FREQUENCY AND THE LEXICAL ORGANIZATION OF VERBS

Daniel Cerato Germann, Aline Villavicencio, Maity Siqueira

Federal University of Rio Grande do Sul (Brazil)
Goals

 To investigate the influence of linguistic factors on lexical acquisition and organization
 To build linguistic profiles to help diagnose pathologies that affect language
Overview

- Related Work
- Materials and Methods
- Results
- Conclusions and Future Work
Related Work

- Some factors that may affect language acquisition:
  - Frequency
  - Polysemy
  - Syntactic structure
  - Concreteness
  - Specificity
  - Conventionality
  - Figurativeness
Related Work

- Goldberg's (1999): crosslinguistically polysemic and frequent verbs acquired early
- Korhonen and Buttery (2007): verb type preferences influence acquisition of subcategorization frames
- Graph theory and network analysis methods useful in mapping differences in associative structures across groups.
- Steyvers and Tenenbaum (2005): model of semantic network growth compatible with effects of learning history variables (age of acquisition and frequency) in semantic processing tasks
Goals

- To investigate the influence of linguistic factors on lexical acquisition and organization
- To build linguistic profiles to help diagnose pathologies that affect language
Materials and Methods

- Longitudinal study
- Data from psycholinguistic Naming Task:

  - 17 videos of division or destruction actions
  - Each action: 1 actor and 1 object
55 Participants per group

- G1: 2;0 to 3;11 year old (M=3;1)
- G2 (G1 after 2 years): 4;1 to 6;6 year old (M=5;5)
- G3: 17;0 to 34;0 (M=21;8)

Native speakers of Brazilian Portuguese

Verbs annotated with frequency:

- Fscore: occurrences in “Florianópolis” corpus (Scliar-Cabral, 1993; MacWhinney, 2000)
Verbs

- Manually pre-processed to remove noise
  - Unrelated (*I don't know*) or low frequency (<=1) answers

<table>
<thead>
<tr>
<th>Answers</th>
<th># per group</th>
<th># per action</th>
<th># types</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>785/935</td>
<td>46.18</td>
<td>6.76</td>
</tr>
<tr>
<td>G2</td>
<td>911/935</td>
<td>53.59</td>
<td>5.53</td>
</tr>
<tr>
<td>G3</td>
<td>917/935</td>
<td>53.94</td>
<td>4</td>
</tr>
</tbody>
</table>
Materials and Methods

- Verb interactions represented in semantic networks
  - Structural properties of lexicon
    - Network analysis and modification
  - Contents of lexicon
    - Similarity measure
Materials and Methods

- Semantic Network
  - For each film, verbs form a clique

Film1: Smash, Crash, Break
Film2: Break, Saw, Cut
Materials and Methods

- Semantic Network
  - For each film, verbs form a clique

- Smash
- Crash
- Break
- Saw
- Cut
Materials and Methods

- Semantic Network
  - For each film, verbs form a clique

  ![Semantic Network Diagram]

  - Smash
  - Crash
  - Break
  - Saw
  - Cut
Materials and Methods

- Semantic Network
  - For each film, verbs form a clique

Diagram:
- Smash
  - Crash
  - Break
  - Break
  - Saw
  - Cut
Materials and Methods

- Semantic Network
  - For each film, verbs form a clique

- Diagram:
  - Smash
  - Crash
  - Break
  - Saw
  - Cut
Materials and Methods

- For each group, one semantic network
- Compare structure and content of groups
  - Will removing less frequent verbs from older group, result in convergence of overall structure to that of a younger group?
  - Will we see increase in similarity of content/structure?
  - Network modification
Materials and Methods

- For each group, one semantic network
- Compare structure and content of groups
  - Will removing less frequent verbs from older group, result in convergence of overall structure to that of a younger group?
    - Will we see increase in similarity of content/structure?
  - Network modification
    - Starting with older group and removing verbs in increasing order of frequency
Materials and Methods

- For each group, one semantic network
- Compare structure and content of groups
  - Will removing less frequent verbs from older group, result in convergence of overall structure to that of a younger group?
    - Will we see increase in similarity of content/structure?
  - Network modification
    - Starting with older group and removing verbs in increasing order of frequency
Materials and Methods

- Measure similarity through:
  - Jaccard coefficient
  - Number of vertices \((n)\)
  - Average minimal path length \((L)\)
  - Density \((D)\)
  - Number of edges \((M)\)
  - Average clustering coefficient \((C/s)\)
  - Average node connectivity \((k)\)
  - Average number of repetitions \((r)\)
Materials and Methods

- Measure similarity through:
  - Jaccard coefficient
  - Number of vertices \((n)\)
  - Average minimal path length \((L)\)
  - \textit{Density} \((D)\)
  - Number of edges \((M)\)
  - Average clustering coefficient \((C/s)\)
  - \textit{Average node connectivity} \((k)\)
  - Average number of repetitions \((r)\)

Measures of proportion of
# of edges vs # of nodes
(semantic share)
Materials and Methods

- Measure similarity through:
  - Jaccard coefficient
  - Number of vertices \( (n) \)
  - Average minimal path length \( (L) \)
  - Density \( (D) \)
  - Number of edges \( (M) \)
  - Average clustering coefficient \( (C/s) \)
  - Average node connectivity \( (k) \)
  - Average number of repetitions \( (r) \)

Measures of vocabulary structuring (small distances, more repetition and high clustering → unstructured lexicon)
**Materials and Methods**

- Frequency decreases as age increases:

<table>
<thead>
<tr>
<th>Answers</th>
<th>Average type</th>
<th>Average token</th>
</tr>
</thead>
<tbody>
<tr>
<td>G1</td>
<td>44.05</td>
<td>43.44</td>
</tr>
<tr>
<td>G2</td>
<td>35.92</td>
<td>35.71</td>
</tr>
<tr>
<td>G3</td>
<td>17.84</td>
<td>21.22</td>
</tr>
</tbody>
</table>
## Results

- Increase in vocabulary size and decrease of repetition with age

<table>
<thead>
<tr>
<th></th>
<th>G1</th>
<th>G2</th>
<th>G3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Number of vertices ((n))</td>
<td>22</td>
<td>25</td>
<td>31</td>
</tr>
<tr>
<td>Average minimal path length ((L))</td>
<td>1.46</td>
<td>1.6</td>
<td>1.98</td>
</tr>
<tr>
<td>Density ((D))</td>
<td>0.55</td>
<td>0.42</td>
<td>0.27</td>
</tr>
<tr>
<td>Number of edges ((M))</td>
<td>128</td>
<td>126</td>
<td>126</td>
</tr>
<tr>
<td>Average clustering coefficient ((C/s))</td>
<td>0.84</td>
<td>0.78</td>
<td>0.78</td>
</tr>
<tr>
<td>Average node connectivity ((k)), (\mu = 11.64, SD = 6.73)</td>
<td>(\mu = 10.08, SD = 4.86)</td>
<td>(\mu = 8.13, SD = 4.76)</td>
<td></td>
</tr>
<tr>
<td>Average number of repetitions ((r)), (\mu = 5.23, SD = 4.41)</td>
<td>(\mu = 3.76, SD = 3.15)</td>
<td>(\mu = 2.19, SD = 1.58)</td>
<td></td>
</tr>
</tbody>
</table>
Results

- Networks get structurally closer.

- Networks get closer in content:
  - Jaccard score: G1-G2 = 0.52 G2-G3 = 0.44
Results

- Frequency decreases with age
- Vocabulary grows and specializes with age
- Children’s groups more similar than adults.
Conclusions and Future Work

- Investigation of influence of linguistic factors on lexical acquisition and organization
  - Frequency
    - Different network topology for different groups
    - Different content for different groups

- Build linguistic profiles to help diagnose pathologies that affect language
Conclusions and Future Work

- Test other factors: polyssemy, concreteness, syntactic complexity
- Investigate lexical dissolution in pathologies (Alzheimer’s disease)
- Obtain larger data sets
- Test with other categories (nouns, expressions,...)
Acknowledgments

Research partly supported by CNPq (Projects 479824/2009-6 and 309569/2009-5), FINEP and SEBRAE (COMUNICA project FINEP/SEBRAE 1194/07).


