Probing language models’ knowledge of position-role mappings with novel word learning

Michael Wilson, Jackson Petty, & Robert Frank

Yale
• How do children learn to interpret verbal argument structure?

(1) a. [The girl]$_{AGENT}$ kicked [the toy]$_{THEME}$. 
How do children learn to interpret verbal argument structure?

(1) a. [The girl]$_{\text{AGENT}}$ kicked [the toy]$_{\text{THEME}}$.

   b. [The girl]$_{\text{AGENT}}$ kicked [the boy]$_{\text{THEME}}$. 
• How do children learn to interpret verbal argument structure?

(1) a. [The girl]$_{\text{AGENT}}$ kicked [the toy]$_{\text{THEME}}$.
b. [The girl]$_{\text{AGENT}}$ kicked [the boy]$_{\text{THEME}}$.
c. [The boy]$_{\text{THEME}}$ was kicked by [the girl]$_{\text{AGENT}}$. 
How do children learn to interpret verbal argument structure?

(1)  

a.  [The girl]$_{AGENT}$ kicked [the toy]$_{THEME}$.

b.  [The girl]$_{AGENT}$ kicked [the boy]$_{THEME}$.

c.  [The boy]$_{THEME}$ was kicked by [the girl]$_{AGENT}$.

d.  [The boy]$_{THEME}$ that [the girl]$_{AGENT}$ kicked...
• Three possible learning strategies:

1. Generalization-light learning, specific to each verb/structure combination (e.g., Tomasello 1992).
2. Innate, Language-specific constraints (e.g., Baker 1988; Gleitman 1990; Pinker 1989).
3. General learning biases + input (e.g., McCoy et al. 2020; Min et al. 2020; Mulligan et al. 2021).

• We investigate strategy 3 by teaching computational language models two new nouns.
  • We can’t control what innate biases humans have, nor their total linguistic input—but we can do this with computers.
Take-aways

• **What** do the models know?
  • Position-role mappings for many **verbs**, **argument structures**, and **structure types**.
  • Patterns **reminiscent of human language acquisition** emerge.

• **How** are the models able to generalize?
  • Learned representations of the new words encode **lexical semantic information**.
Background
• Language models: computational devices that generate predictions about words.

• Here, we can think of them as devices that are trained to become very good at solving “fill-in-the-blank” problems.¹

• We use three kinds of models in our experiments: BERT (Devlin et al. 2018), DistilBERT (Sanh et al. 2019), and RoBERTa (Liu et al. 2019). Of these models, RoBERTa has the largest number of parameters, BERT the second-largest, and DistilBERT the fewest.

¹A.k.a. masked language modeling.
Use the output of the masked word's position to predict the masked word.

Possible classes:
- All English words: 10%
- Improvisation: 0.1%
- Zyzzyva: 0%

Randomly mask 15% of tokens.

Input:

Image credit: https://jalammar.github.io/illustrated-bert/
Methodology
Methodology: Overview

• We add two novel words to the models’ vocabularies in a single argument structure (Kim & Smolensky 2021).

• Each novel word is associated with a single thematic role: RICKET = recipient, THAX = theme, GORX = goal.

I gave the RICKET the box. → I gave the [MASK] the box. → BERT → [MASK] = ???
I gave the teacher the THAX. → I gave the teacher the [MASK]. → [MASK] = ???

• Only one verb (give, send, spray, load) in one structure (PD, DO, theme-object, goal-object) per model.

• 4–6 sentences per new argument.
Methodology: Procedure

- 5 each of BERT, DistilBERT, RoBERTa trained per dataset

- Trained until performance failed to improve for 30 consecutive iterations

- Parameters unrelated to the novel words were frozen.

- Evaluation data: 8 (dative)–12 (spray/load) instances × 78 sentence types × 8 verbs in various lexical semantic subclasses (Pinker 1989; Rappaport Hovav & Levin 2007).
  - dative: \(((\text{give, hand}), (\text{teach, tell})), ((\text{send, mail}), (\text{throw, toss}))\)
  - spray/load: \(((\text{spray, shower}), (\text{rub, dab})), ((\text{load, stock}), (\text{stuff, pack}))\)
Methodology: Evaluation

• If you’re trained on DO active sentences with *give*, we might be interested in:

(2)  
  a. The teacher gave the THAX to the RICKET.  
     (*give* PD active)
Methodology: Evaluation

• If you’re trained on DO active sentences with *give*, we might be interested in:

  (2) a. The teacher gave the THAX to the RICKET.  
      (give PD active)  
     b. The RICKET was given the THAX.  
      (give DO passive)
Methodology: Evaluation

• If you’re trained on DO active sentences with give, we might be interested in:

  (2)  a. The teacher gave the THAX to the RICKET. \(\text{(give PD active)}\)
  b. The RICKET was given the THAX. \(\text{(give DO passive)}\)
  c. The teacher handed the RICKET the THAX. \(\text{(hand DO active)}\)

...
Methodology: Evaluation

- If you’re trained on DO active sentences with *give*, we might be interested in:

  2. a. The teacher gave the THAX to the RICKET. *(give PD active)*
  b. The RICKET was given the THAX. *(give DO passive)*
  c. The teacher handed the RICKET the THAX. *(hand DO active)*
  d. It was the doctor that gave the THAX to the RICKET. *(cleft subject give PD active)*
  e. Which THAX does the teacher seem to have mailed to the RICKET? *(mat-wh-Q object raising mail PD active)*
  f. The THAX that the person told to the RICKET was everyone’s favorite. *(ORC tell PD active)*
  g. It was the THAX that the RICKET was taught. *(cleft 2-object teach DO passive)*
  h. Which THAX was the RICKET given? *(mat-wh-Q 2-object give DO passive)*
  i. I wonder which THAX the worker seems to have given to the RICKET. *(emb-wh-Q object raising give PD active)*
  j. Which woman sent the RICKET the THAX? *(mat-wh-Q subject send DO active)*

... *(8–12 instances × 78 types × 8 verbs = 4,992–7,488 total)*

- Every test sentence had both new words.
Methodology: Evaluation

• Obtain a confidence score for each novel word and sentence.
  
  (3) The RICKET that the THAX was given to was everyone’s favorite.
  (i.e., p-object relative clause give PD passive)

  (4) The teacher gave the RICKET the THAX.
  (baseline for model trained on give DO active)

  >0 if the model accurately predicts, e.g., THAX is more likely than RICKET in THAX position.

• Convert the confidence score to a binary accuracy score, conditioned on performance for that word in the corresponding baseline sentence.

\(^2\)The log odds ratio of expected/alternative novel word in a position
Results
X: The teacher gave the RICKET the THAX.
Y: I wonder which THAX the teacher seems to have given to the RICKET.

X: The teacher gave the RICKET the THAX.
Y: It was the THAX that the RICKET was thrown.
Results: Accuracy by thematic role

<table>
<thead>
<tr>
<th>Fine-tuning structure</th>
<th>PD</th>
<th>DO</th>
<th>theme-object</th>
<th>goal-object</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>RoBERTa</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rec/goal</td>
<td>83.9</td>
<td>84.8</td>
<td>80.6</td>
<td>89.2</td>
</tr>
<tr>
<td>theme</td>
<td>71.5</td>
<td>77.2</td>
<td>79.8</td>
<td>74.2</td>
</tr>
<tr>
<td><strong>BERT</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rec/goal</td>
<td>83.0</td>
<td>91.1</td>
<td>85.7</td>
<td>96.7</td>
</tr>
<tr>
<td>theme</td>
<td>89.5</td>
<td>75.0</td>
<td>92.9</td>
<td>81.3</td>
</tr>
<tr>
<td><strong>DistilBERT</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>rec/goal</td>
<td>66.6</td>
<td>90.6</td>
<td>82.1</td>
<td>96.6</td>
</tr>
<tr>
<td>theme</td>
<td>89.9</td>
<td>58.0</td>
<td>90.1</td>
<td>61.8</td>
</tr>
</tbody>
</table>

- RoBERTa shows a **recipient** bias.
- BERT and DistilBERT show an adjacent **object** bias—they do **better at predicting arguments that were in object position in the fine-tuning sets**.
Results: Accuracy by movement type

<table>
<thead>
<tr>
<th></th>
<th>Dative</th>
<th>spray/load</th>
</tr>
</thead>
<tbody>
<tr>
<td>A-movement?</td>
<td></td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>

- The models perform **worst on sentences with both A- and A-movement.**
- The effect is **superadditive!**
• For spray/load verbs, they perform **better on verbs that match the subclass they were trained on.**
• Overall, the models **generalize well to different verbs and structural contexts**.
  • The only information about the new words was their independent use in a single structural context in 8–12 sentences.

• **Patterns reminiscent of those observed during human language acquisition emerge.**

• How are the models able to generalize as well as they are? In other words, what do they “think” about the new words?
What’s in a word?
The models represent words as points in a 768D space, called embeddings.
  - They get better with the new words by moving them around during training.

Embeddings for words with similar meanings occupy similar locations (Boleda 2020).
  - We use t-SNE to map the high dimensional space to a 2D space that preserves this property for visualization.
• RICKET is near generic animate words, appropriate for a recipient.
• THAX’s position is less clear, but it is farther away from animates.
BERT cosine similarities to syn_give_give_ext target group tokens @ epoch 51/81 (best mean)

- **Animate** targets:
  - Mean cosine similarity: \(0.3090 \pm 0.0056\)
  - Mean cosine similarity: \(0.2629 \pm 0.0052\)
  - Mean cosine similarity: \(0.3090 \pm 0.0056\)
  - Mean cosine similarity: \(0.2778 \pm 0.0058\)

- **Inanimate** targets:
  - Mean cosine similarity: \(0.2722 \pm 0.0048\)
  - Mean cosine similarity: \(0.2629 \pm 0.0052\)
  - Mean cosine similarity: \(0.2722 \pm 0.0048\)
  - Mean cosine similarity: \(0.2629 \pm 0.0052\)

**Tuning:**
- DO give active
- Masking: always, with punctuation

**Epochs:**
- Min epochs: 70
- Max epochs: 5000
- Patience: 30 (\(=0\))
Mean cosine similarity of gorx to location  thax to location targets: 0.3020 (±0.0052)  0.2680 (±0.0049) = 0.0340 (±0.0072)

Mean cosine similarity of thax to mass/plural  gorx to mass/plural targets: 0.3311 (±0.0059)  0.3132 (±0.0062) = 0.0179 (±0.0086)

Mean cosine similarity of gorx to location  gorx to mass/plural targets: 0.3020 (±0.0052)  0.3132 (±0.0062) = 0.0111 (±0.0081)

Mean cosine similarity of thax to mass/plural  thax to location targets: 0.3311 (±0.0059)  0.2680 (±0.0049) = 0.0631 (±0.0077)
Discussion & Conclusion
Discussion: Linguistic bootstrapping

• Syntax guides the acquisition of (typically verbal) semantics.  
  (Gleitman 1990; Gleitman et al. 2005)

• Verbal semantics guides acquisition of syntactic argument structure.  
  (Pinker 1989 et seq.)
  • The models use **syntactic argument structure** to identify semantic information.  
    → **syntactic bootstrapping**
  • They use **verbal lexical semantics** to guide syntactic generalization.  
    → **semantic bootstrapping**

• They display such behavior despite the lack of any obvious linguistic bias in their architecture.
Conclusion: Open questions

- Are domain general biases + input a valid strategy for children?
  - These models are pre-trained on a lot of text.
  - However, the models were able to generalize the use of the novel words on the basis of only 4–6 uses, recalling fast-mapping behavior (Trueswell et al. 2020).

- What do the models know about relationships between different structures? Two possibilities:
  1. They have learned abstract relations between structures (e.g., transformations, etc.).
  2. They represent mappings between positions and thematic roles redundantly for each different kind of structure.
    - This would still require the models to encode information about structure types that goes beyond information about structure tokens.
Thank you!

This work was made possible by support from the National Science Foundation grant BCS-1919321.
References


Appendix
### Appendix: Dative lexical semantic subclasses

<table>
<thead>
<tr>
<th></th>
<th>1\textsuperscript{st} subclass</th>
<th>2\textsuperscript{nd} subclass</th>
<th>Fine-tuning verb</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>giving ({<em>give, hand</em>})</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>communication ({<em>tell, teach</em>})</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>sending ({<em>send, mail</em>})</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>ballistic motion ({<em>throw, toss</em>})</td>
<td></td>
</tr>
<tr>
<td>BERT</td>
<td>give-type</td>
<td>92.5</td>
<td>90.9</td>
</tr>
<tr>
<td></td>
<td>send-type</td>
<td>82.3</td>
<td>82.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>82.1</td>
<td>85.5</td>
</tr>
<tr>
<td></td>
<td></td>
<td>81.5</td>
<td>80.5</td>
</tr>
<tr>
<td>DistilBERT</td>
<td>give-type</td>
<td>88.0</td>
<td>83.9</td>
</tr>
<tr>
<td></td>
<td>send-type</td>
<td>74.8</td>
<td>80.3</td>
</tr>
<tr>
<td></td>
<td></td>
<td>73.1</td>
<td>71.1</td>
</tr>
</tbody>
</table>

- BERT and RoBERTa do **better on give-type subclasses, except when trained on send**.
- DistilBERT does **better on giving and ballistic motion verbs when trained on give**, and **better on communication and sending verbs when trained on send**.
### Results: Spray/load lexical semantic subclasses

<table>
<thead>
<tr>
<th>Model</th>
<th>1st subclass</th>
<th>2nd subclass</th>
<th>Fine-tuning verb</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td>spray</td>
</tr>
<tr>
<td>RoBERTa</td>
<td>spray-type</td>
<td>particulate (spray, shower)</td>
<td>90.1</td>
</tr>
<tr>
<td></td>
<td></td>
<td>goop (dab, rub)</td>
<td>84.7</td>
</tr>
<tr>
<td></td>
<td>load-type</td>
<td>loading (load, stock)</td>
<td>77.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>stuffing (stuff, pack)</td>
<td>75.4</td>
</tr>
<tr>
<td>BERT</td>
<td>spray-type</td>
<td>particulate</td>
<td>93.9</td>
</tr>
<tr>
<td></td>
<td></td>
<td>goop</td>
<td>89.7</td>
</tr>
<tr>
<td></td>
<td>load-type</td>
<td>loading</td>
<td>88.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>stuffing</td>
<td>90.5</td>
</tr>
<tr>
<td>DistilBERT</td>
<td>spray-type</td>
<td>particulate</td>
<td>85.8</td>
</tr>
<tr>
<td></td>
<td></td>
<td>goop</td>
<td>82.6</td>
</tr>
<tr>
<td></td>
<td>load-type</td>
<td>loading</td>
<td>83.4</td>
</tr>
<tr>
<td></td>
<td></td>
<td>stuffing</td>
<td>81.5</td>
</tr>
</tbody>
</table>

- RoBERTa does **better on spray-type verbs given spray tuning, and vice versa for load-type verbs.**
- BERT does **better when trained on spray.**
- DistilBERT is like BERT, except for stuffing verbs.
Multiple models' cosine similarities to syn_dab_spray_ext target group tokens, epochs: multiple (best mean)

min epochs: 70, max epochs: 5000, patience: 30 (=0)
tuning: sl goal-object spray active, masking: always, with punctuation

Mean cosine similarity of GORX to location  THAX to location targets: 0.2384 (±0.0338)  0.2034 (±0.0334) = 0.0350 (±0.0475)

Mean cosine similarity of THAX to mass/plural  GORX to mass/plural targets: 0.2553 (±0.0379)  0.2433 (±0.0392) = 0.0120 (±0.0545)

Mean cosine similarity of GORX to location  GORX to mass/plural targets: 0.2384 (±0.0338)  0.2433 (±0.0392) = 0.0049 (±0.0517)

Mean cosine similarity of THAX to mass/plural  THAX to location targets: 0.2553 (±0.0379)  0.2034 (±0.0334) = 0.0519 (±0.0505)
Multiple models' cosine similarities to syn_give_give_ext target group tokens, epochs: multiple (best mean)

- Tuning: dative DO give active, masking: always, with punctuation

Mean cosine similarity of RICKET to animate  THAX to animate targets: 0.2371 (±0.0328)  0.1963 (±0.0360) = 0.0408 (±0.0487)

Mean cosine similarity of THAX to inanimate  RICKET to inanimate targets: 0.2150 (±0.0350)  0.2101 (±0.0344) = 0.0049 (±0.0491)

Mean cosine similarity of RICKET to animate  RICKET to inanimate targets: 0.2371 (±0.0328)  0.2101 (±0.0344) = 0.0270 (±0.0475)

Mean cosine similarity of THAX to inanimate  THAX to animate targets: 0.2150 (±0.0350)  0.1963 (±0.0360) = 0.0187 (±0.0502)