Chaining and the formation of spatial semantic categories in childhood

Aotao Xu (a26xu@cs.toronto.edu)
Department of Computer Science, University of Toronto

Yang Xu (yangxu@cs.toronto.edu)
Department of Computer Science, Cognitive Science Program, University of Toronto

Abstract

Children face the problem of extending a limited spatial lexicon to potentially infinite spatial situations. Previous work has examined how spatial semantic categories may be formed in child development, but it is unclear how children extend these categories to novel situations over the developmental time course. Drawing on cognitive linguistic theories of category extension, we present a framework that models the incremental extension of spatial relational words to novel situations through time. We describe a longitudinal dataset and computational analyses for investigating the extension of spatial word meanings in a developmental setting. Our preliminary results suggest that the formation of spatial categories takes place through an exemplar-based process of chaining, similar to the process underlying the growth of linguistic categories in history. Our work offers opportunities to explore the connection between ontogeny and phylogeny in the process of word meaning extension.

Keywords: language and development; spatial language; word meaning extension; chaining; exemplar theory

Introduction

Spatial cognition is fundamental to survival for many species. One feature of spatial cognition unique to humans is our ability to describe space using language. While spatial language is limited by a finite lexicon (Landau & Jackendoff, 1993), the continuity of space entails a potentially infinite set of spatial situations. This tension of extending a finite spatial lexicon to unbounded novel spatial situations is prominent in children's language development, since young learners do not possess adult-like spatial language yet have the need to communicate novel spatial scenes. Here we explore in a formal framework how children incrementally extend spatial categories to novel spatial situations over the developmental time course.

In English and many other languages, spatial relations are commonly expressed in a closed class of words known as spatial adpositions (e.g., *on* and *in*). A salient property of these spatial categories is that they can extend to describe a diverse set of scenes. For instance, English *on* may be used to describe scenes such as *picture on wall*, *cup on table*, or *spider on ceiling*. Previous research has suggested that the formation of spatial semantic categories such as *on* and *in* relies on the linguistic input that children receive in their environment (Bowerman & Choi, 2003; Levinson, Meira, Language, & Group, 2003). Existing work has also shown that spatial categories can be learned from stored exemplars of spatial word usages (Bowerman & Choi, 2001) or co-occurrences with verb predicates (Johannes, Wilson, & Landau, 2016; Landau, 2018). We extend these studies to examine how children incrementally adapt spatial categories toward novel situations when the mappings between the spatial words and the novel spatial scenes have not yet been established.

Our starting point is the view by Vygotsky (1962) suggesting that one strategy in children's conceptual formation is through "chain complexes". In a chain complex, a new object is added to a category if it is similar to existing objects in the category, such that this new member, just like the other pre-existing members, acts as a "magnet" for a series of other objects (Vygotsky, 1962). This idea of chaining has been independently discussed in cognitive linguistics as a mechanism for linguistic category extension, best known as radial categories (Lakoff, 1987). Starting from a core meaning or sense, a category extends by adding novel referents related to existing members, and hence forming a chain-like structure over time. As a classic example, Brugman and Lakoff (1988) analyzed how the spatial word *over* may extend to a variety of scenarios through the process of chaining. We believe that chaining might explain how spatial categories like *on* and *in* get extended incrementally to express a diverse set of spatial scenes. To our knowledge no existing work has formally explored and evaluated this idea in the formation of spatial semantic categories through childhood.

We present a computational framework that formulates the formation of spatial semantic categories as chaining. At a specific time during the developmental time course, we model the categorization of a novel spatial situation based on its similarity with existing exemplars of spatial words in semantic space; a spatial word is chosen when the situation is more similar to its category members than to other competing categories. Throughout the time course, spatial categories are updated incrementally with exemplars observed from children's linguistic input. Figure 1 illustrates our framework.

Previous work has applied computational models of chaining to predict the historical growth of semantic and grammatical categories (Xu, Regier, & Malt, 2016; Ramiro, Srinivasan, Malt, & Xu, 2018; Habibi, Kemp, & Xu, 2020) and children's overextension in a cross-sectional setting (Ferreira Pinto Jr. & Xu, 2019). In these cases, it has been suggested that the process of chaining can be best understood as an exemplar model of categorization (Nosofsky, 1986). Our work extends these studies by examining the relation of chaining

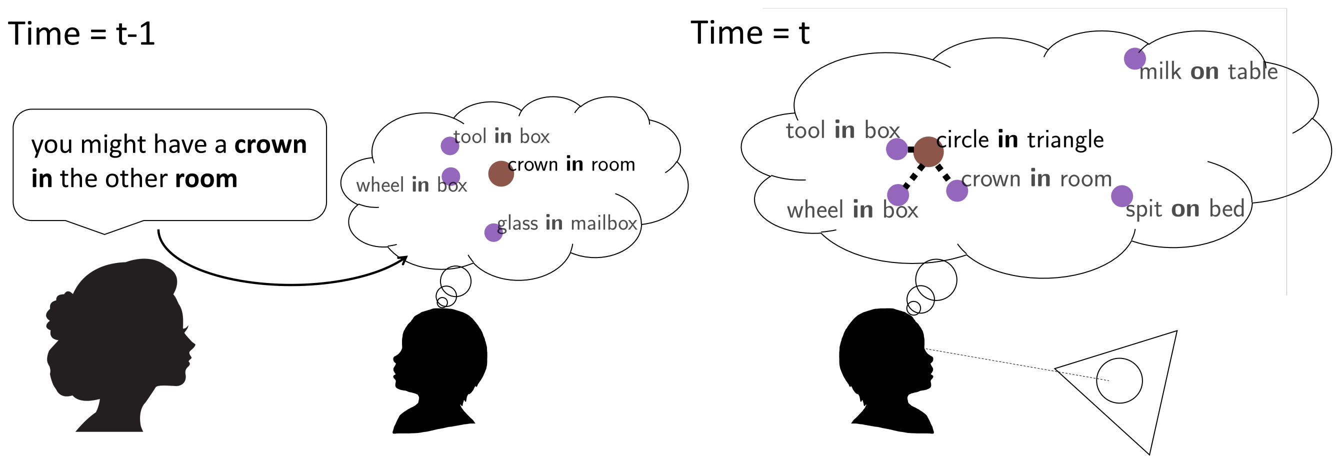

Figure 1: Illustration of chaining in a developmental setting: at time $t-1$, an adult produces an utterance describing the spatial situation *crown in room*, which the child stores as an exemplar for the category *in*; at a future time t when the child categorizes a novel spatial situation *circle in triangle*, stored exemplars similar to the situation attract it to the category *in*. Each purple dot represents a stored exemplar for a spatial situation observed in the past, and each brown dot represents a novel spatial situation.

and children's spatial language development. Our proposal is consistent with recent work suggesting that language acquisition relies on an exemplar-based mechanism as opposed to stored abstractions (Ambridge, 2020). In our analyses, we contrast an exemplar model with a prototype model that extends spatial categories using stored abstractions for the categories. Our work also relates to other computational studies that learn from labelled data to classify spatial situations with spatial words (e.g., Regier, 1996; Golland, Liang, & Klein, 2010; Xu & Kemp, 2010; Beekhuizen, Fazly, & Stevenson, 2014), but it differs critically in that it captures the incremental process in which spatial words extend to novel situations as they emerge in child development.

Computational framework

Our computational framework involves two main components: 1) a class of models that capture the process of chaining and category extension, and 2) a mechanism for updating the models under new observations of spatial word usage. We assume a finite spatial lexicon $w \in W$ and model a spatial situation s as a real-valued vector in some semantic space S.

Models of chaining

We formulate the extension of spatial words to novel situations as probabilistic inference over a finite set of known spatial words. At a future time t, the model has access to previously observed usages of spatial words. Let $x_i = (w_i, s_i)$ be the i-th word-situation pair, and suppose the model has access to $n > 0$ such pairs prior to time t. Further suppose a novel situation s^* has not been observed in $x_1, \ldots, x_n$. Then, the problem can be formulated as selecting the most likely w describing s^* incrementally at each future time t, given past observations prior to t:

$$\arg\max_w p(w|s^*, x_1, x_2, \ldots, x_n) \tag{1}$$

To model this process, we draw on recent work of chaining models (Ramiro et al., 2018; Habibi et al., 2020; Grewal & Xu, 2020) that are based on formal models of categorization (Reed, 1972; Nosofsky, 1986; Ashby & Alfonso-Reese, 1995). In particular, this class of models can be derived by applying Bayes rule to Equation 1:

$$p(w|s^*, x_1, \ldots, x_n) \propto p(s^*|w, x_1, \ldots, x_n) p(w|x_1, \ldots, x_n) \tag{2}$$

Here the left-hand side is the category posterior, and terms on the right-hand side correspond to the likelihood and the category prior, respectively. Intuitively, Equation 2 says the most likely word can be inferred by computing the likelihood of s^* given w and stored exemplars, in joint consideration with the prior probability of using w.

Exemplar model. Exemplar models of categorization assign category labels to stimuli by computing their similarity with the exemplars of a category (Nosofsky, 1986). The exemplar model can be reformulated in terms of Equation 2 when the likelihood is given by a specific kernel density estimator (Ashby & Alfonso-Reese, 1995), such that the likelihood of s^* given w is estimated using a weighted sum of distances between s^* and previous situations described by w:

$$p(s^*|w, x_1, \ldots, x_n) = \frac{1}{n_w} \sum_{i:w_i=w} K_h(s^* - s_i) \tag{3}$$

Here $K(\cdot)$ is the kernel function and h is its kernel width. We drew on the generalized context model with a Gaussian similarity function (Nosofsky, 1986), which corresponds to a

Gaussian kernel function:

$$K_h(s^* - s_i) \propto \frac{1}{h^m} \exp\left(-\frac{d(s^*, s_i)^2}{2h^2}\right) \quad (4)$$

Here $d(\cdot,\cdot)$ is the Euclidean distance between two situations of dimension m. In the original general context model, h corresponds to a sensitivity parameter that determines the degree to which the model generalizes from exemplars: a large h implies broad generalization, and vice versa. Originally, h is the same for all categories. We also consider a variation where h differs by category, following related work in machine learning (John & Langley, 1995). We will refer to the original model as the *exemplar* approach, and the variation as the *extended exemplar* approach. In both cases, the parameter h is estimated from previous observations.

Prototype model. An alternative to exemplar-based chaining models is based on the central tendency of a category, specifically the category prototypes. Prototype models of categorization assign category labels to stimuli by computing their distances from category prototypes in some feature space (Reed, 1972). This can be reformulated in terms of Equation 2 when we assume the category prior is uniform and the likelihood is a certain Gaussian (Ashby & Alfonso-Reese, 1995). Following established work in machine learning (John & Langley, 1995), we considered a variation of classic prototype models such that the prior is not uniform (see Equation 7). This yields a Gaussian likelihood for each spatial word w:

$$p(s^*|w, x_1, \ldots, x_n) \propto \frac{1}{\sigma_w^m} \exp\left(-\frac{d(s^*, \mu_w)^2}{2\sigma_w^2}\right) \quad (5)$$

where σ_w is the standard deviation of the Gaussian, and μ_w is the prototype of the category. Following previous work (Reed, 1972; Habibi et al., 2020), we also define μ_w as the average of exemplar situations described by w:

$$\mu_w = \frac{1}{n_w} \sum_{i: w_i = w} s_i \quad (6)$$

Similar to h in Equation 4, σ_w determines the degree to which the model generalizes from the category prototype. Like the exemplar models, σ_w was estimated from previous observations. We will refer to this model as the *prototype* approach.

Category prior and baseline model. Similar to existing work (Beekhuizen et al., 2014; Ferreira Pinto Jr. & Xu, 2021; Habibi et al., 2020; Grewal & Xu, 2020), we define the prior distribution based on word frequency. Specifically, we compute the prior probability of word w as the proportion of its occurrences over past observations:

$$p(w|x_1, \ldots, x_n) = \frac{n_w}{n} \quad (7)$$

Here $n_w = \sum_{i=1}^{n} \mathbb{1}(w_i = w)$ is the number of past observations described by w, and $\mathbb{1}(\cdot)$ is the indicator function.

For comparison with our models of chaining, we considered a simple baseline model where category extension is only based on the frequency of stored usages of spatial words. Specifically, this baseline models category extension for any novel situation s^* by simply selecting the most likely spatial word w computed using Equation 7. We will refer to this model as the *frequency baseline*.

Model update

Over time, more usages of spatial words are observed, which need to be incrementally integrated into the chaining models we described. Following previous exemplar models (Estes, 1986; Ashby & Alfonso-Reese, 1995), we assumed each new observation x_i^* is stored in memory with a fixed probability p, independently of other exemplars. Thus the number of new exemplars stored at emerging time t, denoted k, follows a binomial distribution:

$$k \sim B(c, p) \quad (8)$$

When a novel scene s^* needs to be referred to at t, the model simply computes Equation 1 using the updated list of stored exemplars $x_1, \ldots, x_n, x_{n+1}, \ldots, n_{n+k}$. Although $c - k$ observations are not stored in memory, we hypothesized that they still potentially influence future inference to some extent (e.g., affecting the kernel width parameter).

Data

In our computational formulation, each instance of observation or extension made by a child at a specific time t is a word-situation pair, $(w, s) \in W \times S$. To obtain these instances, we 1) selected a subset of common spatial words W from the literature, 2) collected usages of these words and the co-occurring figure (f) and ground (g) objects from time-stamped text data, and 3) used semantic representation of these objects to construct a semantic space S.

Spatial words. We focused on two most common spatial prepositions, *in* and *on*, as they are among the earliest spatial categories children encounter (Baillargeon, Needham, & DeVos, 1992; Casasola & Cohen, 2002) and acquire (Clark, 1973; Bowerman & Choi, 2003). We treated each preposition as a relational predicate, such that it takes two arguments: a figure object (e.g., cup) and a ground object (e.g., table).

Objects. To obtain naturalistic usages of spatial words, we collected their co-occurrence with figure and ground objects (i.e., nouns). We grounded our data in early linguistic environments using North American English data from CHILDES (MacWhinney, 2000), a large collection of transcribed child speech (CS) and child-directed speech (CDS). The collection includes 6,647 conversations involving 690 children, each containing a sequence of sentences and a record of the age of the child involved. Each sentence is labelled by the identity of the speaker and is annotated with its dependency parse tree and part of speech tags.

For a spatial word $w \in W$ in a sentence, we set f as the rightmost noun to the left of w, and we set g as the child of w if it is a noun, or otherwise we selected the leftmost noun to the right of w. We discarded cases where at least one of the two

Sentence	Spatial word	Figure object	Ground object
do you have a scratch on your arm	on	scratch	arm
because she always put the nipple in into her mouth	in	nipple	mouth
you help me put the clothes in the dryer	in	clothes	dryer
I want bread and butter and jelly on my bread	on	jelly	bread
she has spider on her arm	on	spider	arm
having sugar in her meat	in	sugar	meat

Table 1: Examples of word-figure-ground triples extracted from CHILDES. The top three sentences are examples of child-directed speech, and the bottom three are examples of child speech.

objects is missing. We further filtered the objects to exclude idioms (e.g., *in case*, *on time*) using multi-word entries from the Historical Thesaurus of English (Kay, Roberts, Samuels, & Wotherspoon, 2017). All entities were lemmatized by the WordNet lemmatizer in NLTK (Loper & Bird, 2002). Across all corpora, this provided us with 23,012 unique triples of (w, f, g) from CDS and 6,296 unique triples from CS.

Visual semantic space. We approximated the semantic representation of situation s described by w using vector representations of the figure and ground objects that co-occur with w. Since the geometry of the objects are relevant to the semantics of spatial words to some extent (Landau & Jackendoff, 1993), we used visual embeddings from a convolution neural network, VGG-19 (Simonyan & Zisserman, 2015), which is pretrained on the ImageNet database (Deng et al., 2009). We mapped each object to a 4096-dimension vector following procedures in existing work (Ferreira Pinto Jr. & Xu, 2021). To obtain a single vector for situation s, we concatenated the vectors of f and g; the semantic representation of each situation s is thus $m = 8182$-dimensional. In *Discussion*, we outline the merits and limitations of this approach.

We intersected our set of (w, f, g) triples with ImageNet which provides images for sets of synonyms (synsets) based on WordNet (Fellbaum, 1998). Specifically, we only kept triples where the figure and ground exist in WordNet and the majority of their synsets are nouns. This yielded 3,886 unique word-situation pairs from CDS and 975 unique pairs from CS. Table 1 shows examples of the extracted triples after intersection. Note that our concatenation method implies all situations sharing the same figure and ground (e.g., *hand on mouth* and *hand in mouth*) had the same representation s; these situations covered 3.2% ($n = 112$) of the CDS pairs and 0.62% ($n = 6$) of the CS pairs.

Results

We first describe computational analyses to predict the extension of spatial words observed in child language development. We then interpret the results of our analyses.

Computational analyses. Assuming all samples of spatial word usages are drawn from the same real-world distribution, we constructed an early linguistic environment from which a child observes examples of spatial word usage by pooling word-situation pairs across CDS in CHILDES. We then obtained 6 sets of data from CS, where 4 sets corresponded to individuals and 2 sets were pooled from multiple children to approximate the extended, fine-grained developmental trajectories. In the CDS set and in each CS set, every word-situation pair was time-stamped based on its earliest occurrence in the set. Pairs in CS sets were removed if they occurred earlier in the CDS set, since our analyses focused on children's extension of spatial categories to novel situations. Note although word-situation pairs are unique in each set, they may still partially overlap (e.g., *pillow on bed* and *jammies on bed*).

Over the developmental time course, each of our models was evaluated based on its ability to predict the extension of spatial words in CS (child production) given observations in CDS (child input). At time t, by maintaining a list of stored exemplars up to t, we evaluated every model as follows: if a data point (w, s^*) appears at t, we used the stored exemplars and s^* as inputs to compute a prediction w' from Equation 1, which was compared against w. Then, we updated the stored exemplars following a simplified version of Equation 8: if c observations were made at t, we randomly added a fixed 80% of them to the list.[1] We repeated these steps over the entire developmental time course; the length of the time course was determined by the CS set being used. Due to the randomness in storing exemplars, we repeated the procedure 10 times.

To predict spatial word choice at each time t (Equation 1), we also need to estimate model parameters using observations prior to time t. For the two exemplar models, we estimate the kernel width h by maximizing the following log likelihood function:

$$\log \mathcal{L}(h) = \sum_{(w,s^*)\in X_t} \log p(w|s^*, x_1, ..., x_n; h) \tag{9}$$

Here X_t is the set of discarded observations complementary to the exemplars $x_1, ..., x_n$ stored prior to time t. Since $h > 0$, we optimized with respect to the log transformation of h, using mini-batch gradient descent and the Adam optimizer (Kingma & Ba, 2015) with the following settings: batch size $N = 64$, learning rate $\alpha = 0.01$, decay rates $(\beta_1, \beta_2) = (0.9, 0.999)$. For tractable computation, we applied early stopping after 20 epochs, and the parameters were updated only if the total number of stored exemplars changed by 100 since the last estimate. For the prototype approach, we estimated σ_w using the same method.

[1]We chose 80% to ensure the model has a sufficient number of exemplars to use for prediction.

Child data	Frequency baseline	Exemplar	Exemplar extended	Prototype	N
Pooled	0.629	0.737 (0.009)	**0.748** (0.007)	0.672 (0.008)	975
Target Child (HSLLD)	0.651	0.712 (0.015)	**0.714** (0.011)	0.644 (0.013)	146
Nina (Suppes)	0.517	0.697 (0.016)	**0.710** (0.013)	0.646 (0.013)	145
Abe (Kuczaj)	0.538	0.740 (0.018)	**0.762** (0.016)	0.635 (0.011)	91
Adam (Brown)	0.583	**0.781** (0.015)	0.771 (0.011)	0.688 (0.009)	72
Mark (MacWhinney)	0.583	0.697 (0.025)	**0.699** (0.021)	0.611 (0.022)	72

Table 2: Summary of model predictive accuracy. The first column shows the name of each child and the relevant corpus in brackets. The last columns shows the size of each CS set. Each other cell shows the average accuracy and its standard deviation over 10 runs of experiment; bold font indicates the best performing model for each instance of child data.

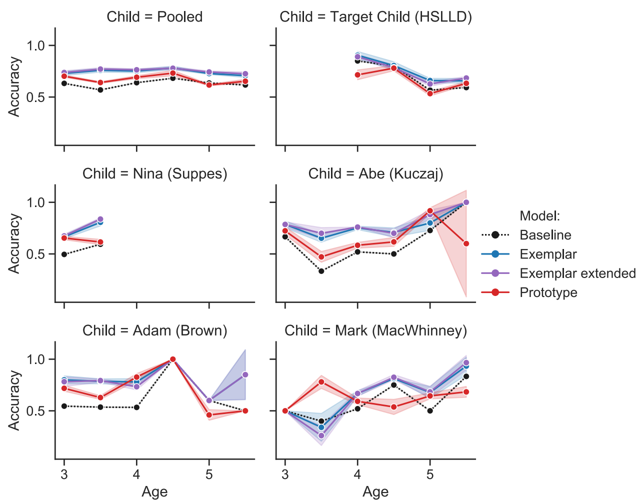

Figure 2: Summary of model predictive accuracy over time. Each point at age t shows the accuracy of the model for test cases between the $t-0.5$ and t, averaged over 10 runs of our experiment. Ribbons indicate standard deviation.

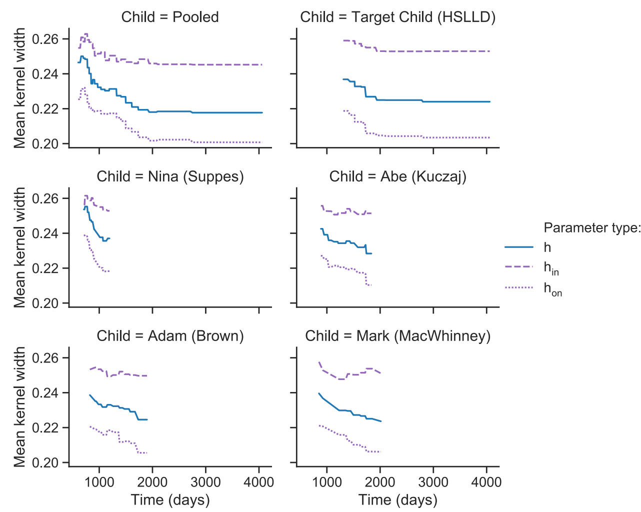

Figure 3: Changes in the value of optimal kernel parameters over time. Values for h were from the exemplar model, and those for h_{in}, h_{on} were from the extended exemplar model. Each point is an average kernel value over 10 runs of our experiments, which was estimated at a certain age of the child (measured in days).

Interpretation of results. Table 2 summarizes the performance of our models. Across all CS sets, we observe that the exemplar models outperform the frequency baseline and the prototype alternative. This is consistent with previous work which finds the exemplar model of chaining to best predict how numeral classifiers and adjectives apply to novel nouns over time (Habibi et al., 2020; Grewal & Xu, 2020). We hypothesize the inferior performance of our prototype approach is because a single prototype is unable to capture polysemous usages of spatial words. For example, the spatial word *on* in *cup on table* describes support, but it describes surface attachment in *picture on wall*; a single prototype is unlikely to be prototypical for both support and attachment. However, under an exemplar model, separate clusters of exemplars enable a category to occupy distinct subsets of the semantic space, thus providing a more fine-grained representation of the category. We also observe the extended exemplar model tends to provide more accurate predictions than the original version, but the difference is marginal.

Figure 2 provides a detailed temporal perspective on the results summarized in Table 2. There is some degree of random fluctuation in model performance, which seems to be the result of small sample sizes (see Table 2) as the fluctuations are stronger in the 3 CS sets with fewer test cases than the others. Nonetheless, in both pooled CS sets and individual cases, we observe the patterns we saw for the aggregate results—with the exemplar model best accounts for the data—hold over time. This provides evidence that children's formation of spatial relational categories can be explained by an exemplar-based process of chaining.

Since the kernel width of our exemplar models were re-estimated over time, we assess the temporal trend in how the models generalize from stored exemplars. Figure 3 shows that the kernel width became smaller over time. To quantify this trend, we fitted linear mixed models that regress kernel widths on instants of the developmental time course t (in days); for each CS set, the model considers a random intercept for each run of our experiments to control for multiple measurements. Table 3 summarizes the results for the exemplar model (and the results for the extended exemplar model were similar). We observe that across all CS sets, the regression coefficient β_{time} is significantly negative, which suggests that the model learned not to over-generalize over time, mimicking the process in which children learn from overextension (large h) to narrowing (small h) in the usage of spatial words such as *in, on,* and *off* (Clark, 1978; Bowerman, 1978).

Child data	β_{time}	p	N
Pooled	-1.1e-5	< 0.001	3,990
Target Child (HSLLD)	-5e-6	< 0.001	940
Nina (Suppes)	-4.5e-5	< 0.001	450
Abe (Kuczaj)	-1e-5	< 0.001	650
Adam (Brown)	-1e-5	< 0.001	390
Mark (Macwhinney)	-1.2e-5	< 0.001	260

Table 3: Regression statistics between kernel values and time.

Discussion and conclusion

The continuity of space presents children with the problem of extending a limited spatial lexicon to an infinite set of spatial situations. Synthesizing work on spatial semantic categories and category extension, we have presented a framework of chaining based on stored exemplars that captures the incremental extension of spatial words to novel spatial situations. To test our models of chaining, we reconstructed developmental trajectories using linguistic data from CHILDES and we described a method to construct the semantic representations of observed exemplars. Our results provide preliminary evidence that the formation of early spatial categories takes place through a process of chaining that resembles the historical process of category growth.

In our work, we represented spatial situations using visual embeddings for the figure and ground objects involved. The motivation for our construction scheme was that it facilitates the link between linguistic signals and visual representations. In practice, it allowed us to approximate spatial situations in a child's environment directly using linguistic data from CHILDES. In theory, we only had to assume the child understands the visual properties of the figure and ground objects in usages of spatial words, and we did not have to assume the situation was visible when they were observing spatial word usages. Spatial situations described by spatial words are not always visible; for example, consider the exemplar from CHILDES, *you might have a crown in the other room.*

However, since our vector representation of spatial situations was a concatenation of object embeddings, the vector for the situation did not encode any relation between the two objects. This stands in contrast with previous computational work where spatial relations were encoded in representations of spatial situations (Regier, 1996; Xu & Kemp, 2010; Beekhuizen et al., 2014). One consequence of figure-ground independence is any pair of distinct situations that happen to contain the same figure and ground would be represented identically by our approximation. For example, let v_{hand}, v_{mouth} be the visual embeddings for *hand* and *mouth*, respectively, then *hand on mouth* and *hand in mouth* will be encoded identically as (v_{hand}, v_{mouth}). Although they seem relatively rare at least for *in* and *on* (see *Data*), it was impossible for our models to predict all such CS utterances correctly since they only had information on figure and ground. In future work, we would like to test if our models can be improved by using exemplars of full visual scenes beyond independent objects.

Here we focused on analyzing spatial adpositions and their spatial semantic categories. Thus, we sought to exclude non-spatial usages of these adpositions from our analyses by intersecting the extracted figure and ground objects with ImageNet. Although this excluded non-spatial usages where the objects are not physical (e.g., *dream in mind*), it still left us with false negatives where the objects are physical (e.g., *book on animal*, where *on* means *about*). In future work, one way to address this limitation is to further intersect our dataset with figure-ground-preposition triples grounded in real spatial situations, such as annotated naturalistic images. Another potential future direction would be to extend our framework to non-spatial usages of adpositions, where our current visual embeddings would likely be insufficient for modelling abstract usages.

Our computational framework is inspired by previous computational formalisms of chaining, but it differs from previous approaches in two main aspects. First, our framework formalizes the process of chaining at the agent level over a developmental trajectory, whereas previous frameworks have either formalized chaining as a (population-level) historical process (Ramiro et al., 2018; Habibi et al., 2020; Grewal & Xu, 2020) or not yet described how chaining is updated by linguistic inputs over the developmental time course (Ferreira Pinto Jr. & Xu, 2019). Second, our framework extends previous work by enabling chaining models to vary the degree to which they generalize from exemplars across different categories. We might expect children to overextend or underextend unfamiliar words more often than familiar ones. This extension offers the possibility to examine varying degrees of overextension and underextension across different words in the framework of chaining.

Acknowledgments

We would like to thank Myrto Grigoroglou, Barbara Malt, Jade Yu, Zhewei Sun, and Aida Ramezani for constructive comments on an earlier version of the paper. AX is funded partly by the U of T–UoM IRTG program. YX is funded through an NSERC Discovery Grant, a SSHRC Insight Grant, and an Ontario ERA award.

References

Ambridge, B. (2020). Against stored abstractions: A radical exemplar model of language acquisition. *First Language*, *40*(5-6), 509–559.

Ashby, F. G., & Alfonso-Reese, L. A. (1995). Categorization as probability density estimation. *Journal of mathematical psychology*, *39*(2), 216–233.

Baillargeon, R., Needham, A., & DeVos, J. (1992). The development of young infants' intuitions about support. *Early development and parenting*, *1*(2), 69–78.

Beekhuizen, B., Fazly, A., & Stevenson, S. (2014). Learning meaning without primitives: Typology predicts developmental patterns. In *CogSci.*

Bowerman, M. (1978). The acquisition of word meaning: An investigation into some current conflicts. In *The development of communication* (pp. 263–287). Wiley.

Bowerman, M., & Choi, S. (2001). language-specific in the acquisition of spatial semantic categories. *Language acquisition and conceptual development*(3), 475.

Bowerman, M., & Choi, S. (2003). Space under construction: Language-specific spatial categorization in first language acquisition. *Language in mind: Advances in the study of language and thought*, 387–427.

Brugman, C., & Lakoff, G. (1988). Cognitive topology and lexical networks. In *Lexical ambiguity resolution* (pp. 477–508).

Casasola, M., & Cohen, L. B. (2002). Infant categorization of containment, support and tight-fit spatial relationships. *Developmental science*, *5*(2), 247–264.

Clark, E. V. (1973). Non-linguistic strategies and the acquisition of word meanings. *Cognition*, *2*(2), 161–182.

Clark, E. V. (1978). Strategies for communicating. *Child Development*, 953–959.

Deng, J., Dong, W., Socher, R., Li, L.-J., Li, K., & Fei-Fei, L. (2009). Imagenet: A large-scale hierarchical image database. In *CVPR* (pp. 248–255).

Estes, W. K. (1986). Array models for category learning. *Cognitive Psychology*, *18*(4), 500–549.

Fellbaum, C. (1998). *Wordnet: An electronic lexical database*. Cambridge, MA: MIT Press.

Ferreira Pinto Jr., R., & Xu, Y. (2019). Children's overextension as communication by multimodal chaining. In *CogSci.*

Ferreira Pinto Jr., R., & Xu, Y. (2021). A computational theory of child overextension. *Cognition*, *206*, 104472.

Golland, D., Liang, P., & Klein, D. (2010). A game-theoretic approach to generating spatial descriptions. In *EMNLP.*

Grewal, K., & Xu, Y. (2020). Chaining and historical adjective extension. In *Cogsci.*

Habibi, A. A., Kemp, C., & Xu, Y. (2020). Chaining and the growth of linguistic categories. *Cognition*, *202*, 104323.

Johannes, K., Wilson, C., & Landau, B. (2016). The importance of lexical verbs in the acquisition of spatial prepositions: The case of in and on. *Cognition*, *157*, 174–189.

John, G. H., & Langley, P. (1995). Estimating continuous distributions in Bayesian classifiers. In *UAI.*

Kay, C., Roberts, J., Samuels, M., & Wotherspoon, I. (2017). *The Historical Thesaurus of English, version 4.21.* Glasgow, UK: University of Glasgow. Retrieved from `http://historicalthesaurus.arts.gla.ac.uk/`

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *ICLR.*

Lakoff, G. (1987). *Women, fire, and dangerous things: What categories reveal about the mind*. University of Chicago press.

Landau, B. (2018). Learning simple spatial terms: Core and more. *Topics in Cognitive Science*, *12*(1), 91–114.

Landau, B., & Jackendoff, R. (1993). "What" and "where" in spatial language and spatial cognition. *Behavioral and Brain Sciences*, *16*, 217–217.

Levinson, S., Meira, S., Language, T., & Group, C. (2003). 'natural concepts' in the spatial topological domain-adpositional meanings in crosslinguistic perspective: An exercise in semantic typology. *Language*, 485–516.

Loper, E., & Bird, S. (2002). Nltk: The natural language toolkit. In *Proceedings of the ACL workshop on effective tools and methodologies for teaching natural language processing and computational linguistics.*

MacWhinney, B. (2000). *The CHILDES project: Tools for analyzing talk. transcription format and programs* (Vol. 1). Psychology Press.

Nosofsky, R. M. (1986). Attention, similarity, and the identification–categorization relationship. *Journal of Experimental Psychology: General*, *115*(1), 39.

Ramiro, C., Srinivasan, M., Malt, B. C., & Xu, Y. (2018). Algorithms in the historical emergence of word senses. *Proceedings of the National Academy of Sciences*, *115*(10), 2323–2328.

Reed, S. K. (1972). Pattern recognition and categorization. *Cognitive Psychology*, *3*(3), 382–407.

Regier, T. (1996). *The human semantic potential: Spatial language and constrained connectionism*. MIT Press.

Simonyan, K., & Zisserman, A. (2015). Very deep convolutional networks for large-scale image recognition. In *ICLR.*

Vygotsky, L. S. (1962). *Language and thought*. MIT press.

Xu, Y., & Kemp, C. (2010). Constructing spatial concepts from universal primitives. In *CogSci.*

Xu, Y., Regier, T., & Malt, B. C. (2016). Historical semantic chaining and efficient communication: The case of container names. *Cognitive Science*, *40*(8), 2081–2094.