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

Jennifer Weber and Eliana Colunga
Department of Psychology and Neuroscience

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)

<table>
<thead>
<tr>
<th>CHILDES</th>
<th>Books</th>
<th>Movies</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number</td>
<td>many</td>
<td>1,039</td>
</tr>
<tr>
<td>Sentences</td>
<td>1,105,870</td>
<td>54,213</td>
</tr>
<tr>
<td>Tokens</td>
<td>4,716,063</td>
<td>50,312</td>
</tr>
<tr>
<td>Types</td>
<td>27,337</td>
<td>2,925</td>
</tr>
</tbody>
</table>

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 (G & 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 weights between them

Results – Comparing Models to Each Other
Toddler > GoogleNews & Combined for load centrality
- Marginal differences: T > G for Eigenvector Centrality & Edge Weight
- Comparing to sliding window (5): T > S for every measure
- The Toddler corpus not only performed better than random, but better than the other models!

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

References