The development of abstract concepts in children's early lexical networks

Anonymous NAACL submission

Abstract

How do children learn abstract concepts such as animal vs. artifact? Previous research has suggested that such concepts can partly be derived using cues from the language children hear around them. Following this suggestion, we propose a model where we represent the children's developing lexicon as an evolving network. The nodes of this network are based on vocabulary knowledge as reported by parents, and the edges between pairs of nodes are based on the probability of their co-occurrence in a corpus of child-directed speech. We found that several abstract categories can be identified as the dense regions in such networks. In addition, our simulations suggest that these categories develop simultaneously, rather than sequentially, thanks to the children's word learning trajectory which favors the exploration of the global conceptual space.

1 Introduction:

One of the central challenges in cognitive development is to understand how concepts develop (Carey, 2009; Keil, 1992; Gopnik and Meltzoff, 1997). Of particular interest is the case of abstract concepts which have non-obvious shared properties such as "animal" and "artifact". For example, a cat and a bird are perceptually quite different but they share some fundamental properties (e.g., breathing, feeding, and reproducing) which make them animals (as opposed to artifacts). In such cases, learning requires in part cultural/linguistic cues which provide information beyond what can be obtained through the senses (Gelman, 2009; Harris, 2012; Csibra and Gergely, 2009). One way children's conceptual learning can benefit from the language they hear around them is through word co-occurrence. For example, one can learn an abstract concept (e.g., animal) simply by observing how its instances (e.g., "cat" and "bird") go together in speech. Indeed, previous work has shown that the caregiver's input contains rich co-occurrence information about various abstract concepts (Huebner and Willits, 2018). This work, however, has explored the conceptual space from an adult perspective (using the words uttered by the caregivers). Here we explore how abstract concepts may develop from the children's perspective, investigating how their word learning trajectory influences the higher-level organization of their developing lexicon.

We study development in light of two hypothetical models. On the first, called the simultaneous model, learning starts by exploring all parts of the conceptual space; categories are refined simultaneously over development. On the second, called the sequential model, learning starts by exploring a small region of the conceptual space (e.g., the category "animals") and only after the refinement of this category, does the learner move to another.

The paper is organized as follows. First, we describe the research strategy. In brief, we represented the developing lexicon as an evolving network and we used word co-occurrence in parent speech as a measure of words' relatedness. We operationalized abstract concepts as the highly interconnected regions of the network. Second, we explore how the pattern of children's word learning influences higher-level conceptual development, and whether this development corresponds to a simultaneous or a sequential conceptual growth.
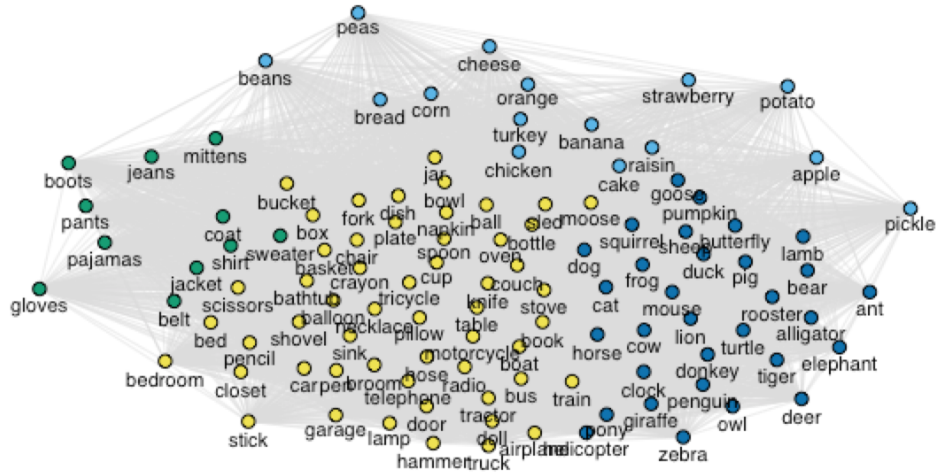
Figure 1: Network obtained using a sample of nouns in CDI data (nodes), and co-occurrence-based similarity from a corpus of child-directed speech (edges). Colors indicate highly interconnected clusters identified using unsupervised network community detection. The clusters correspond, overall, to four higher-level concepts: animal, food, clothes, and artifacts.

2 Data and Methods

2.1 Constructing lexical networks

The networks' nodes were nouns from Word-bank (Frank et al., 2017), an open repository aggregating cross-linguistic developmental data of the MacArthur-Bates Communicative Development Inventory (CDI), a parent report vocabulary checklist, Toddler version (Fenson et al., 1994). Pairs of nouns were linked by weighted edges representing their semantic similarity derived based on co-occurrence in the corpus of child-directed speech CHILDES (MacWhinney, 2014), using Word2Vec algorithm (Mikolov et al., 2013).

First, we constructed the end-state network based on all nouns learned by the last age of acquisition. We used a subset of CDI nouns for which cross-linguistic translations are present, allowing us to explore cross-linguistic variability. We used data from the ten following languages: Croatian, Danish, English, French, Italian, Norwegian, Russian, Spanish, Swedish, and Turkish. The size of this subset varied from 314 in Russian (representing 100% of total nouns in this language) to 176 in Turkish (representing 59.26% of total nouns). Second, in order to study development towards the end-state, we constructed a different network at each month, based on the nouns that have been learned by that month.

2.2 Identifying abstract concepts in a network

We assume that abstract concepts correspond to clusters of highly interconnected nodes in the networks. We identified such clusters using Walk-Trap (Pons and Latapy, 2006), an unsupervised community detection algorithm based on the fact that a random walker tends to be trapped in dense parts of a network. Figure 1 shows the outcome of cluster identification in the end-state network in English. The algorithm obtained four major clusters corresponding to the categories of clothes, food, animal and artifacts. We refer to this end-state clustering as $\mathcal{C}^*$. To examine developmental change in the conceptual organization, we ran the cluster identification algorithm at each month of acquisition t, and we compared the resulting clustering, noted $\mathcal{C}_t$, to that of the end-state $\mathcal{C}^*$. The method of this comparison is detailed below.

2.3 Measuring conceptual development

We measure conceptual development by comparing $\mathcal{C}_t$ to $\mathcal{C}^*$ across time. We used a standard method in clustering comparison, which is based on word pairs on which the two clusterings agree or disagree (Rand, 1971; Hubert and Arabie, 1985). We quantify clustering comparison using precision $P(\mathcal{C}_t)$ and recall $R(\mathcal{C}_t)$, defined as follows:
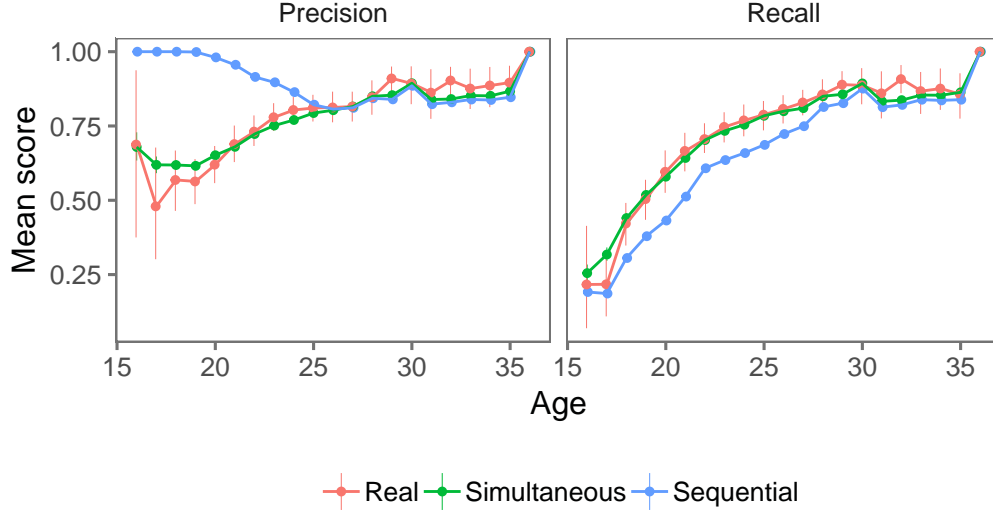
Figure 2: Mean precision and recall scores obtained through comparing the end-state clustering to clusterings at different months of acquisition, averaged across languages and numbers of clusters. Colors indicates real and hypothetical word sampling mechanisms. Errors bars represent 95% confidence intervals.

$$P(\mathcal{C}_t) = \frac{|tp(\mathcal{C}_t)|}{|tp(\mathcal{C}_t)| + |fp(\mathcal{C}_t)|}$$

$$R(\mathcal{C}_t) = \frac{|tp(\mathcal{C}_t)|}{|tp(\mathcal{C}_t)| + |fn(\mathcal{C}_t)|}$$

Where $tp(\mathcal{C}_t)$ are the true positives, defined as the word pairs that are placed in the same cluster under $\mathcal{C}_t$ and in the same cluster under $\mathcal{C}^*$. $fp(\mathcal{C}_t)$ are the false positives defined as the pairs placed in the same cluster under $\mathcal{C}_t$ and in different clusters under $\mathcal{C}^*$. Finally, $fn(\mathcal{C}_t)$ are the false positives defined as the pairs placed in different clusters under $\mathcal{C}_t$ and in the same cluster under $\mathcal{C}^*$.

We made this comparison using different degrees of clustering granularity. More precisely, we fixed the same number of clusters for *both* $\mathcal{C}_t$ and $\mathcal{C}^*$, and we varied this number from two to four clusters. We did not use the trivial case of one cluster, nor did we did use more than four clusters, since this number was optimal for the largest network (i.e., the end-state network) based on the modularity maximization criterion (Newman, 2006).

2.4 Learning mechanisms

We examined how abstract concepts develop under an average word learning trajectory derived from real developmental data. To construct this trajectory, we used the normative age of acquisition, that is, the age at which a word is produced by at least 50% of children in each language

(Goodman et al., 2008). We compared this development to the development induced by a simultaneous model and the development induced by a sequential model.

The simultaneous model was instantiated as a uniform sampling across time from the end-state vocabulary. The sequential model had the additional constraint of sampling from one category at a time: the first word is selected randomly from one cluster, subsequent words are sampled from the same cluster. After all words from this cluster are used, a word from a different cluster is chosen, and the same process is repeated until all clusters are covered.

3 Results

Figure 2 shows the scores obtained through comparing $\mathcal{C}^*$ to $\mathcal{C}_t$ at different points in time t. For the real word learning trajectory, both precision and recall start relatively low, indicating that the induced conceptual organization is initially quite different from that of the end-state. Both measures converge towards 1 (i.e., perfect score) as $\mathcal{C}_t$ becomes more and more similar to $\mathcal{C}^*$.

The simultaneous model mimics closely the patterns of real conceptual development, explaining almost all the variance in mean precision ($R^2 = 0.94$) and recall ($R^2 = 0.99$). In contrast, the sequential model had generally a higher precision, i.e., it induced fewer false positive pairs. This result is due to the fact that we sampled instances

from the same category. However, the same model had generally lower recall scores, i.e., it induced more false negative pairs. This second result was due to the fact that sampling from the same category leads to clusterings that are finer in their conceptual granularity than the end-state. As a consequence of this discrepancy with respect to real development, the sequential model explained less variance than the simultaneous model did in both its mean precision ($R^2 = 0.44$) and recall ($R^2 = 0.96$).

4 Discussion

Can children learn abstract concepts based on word co-occurrence in the language they hear around them? Previous work has shown that child-directed speech contains information about several abstract concepts (Huebner and Willits, 2018). Here we investigated when and how this information becomes available to children as their lexical network grows. We found that even with a small lexicon, several high-level concepts such as "animal", "artifact", "food" and "clothes" emerge bottom-up as clusters of highly interconnected nodes in the network. Furthermore, compared with a model that posited sequential learning, we found that these categories tended to emerge in concert with one another.

The development of the higher-level conceptual structure seems to be unaffected by the order with which words are acquired (as long as this order approximates a uniform sampling from the end-state lexicon), suggesting that the process of conceptual development can accommodate a wide range of word learning trajectories without a qualitative change in the higher-level organization. For example, whether acquisition starts first with the words "cat" and "banana" or with the words "cow" and "potato" does not qualitatively affect the higher-level organization involving "animal" and "food". This property is important as it suggests, for instance, that development is resilient to variability in the children's linguistic input (Slobin, 2014; Hart and Risley, 1995).

Developmental changes were captured by precision and recall. The increase in precision means that false positives decrease over time: some word pairs that are initially lumped together in a same category, are eventually differentiated. Similarly, the increase in recall means that false negatives decrease, that is, some word pairs that are initially

distinct, become eventually subsumed by a same category. These patterns suggest a process of conceptual reorganization involving both "differentiation" and "coalescence" as has been suggested in the developmental literature (Carey, 2009).

That said, these developmental changes were not necessarily related to specific concepts (since the patterns were similar in the simultaneous model where we randomized the order of word learning). Instead, this finding suggests that differentiation and coalescence of word pairs in our data are related to the change in the vocabulary size across development: As more words are added to their lexical network, learners may approximate better the underlying conceptual organization of the mature lexicon and would make fewer categorization errors. Indeed, research in network science indicates that properties of a real network become more distorted as the size of a sampled sub-network decreases (Leskovec and Faloutsos, 2006).

One limitation of this study is that we used the normative age of acquisition, computed using different children at different age groups. This choice was due to the cross-sectional nature of available CDI data. Though such a measure has been widely used to study important aspects of the early lexical networks (Hills et al., 2009; Stella et al., 2017; Storkel, 2009), it only applies at the population level. In our case, though we found that concepts develop simultaneously, individual children may display, at least locally, a sequential-like behavior. For example, prior knowledge about dinosaurs may enable the learning of new dinosaur-related words more easily (Chi and Koeske, 1983).

In sum, this work provided a quantitative account of how abstract concepts can emerge from the interaction of the children's emerging vocabulary and the properties of their linguistic input. One important direction for future work is to investigate the extent to which the correlational findings obtained in this study (e.g., the identity of categories formed across development or the fact that categorization errors decrease with the size of the lexicon) can be corroborated by controlled behavioral experiments.

All data and code will be available online

References

Susan Carey. 2009. *The origin of concepts*. Oxford University Press.

Michelene TH Chi and Randi Daimon Koeske. 1983. Network representation of a child's dinosaur knowledge. *Developmental psychology*, 19(1).

Gergely Csibra and György Gergely. 2009. Natural pedagogy. *Trends in cognitive sciences*, 13(4).

Larry Fenson, Philip S. Dale, J. Steven Reznick, Elizabeth Bates, Donna J. Thal, Stephen J. Pethick, Michael Tomasello, Carolyn B. Mervis, and Joan Stiles. 1994. Variability in early communicative development. *Monographs of the Society for Research in Child Development*, 59(5).

Michael C. Frank, Mika Braginsky, Daniel Yurovsky, and Virginia A. Marchman. 2017. Wordbank: an open repository for developmental vocabulary data. *Journal of Child Language*, 44(3):677–694.

Susan A Gelman. 2009. Learning from others: Children's construction of concepts. *Annual review of psychology*, 60.

Judith C. Goodman, Philip S. Dale, and Ping Li. 2008. Does frequency count? parental input and the acquisition of vocabulary. *Journal of Child Language*, 35(3):515–531.

Alison Gopnik and Andrew N Meltzoff. 1997. *Words, thoughts, and theories*. MIT Press.

Paul L Harris. 2012. *Trusting what you're told: How children learn from others*. Harvard University Press.

Betty Hart and Todd R Risley. 1995. *Meaningful differences in the everyday experience of young American children*. Paul H Brookes Publishing.

Thomas T. Hills, Mounir Maouene, Josita Maouene, Adam Sheya, and Linda Smith. 2009. Longitudinal analysis of early semantic networks: Preferential attachment or preferential acquisition? *Psychological Science*, 20(6):729–739.

Lawrence Hubert and Phipps Arabie. 1985. Comparing partitions. *Journal of classification*, 2(1).

Philip A Huebner and Jon A Willits. 2018. Structured semantic knowledge can emerge automatically from predicting word sequences in child-directed speech. *Frontiers in Psychology*, 9:133.

Frank C Keil. 1992. *Concepts, kinds, and cognitive development*. MIT Press.

Jure Leskovec and Christos Faloutsos. 2006. Sampling from large graphs. In *Proceedings of the 12th ACM SIGKDD international conference on Knowledge discovery and data mining*. ACM.

Brian MacWhinney. 2014. *The CHILDES project: Tools for analyzing talk, Volume II*. Psychology Press.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*.

Mark EJ Newman. 2006. Modularity and community structure in networks. *Proceedings of the national academy of sciences*, 103(23).

Pascal Pons and Matthieu Latapy. 2006. Computing communities in large networks using random walks. *Journal of Graph Algorithms and Applications*, 10(2).

William M Rand. 1971. Objective criteria for the evaluation of clustering methods. *Journal of the American Statistical association*, 66(336).

Dan Isaac Slobin. 2014. *The crosslinguistic study of language acquisition*, volume 4. Psychology Press.

Massimo Stella, Nicole M Beckage, and Markus Brede. 2017. Multiplex lexical networks reveal patterns in early word acquisition in children. *Scientific Reports*, 7.

Holly L. Storkel. 2009. Developmental differences in the effects of phonological, lexical and semantic variables on word learning by infants. *Journal of Child Language*, 36(2):29–321.