Modelling the acquisition of lexical meaning from
caregiver-child interaction

Getting the semantics straight

Barend Beekhuizen¹, Afsaneh Fazly², Aida Nematzadeh² &
Suzanne Stevenston²

¹Leiden University  ²University of Toronto

18 January 2013
Goals

Topic
Cognitive models of acquiring word-meaning mappings

Goal #1
Discuss sources of semantic data for models and present a new one

Goal #2
Show how this data can be used to re-evaluate old claims
Cross-situational models of acquiring word meanings\(^1\)
Source of meaning: situational context.
Your average CHILDES corpus **does not contain** that.

Cross-situational models of acquiring word meanings\(^1\)

Source of meaning: situational context.

Your average CHILDES corpus does not contain that.

So: method of synthesizing semantics.

- Every word is a semantic symbol (Fazly, Alishahi & Stevenson 2010)
- Obtain lexical semantics from WordNet (id., 2008)

Allows you to make large quantities of data.

Cross-situational models of acquiring word meanings\(^1\)

Source of meaning: situational context.

Your average CHILDES corpus does not contain that.

So: method of synthesizing semantics.

- Every word is a semantic symbol (Fazly, Alishahi & Stevenson 2010)
- Obtain lexical semantics from WordNet (id., 2008)

Allows you to make large quantities of data.

But: quality of data?

- Cognitive availability of meaning?
- Situational availability? (noise, referential uncertainty)

---

Cross-situational models of acquiring word meanings\(^1\)

Source of meaning: situational context.

Your average CHILDES corpus does not contain that.

So: method of synthesizing semantics.

- Every word is a semantic symbol (Fazly, Alishahi & Stevenson 2010)
- Obtain lexical semantics from WordNet (id., 2008)

Allows you to make large quantities of data.

But: quality of data?

- Cognitive availability of meaning?
- Situational availability? (noise, referential uncertainty)

Recent method: annotating video material (Yu, Roy, Frank)

But: either limited to basic-level objects or in the pragmatic realism (explicit labeling task).

Data!

Goal #1

Provide **situational descriptions** (of properties, objects, relations, actions) for a dataset of videotaped caregiver-child interaction that can function as a source for acquiring (first) word meanings.
Goal #1

Provide situational descriptions (of properties, objects, relations, actions) for a dataset of videotaped caregiver-child interaction that can function as a source for acquiring (first) word meanings.

- Some desiderata:
  - Children should be *young enough* not to know too much already.
  - Coded descriptions should be *cognitively available*.
  - Coded descriptions should stay close to what’s *observable*; the coders should not have to infer too much.
Goal #1

Provide situational descriptions (of properties, objects, relations, actions) for a dataset of videotaped caregiver-child interaction that can function as a source for acquiring (first) word meanings.

- Some desiderata:
  - Children should be young enough not to know too much already.
  - Coded descriptions should be cognitively available.
  - Coded descriptions should stay close to what’s observable; the coders should not have to infer too much.

- Realizations:
  - High-quality data can only complement high-quantity data, not replace it.
  - Little earlier work: the specifics may contain serious methodological flaws (more than happy to find out!)
The block game corpus

- ± 120 90-min videos of mother-daughter (16mo) interaction, gathered by Child Studies in Leiden
- Every dyad played a game of putting differently-shaped blocks in a bucket through corresponding holes
- 32 dyads (± 5 min. each) were situationally coded by two coders using ELAN and transcribed by first author
The block game corpus

- ± 120 90-min videos of mother-daughter (16mo) interaction, gathered by Child Studies in Leiden
- Every dyad played a game of putting differently-shaped blocks in a bucket through corresponding holes
- 32 dyads (± 5 min. each) were situationally coded by two coders using ELAN and transcribed by first author
- **175 minutes** of material, **7842 word tokens**, **2492 utterances**.
- **Situational coding**. For every interval of 3 seconds, code:
  - simple behavior \(\text{grab, move, position, letgo}\)
  - changes in spatial relations \(\text{in, on, out, off, match}\)
  - objects \(\text{block, bucket, mother, table}\)
  - properties \(\text{triangular, square, red, blue}\)
- **Structured**: \(\text{grab(mother, (red, square, block))}\)
- **High intra- & interannotator agreement** (almost all \(\kappa > 0.8\))
Table: A sample of the dataset. The dash-separated abbreviations denote blocks and holes and their properties (colors & shapes)

<table>
<thead>
<tr>
<th>time</th>
<th>type</th>
<th>coding/transcription</th>
</tr>
</thead>
<tbody>
<tr>
<td>0m0s</td>
<td>situation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>language</td>
<td>een. nou jij een.</td>
</tr>
<tr>
<td></td>
<td>translation</td>
<td>“One. Now you try one.”</td>
</tr>
<tr>
<td>0m3s</td>
<td>situation</td>
<td>position(mother, toy, on(toy, floor)) grab(child, b-ye-tr) move(child, b-ye-tr, on(b-ye-tr, floor), near(b-ye-tr, ho-ro)), mismatch(b-ye-tr, ho-ro)</td>
</tr>
<tr>
<td></td>
<td>language</td>
<td>nee daar.</td>
</tr>
<tr>
<td></td>
<td>translation</td>
<td>“No, there.”</td>
</tr>
<tr>
<td>0m6s</td>
<td>situation</td>
<td>point(mother, ho-tr, child) position(child, b-ye-tr, near(b-ye-tr, ho-ro)) mismatch(b-ye-tr, ho-ro)</td>
</tr>
<tr>
<td></td>
<td>language</td>
<td>nee lieverd hier past ie niet.</td>
</tr>
<tr>
<td></td>
<td>translation</td>
<td>“No sweetie, it won’t fit in here.”</td>
</tr>
</tbody>
</table>
How to learn the meaning of a word?

- Cross-situationally observing objects, relations, events, properties.
- Seems insufficient (esp. for relational terms; verbs, prepositions)
  - Number of possibilities is vast (Gentner 1978)
  - Many actions and relations do not take place at the moment of utterance (Gleitman 1990)
- Bootstrapping by using linguistic structure (Gleitman 1990), intentionality (Tomasello 2003), ...
Acquiring lexical meaning

- How to learn the meaning of a word?
  - Cross-situationally observing objects, relations, events, properties.
  - Seems insufficient (esp. for relational terms; verbs, prepositions)
    - Number of possibilities is vast (Gentner 1978)
    - Many actions and relations do not take place at the moment of utterance (Gleitman 1990)
  - Bootstrapping by using linguistic structure (Gleitman 1990), intentionality (Tomasello 2003), . . .

Goal #2

Using this data set to **re-evaluate the claim** that relational terms are more difficult than non-relational terms.
Fazly, Alishahi & Stevenson (2010) incremental model of aligning words in utterance $U = \{w_1, \ldots, w_n\}$ with features in situation $S = \{f_1, \ldots, f_n\}$. 

Calculating alignment on the basis of conditional probabilities:

$$a(w|f, U(t), S(t)) = p(t-1)(f|w) \sum_{w' \in U(t)} p(t-1)(f|w')$$

Updating the association score (initialized at 0):

$$assoc(t)(w, f) = assoc(t-1)(w, f) + a(w|f, U(t), S(t))$$

Recalculating the conditional probabilities on the basis of the association scores:

$$p(t)(f|w) = assoc(t)(w, f) + \lambda \sum_{f' \in F} assoc(t)(w, f') + \beta \times \lambda$$
The model

- Fazly, Alishahi & Stevenson (2010) incremental model of aligning words in utterance $U = \{w_1, \ldots, w_n\}$ with features in situation $S = \{f_1, \ldots, f_n\}$.

- Calculating alignment on the basis of conditional probabilities:

$$a(w|f, U(t), S(t)) = \frac{p^{(t-1)}(f|w)}{\sum_{w' \in U(t)} p^{(t-1)}(f|w')}$$ (1)
The model

- Fazly, Alishahi & Stevenson (2010) incremental model of aligning words in utterance $U = \{w_1, \ldots, w_n\}$ with features in situation $S = \{f_1, \ldots, f_n\}$.
- Calculating alignment on the basis of conditional probabilities:

$$a(w|f, U^{(t)}, S^{(t)}) = \frac{p^{(t-1)}(f|w)}{\sum_{w' \in U^{(t)}} p^{(t-1)}(f|w')}$$

- Updating the association score (initialized at 0):

$$\text{assoc}^{(t)}(w, f) = \text{assoc}^{(t-1)}(w, f) + a(w|f, U^{(t)}, S^{(t)})$$
The model

- Fazly, Alishahi & Stevenson (2010) incremental model of aligning words in utterance $U = \{w_1, \ldots, w_n\}$ with features in situation $S = \{f_1, \ldots, f_n\}$.
- Calculating alignment on the basis of conditional probabilities:

$$a(w | f, U^{(t)}, S^{(t)}) = \frac{p^{(t-1)}(f | w)}{\sum_{w' \in U^{(t)}} p^{(t-1)}(f | w')}$$

- Updating the association score (initialized at 0):

$$\text{assoc}^{(t)}(w, f) = \text{assoc}^{(t-1)}(w, f) + a(w | f, U^{(t)}, S^{(t)})$$

- Recalculating the conditional probabilities on the basis of the association scores:

$$p^{(t)}(f | w) = \frac{\text{assoc}^{(t)}(w, f) + \lambda}{\sum_{f' \in F} \text{assoc}^{(t)}(w, f') + \beta \times \lambda}$$
Data preparation

- Representations are structured, so flatten them:
  \[ \text{grab(mother,(red,square,block))} \rightarrow \{\text{grab,mother,red,square,block}\} \]
- Take the set of all flattened representations of the situation taking place in the interval in which the utterance was beginning to be produced.
- We used lemma representations for the words Beekhuizen, Fazly, Nematzadeh & Stevenson.
Evaluation

- No golden lexicon, so hand-built one for ‘meaningful’ words ($n = 41$):
  - Object labels: *blok* meaning *block*
  - Properties: *rood* meaning *red*
  - Spatial relations: *op* meaning *on*
  - Actions: *passen* meaning *match*, *stoppen* meaning \{move, in\}
Evaluation

- No golden lexicon, so hand-built one for ‘meaningful’ words ($n = 41$):
  - Object labels: blok meaning block
  - Properties: rood meaning red
  - Spatial relations: op meaning on
  - Actions: passen meaning match, stoppen meaning \{move, in\}
- Two (partially complementary) measures:
  - Summed Conditional Probability (SCP): how much probability mass is assigned to the true meanings given a word?
  - Average Precision (AP): how are the true meanings ranked (on conditional probability) w.r.t. the other meanings.
Table: Results of experiment 1. Given are mean $SCP$ and $AP$ values per class

<table>
<thead>
<tr>
<th></th>
<th>property</th>
<th>object</th>
<th>spatial</th>
<th>action</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>$SCP$</td>
<td>0.10</td>
<td>0.05</td>
<td>0.09</td>
<td>0.07</td>
<td>0.08</td>
</tr>
<tr>
<td>$AP$</td>
<td>0.81</td>
<td>0.25</td>
<td>0.19</td>
<td>0.15</td>
<td>0.31</td>
</tr>
</tbody>
</table>

- Conditional probability distributions do not get very peaky in general
- Ranking is good for properties (colors, shapes), but rather bad for other classes.
Model dependence?

- Compared with one other model: Jon Stevens (2011)’ hypothesis testing model.
- **Same direction of results**: properties > objects > spatial relations > actions

<table>
<thead>
<tr>
<th></th>
<th>property</th>
<th>object</th>
<th>spatial</th>
<th>action</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>FAS10</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCP</td>
<td>0.10</td>
<td>0.05</td>
<td>0.09</td>
<td>0.07</td>
<td>0.08</td>
</tr>
<tr>
<td>AP</td>
<td>0.81</td>
<td>0.25</td>
<td>0.19</td>
<td>0.15</td>
<td>0.31</td>
</tr>
<tr>
<td><strong>S11</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>SCP</td>
<td>0.09</td>
<td>0.06</td>
<td>0.06</td>
<td>0.02</td>
<td>0.05</td>
</tr>
<tr>
<td>AP</td>
<td>0.28</td>
<td>0.20</td>
<td>0.13</td>
<td>0.09</td>
<td>0.17</td>
</tr>
</tbody>
</table>
**Interpretation**

<table>
<thead>
<tr>
<th></th>
<th>property</th>
<th>object</th>
<th>spatial</th>
<th>action</th>
<th>total</th>
</tr>
</thead>
<tbody>
<tr>
<td>SCP</td>
<td>0.10</td>
<td>0.05</td>
<td>0.09</td>
<td>0.07</td>
<td>0.08</td>
</tr>
<tr>
<td>AP</td>
<td>0.81</td>
<td>0.25</td>
<td>0.19</td>
<td>0.15</td>
<td>0.31</td>
</tr>
</tbody>
</table>

- Re-evaluation *corroborates* Gleitman’s finding:
  - Properties $>_{\text{object labels}}>_{\text{spatial relations and actions}}$

- Why are the latter three harder to learn?
  1. True meaning is *absent* from $S$
  2. Foil features are structurally *present* in $S$
  3. True meaning is also *present* in many other $S$s

- Combination of these! For *properties*, 2) and 3) hold as well.
Focussing on **absent true meanings**

Perhaps the temporal scope is **too narrow?**

Learners may focus on situations slightly temporally displaced

**Pragmatically defined window:** \( S = \) all coded material in intervals between the previous utterance, \( U^{(t-1)} \), and the next one, \( U^{(t+1)} \).

Variable: sometimes a large window of situations, sometimes just the time of the utterance itself.
W | prop. | object | spatial | action | total
---|-------|--------|---------|--------|--------
0 : 0 | SCP | 0.10 | 0.05 | 0.09 | 0.07 | **0.08**
   | AP   | 0.81 | 0.25 | 0.19 | 0.15 | **0.31**
U(t-1):U(t+1) | SCP | 0.08 | 0.05 | 0.10 | 0.08 | **0.07**
   | AP   | 0.79 | 0.41 | 0.22 | 0.20 | **0.39**

- **Slight increase** for three less-learned categories:
  - wider context is informative, more true meanings found
  - while not producing more referential uncertainty (as expected).

- **Pragmatics**: people talk about what should happen, or what has happened.
- Difficulty of getting **good data**; perhaps more tedious than developing a realistic model.
- Manual coding of situational contexts can be done
  - to complement synthesization methods (how much noise and uncertainty is realistic for which meaning category?)
  - to perform small-scale evaluations experiments
- However, ideally: **wider situational contexts**
• Difficulty of getting good data; perhaps more tedious than
developing a realistic model.

• Manual coding of situational contexts can be done
  • to complement synthesisization methods (how much noise and
    uncertainty is realistic for which meaning category?)
  • to perform small-scale evaluations experiments

• However, ideally: wider situational contexts

• Verbs and Prepositions are harder to learn than Nouns, which
  are harder than Color & Shape terms

• A wider scope helps a bit

• Structured learning? (Bootstrapping on syntax, using
  structure of semantics)
• Difficulty of getting good data; perhaps more tedious than developing a realistic model.

• Manual coding of situational contexts can be done
  • to complement synthesisation methods (how much noise and uncertainty is realistic for which meaning category?)
  • to perform small-scale evaluations experiments

• However, ideally: wider situational contexts

• Verbs and Prepositions are harder to learn than Nouns, which are harder than Color & Shape terms

• A wider scope helps a bit

• Structured learning? (Bootstrapping on syntax, using structure of semantics)

• Realistic data is important!