdellah Fourtassi

afourtas@stanford.edu

Department of Psychology
Stanford University

Sophie Regan

sregan20@stanford.edu

Department of Psychology
Stanford University

Michael C. Frank

mcf Frank@stanford.edu

Department of Psychology
Stanford University

Abstract

Stager and Werker (1997) first showed that children's early ability to distinguish similar sounds do not automatically translate into word learning skills. Early explanations proposed that children may not be able to encode subtle phonetic contrasts when learning novel word meanings, thus suggesting a discontinuous/stage-like pattern of development. However, later work has revealed (e.g., through using simpler testing methods) that children do encode such contrasts, thus favoring a rather a continuous pattern of development. The present study proposes a precise probabilistic model describing how development may proceed in a continuous fashion across the lifespan. The purpose of the model is to account for previously documented facts while providing new predictions. We collected data from both preschool children and adults, and we showed that the model can explain various patterns of learning both within the same age and across development. Our work highlights the role of computational modeling in advancing our understanding of development both through organizing existing knowledge and through generating new principled hypotheses.

Keywords: word learning, cognitive development, computational modeling

Introduction

Over the first year of life, children become sensitive to the phonetic variations that are used to distinguish meanings in their native language (Werker & Tees, 1984). One could imagine that these perceptual skills would be automatically applied to the task of word learning. However, developmental data show that 14 m.o children find it challenging to associate minimally different (but perceptually discriminated) sounds such as “bin” and “din” to different objects (Stager & Werker, 1997).

Several factors can explain this finding. For example it is possible that the task of meaning learning increases cognitive demands on children (compared to a simple perceptual discrimination). In particular, it requires paying attention to both the sounds and the corresponding objects, which may hinder precise encoding in memory of some phonetic details (Hofer & Levy, 2017; Stager & Werker, 1997). Additional difficulty might arise from ambiguous phonological boundaries at this stage of development (e.g., Rost & McMurray, 2009), or from uncertainty about the referential status of the novel word (Fennell & Waxman, 2010).

Regardless of the exact explanation, it is generally accepted that by around 17 m.o, children succeed in the same task (Werker, Fennell, Corcoran, & Stager, 2002). What could be the mechanism of development? Early accounts assumed—whether implicitly or explicitly—that children encode words in a binary way: they either fail or succeed in encoding the relevant phonetic details (simultaneously with

the meanings). This account suggested a discontinuous/stage-like pattern of development whereby younger children fail to encode the contrastive phonetic detail, whereas older children succeed.

Nevertheless, subsequent findings suggest otherwise. On the one hand, 14-month-olds—who typically fail in the original task—succeed when an easier *testing* method is used, even under the same *learning* conditions (Yoshida, Fennell, Swingley, & Werker, 2009). They also succeed when uncertainty is mitigated via disambiguating cues (e.g., Thiessen, 2007). On the other hand, adults show patterns of learning similar to those shown by 14-month-olds when the task is more challenging and when similarity between words increases (Pajak, Creel, & Levy, 2016; White, Yee, Blumstein, & Morgan, 2013).

These evidence point towards another scenario, where the representations are encoded in a probabilistic (rather than binary) way, and where development is continuous, rather than stage-like (Pajak et al., 2016; Swingley, 2007; Yoshida et al., 2009). On this account, correct representations are learned early in development, but these representations are encoded with higher uncertainty in younger children, leading to apparent failure in relatively demanding tasks. Development is a continuous process whereby the initial noisy representations become more precise. Crucially, more precise representations are still imperfect: Even with adults show low accuracy learning when the sounds are subtle enough, e.g., non-native sounds (Pajak et al., 2016).

We provide an intuitive illustration of how such an account explains patterns of learning and development in Figure 1. We observe low accuracy in word learning when the perceptual distance between the labels is small relative to the uncertainty/tolerance with which these labels are encoded. For example, in Stager and Werker's original experiment, children are supposed to associate label 1 (“bih”) and label 2 (“dih”) with object 1 and object 2, respectively. Though children could learn that the label “bih” is a better match to object 1 than “dih”, they could still judge the sound “dih” as a plausible instance of the label “bih”, thanks to the relatively large tolerance of the encoding.

Accuracy in word learning increases when the perceived distance between the labels is large relative to the tolerance of their encoding. This improvement can occur in the same developmental stage if the perceptual distance between labels is enhanced through using different-sounding labels (e.g., “lif” vs. “neem” instead of “bih” and “dih”) or through using additional cues (Rost & McMurray, 2009; Thiessen, 2007). Besides, accuracy can improve across development through the

refinement of the encoding.

Building on this intuition, the current work proposes a precise probabilistic model, which we use to both account for previous experimental findings, and to make new predictions that have not been tested before. Using new data collected from both preschool children and adults, we show that the model can explain various patterns of learning both within the same age and across development.

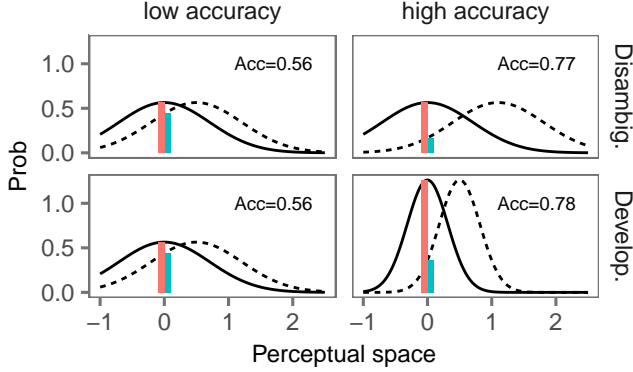


Figure 1: An illustration of the probabilistic/continuous account using simulated data. Two words are represented as distributions over the perceptual space. Here accuracy expresses the extent to which a given sound instance indicates the word on the left (red bar), as opposed to the word on the right (green bar). The values vary between 0.5 (total overlap between the distributions) and 1 (no overlap). Accuracy is low when the perceptual distance between labels is small relative to the variance. Accuracy increases when the perceptual distance is enhanced (through disambiguation), or when the variance decreases (e.g., through development).

Model

Task

We model the word learning task introduced by Stager & Werker (1997), and a testing method similar to the one used by Yoshida et al. (2009). In this task, participants are first exposed to the association between pairs of nonsense words (e.g., “lif”/“neem”) and pairs of objects. After this exposure phase, participants perform a series of two-alternative forced choices. In each testing trial, one of the two sounds is uttered (e.g., “lif”) and participants choose the corresponding object from the two alternatives. An overview of the task is shown in Figure 2.

Probabilistic structure

Our model consists of a set of variables describing the general process of spoken word recognition in a referential situation. These variables are related in a way that reflects the simple generative scenario represented graphically in Figure 3. When a speaker utters a sound in the presence of an object, the observer assumes that the object o activated the concept C

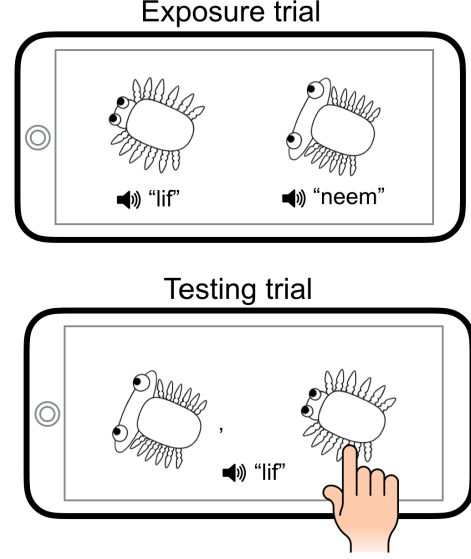


Figure 2: An overview of the task used in this study.

in the speaker’s mind. The concept prompted the corresponding label L . Finally, the label was physically instantiated by the sound s .

A similar probabilistic structure was used by Lewis & Frank (2013) to model concept learning, and by Hofer & Levy (2017) to model spoken word learning. However, the first study assumed that the sounds are heard unambiguously, and the second assumed the concepts are observed unambiguously. In our model, we assume there to be ambiguity at the level of both the labels and the concepts. For simplicity, we assume that the probability of membership of objects and sounds to concepts and labels, respectively, are normally distributed:

$$p(o|C) \sim \mathcal{N}(\mu_C, \sigma_C^2)$$

$$p(s|L) \sim \mathcal{N}(\mu_L, \sigma_L^2)$$

Finally, we assume there to be one-to-one mappings between concepts and labels, and that observers have successfully learned these mappings during the exposure phase:

$$P(L_i|C_j) = \begin{cases} 1 & \text{if } i = j \\ 0 & \text{if } i \neq j \end{cases}$$

Simulations

During the testing phase (see Figure 2), participants are presented with one target sound $s_T \in \{s_1, s_2\}$ and with both objects o_1 and o_2 . In order to make a choice, we need to determine which object is more probable under the target sound s_T . In other words, we need to compare the probabilities $P(o_1|s_T)$ and $P(o_2|s_T)$. The values of these probabilities can be computed by summing over all possible concepts and labels:

$$P(o|s) = \sum_{C,L} P(o, C, L|s) \propto \sum_{C,L} P(o, C, L, s)$$

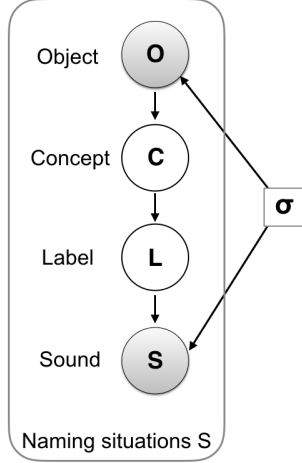


Figure 3: Graphical representation of our model. Circles indicate random variables (shading indicates observed variables). The squares indicates fixed model parameters.

The joint probability $P(o, C, L, s)$ is obtained by factoring the Bayesian network in Figure 3:

$$P(o, C, L, s) = P(s|L)P(L|C)P(C|o)P(o)$$

which could be transformed using Bayes rule into:

$$P(o, C, L, s) = P(s|L)P(L|C)P(o|C)P(C)$$

Finally, assuming that the concepts' prior probability is uniformly distributed¹, we obtain the following expression, where all conditional dependencies are now well defined:

$$P(o|s) = \frac{\sum_{C,L} P(s|L)P(o|C)P(L|C)}{\sum_o \sum_{C,L} P(s|L)P(o|C)P(L|C)}$$

From this general expression we derive the exact analytical formula, which expresses the probability of accurate responses in the testing phase of the task:

$$P(o_T|s_T) = \frac{1 + e^{-(\Delta s^2 + \Delta o^2)/2\sigma^2}}{1 + e^{-(\Delta s^2 + \Delta o^2)/2\sigma^2} + e^{-\Delta s^2/2\sigma^2} + e^{-\Delta o^2/2\sigma^2}} \quad (1)$$

In order to have a more quantitative understanding of the model, we simulate the values of the predicted accuracy (Expression 1) as a function of the perceptual distance between the sounds Δs . We used as parameters the two remaining variables, i.e., the visual distance between the semantic referents Δo and the standard deviation of the distributions $p(s|L)$ and $p(o|C)$ (which, in this simulation, are assumed to have similar values, i.e., $\sigma = \sigma_C \approx \sigma_L$). The simulations are shown in Figure 4.

The simulations explain some previously documented facts, and make new predictions:

¹This is a reasonable assumption given the similarity of the concepts used in each naming situation. See the stimuli sub-section in the Experiment below.

- 1) For fixed values of Δs and σ , the probability of accurate responses increases as a function of Δs . This pattern accounts for the fact that similar sounds are generally more challenging to learn than different sounds for both children (Stager & Werker, 1997) and adults (Pajak et al., 2016).
- 2) For fixed values of Δs and Δo , accuracy increases when the representational uncertainty (characterized with σ) decreases. This fact may explain development, i.e., younger children have noisier representations (see Swingley, 2007; Yoshida et al., 2009), which leads to lower word recognition accuracy, especially for similar-sounding words.
- 3) For fixed values of Δs and σ , accuracy increases with the visual distance between the semantic referents Δo . This is a new prediction that our model makes. Previous work studied the effect of several bottom-up and top-down properties in disambiguating similar sounding words (Dautriche, Swingley, & Christophe, 2015; Fennell & Waxman, 2010; Rost & McMurray, 2009; Thiessen, 2007), but no previous study tested the effect of the visual distance between the semantic referents.

To sum, we introduced a model that accounts for some qualitative learning patterns observed in previous studies, and makes a new prediction. In the experiment below, we test whether the model makes accurate *quantitative* predictions by fitting it to new experimental data collected from preschool children and adults.

Experiment

In this experiment, we tested participants in the word learning task introduced above (Figure 2). We explored all three parameters of the model. In particular, both the sound similarity (Δs) and object similarity (Δo) were varied simultaneously in a within-subject design. Besides, two age groups (preschool children and adults) were tested on the same task to explore whether development can be characterized with the uncertainty parameter, σ , in the probabilistic representations.

Methods

Participants We planned to recruit a sample of 60 children ages 4-5 years from the Bing Nursery School on Stanford University's campus. So far, we collected data from N=47 children (mean age= months, F=). An additional 28 children participated but were removed from analyses because they were not above chance on the catch trials (as was specified in the pre-registration²). We also collected a sample of N=20 adults on Amazon Mechanical Turk (this is a pilot, I will plan for 60 participants). N=2 adult participants were excluded because of low scores on the catch trials (see pre-registration). (I still need to write the code that derives these numbers automatically from the data).

²<https://osf.io/jrh38/>

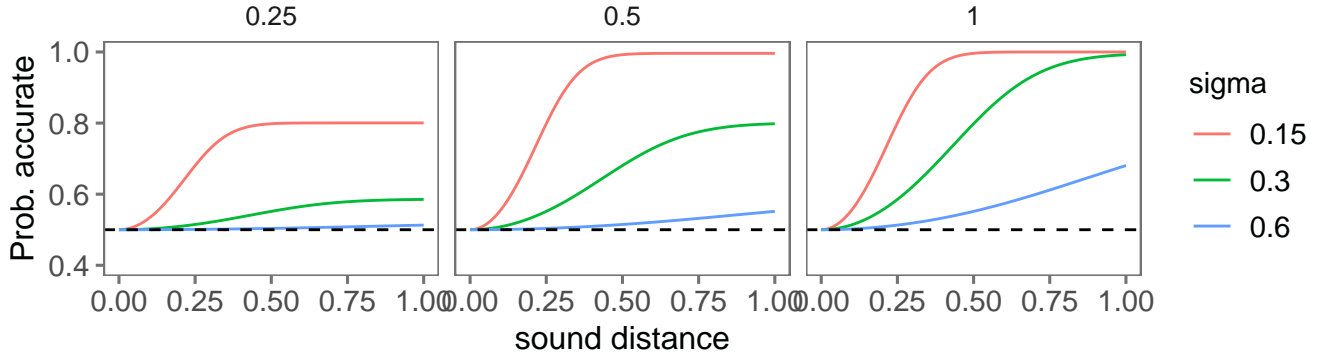


Figure 4: The predicted probability of accurate responses in the testing phase, as a function the perceptual distance between the sounds. Colors indicate different values of the standard deviation which we assume is common to both the labels’ and concepts’ probabilistic representations. Panels represent graphs using different values of the visual distance between the objects.

Stimuli and similarity rating The sound stimuli were generated using the MBROLA Speech Synthesizer (Dutoit, Pagel, Pierret, Bataille, & Van der Vrecken, 1996). We generated three kinds of nonsense word pairs which varied in their degree of similarity to English speakers: 1) “different”: “lif”/“neem” and “zem”/“doof”, 2) “intermediate”: “aka”/“ama” and “ada”/“aba”, and 3) “similar” non-English minimal pairs: “ada”/“ad^ha” (in hindi) and “aʃa”/“a^ha” (in arabic).

As for the objects, we used the Dynamic Stimuli javascript library³ which allowed us to generate objects in four different categories: “tree”, “bird”, “bug”, and “fish”. These categories are supposed to be naturally occurring kinds that might be seen on an alien planet. In each category, we generated “different”, “intermediate” and “similar” pairs by manipulating a continuous property controlling features of the category’s shape (e.g, body stretch or head fatness).

In a separate survey, $N = 20$ participants recruited on Amazon Mechanical Turk evaluated the similarity of each sound and object pair on a 7-point scale. We processed the resulting data so that it can be used in the model. We normalized the values with respect to the most distant level, and we scaled them within the range $[0, 1]$. Processed data are shown in Figure 5, for each stimuli group. This data will be used in the models as the perceptual distance of sound pairs (Δ_s) and object pairs (Δ_o).

Design Each age group saw only two of the three levels of similarity described in the previous sub-section: “different” vs. “intermediate” for preschoolers, and “intermediate” vs. “similar” for adults. We made this choice because, on the one hand, adults were at ceiling with “different” sounds/objects, and on the other hand, children were at chance with the “similar” sounds/objects. That said, this difference in the level of similarity is accounted for in the model through using the appropriate perceptual distance used in each age group (Figure 5).

The experiment consisted of four conditions which in-

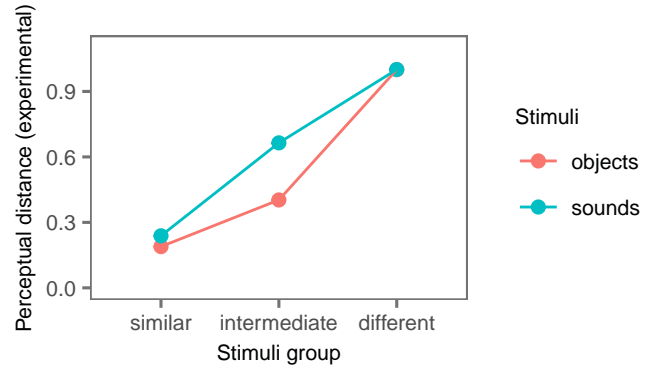


Figure 5: Normalized distances for both sound and object pairs used in this study.

involved, each, one pair of sounds-objects associations. These conditions were constructed by crossing the sound’s degree of similarity with the object’s degree of similarity leading to a 2×2 factorial design in each age group. Besides the 4 conditions, we also tested participants on a fifth catch condition which was similar in its structure to the other ones, but was used only to select participants who were able to follow the instructions and show minimal learning.

Procedure Preschoolers were asked if they would be willing to play a game on a tablet with the experimenter and were informed that they could stop playing at any time. The experimenter explained that the game consisted in learning some words spoken in an alien planet. The experiment began with two simple examples (not included in the analysis), and in these examples children were given feedback from the experimenter so as to make sure they correctly understood the structure of the task. After the introduction and examples, children were tested in a sequence of five conditions: the four experimental conditions plus the catch condition. In each condition, participants saw a first block of four exposure trials followed by four testing trials, and a second block of two exposure trials (for memory refreshment) followed by an additional four

³<https://github.com/erindb/stimuli>

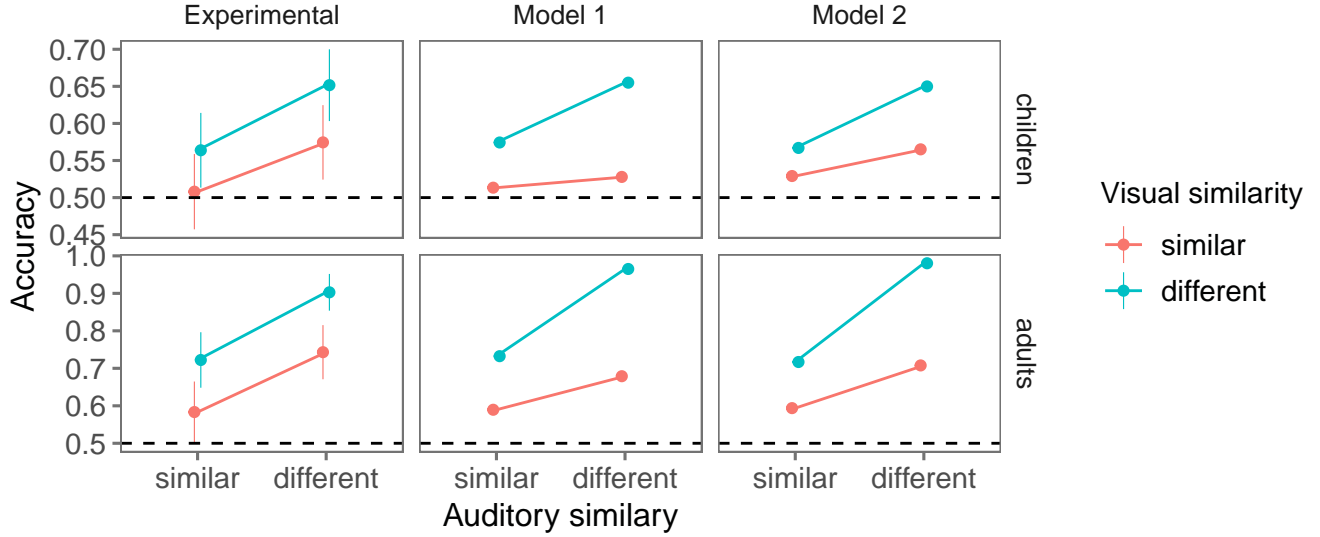


Figure 6: Accuracy of novel word recognition as a function of the sound distance, the object distance, and the age group (preschool children vs. Adults). Experimental results are shown on the left. Predictions from Model 1 (one free parameter) and Model 2 (two free parameters) are shown in the middle and on the right, respectively.

testing trials.

In the exposure trials, children saw two objects associated with their corresponding sounds. We presented the first object on the left side of the tablet’s screen simultaneously with the corresponding sound. The second sound-object association followed on the other side of the screen after 500ms. For both objects, visual stimuli were present for the duration of the sound clip (800ms). In the testing trials, children saw both objects simultaneously and heard only one sound. They completed the trial by selecting which of the two objects corresponded to the sound. They responded by touching one of the pictures on the tablet.

The object-sound pairings were randomized across participants, as was the order of the conditions (except for the catch condition which was always placed in the middle of the testing sequence). We also randomized the on-screen position (left vs. right) of the two pictures on each testing trial.

The procedure in the adult experiment was identical to that of preschool children, except that preschoolers were accompanied by an experimenter and used a tablet, whereas adults used their local computers to complete the experiment online.

Model fitting We fit the probability function (equation 2) to the participants’ responses in each age group. The values of Δs and Δo were set based on data from the similarity judgment task (Figure 5). We used two models: **model 1** fit only one parameter ($\sigma = \sigma_C = \sigma_L$), and **model 2** fit two parameters ($\sigma_C \neq \sigma_L$). The values of the parameters were derived using weighted least-squares estimates.

Results

We first analyzed the experimental results shown in Figure (6, left), using a mixed-effects logistic regression with sound and object distances as fixed effects, and with a maximal ran-

dom effects structure (Barr, Levy, Scheepers, & Tily, 2013). Results are shown in Table 1. We found a main effect of sound distance on the accuracy of learning in both children and adults, thus replicating previous findings. We also found a main effect of object distance, thus confirming the new prediction of our model.

Table 1: Estimates of predictor coefficients (and their standard errors) by age group in the regression model

	Children	Adults
(Intercept)	0.426* (0.199)	3.114** (1.015)
Sound	0.272** (0.100)	2.320* (0.981)
Object	0.315* (0.137)	2.133* (0.952)
Sound x Object	0.151 (0.097)	1.821 (0.976)

Note: * $p < 0.05$; ** $p < 0.01$; *** $p < 0.001$

Figures (6, middle and right graphs) show the predictions of the models. Both model 1 and model 2 fit reasonably well the experimental data in both children and adults. In particular, they both correctly predict the relative recognition accuracy across conditions: the pair of words that differ on both the object and sound levels were the easiest to learn, followed by the pairs of words that differ on only one level, then the pair of words that are similar on both levels. Note, however, that both models predict a slight interaction between sounds and objects, but such an interaction was not significant in the experimental data (Table 1).

For Model 1, the fitted uncertainty parameter was $\sigma = 0.63$ [0.53, 0.73]⁴ for preschoolers, and $\sigma = 0.16$ [0.12, 0.19] for adults. As predicted, the uncertainty parameter was larger

⁴95 % Confidence intervals.

for children than it was for adults. The model explained the majority of the variance ($R^2 = 0.94$). For model 2, children had a label-specific uncertainty of $\sigma_S = 0.9$ [0.68, 1.11], and a concept-specific uncertainty of $\sigma_C = 0.29$ [0.1, 0.49]. Adults had a label-specific noise of $\sigma_S = 0.16$ [0.05, 0.28], and a concept-specific noise of $\sigma_C = 0.14$ [0.05, 0.23]. The model explained almost all the variance ($R^2 = 0.96$). Interestingly, despite the fact that Model 1 used only one degree of freedom, it captures the data variance almost as well as Model 2 (which used two degrees of freedom).

General Discussion

Since the seminal work of Stager & Werker (1997) showing a discrepancy between children's speech perception abilities and their word learning skills, researchers have debated about the source of such a discrepancy. In this work, we were more interested in the development of this phenomenon, regardless of its exact origin. Building on some previous findings and discussions (Swingley, 2007; Yoshida et al., 2009), we proposed a model which assumes that subtle perceptual features are correctly encoded during word meaning learning, but that the precision of this encoding vary across development. We compared the predictions to new data which we collected from preschool children and adults. We found that the probabilistic model captured not only the patterns of development, but also various learning patterns within the same age group.

The model made a new prediction which we tested experimentally for the first time: learning similar words is not only modulated by to similarity of their phonological forms, but also by the visual similarity of their semantic referents. More generally, since visual similarity is an early organizing feature in the semantic domain (e.g., Wojcik & Saffran, 2013), our finding suggests that children may prioritize the acquisition of words that are quite distant in the semantic space. This suggestion is supported by recent findings based on the investigation of early vocabulary growth (Engelthaler & Hills, 2017; Sizemore, Karuza, Giusti, & Bassett, 2018). That said, further work is needed to explore the effect on word learning of other semantic dimensions that could be encoded by children (e.g., conceptual/functional features).

To conclude, this paper proposes a computational model for the probabilistic/continuous account, and the findings do endorse this account as a valid framework for the study of similar word learning and the mechanism of its development. More generally, together with previous modeling efforts (e.g., Apfelbaum & McMurray, 2011; Hofer & Levy, 2017), our work highlight the crucial role computation models can play in theory building through both a compact organization of existing knowledge and through generating new principled hypothesis.

All data and code for these analyses are available at
<https://github.com/afourtassi/networks>

Acknowledgements

This work was supported by a post-doctoral grant from the Fyssen Foundation, NSF #1528526, and NSF #1659585.

References

- Apfelbaum, K., & McMurray, B. (2011). Using variability to guide dimensional weighting: Associative mechanisms in early word learning. *Cognitive Science*, 35(6).
- Barr, D., Levy, R., Scheepers, C., & Tily, H. (2013). Random effects structure for confirmatory hypothesis testing: Keep it maximal. *Journal of Memory and Language*, 68(3).
- Dautriche, I., Swingley, D., & Christophe, A. (2015). Learning novel phonological neighbors: Syntactic category matters. *Cognition*, 143.
- Dutoit, T., Pagel, V., Pierret, N., Bataille, F., & Van der Vrecken, O. (1996). The mbrola project: Towards a set of high quality speech synthesizers free of use for non commercial purposes. In *Proceedings of ICSLP* (Vol. 3). IEEE.
- Engelthaler, T., & Hills, T. T. (2017). Feature biases in early word learning: Network distinctiveness predicts age of acquisition. *Cognitive Science*, 41.
- Fennell, C., & Waxman, S. (2010). What paradox? Referential cues allow for infant use of phonetic detail in word learning. *Child Development*, 81.
- Hofer, M., & Levy, R. (2017). Modeling Sources of Uncertainty in Spoken Word Learning. In *Proceedings of the 39th Annual Meeting of the Cognitive Science Society*.
- Lewis, M., & Frank, M. (2013). An integrated model of concept learning and word-concept mapping. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 35).
- Pajak, B., Creel, S., & Levy, R. (2016). Difficulty in learning similar-sounding words: A developmental stage or a general property of learning? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42(9).
- Rost, G., & McMurray, B. (2009). Speaker variability augments phonological processing in early word learning. *Developmental Science*, 12.
- Sizemore, A. E., Karuza, E. A., Giusti, C., & Bassett, D. S. (2018). Knowledge gaps in the early growth of semantic feature networks. *Nature Human Behaviour*, 2(9).
- Stager, C., & Werker, J. (1997). Infants listen for more phonetic detail in speech perception than in word-learning tasks. *Nature*, 388(6640).
- Swingley, D. (2007). Lexical exposure and word-form encoding in 1.5-year-olds. *Developmental Psychology*, 43(2).
- Thiessen, E. (2007). The effect of distributional information on children's use of phonemic contrasts. *Journal of Memory and Language*, 56.
- Werker, J., & Tees, R. (1984). Cross-language speech perception: Evidence for perceptual reorganization during the first year of life. *Infant Behavior and Development*, 7.
- Werker, J., Fennell, C., Corcoran, K., & Stager, C. (2002). Infants' ability to learn phonetically similar words: Effects

- of age and vocabulary size. *Infancy*, 3.
- White, K., Yee, E., Blumstein, S., & Morgan, J. (2013). Adults show less sensitivity to phonetic detail in unfamiliar words, too. *Journal of Memory and Language*, 68(4).
- Wojcik, E., & Saffran, J. (2013). The ontogeny of lexical networks: Toddlers encode the relationships among referents when learning novel words. *Psychological Science*, 24(10).
- Yoshida, K., Fennell, C., Swingley, D., & Werker, J. (2009). 14-month-olds learn similar-sounding words. *Developmental Science*, 12.