Representing the Toddler Lexicon: Do the Corpus and Semantics Matter?

University of Colorado Boulder

Jennifer Weber and Eliana Colunga

Department of Psychology and Neuroscience

"Meadow at Chautauqua" by Anne Gifford (http://www.annegifford.com)

Introduction

Modeling early vocabulary growth trajectories requires accurately representing the lexical structure of toddlers

But…

- Many metrics utilize adult association norms, judgements, or metrics from adult-language corpora
  - Recently use co-occurrence on CHILDES
- Adult vocabulary acquisition norms over child acquisition data
  - Recently use parent-report vocabulary checklists
- Past work used co-occurrence metrics, though distributional metrics such as Word2Vec have not been tested
- Could use network centrality measures to predict future unknown vocabulary
  - Verify using longitudinal vocabulary data

Research Questions

1. *Can we understand early language development better by approximating the language a young child growing up in an English-speaking environment might typically encounter?*
2. *Can we use a predictive neural network model to derive more accurate network representations than sliding window co-occurrence models?*

Methods – The Corpus

The first step is to create a broad corpus of toddler language input from which to derive semantic similarities

- Includes the parent input during parent-child conversations (CHILDES – MacWhinney, 2000), lab-transcribed young children's picture books, and fan-created G-rated movie transcripts (*see Table 1*)

	CHILDES	Books	Movies
Number	many	1,039	81
Sentences	1,105,870	54,213	92,919
Tokens	4,716,063	510,312	507,625
Types	27,337	5,895	5,822

Table 1: Toddler Corpus Statistics

Methods – Lexical Network Creation

The technique of algorithm used to calculate similarity statistics from the corpus could be important to representativeness as well

- **Vocabulary** (*parent-checklist*): McArthur Bates Communicative Development Inventory (MCDI)
  - Average child from 16-30 months of age
  - Averaged over ~200 children collected in-lab
- Compare sliding window similarities **(5)**
  - connected if both words occur together in 5-word window (Hills et al., 2010) - to embeddings derived from Word2Vec (co-occur in same window & in same context) for the Toddler corpus
- Further compare Word2Vec derived from GoogleNews corpus (Adult - **G**), derived from our created Toddler corpus (child - **T**), or a combined Word2Vec trained on GoogleNews and fine-tuned on our Toddler corpus (**C**).
- **Create lexical networks** – each node = 1 word, and the connections between words = strength of the semantic similarity (based on co-occurrence or Word2Vec)

Centrality Measures

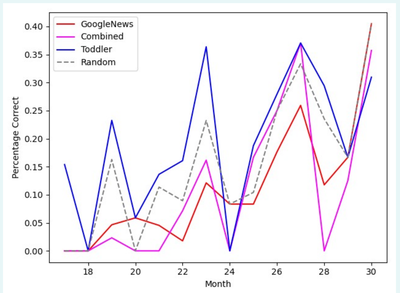
Create lexical networks representing the similarity between words known by a typical child

- connections represent either the number of times the word co-occurred with another or the cosine similarity from Word2Vec

- **PageRank:** uses quantity and quality to determine importance of any one node
- **Degree:** weighted number of connections to each known node
- **Clustering Coefficient:** degree to which a node clusters to others
- **Load Centrality:** fraction of shortest paths which pass through a particular node
- **Eigenvector Centrality:** measure of influence of a node
- **Edge Weight:** more similar words have stronger wights between them

Results – Comparing Models to Each Other

Toddler > GoogleNews & Combined for load centrality



Load Centrality Network Comparison

Marginal differences:

T > G for Eigenvector Centrality & Edge Weight

Comparing to sliding window (5): **T > 5** for every measure

The Toddler corpus not only performed better than random, but better than the other models!

Network Overlap

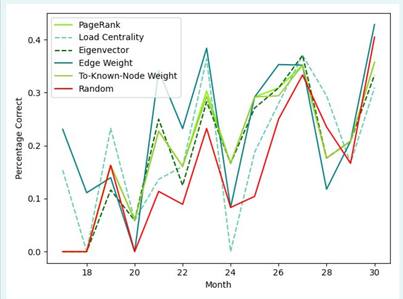
Top 50					Bottom 50				
Centrality Measure	**TvG**	**TvC**	**GvC**	**Tv5**	**Centrality Measure**	**TvG**	**TvC**	**GvC**	**Tv5**
PageRank	3	9	7	7	*PageRank*	5	5	0	5
(Weighted) Degree	6	9	7	0	*(Weighted) Degree*	5	3	0	0
Clustering Coefficient	5	11	9	0	*Clustering Coefficient*	6	2	0	1
Load Centrality	3	7	9	1	*Load Centrality*	4	9	2	10
Eigenvector Centrality	5	11	9	0	*Eigenvector Centrality*	6	2	0	1

Overlap measures if any unique word appears in one of the other two corpora (or both), but the word does not have to appear in all three to be counted in this overlap measure. Expected examples of words that appeared in multiple top 50 lists include: *monkey, balloon, cookie, blanket, puppy*, and *spoon*. Some words were more unexpected, such as *tractor, pumpkin, and rooster.*

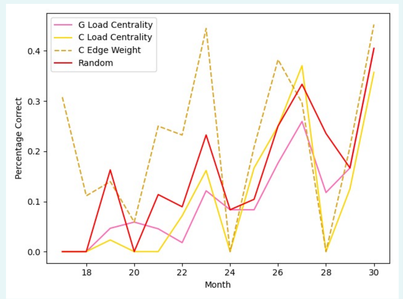
Results – Comparing to Random

For each unknown word, we added it and all its connections into the network of known words. We calculated the unknown words influence through each measure before removing it and moving to the next unknown word. Based on the number of words actually learned during the month, we predicted the top influencing words as those the network would learn. Accuracy was based on the percentage of words the network predicted that were words the average child actually learned that month.

Toddler > Random for every measure used (2 only marginally)



Toddler Network Compared to Random

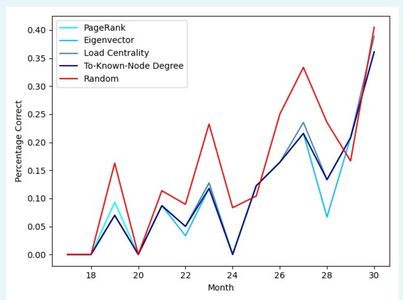


GoogleNews and Combined Versus Random

GoogleNews & Combined < Random for load centrality
G & C = Random for all others

Sliding Window (5) < Random for every measure

Only Word2Vec embeddings derived from the Toddler corpus predicted language development better than random, and sliding window co-occurrences performed worse!



Sliding Window (5) Versus Random

Discussion

Using toddler input corpora, word embeddings and similarities drawn from neural network models such as Word2Vec, and fully-connected, weighted networks can provide a level of accurate word-learning prediction better than random chance, embeddings trained on adult-language corpora, and toddler sliding-window co-occurrence similarities.

- Expand and generalize the present Toddler corpus
  - Cultural, language differences (presently North-American English)
    - Children growing up in multi-lingual environments
    - Children growing up with different child-rearing practices
      - No screen media, differing amounts of conversational input
  - Children with language, cognitive or sensory disorders
- Other predictive models using same Word2Vec embeddings and network measures
  - Logistic regression, predictive neural networks
  - Compare to other predictive models
    - preferential attachment growth
- Use these models to theorize about developmental mechanisms
  - Create learning materials and help inform interventions

Our analyses not only suggest the need to be mindful when choosing similarity metrics or semantic network structure, but highlight the importance of achieving a high degree of representativeness for different populations.

References

Beckage, N., Smith, L., & Hills, T. (2011). Small worlds and semantic network growth in typical and late talkers. PloS one, 6(5), e19348.

Fenson, L., Dale, P. S., Reznick, J. S., Bates, E., Thal, D. J., Pethick, S. J., ... & Stiles, J. (1994). Variability in early communicative development. Monographs of the society for research in child development, i-185.

Hills, T. T., Maouene, J., Riordan, B., & Smith, L. B. (2010). The associative structure of language: Contextual diversity in early word learning. Journal of memory and language, 63(3), 259-273.

MacWhinney, B. (2000). The CHILDES Project: Tools for analyzing talk. Third Edition. Mahwah, NJ: Lawrence Erlbaum Associates.