Computational modeling of second language acquisition: An integrative usage-based account

Yevgen Matushevych, Afra Alishahi, & Ad Backus

Tilburg School of Humanities
Why computational modeling?

- Lots of variables involved into SLA process!
Why computational modeling?

- Lots of variables involved into SLA process!
Why computational modeling?

- Lots of variables involved into SLA process!
Why computational modeling?

- Lots of variables involved into SLA process!
Why computational modeling?

-Lots of variables involved into SLA process!
Why computational modeling?

How to account for all these?
Why computational modeling?

How to account for all these?

Computational modeling → separate and study each variable!
Task: learning argument structure constructions (VACs)

Adele Goldberg: “Argument structure constructions are a special subclass of constructions that provides the basic means of clausal expression in a language.”

They are present in nearly every utterance.
A good estimate of overall language proficiency.
An existing Bayesian model (Alishahi & Stevenson, 2008).

Conceptually frames are utterance–scene pairs.
Frame for the utterance *Mom put toys in boxes*:

<table>
<thead>
<tr>
<th>Feature</th>
<th>Value</th>
</tr>
</thead>
<tbody>
<tr>
<td>Head verb</td>
<td>put</td>
</tr>
<tr>
<td>Verb semantic primitives</td>
<td>&lt;cause, move&gt;</td>
</tr>
<tr>
<td>Number of arguments</td>
<td>3</td>
</tr>
<tr>
<td>Arguments</td>
<td>mom, toys, boxes</td>
</tr>
<tr>
<td>Argument 1 semantics</td>
<td>&lt;parent, person, female, ... &gt;</td>
</tr>
<tr>
<td>Argument 1 event properties</td>
<td>&lt;volitional, cause-location, ... &gt;</td>
</tr>
<tr>
<td>Syntactic pattern</td>
<td>arg1 verb arg2 arg3</td>
</tr>
</tbody>
</table>
Constructions in this model are groups of frames:
The model

\[ \text{BestConstruction}(F) = \arg \max_k P(k \mid F) \]

\[ P(k \mid F) = \frac{P(k)P(F \mid k)}{P(F)} \sim P(k)P(F \mid k) \]

\[ P(0) = \frac{1}{N + 1} \]

\[ P(k) = \frac{\binom{N}{k}}{N + 1} \]

\[ P(F \mid k) = \prod_{i \in \text{Features}(F)} P_i(v \mid k) \]

\( F \) – frame, \( k \) – construction, \( v \) – feature value
\( N \) – number of frames observed
Evaluation is based on language use and can include various comprehension/production tasks (predicting missing features in a frame).

Sentence production task: predict the syntactic pattern for a given frame.
The model

Manupilate input parameters → see how the performance changes

Parameters:
- age of onset (= L2 delay)
- language ratio (= |L1| / |L2|)
- frequency distributions in the input
Data

Testing the model on a small naturalistic dataset
The Flensburg English Classroom corpus

L2

6 frequent verbs
The Flensburg **English** Classroom corpus

6 frequent verbs

CHILDES: 3 **German** children

6 frequent verbs
Data

The Flensburg **English** Classroom corpus

6 frequent verbs

**CHILDES:** 3 **German** children

6 frequent verbs

100 instances for each verb, manual annotating
L2 development: baseline

Manipulating the input using different scenarios. For this case:

- L1 training: 20 frames
- L1 testing: 20 frames

25 times 20 = 500 frames
L2 development: baseline

Manipulating the input using different scenarios. For this case:

- L1 training: 20 frames
- L1 testing: 20 frames
- L1+L2 training (Ratio 3:1): 20 frames
- L1 & L2 testing: 20 frames

25 times 20 = 500 frames
Manipulating the input using different scenarios. For this case:

- **L1 training**: 20 frames
- **L1 testing**: 20 frames
- **L1+L2 training**: Ratio 3:1, 20 frames
- **L1 & L2 testing**: 20 frames
- ...
Manipulating the input using different scenarios. For this case:

- L1 training: 20 frames
- L1 testing: 20 frames
- L1+L2 training: Ratio 3:1, 20 frames
- L1 & L2 testing: 20 frames
- ...
Manipulating the input using different scenarios. For this case:

- L1 training: 20 frames
- L1 testing: 20 frames
- L1+L2 training: Ratio 3:1, 20 frames
- L1 & L2 testing: 20 frames

1: $p < .05$
2: $p < .05$

L2 proficiency $\neq$ L1 proficiency
L2 development: baseline

Manipulating the input using different scenarios. For this case:

- **L1 training**: 20 frames
- **L1 testing**: 20 frames
- **L1+L2 training**: Ratio 3:1, 20 frames
- **L1 & L2 testing**: 20 frames

<table>
<thead>
<tr>
<th>Step</th>
<th>Accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>[Graph showing accuracy over steps for L1 and L2]</td>
</tr>
<tr>
<td>2</td>
<td>[Graph showing accuracy over steps for L1 and L2]</td>
</tr>
</tbody>
</table>

1: $p < .05$

2: $p < .05$

L2 proficiency $\neq$ L1 proficiency
L2 development: manipulating age

![Graph showing the development of L2 with manipulating age. The x-axis represents steps, and the y-axis represents accuracy. The graph compares L1 and L2 development, with L2 showing a slower increase in accuracy compared to L1.]
L2 development: manipulating age
L2 development: manipulating age
L2 development: manipulating age
L2 development: manipulating ratio
L2 development: manipulating ratio
L2 development: all equal

Equal learning conditions for L1 and L2:
Results

Effect of skewed input frequencies
Effect of skewed input frequencies

- Language learners are sensitive to input frequencies.
- Certain verbs occur in a certain construction much more often than other verbs (Ellis & Ferreira-Junior, 2009).
- Skewed distributions may have facilitatory effect on learning.
- The effect shown for artificial L2 verbs (Boyd & Goldberg, 2009), but no results for natural L2 yet.
Effect of skewed input frequencies
Larger dataset
Data

PropBank + Penn Treebank + SemLink + FrameNet + WordNet
~ 3,800 frame instances
Data

EN

PropBank + Penn Treebank + SemLink + FrameNet + WordNet
~ 3,800 frame instances

GE

SALSA + TIGER corpus + dict.cc + WordNet
~ 3,400 frame instances
PropBank + Penn Treebank + SemLink + FrameNet + WordNet
~ 3,800 frame instances

SALSA + TIGER corpus + dict.cc + WordNet
~ 3,400 frame instances

Both datasets can be used as either L1 or L2.
## Data

<table>
<thead>
<tr>
<th>Small datasets</th>
<th>Big datasets</th>
</tr>
</thead>
<tbody>
<tr>
<td>+ spoken language</td>
<td>- newspaper language</td>
</tr>
<tr>
<td>+ manually annotated</td>
<td>- semi-automatic annotation</td>
</tr>
<tr>
<td>- small size</td>
<td>+ larger size</td>
</tr>
<tr>
<td>- unpredictable distributions</td>
<td>+ Zipfian distributions</td>
</tr>
</tbody>
</table>
Data

Small datasets

+ spoken language
+ manually annotated
- small size
- unpredictable distributions

Big datasets

- newspaper language
- semi-automatic annotation
+ larger size
+ Zipfian distributions
Future work

- Model performs as expected on small datasets, but needs to be tested on larger ones.
- The datasets per se have to be carefully compared.
- This can be a good framework for studying different L2 phenomena.
- More languages to be added. Spanish (Ancora), Czech (Prague Dependency Treebank).
Questions?

Thank you for your attention!