Polysemy: a new bias in child language acquisition?

Bernardino Casas
Neus Català
Antoni Hernández-Fernández
Ramon Ferrer-i-Cancho
Jaume Baixeries

QUALICO 2018
July 5-8, Wroclaw, Poland
Polysemy: a new bias in child language acquisition?
Contents

1. Introduction
2. Materials and Methods
3. Results
4. Conclusions and future work
Learning vocabulary

- Children are exposed to thousands and thousands of word tokens through an accumulation of small interactions grounded in context.
- During the word acquisition process of a child, some words are learned first instead of others.
- What is the reason for some word is learned before than another one?
Biases in the learning process

Many biases have been hypothesized in order to explain why some words are learned earlier:

- frequency (Goodman, Dale, & Li, 2008)
- basic taxonomic level, i.e. less generic words (Mervis, 1987; Tomasello, 2000)
- neighborhood density, i.e. number of words that sound similar to a given word (Storkel, 2004; Storkel, Armbruster, & Hogan, 2006)
- nouns, e.g., children learn first nouns before verbs (Gentner, 1982, 2006)
- …

We hypothesize that polysemy is a factor that influences the vocabulary learning process of children.
What is polysemy?

Definition 1: Polysemy

*Polysemy* is the capacity of a word to have multiple meanings.

- Our measure of polysemy of a word is the number of meanings that have this word according the syntactic category.
- For example, the word *cat*:
  - as noun has 8 meanings → polysemy 8
  - as verb has 2 meanings → polysemy 2
Main goals

First goal
Investigate the relationship between the polysemy of a word and child age for English.

Second goal
Validate the nouns bias in language acquisition for English.
When exploring biases in child speech, it is crucial to distinguish between biases which are genuine and biases which may just be the result of mimicking some form of adult speech.

Motherese is the special language that adults use to speak with children.

The adult speech is the target stage in the evolution of the verbal production of a child.

If we take as usual adult speech as control then it is crucial to know what kind of input is receiving the child and if the adult is performing adaptive efforts.
Fighting versus Motherese

IMPORTANT!!
We study each adult role that appears in a conversation with a child separately. Thus, we improve the perspective of the use of language in the adults for every role: MOTHER, FATHER and INVESTIGATOR.
Contents

1. Introduction

2. Materials and Methods

3. Results

4. Conclusions and future work
Princeton WordNet is a large lexical database that we use to obtain a measure of potential polysemy (number of meanings of a word taking into account the syntactic category of that word). WordNet has 155,287 lemmas and 117,659 synsets (meanings).

SemCor corpus is a syntactically tagged corpus that we use to obtain a more real measure of polysemy (used meanings).

Treetagger is a tool that annotates text with part-of-speech (POS) and lemma information (canonical form).
Research Materials II

**CHILDES** database contains transcriptions of conversations between children and adults (usually mothers and sons). The data source of this study are CHILDES conversations in English. Children age between 10 and 70 months aprox.

<table>
<thead>
<tr>
<th>Role</th>
<th>Types</th>
<th>Tokens</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>WN</td>
</tr>
<tr>
<td>Children</td>
<td>9,187</td>
<td>4,702</td>
</tr>
<tr>
<td>Mother</td>
<td>13,225</td>
<td>6,996</td>
</tr>
<tr>
<td>Father</td>
<td>6,808</td>
<td>4,221</td>
</tr>
<tr>
<td>Investig</td>
<td>3,977</td>
<td>2,793</td>
</tr>
</tbody>
</table>
Research Methods

- We analyze syntactically every conversation (using Treetagger) and we extract the words that someone "knows" in every point of time (age of children).
- We take into account only content words (nouns, verbs, adjectives and adverbs).
- The distribution of the data over time is not homogeneous (some time points have more data than other).
- Two ways to calculate the polysemy of a word: by using WordNet or SemCor.
Evolution of mean polysemy vs child age

breakpoint:

31.6 ± 0.1 months  
31.4 ± 0.1 months
Correlations between mean polysemy and child age

WordNet

SemCor

S+ : Significant positive correlations
S- : Significant negative correlations
S? : Non Significant correlations

Bernardino Casas  UPC-BarcelonaTech
Polysemy: a new bias in child language acquisition?
Validating results using ANOVA I

In order to validate the results, we have divided the timeline into ten segments of the same time size.

In each segment, we compare the mean polysemy (both WordNet and SemCor) of children with that of adults.

The analysis is based on a one-way independent samples ANOVA with role (two levels: children and adults) as fixed factor and the identifier of the individual as random factor.
## ANOVA test results

<table>
<thead>
<tr>
<th>segment</th>
<th>from / to</th>
<th>N</th>
<th>F</th>
<th>p-val</th>
<th>F</th>
<th>p-val</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11.9 / 17.4</td>
<td>14</td>
<td>13.331</td>
<td>0.002</td>
<td>18.141</td>
<td>$&lt; 10^{-3}$</td>
</tr>
<tr>
<td>2</td>
<td>17.4 / 22.9</td>
<td>99</td>
<td>158.401</td>
<td>$&lt; 10^{-21}$</td>
<td>241.664</td>
<td>$&lt; 10^{-27}$</td>
</tr>
<tr>
<td>3</td>
<td>22.9 / 28.4</td>
<td>139</td>
<td>90.728</td>
<td>$&lt; 10^{-16}$</td>
<td>173.963</td>
<td>$&lt; 10^{-25}$</td>
</tr>
<tr>
<td>4</td>
<td>28.4 / 33.9</td>
<td>139</td>
<td>35.397</td>
<td>$&lt; 10^{-7}$</td>
<td>82.031</td>
<td>$&lt; 10^{-15}$</td>
</tr>
<tr>
<td>5</td>
<td>33.9 / 39.4</td>
<td>127</td>
<td>11.723</td>
<td>$&lt; 10^{-3}$</td>
<td>31.37</td>
<td>$&lt; 10^{-7}$</td>
</tr>
<tr>
<td>6</td>
<td>39.4 / 44.9</td>
<td>96</td>
<td>7.659</td>
<td>0.003</td>
<td>15.817</td>
<td>$&lt; 10^{-4}$</td>
</tr>
<tr>
<td>7</td>
<td>44.9 / 50.3</td>
<td>25</td>
<td>3.827</td>
<td>0.031</td>
<td>12.408</td>
<td>$&lt; 10^{-3}$</td>
</tr>
<tr>
<td>8</td>
<td>50.3 / 55.8</td>
<td>18</td>
<td>1.751</td>
<td>0.102</td>
<td>8.731</td>
<td>0.005</td>
</tr>
<tr>
<td>9</td>
<td>55.8 / 61.3</td>
<td>48</td>
<td>0.859</td>
<td>0.179</td>
<td>3.313</td>
<td>0.038</td>
</tr>
<tr>
<td>10</td>
<td>61.3 / 66.8</td>
<td>9</td>
<td>0.005</td>
<td>0.473</td>
<td>0.602</td>
<td>0.232</td>
</tr>
</tbody>
</table>

\[ H_0 = "" (null hypothesis) and \ H_1 = "" (alternative hypothesis) \]
Validating results using ANOVA II

- The results of this test show that the average WordNet polysemy in children is significantly lower than that of adults until the 7th segment for WordNet polysemy, with an F-value that decreases gradually at each segment.
- The results of this test show the same pattern for SemCor polysemy: the mean SemCor polysemy in children is significantly lower than adults until the 9th segment.
We have shown that polysemy tends to increase markedly over time in children compared to adults.

Children exhibit a two-phase (fast- slow) growth of the mean polysemy, delimited by a breakpoint (31th month).

In adults there is neither clear positive nor negative tendency.

WordNet and SemCor polysemy show a similar trend.
Percentage of word tokens over time

**CHILDREN**

- adjectives
- nouns
- adverbs
- verbs

**MOTHER**

- adjectives
- nouns
- adverbs
- verbs

breakpoint:
- nouns: 30.0 ± 0.1 months
- verbs: 33.0 ± 0.1 months
The percentage of nouns decreases over time in children: it starts at 80%, drops to 40% and then, stabilizes.

Verbs exhibit an opposite tendency in children: they start at 10% and increase their contribution to 40%, and finally they stabilize.

These results are consistent with a well-known phenomenon: children tend to learn nouns earlier than verbs (Saxton, 2010).
Controlling by syntactic category

We have analyzed the conversations considering only those tokens that belong to a specific morpho-syntactic category: noun, adjective, verb or adverb.
Evolution of nouns vs child age

WordNet

SemCor

Age (in months)

Mean polysemy (WordNet)

Father
Investig
Mother
Children

Age (in months)

Mean polysemy (SemCor)

Father
Investig
Mother
Children

Polysemy: a new bias in child language acquisition?
Correlations between nouns vs child age

WordNet

SemCor

S+ : Significant positive correlations
S- : Significant negative correlations
S? : Non Significant correlations
Evolution of verbs vs child age

WordNet

SemCor

Polysemy: a new bias in child language acquisition?
Correlations between verbs and child age

WordNet

SemCor

S+ : Significant positive correlations
S- : Significant negative correlations
S? : Non Significant correlations
The tendency for mean polysemy to increase in children seems to blur if we control by syntactic category (nouns, verbs, adjectives and adverbs).

Verbs polysemy is higher than nouns polysemy.
Contents

1. Introduction

2. Materials and Methods

3. Results

4. Conclusions and future work

Polysemy: a new bias in child language acquisition?
Preference for non-polysemous words

The potential preference for non-polysemous words may be:

- **a standalone bias.** This hypothesis would be consistent with the lower uncertainty for those words with respect to their meaning, a factor that may reduce the cognitive cost of learning them as the cost for the listener would be smaller (Zipf, 1949).

- **a side-effect of another bias,** for instance an initial preference for nouns, the so-called noun-bias (Gentner, 1982, 2006; McDonough, Song, Hirsh-Pasek, Golinkoff, & Lannon, 2011) combined with the fact that nouns have lower polysemy (Fausey, Yoshida, Asmuth, & Gentner, 2006).
Preference for nouns or for low polysemy?

Children learn firstly nouns because they prefer nouns or maybe are they choosing nouns by the low polysemy of this category?
Conclusions I

- There is a non-trivial pattern in the evolution of polysemy over time.
- Children increase their mean polysemy in two phases: an initial phase with a fast growth of polysemy and a second phase with a slower growth of polysemy.
- In contrast, adults interacting with them do not show this tendency.
- This non-trivial pattern weakens when the analysis is segmented by syntactic category.
Conclusions II

- Children show a tendency to learn nouns first and then verbs, which is consistent with previous research (Gentner, 1982, 2006; Goodman et al., 2008).

- Verbs have a significantly higher mean polysemy than nouns in all roles: children and adults.

- The last two facts could explain the pattern of the evolution of polysemy over time to some extent.
That role of a standalone bias for low polysemy cannot be neglected. Our findings and (Crossley, Salsbury, & McNamara, 2010) suggest that L1 and L2 learners resemble each other in their dominant biases on polysemy: a preference for low polysemy words prevails in both kinds of learners.
Future work

- A deeper understanding of why the bias for low polysemy weakens within specific syntactic categories and how it interacts with frequency is a challenge for future research.
- The relationship between the senses that a speaker really knows about a word and its potential number of synsets should be investigated in detail.
Thank you for your attention!
Questions?


This work is licensed under a Creative Commons “Attribution-NonCommercial-ShareAlike 4.0 International” license.