The role of continuous developmental change in explaining discontinuities in word learning

Abdellah Fourtassi
afourtas@stanford.edu
Department of Psychology
Stanford University

Sophie Regan
sregan20@stanford.edu
Department of Psychology
Stanford University

Michael C. Frank
mcfrank@stanford.edu
Department of Psychology
Stanford University

Abstract

Cognitive development is often characterized in term of discontinuities. Nevertheless, in some cases, this characterization can be problematic and unnecessary. In this paper, we show that a phenomenon that appears stage-like, can be well described by a continuous account. We used as a case study the finding by Stager and Werker (1997) that children's early ability to distinguish similar sounds does 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 continuous pattern of development. Here we propose a precise probabilistic model describing how development may proceed in a continuous fashion across the lifespan. The model accounts for previously documented facts and provides new predictions. We collected data from preschool children and adults, and we showed that the model can explain various patterns of learning both within the same age and across development. The findings suggest that other aspects of cognitive development which are typically thought of as discontinuities, may emerge from simpler, continuous mechanisms.

Keywords: word learning, cognitive development, computational modeling

Introduction

Cognitive development is sometimes characterized in terms of a succession of discontinuous stages (Piaget, 1954). Such characterization may be problematic as it often renders the study of the learning mechanism quite challenging. Indeed, it is unclear how a precise algorithm—-whose steps are gradual and logically related—-can account for transitions between stages involving qualitatively different representations (e.g., Carey, 2009).

Nevertheless, at least in some cases, development may only appear to be stage-like. This appearance can be due, for example, to the use of a cognitively-demanding task which may mask learning, or to the use of hard statistical thresholding (in particular, p-value < 0.05) which creates a spurious dichotomy between success and failure in observing a given behavior. In such cases, positing discontinuous stages is unnecessary. Instead, as we shall see in this paper, a continuous model—-involving similar representations across the lifespan—-may provide a simpler and more transparent account of development.

We use a case study from word learning literature. Stager & Werker (1997) first showed that children's early ability to distinguish similar sounds ("dih" vs. "bih") does not automatically translate into word learning skills (i.e., associating these sounds with different meanings). Subsequent research has focused on proposing possible explanations for this observed gap between speech perception and word learning (e.g., Fennell & Waxman, 2010; Hofer & Levy, 2017; Rost & McMurray, 2009; Stager & Werker, 1997).

In this work we focus, rather, on the development of this phenomenon. In fact, 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). How does development proceed? Early accounts assumed 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 have suggested 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 representation are still imperfect: Even 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 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 uncertainty of the encoding.

According to this account, accuracy in word learning improves if we increase either the perceptual distinctiveness of the stimuli (e.g., through using different-sounding labels), or the precision of the encoding itself (e.g., across development). 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.

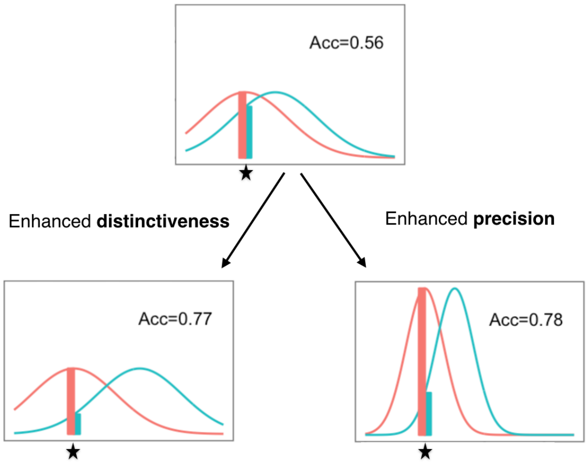

Figure 1: An illustration of the probabilistic/continuous account using simulated data. A word is represented with a distribution over the perceptual space (indicated in red or green). When the uncertainty of the representation is large relative to the distance between the stimuli (top panel), an instance of the red category (indicated with a star) could also be a plausible instance of the green category, hence the low recognition accuracy score. The accuracy increases when the stimuli are less similar (left panel), or when the representation are more precise (right panel).

Model

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 2. When a speaker utters a sound in the presence of an object, the observer assumes that the object o activated the concept C in the speaker’s mind. The concept prompted the corresponding label L. Finally, the label was physically instantiated by the sound s.

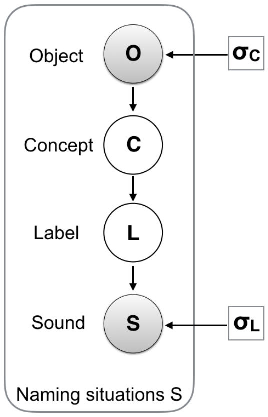

Figure 2: Graphical representation of our model. Circles indicate random variables (shading indicates observed variables). The squares indicates fixed model parameters.

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:

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

Inference

The learner hears a sound s and has to decided which object o provides an optimal match to this sound. To this end, the learner has to compute the probability $P(o|s)$ for all possible objects. This probability can be computed by summing over all possible concepts and labels:

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

The joint probability $P(o,C,L,s_T)$ is obtained by factoring the Bayesian network in Figure 2:

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

which could be transformed using Bayes rule into:

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

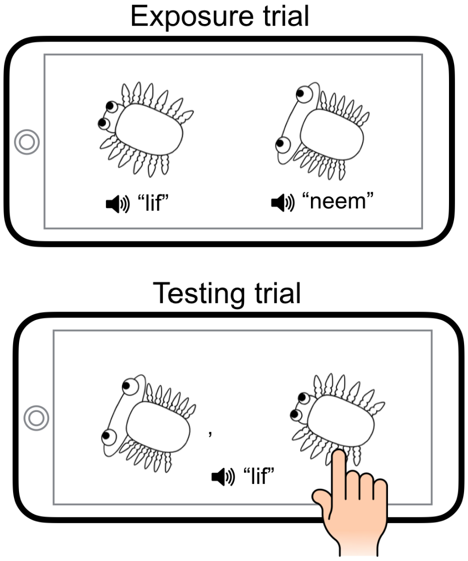

Figure 3: An overview of the task used in this study.

Finally, assuming that the concepts' prior probability is uniformly distributed[1], we obtain the following expression, where all conditional dependencies are now well defined:

$$P(o|s) = \frac{\sum_{C,L} P(s|L)P(o|C)P(L|C)}{\sum_o \sum_{C,L} P(s|L)P(o|C)P(L|C)} \quad (1)$$

Task and model predictions

We use the model to predict performance in 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 test trials. In each of these trials, 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 3.

We use the general Equation 1 to derive the exact analytical expression for the probability of accurate responses $p(o_T|s)$ (target object o_T given a sound s) in the simple case of two-alternative forced choice in the testing phase of our experimental task:

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

Figure 4 show simulations of the predicted accuracy (Expression 2) as a function of the distinctiveness parameters (Δs and Δo) and the precision parameter, i.e., the variance of the distributions $p(s|L)$ and $p(o|C)$ (which, for ease of the simulations' interpretation, are assumed to have similar values, i.e., $\sigma = \sigma_C \approx \sigma_L$).

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

1) For fixed values of Δo 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 (e.g., Fennell & Waxman, 2010; Rost & McMurray, 2009; Thiessen, 2007), but no previous study tested the effect of the visual distance between the semantic referents.

In 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 3). 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. 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. The Experiment and the predictions were pre-registered.

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-registration[2]). We also collected a sample of N=20 adults on Amazon Mechanical Turk (this is a pilot, I will plan

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

[2] https://osf.io/jrh38/

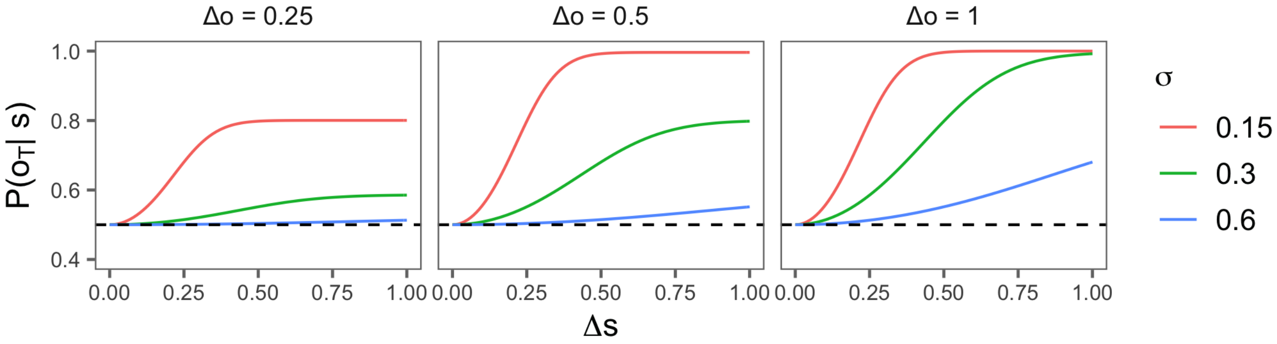

Figure 4: The predicted probability of accurate responses in the testing phase as a function of stimuli distinctiveness Δ*s* (x-axis) and Δ*o* (different panels) and representation precision σ (color). Dashed line represents chance.

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).

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^{h}a" (in hindi) and "aʕa"/"aħa" (in arabic).

As for the objects, we used the Dynamic Stimuli javascript library[3] 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).

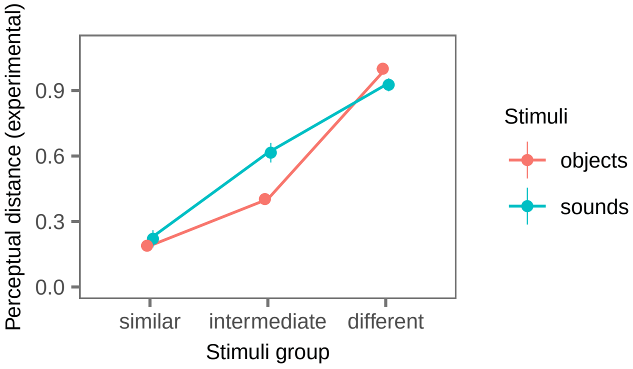

Figure 5: Distances for both sound and object pairs from an adult norming study. Error bars represent 95% confidence intervals.

The experiment consisted of four conditions which 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 2x2 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 tested at the nursery school using a tablet, whereas adults used their local computers to complete the same experiment online. Participants 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

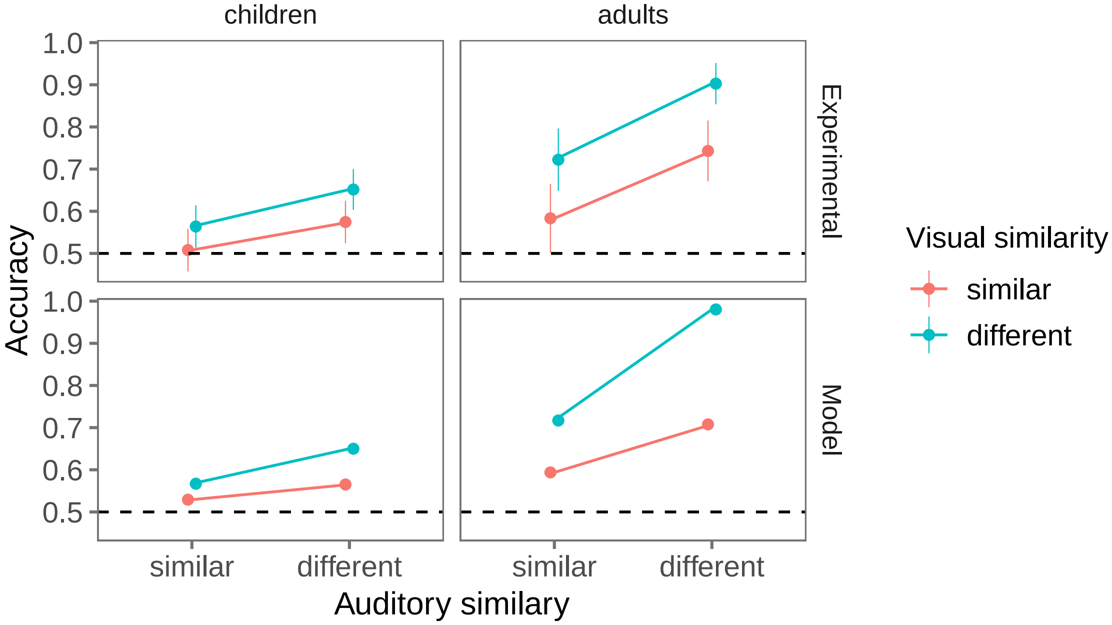

Figure 6: Accuracy of novel word recognition as as a function of the sound distance, the object distance, and the age group (preschool children vs. Adults). We show both experimental results (top) and model predictions (bottom). Error bars represent 95% confidence intervals.

testing trials.

In the exposure trials, participants 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, participants saw both objects simultaneously and heard only one sound. They completed the trial by selecting which of the two objects corresponded to the sound. 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.

Results

We first analyzed the experimental results shown in Figure (6, top), using a mixed-effects logistic regression with sound distance, object distance and age group as fixed effects, and with a maximal random effects structure (allowing us to take into account the full nested structure of our data) (Barr, Levy, Scheepers, & Tily, 2013). We found main effects for all the fixed effects, i.e., the sound distance ($\beta = 0.68$, $p < 0.001$)—replicating previous findings, the object distance ($\beta = 0.6$, $p < 0.001$)—confirming the new prediction of our model, and age group ($\beta = 0.59$, $p < 0.001$)—showing that performance improves with age. Besides, we found two-way interactions between sound and object distances ($\beta = 0.36$, $p < 0.05$) and between sound distance and age ($\beta = 0.37$, $p < 0.01$).

Second, 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). Thus, the model has two degrees of freedom in each group, i.e., σ_C and σ_L. Figure 6 (bottom) show the predictions of the model. The model correctly predicts 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. At the quantitative level, the models explained almost all the variance in the participants' mean responses ($R^2 = 0.96$).

We also fit a simplified model with only one degree of freedom ($\sigma = \sigma_C = \sigma_L$). This second model is not intended to test any theoretical hypothesis. We use it, rather, as a baseline to understand how much of the explanatory power of the model is due to the extra degree of freedom. This baseline model still captured the qualitative pattern (graph not shown), and explained the majority of the variance ($R^2 = 0.94$). Besides, model comparison using Akaike information criterion yielded $AIC(model) - AIC(baseline) =$ -3.79 for children data and $AIC(model) - AIC(baseline) = 1.62$ for adult data. These numbers suggest that the model remains highly predictive even with one degree of freedom.

As for the values of the parameters, children had a label-specific uncertainty of $\sigma_S = 0.9$ [0.68, 1.11][4], and a concept-specific uncertainty of $\sigma_C = 0.29$ [0.1, 0.49]. Adults had a

[4] All uncertainty intervals in this paper represent 95% Confidence Intervals.

label-specific uncertainty of $\sigma_S = 0.16$ [0.05, 0.28], and a concept-specific uncertainty of $\sigma_C = 0.14$ [0.05, 0.23]. As predicted, the uncertainty parameters were larger for children than they were for adults.

General Discussion

This paper studies the hypothesis that some seemingly stage-like patterns in cognitive development can be characterized in a continuous fashion. We used as a case study the seminal work of Stager & Werker (1997) showing a discrepancy between children's speech perception abilities and their word learning skills. While most previous work debated the source of this discrepancy, here we were more interested in its mechanism of development.

Building on some previous findings and discussions (e.g., 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 improves across development. We tested the model's predictions against data collected from preschool children and adults and we showed that development can be characterized as a continuous refinement in qualitatively similar representations across the life span.

The model made a new prediction which we tested experimentally for the first time: learning similar words is not only modulated by the 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).

One limitation of this work is that the model was fit to data from children at a relatively older age (4-5 years old) than what is typically studied in the literature (around 14 month-old). While an important conclusion of this paper is that the same process may apply to different developmental stages, it would be interesting to experimentally verify the generality of this conclusion in the context of early language acquisition. A challenge to this next step would be to construct a computational framework which would compare performance across heterogeneous testing methods (e.g., looking time vs. forced-choice).

To sum, this paper proposes a model that accounts for the development of an important aspect of word learning. The findings show that data can be explained based on a continuous process operating over similar representations across development. We used a case from word learning as an example, but the same idea might apply to other aspects of cognitive development which are tyipcally thought of as stage-like (e.g., acquisition of a theory of mind). A crucial task would be to investigate the extent to which such discontinuities emerge due to genuine qualitative changes, and the extent to which they reflect the granularity of the researchers' own measurement tools.

References

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).

Carey, S. (2009). *The origin of concepts*. Oxford University Press.

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).

Piaget, J. (1954). *The construction of reality in the child*. New York, NY, US: Basic Books.

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., 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*.