Testing the learnability of grammar for humans and machines: Investigations with artificial neural networks

Alex Warstadt
New York University Linguistics
21 October 2021
Text as Data
What currently holds the state-of-the-art in language learning?
What currently holds the state-of-the-art in language learning?
Our linguistic environments color learning.
What about the canvas?
SLOW
PROCEED WITH CAUTION
What currently holds the state-of-the-art in language learning?
Roadmap

1. Tests for grammatical knowledge
   a. Acceptability judgments
   b. CoLA
   c. BLiMP

2. More human-like training
   a. Probing
   b. Acquiring inductive bias

3. What neural networks can teach us about humans
   a. The idea experiment
   b. Obstacles and opportunities
Part 1:
Tests for grammatical knowledge

Acceptability Judgments
Acceptability Judgments

Is this sentence OK?
Acceptability Judgments

What did Betsy paint a picture of?

Is this sentence OK?
Acceptability Judgments

What did Betsy paint a picture of?

What was a picture of painted by Betsy?

Is this sentence OK?
Acceptability Judgments

What did Betsy paint a picture of?  ✓

What was a picture of painted by Betsy?  ✗
What’s the relation between acceptability judgments and grammar?
The fundamental aim in the linguistic analysis of a language $L$ is to separate the grammatical sequences which are the sentences of $L$ from the ungrammatical sequences which are not sentences of $L$ and to study the structure of the grammatical sequences.

One way to **test the adequacy of a grammar proposed for [language] L** is to determine whether or not the sequences that it generates are actually grammatical, i.e., acceptable to a native speaker."

Human grammatical knowledge is:

- Complex
- Strongly held
- Implicit (not taught)
- Widely shared
Linguistic Competence of NNs?

We can compare NNs to humans by recasting acceptability judgments as an NLP task.

An NN with knowledge of grammar should easily learn to make human-like acceptability judgments.
CoLA
The Corpus of Linguistic Acceptability

Neural Network Acceptability Judgments

Alex Warstadt
New York University
warstadt@nyu.edu

Amanpreet Singh
New York University
Facebook AI Research
amanpreet@nyu.edu

Samuel R. Bowman
New York University
bowman@nyu.edu
Table 1: Breakdown of CoLA by source.

<table>
<thead>
<tr>
<th>Source</th>
<th>n</th>
<th>%label=1</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Total</strong></td>
<td>10657</td>
<td>70.5</td>
<td></td>
</tr>
<tr>
<td><strong>In Domain</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Adger (2003)</td>
<td>948</td>
<td>71.9</td>
<td>Syntax textbook</td>
</tr>
<tr>
<td>Baltin (1982)</td>
<td>96</td>
<td>66.7</td>
<td>Movement</td>
</tr>
<tr>
<td>Baltin and Collins (2001)</td>
<td>880</td>
<td>66.7</td>
<td>Handbook</td>
</tr>
<tr>
<td>Bresnan (1973)</td>
<td>259</td>
<td>69.1</td>
<td>Comparatives</td>
</tr>
<tr>
<td>Carne (2013)</td>
<td>870</td>
<td>80.3</td>
<td>Syntax textbook</td>
</tr>
<tr>
<td>Culicover and Jackendoff (1999)</td>
<td>233</td>
<td>59.2</td>
<td>Comparatives</td>
</tr>
<tr>
<td>Dayal (1998)</td>
<td>179</td>
<td>75.4</td>
<td>Modality</td>
</tr>
<tr>
<td>Gazdar (1981)</td>
<td>110</td>
<td>65.5</td>
<td>Coordination</td>
</tr>
<tr>
<td>Goldberg and Jackendoff (2004)</td>
<td>106</td>
<td>77.4</td>
<td>Resultative</td>
</tr>
<tr>
<td>Kadmon and Landman (1993)</td>
<td>93</td>
<td>81.7</td>
<td>Negative Polarity</td>
</tr>
<tr>
<td>Kim and Sells (2008)</td>
<td>1965</td>
<td>71.2</td>
<td>Syntax Textbook</td>
</tr>
<tr>
<td>Levin (1993)</td>
<td>1459</td>
<td>69.0</td>
<td>Verb alternations</td>
</tr>
<tr>
<td>Miller (2002)</td>
<td>426</td>
<td>84.5</td>
<td>Syntax textbook</td>
</tr>
<tr>
<td>Rappaport Hovav and Levin (2008)</td>
<td>151</td>
<td>69.5</td>
<td>Dative alternation</td>
</tr>
<tr>
<td>Ross (1967)</td>
<td>1029</td>
<td>61.8</td>
<td>Islands</td>
</tr>
<tr>
<td>Sag et al. (1985)</td>
<td>153</td>
<td>68.6</td>
<td>Coordination</td>
</tr>
<tr>
<td>Sportiche et al. (2013)</td>
<td>651</td>
<td>70.4</td>
<td>Syntax textbook</td>
</tr>
<tr>
<td><strong>Out of Domain</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Chung et al. (1995)</td>
<td>148</td>
<td>66.9</td>
<td>Sluicing</td>
</tr>
<tr>
<td>Collins (2005)</td>
<td>66</td>
<td>68.2</td>
<td>Passive</td>
</tr>
<tr>
<td>Jackendoff (1971)</td>
<td>94</td>
<td>67.0</td>
<td>Gapping</td>
</tr>
<tr>
<td>Sag (1997)</td>
<td>112</td>
<td>57.1</td>
<td>Relative clauses</td>
</tr>
<tr>
<td>Sag et al. (2003)</td>
<td>460</td>
<td>70.9</td>
<td>Syntax textbook</td>
</tr>
<tr>
<td>Williams (1980)</td>
<td>169</td>
<td>76.3</td>
<td>Predication</td>
</tr>
</tbody>
</table>

CoLA

- >10k sentences from the syntax/semantics literature.
- Expert boolean acceptability judgments.
- Broad domain of phenomena
- >20x larger than similar resources.
# CoLA: Phenomena covered

<table>
<thead>
<tr>
<th>Included</th>
<th>excluded</th>
</tr>
</thead>
<tbody>
<tr>
<td>Morphological Violation (a)</td>
<td>*Maryann should leaving.</td>
</tr>
<tr>
<td>Syntactic Violation (b)</td>
<td>*What did Bill buy potatoes and _?</td>
</tr>
<tr>
<td>Semantic Violation (c)</td>
<td>*Kim persuaded it to rain.</td>
</tr>
<tr>
<td>Pragmatical Anomalies (d)</td>
<td>*Bill fell off the ladder in an hour.</td>
</tr>
<tr>
<td>Unavailable Meanings (e)</td>
<td>*He\textsubscript{i} loves John\textsubscript{i}. (<strong>intended:</strong> John loves himself.)</td>
</tr>
<tr>
<td>Prescriptive Rules (f)</td>
<td>Prepositions are good to end sentences with.</td>
</tr>
<tr>
<td>Nonce Words (g)</td>
<td>*This train is arrivable.</td>
</tr>
</tbody>
</table>
## CoLA Sample

<table>
<thead>
<tr>
<th>Label</th>
<th>Sentence</th>
<th>Source</th>
</tr>
</thead>
<tbody>
<tr>
<td>0</td>
<td>The ball wiggled itself loose.</td>
<td>gj04</td>
</tr>
<tr>
<td>0</td>
<td>The more books I ask to whom he will give, the more he reads.</td>
<td>cj99</td>
</tr>
<tr>
<td>1</td>
<td>I said that my father, he was tight as a hoot-owl.</td>
<td>r-67</td>
</tr>
<tr>
<td>1</td>
<td>The jeweller inscribed the ring with the name.</td>
<td>l-93</td>
</tr>
<tr>
<td>0</td>
<td>We rummaged papers through the desk.</td>
<td>l-93</td>
</tr>
<tr>
<td>0</td>
<td>many evidence was provided.</td>
<td>ks08</td>
</tr>
<tr>
<td>1</td>
<td>They can sing.</td>
<td>ks08</td>
</tr>
<tr>
<td>1</td>
<td>This theorem will take only five minutes to establish that he proved in 1930.</td>
<td>ks08</td>
</tr>
<tr>
<td>1</td>
<td>The men would have been all working.</td>
<td>b-82</td>
</tr>
<tr>
<td>1</td>
<td>Would John hate that?</td>
<td>b-82</td>
</tr>
<tr>
<td>0</td>
<td>Who do you think that will question Seamus first?</td>
<td>c-13</td>
</tr>
<tr>
<td>0</td>
<td>Usually, any lion is majestic.</td>
<td>d-98</td>
</tr>
<tr>
<td>1</td>
<td>Larry Twentyman hunted all the foxes.</td>
<td>m-02</td>
</tr>
<tr>
<td>1</td>
<td>I wrote Blair a letter, but I tore it up before I sent it.</td>
<td>rhl07</td>
</tr>
<tr>
<td>1</td>
<td>That’s the kindest answer that I ever heard.</td>
<td>b-73</td>
</tr>
</tbody>
</table>
Measuring Human Performance

Human Agreement with CoLA

Matthews Correlation (MCC)

- Non-Linguist*
- Linguist Average
- Linguist Aggregate
Baselines

CoLA Baselines Results

Matthews Correlation (MCC)

- BOW
- 3-gram LM
- LSTM LM
- ELMo-Style LSTM
- Non-Linguist
- Linguist Average
- Linguist Aggregate
Early Transformers

CoLA Performance (Post GLUE)

Mathews Correlation (MCC)

- BOW
- 3-gram LM
- LSTM LM
- ELMo-Style LSTM
- GPT
- BERT
- Non-Linguist
- Linguist Average
- Linguist Aggregate
Superhuman Results?

CoLA Performance (SoTA)

Mathews Correlation (MCC)

- BOW
- 3-gram LM
- LSTM LM
- ELMo-Style LSTM
- GPT
- BERT
- Non-Linguist
- XLNet
- Linguist Average
- T5 (Google)
- ERNIE (Baidu)
- Linguist Aggregate
Not so fast...
Evaluating on CoLA requires supervised training, which exposes the model to explicit information about acceptability.
Enter: Minimal Pairs

A pair of two nearly identical sentences which differ in acceptability.

Betsy is *eager* to sleep.

Betsy is *easy* to sleep.
Why Minimal Pairs?

If $P_{LM}(S✓) > P_{LM}(S✗)$, then LM detects a contrast in acceptability.
Recently, there’s been an abundance of work testing LMs on minimal pairs.
## Sample of Work Using Minimal Pairs

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>Relevant work</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anaphor/binding</td>
<td>Marvin &amp; Linzen (2018); Futrell et al. (2018); Warstadt et al. (2019b)</td>
</tr>
<tr>
<td>Subject-verb agreement</td>
<td>Linzen et al. (2016); Futrell et al. (2018); Gulordava et al. (2019); Marvin &amp; Linzen (2018); An et al. (2019); Warstadt et al. (2019b)</td>
</tr>
<tr>
<td>Negative polarity items</td>
<td>Marvin &amp; Linzen (2018); Futrell et al. (2018); Jumelet &amp; Hupkes (2018); Wilcox et al. (2019); Warstadt et al. (2019a)</td>
</tr>
<tr>
<td>Filler-gap dependencies &amp; islands</td>
<td>Wilcox et al. (2018); Warstadt et al. (2019b); Chowdhury &amp; Zamparelli (2018, 2019); Chaves (to appear); Da Costa &amp; Chaves (to appear)</td>
</tr>
<tr>
<td>Argument structure</td>
<td>Kann et al. (2019); Warstadt et al. (2019b); Chowdhury &amp; Zamparelli (2019)</td>
</tr>
</tbody>
</table>
Things are getting a bit complicated...
We need... 1 dataset to rule them all.
BLiMP: The Benchmark of Linguistic Minimal Pairs

Alex Warstadt\textsuperscript{1}, Alicia Parrish\textsuperscript{1}, Haokun Liu\textsuperscript{2}, Anhad Mohananey\textsuperscript{2}, Wei Peng\textsuperscript{2}, Sheng-Fu Wang\textsuperscript{1}, Samuel R. Bowman\textsuperscript{1,2,3}

\textsuperscript{1}Department of Linguistics, New York University
\textsuperscript{2}Department of Computer Science, New York University
\textsuperscript{3}Center for Data Science, New York University
Enter: BLiMP

A wide-coverage dataset of targeted minimal pairs.

67 unique paradigms with 1000 minimal pairs each, organized into 12 categories.

Evaluation is simple: just compare LM probabilities on the good and bad sentences.

All minimal pairs in BLiMP:

(a) Are equal in length.
(b) Differ in at most 1 vocabulary item.
<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>N</th>
<th>Acceptable example</th>
<th>Unacceptable example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anaphor agreement</td>
<td>2</td>
<td>Many girls insulted themselves.</td>
<td>Many girls insulted herself.</td>
</tr>
<tr>
<td>Argument structure</td>
<td>9</td>
<td>Rose wasn’t disturbing Mark.</td>
<td>Rose wasn’t boasting Mark.</td>
</tr>
<tr>
<td>Binding</td>
<td>7</td>
<td>It’s himself who Robert attacked.</td>
<td>It’s himself who attacked Robert.</td>
</tr>
<tr>
<td>Control/Raising</td>
<td>5</td>
<td>Kevin isn’t irritating to work with.</td>
<td>Kevin isn’t bound to work with.</td>
</tr>
<tr>
<td>Determiner-N agr.</td>
<td>8</td>
<td>Rachelle had bought that chair.</td>
<td>Rachelle had bought that chairs.</td>
</tr>
<tr>
<td>Ellipsis</td>
<td>2</td>
<td>Anne’s doctor cleans one important book and Stacey cleans a few.</td>
<td>Anne’s doctor cleans one book and Stacey cleans a few important.</td>
</tr>
<tr>
<td>Filler-gap</td>
<td>7</td>
<td>Brett knew what many waiters find.</td>
<td>Brett knew that many waiters find.</td>
</tr>
<tr>
<td>Irregular forms</td>
<td>2</td>
<td>Aaron broke the unicycle.</td>
<td>Aaron broken the unicycle.</td>
</tr>
<tr>
<td>Island effects</td>
<td>8</td>
<td>Which bikes is John fixing?</td>
<td>Which is John fixing bikes?</td>
</tr>
<tr>
<td>NPI licensing</td>
<td>7</td>
<td>The truck has clearly tipped over.</td>
<td>The truck has ever tipped over.</td>
</tr>
<tr>
<td>Quantifiers</td>
<td>4</td>
<td>There was a cat annoying Alice.</td>
<td>There was each cat annoying Alice.</td>
</tr>
<tr>
<td>Subject-Verb agr.</td>
<td>6</td>
<td>These casseroles disgust Kayla.</td>
<td>These casseroles disgusts Kayla.</td>
</tr>
</tbody>
</table>
Data Generation

Data generation allows for large, syntactically controlled datasets.

We use a hand-crafted vocabulary of >3K items.

- More comprehensive than similar resources.
- >70 morphological, syntactic, and semantic features.
Data -- Generation procedure

Sentences are generated according to simple templates

```python
def sample(self):
    # What did John read before filing the book?
    # Wh Aux_mat Subj V_mat ADV V_emb Obj
    # What did John read the book before filing?
    # Wh Aux_mat Subj V_mat Obj ADV V_emb

    V_mat = choice(all_non_finite_transitive_verbs)
    Subj = N_to_DP_mutate(choice(get_matches_of(V_mat, "arg_2", all_nouns)))
    Aux_mat = return_aux(V_mat, Subj, allow_negated=False)
    Obj = N_to_DP_mutate(choice(get_matches_of(V_mat, "arg_2", all_nouns)))
    V_emb = choice(get_matched_by(Obj, "arg_2", get_matched_by(Subj, "arg_1", self.all_ing_transitives)))
    Wh = choice(get_matched_by(Obj, "arg_1", all_wh_words))
    Adv = choice(self.adverbs)

    data = {
        "sentence_good": "%s %s %s %s %s %s %s?" % (Wh[0], Aux_mat[0], Subj[0], V_mat[0], Adv, V_emb[0], Obj[0]),
        "sentence_bad": "%s %s %s %s %s %s %s?" % (Wh[0], Aux_mat[0], Subj[0], V_mat[0], Obj[0], Adv, V_emb[0])
    }

    return data, data["sentence_good"]
```
Via Amazon Mechanical Turk, 20 English speaking annotators evaluate 5 pairs from each paradigm (6700 total judgments).

Forced choice task: annotators select the more acceptable sentence from a pair.

Inclusion criteria: Majority vote agreement with 4/5 pairs in the paradigm.

Majority vote human agreement with our annotations is 96.4% overall; individual human agreement is 88.6%.
Models

1. **N-gram (5-gram)**
   - English Gigaword (3.07B tokens)

2. **LSTM**
   - English Wikipedia (83M tokens), trained by Gulordava et al. (2018)

3. **Transformer**
   - Transformer-XL: Trained on WikiText-103 (103M tokens) by Dai et al. (2019)
   - GPT-2: Trained on WebText (~8B tokens) by Radford et al. (2019)
   - RoBERTa: Trained on Wikipedia, web data, and books (30B tokens) by Liu et al. (2020)*

*results from Salazar et al (2020)
Overall Results

BLiMP Performance Overall: Human comparison

Model-Human Difference

Overall

- N-gram
- LSTM
- TXL
- GPT2_large
- RoBERTa Large
## Agreement Results

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>N</th>
<th>Acceptable example</th>
<th>Unacceptable example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anaphor agreement</td>
<td>2</td>
<td>Many girls insulted themselves.</td>
<td>Many girls insulted herself.</td>
</tr>
<tr>
<td>Argument structure</td>
<td>9</td>
<td>Rose wasn’t disturbing Mark.</td>
<td>Rose wasn’t boasting Mark.</td>
</tr>
<tr>
<td>Binding</td>
<td>7</td>
<td>It’s himself who Robert attacked.</td>
<td>It’s himself who attacked Robert.</td>
</tr>
<tr>
<td>Control/Raising</td>
<td>5</td>
<td>Kevin isn’t irritating to work with.</td>
<td>Kevin isn’t bound to work with.</td>
</tr>
<tr>
<td>Determiner-N agr.</td>
<td>8</td>
<td>Rachelle had bought that chair.</td>
<td>Rachelle had bought that chairs.</td>
</tr>
<tr>
<td>Ellipsis</td>
<td>2</td>
<td>Anne’s doctor cleans one important book and Stacey cleans a few.</td>
<td>Anne’s doctor cleans one book and Stacey cleans a few important.</td>
</tr>
<tr>
<td>Filler-gap</td>
<td>7</td>
<td>Brett knew what many waiters find.</td>
<td>Brett knew that many waiters find.</td>
</tr>
<tr>
<td>Irregular forms</td>
<td>2</td>
<td>Aaron broke the unicycle.</td>
<td>Aaron broken the unicycle.</td>
</tr>
<tr>
<td>Island effects</td>
<td>8</td>
<td>Which bikes is John fixing?</td>
<td>Which is John fixing bikes?</td>
</tr>
<tr>
<td>NPI licensing</td>
<td>7</td>
<td>The truck has clearly tipped over.</td>
<td>The truck has ever tipped over.</td>
</tr>
<tr>
<td>Quantifiers</td>
<td>4</td>
<td>There was a cat annoying Alice.</td>
<td>There was each cat annoying Alice.</td>
</tr>
<tr>
<td>Subject-Verb agr.</td>
<td>6</td>
<td>These casseroles disgust Kayla.</td>
<td>These casseroles disgusts Kayla.</td>
</tr>
</tbody>
</table>
Agreement Results

BLiMP Performance by Phenomenon: Human comparison

Agreement phenomena tend to show the highest performance across models.
# Argument Structure Results

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>N</th>
<th>Acceptable example</th>
<th>Unacceptable example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anaphor agreement</td>
<td>2</td>
<td>Many girls insulted themselves.</td>
<td>Many girls insulted herself.</td>
</tr>
<tr>
<td>Argument structure</td>
<td>9</td>
<td>Rose wasn’t disturbing Mark.</td>
<td>Rose wasn’t boasting Mark.</td>
</tr>
<tr>
<td>Binding</td>
<td>7</td>
<td>It’s himself who Robert attacked.</td>
<td>It’s himself who attacked Robert.</td>
</tr>
<tr>
<td>Control/Raising</td>
<td>5</td>
<td>Kevin isn’t irritating to work with.</td>
<td>Kevin isn’t bound to work with.</td>
</tr>
<tr>
<td>Determiner-N agr.</td>
<td>8</td>
<td>Rachelle had bought that chair.</td>
<td>Rachelle had bought that chairs.</td>
</tr>
<tr>
<td>Ellipsis</td>
<td>2</td>
<td>Anne’s doctor cleans one important book and Stacey cleans a few.</td>
<td>Anne’s doctor cleans one book and Stacey cleans a few important.</td>
</tr>
<tr>
<td>Filler-gap</td>
<td>7</td>
<td>Brett knew what many waiters find.</td>
<td>Brett knew that many waiters find.</td>
</tr>
<tr>
<td>Irregular forms</td>
<td>2</td>
<td>Aaron broke the unicycle.</td>
<td>Aaron broken the unicycle.</td>
</tr>
<tr>
<td>Island effects</td>
<td>8</td>
<td>Which bikes is John fixing?</td>
<td>Which is John fixing bikes?</td>
</tr>
<tr>
<td>NPI licensing</td>
<td>7</td>
<td>The truck has clearly tipped over.</td>
<td>The truck has ever tipped over.</td>
</tr>
<tr>
<td>Quantifiers</td>
<td>4</td>
<td>There was a cat annoying Alice.</td>
<td>There was each cat annoying Alice.</td>
</tr>
<tr>
<td>Subject-Verb agr.</td>
<td>6</td>
<td>These casseroles disgust Kayla.</td>
<td>These casseroles disgusts Kayla.</td>
</tr>
</tbody>
</table>
Most models perform well below humans on argument structure.

Even GPT-2 is **not much better than the n-gram LM.**
# Filler-Gap Dependency Results

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>N</th>
<th>Acceptable example</th>
<th>Unacceptable example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anaphor agreement</td>
<td>2</td>
<td>Many girls insulted themselves.</td>
<td>Many girls insulted herself.</td>
</tr>
<tr>
<td>Argument structure</td>
<td>9</td>
<td>Rose wasn’t disturbing Mark.</td>
<td>Rose wasn’t boasting Mark.</td>
</tr>
<tr>
<td>Binding</td>
<td>7</td>
<td>It’s himself who Robert attacked.</td>
<td>It’s himself who attacked Robert.</td>
</tr>
<tr>
<td>Control/Raising</td>
<td>5</td>
<td>Kevin isn’t irritating to work with.</td>
<td>Kevin isn’t bound to work with.</td>
</tr>
<tr>
<td>Determiner-N agr.</td>
<td>8</td>
<td>Rachelle had bought that chair.</td>
<td>Rachelle had bought that chairs.</td>
</tr>
<tr>
<td>Ellipsis</td>
<td>2</td>
<td>Anne’s doctor cleans one important book and Stacey cleans a few.</td>
<td>Anne’s doctor cleans one book and Stacey cleans a few important.</td>
</tr>
<tr>
<td>Filler-gap</td>
<td>7</td>
<td>Brett knew what many waiters find.</td>
<td>Brett knew that many waiters find.</td>
</tr>
<tr>
<td>Irregular forms</td>
<td>2</td>
<td>Aaron broke the unicycle.</td>
<td>Aaron broken the unicycle.</td>
</tr>
<tr>
<td>Island effects</td>
<td>8</td>
<td>Which bikes is John fixing?</td>
<td>Which is John fixing bikes?</td>
</tr>
<tr>
<td>NPI licensing</td>
<td>7</td>
<td>The truck has clearly tipped over.</td>
<td>The truck has ever tipped over.</td>
</tr>
<tr>
<td>Quantifiers</td>
<td>4</td>
<td>There was a cat annoying Alice.</td>
<td>There was each cat annoying Alice.</td>
</tr>
<tr>
<td>Subject-Verb agr.</td>
<td>6</td>
<td>These casseroles disgust Kayla.</td>
<td>These casseroles disgusts Kayla.</td>
</tr>
</tbody>
</table>
Wh-phenomena are not hard in general, but island effects are hard for most neural models.
## Quantifiers and NPIs results

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>N</th>
<th>Acceptable example</th>
<th>Unacceptable example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anaphor agreement</td>
<td>2</td>
<td>Many girls insulted <em>themselves</em>.</td>
<td>Many girls insulted <em>herself</em>.</td>
</tr>
<tr>
<td>Argument structure</td>
<td>9</td>
<td>Rose wasn’t <em>disturbing</em> Mark.</td>
<td>Rose wasn’t <em>boasting</em> Mark.</td>
</tr>
<tr>
<td>Binding</td>
<td>7</td>
<td>It’s himself who Robert attacked.</td>
<td>It’s himself who <em>attacked</em> Robert.</td>
</tr>
<tr>
<td>Control/Raising</td>
<td>5</td>
<td>Kevin isn’t <em>irritating</em> to work with.</td>
<td>Kevin isn’t <em>bound</em> to work with.</td>
</tr>
<tr>
<td>Determiner-N agr.</td>
<td>8</td>
<td>Rachelle had bought that <em>chair</em>.</td>
<td>Rachelle had bought that <em>chairs</em>.</td>
</tr>
<tr>
<td>Ellipsis</td>
<td>2</td>
<td>Anne’s doctor cleans one <em>important</em> book and Stacey cleans a few.</td>
<td>Anne’s doctor cleans one <em>book</em> and Stacey cleans a few <em>important</em>.</td>
</tr>
<tr>
<td>Filler-gap</td>
<td>7</td>
<td>Brett knew <em>what</em> many waiters find.</td>
<td>Brett knew <em>that</em> many waiters find.</td>
</tr>
<tr>
<td>Irregular forms</td>
<td>2</td>
<td>Aaron <em>broke</em> the unicycle.</td>
<td>Aaron <em>broken</em> the unicycle.</td>
</tr>
<tr>
<td>Island effects</td>
<td>8</td>
<td>Which bikes is John fixing?</td>
<td>Which is John fixing bikes?</td>
</tr>
<tr>
<td>NPI licensing</td>
<td>7</td>
<td>The truck has <em>clearly</em> tipped over.</td>
<td>The truck has <em>ever</em> tipped over.</td>
</tr>
<tr>
<td>Quantifiers</td>
<td>4</td>
<td>There was <em>a</em> cat annoying Alice.</td>
<td>There was <em>each</em> cat annoying Alice.</td>
</tr>
<tr>
<td>Subject-Verb agr.</td>
<td>6</td>
<td>These casseroles <em>disgust</em> Kayla.</td>
<td>These casseroles <em>disgusts</em> Kayla.</td>
</tr>
</tbody>
</table>
Semantic restrictions on quantifiers and NPIs are challenging for most models. Quantifier distributions are the hardest phenomenon for RoBERTa
Part 2
More human like learning environments
Near-human results on BLiMP from RoBERTa are impressive.
But how does RoBERTa’s learning environment compare to humans’?
Growth in LM Training Sets (2018-2020)
MiniBERTas

30B words

1M words

10M words

100M words

1B words
Training

- 1M, 10M, 100M, 1B words of training data
- We simulate the original BERT training set:
  - ~¾ English Wikipedia
  - ~¼ self-published books from Smashwords
- We mostly follow the original RoBERTa training procedure.
- For each size, we train >= 10 models & select 3 with best PPL.
The 12 MiniBERTas on Transformers

https://huggingface.co/nyu-mll
Probing for features

When Do You Need Billions of Words of Pretraining Data?

Yian Zhang,*,1 Alex Warstadt,*,2 Haau-Sing Li,3 and Samuel R. Bowman1,2,3
1Dept. of Computer Science, 2Dept. of Linguistics, 3Center for Data Science
New York University
{yian.zhang, warstadt, xi3119, bowman}@nyu.edu
Five Sets of Probing Methods

1. “Standard” classifier probing
2. “Information theoretic” probing
3. Unsupervised acceptability judgments
4. Unsupervised commonsense knowledge test
5. Fine-tuning on downstream NLU tasks
Five Sets of Probing Methods

1. “Standard” classifier probing
2. “Information theoretic” probing
3. Unsupervised acceptability judgments
4. Unsupervised commonsense knowledge test
5. Fine-tuning on downstream NLU tasks
3. BLiMP: Unsupervised Acceptability Judgments
3. BLiMP: Unsupervised Acceptability Judgments

Overall grammatical knowledge increases mainly between 1M and 100M words.
3. BLiMP: Unsupervised Acceptability Judgments

Agreement phenomena are learned with only ~10M words (and often with very high accuracy)
3. BLiMP: Unsupervised Acceptability Judgments

Long-distance wh-dependencies are still improving with >1B words.
3. BLiMP: Unsupervised Acceptability Judgments
Overall Comparison

Core NLP features are learned with 10M-100M words.
Information theoretic probing looks the same as “standard” classifier probing.
Overall Comparison

Grammaticality knowledge requires more data to start acquiring.
Overall Comparison

World knowledge/commonsense reasoning requires ~1B words.
Overall Comparison

Strong performance on downstream tasks requires billions of words.
Overall Comparison

Syntactic, semantic, and commonsense knowledge are (probably) all *necessary* for good language understanding,

...but not *sufficient*.
Acquiring Inductive Bias

Learning Which Features Matter: RoBERTa Acquires a Preference for Linguistic Generalizations (Eventually)

Alex Warstadt,1 Yian Zhang,2 Haau-Sing Li,3 Haokun Liu,3 Samuel R. Bowman1,2,3
1Dept. of Linguistics, 2Dept. of Computer Science, 3Center for Data Science
New York University
Correspondence: warstadt@nyu.edu
Feature learning isn’t everything.
Feature learning isn’t everything.

...You have to know how/when to use ‘em.
Learning Inductive Biases

Inductive biases limit the learner’s hypothesis space.

Language model pretraining “induces a hypothesis space $H$ that should be useful for many other NLP tasks” (Howard & Ruder, 2018)
It is possible [in human language] to formulate a transformation [...] independently of what the length or internal complexity of the strings belonging to these categories may be. It is impossible, however, to formulate as a transformation such a simple operation as reflection of an arbitrary string [...] or interchange of the \((2n - i)\)th word with the \(2n\)th word throughout a string of arbitrary length [...].
Inductive biases

Class A

Class B

A

B

C

D
Inductive biases
Inductive biases

Class A

Class B

A

B

C

D

E

Concave

Convex

Concave

Concave
Inductive biases
Representing $F \neq \text{Using } F$
Our questions

1. Can a preference for linguistic features over surface features be acquired with sufficient data?
Our questions

1. Can a preference for linguistic features over surface features be acquired with sufficient data?

2. How do feature preferences change as the volume of pretraining data increases?
Our questions

1. Can a preference for linguistic features over surface features be acquired with sufficient data?

2. How do feature preferences change as the volume of pretraining data increases?

3. How does the acquisition of feature preferences differ from the acquisition of (mere) feature representations.
Ambiguous Experiments

Does model X ever prefer linguistic feature A or surface feature B?
Ambiguous Experiments

*Does model X ever prefer linguistic feature A or surface feature B?*

We fine-tune X on an ambiguous binary classification task.
Poverty of the Stimulus Design

Example from the SYNTACTIC POSITION × RELATIVE (LINEAR) POSITION task
Poverty of the Stimulus Design

Example from the SYNTACTIC POSITION x RELATIVE (LINEAR) POSITION task
## Surface vs. Linguistic Features

<table>
<thead>
<tr>
<th>Feature type</th>
<th>Feature description</th>
<th>Positive example</th>
<th>Negative example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Absolute position</td>
<td>Is the first token of S “the”?</td>
<td>The cat chased a mouse.</td>
<td>A cat chased a mouse.</td>
</tr>
<tr>
<td>Length</td>
<td>Is S longer than ( n ) (e.g., 3) words?</td>
<td>The cat chased a mouse.</td>
<td>The cat meowed.</td>
</tr>
<tr>
<td>Lexical content</td>
<td>Does S contain “the”?</td>
<td>That cat chased the mouse.</td>
<td>That cat chased a mouse.</td>
</tr>
<tr>
<td>Relative position</td>
<td>Does “the” precede “a”?</td>
<td>The cat chased a mouse.</td>
<td>A cat chased the mouse.</td>
</tr>
<tr>
<td>Orthography</td>
<td>Does S appear in title case?</td>
<td>The Cat Chased a Mouse.</td>
<td>The cat chased a mouse.</td>
</tr>
<tr>
<td><strong>Surface</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morphology</td>
<td>Does S have an irregular past verb?</td>
<td>The cats slept.</td>
<td>The cats meow.</td>
</tr>
<tr>
<td>Syn. category</td>
<td>Does S have an adjective?</td>
<td>Lincoln was tall.</td>
<td>Lincoln was president.</td>
</tr>
<tr>
<td>Syn. construction</td>
<td>Is S the control construction?</td>
<td>Sue is eager to sleep.</td>
<td>Sue is likely to sleep.</td>
</tr>
<tr>
<td>Syn. position</td>
<td>Is the main verb in “ing” form?</td>
<td>Cats who eat mice are purring.</td>
<td>Cats who are eating mice purr.</td>
</tr>
<tr>
<td><strong>Linguistic</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

5 surface \( \times \) 4 linguistic features = 20 ambiguous tasks
Results: Ambiguous Experiment

Aggregate results over all tasks, separated by pretraining dataset size.

Pretraining data quantity

Linguistic Bias Score

1M 10M 100M 1B 30B
Results: Ambiguous Experiment

Models trained on 1B words or less almost always choose the surface generalization.
Results: Ambiguous Experiment

RoBERTa-base (trained on ~30B words) chooses the linguistic generalization about half the time.
Results: Ambiguous Experiment

The remaining models show similar results. Does this mean they have similar inductive biases?
Inoculation Experiments

- We replace 0.1%, 0.3%, or 1% of the training data with inoculation data.
- We can quantify how strong a bias is by how much counter-evidence is needed to override it.
Results: Inoculation Experiments

Add 0.1% inoculation (10 examples/10k)

RoBERTa base shows a more systematic linguistic bias.
Results: Inoculation Experiments

Add 0.1% inoculation (10 examples/10k)

RoBERTa base shows a more systematic linguistic bias.

The 1B models start to adopt the linguistic generalization fairly often.
Results: Inoculation Experiments

Add 0.3% inoculation (30 examples/10k)

1B model shows a systematic linguistic bias.
Results: Inoculation Experiments

Add 0.3% inoculation (30 examples/10k)

1B model shows a systematic linguistic bias.

The 10M and 100M models start to consistently make the linguistic generalization.
Results: Inoculation Experiments

Add 1% inoculation (100 examples/10k)

The 10M and 100M models systematically make the linguistic generalization.
Results: Inoculation Experiments

A “phase shift” where inoculation starts to change the model behavior happens more easily for models with more pretraining data.
Part 3
What can neural networks teach us about humans?
The ideal experiment
The ideal experiment

What are the necessary conditions for human language acquisition?
Deprivation experiments

What are the necessary conditions for human language acquisition?

Pharaoh Psamtik (664 – 610 BCE)

Frederick II (1194-1250)

James IV (1473-1513)
Deprivation experiments

What are the necessary conditions for human language acquisition?

Is hypothesized advantage A necessary for acquiring linguistic fact F.

... 

Is hypothesized advantage B necessary for acquiring linguistic fact G.
Is hypothesized advantage \( B \) necessary for acquiring linguistic fact \( G \)?

1. Train artificial learner \( L \) without advantage \( A \).
2. Check if \( L \) can acquire fact \( F \).
3. If \( L \) succeeds, and doesn’t have any additional advantage over humans, then \( A \) is not necessary to explain human acquisition of \( F \).
Is hypothesized advantage $B$ necessary for acquiring linguistic fact $G$?

1. Train BERT without advantage $A$.
2. Check if BERT can acquire fact $F$.
3. If BERT succeeds, and doesn’t have any additional advantage over humans, then $A$ is not necessary to explain human acquisition of $F$.
Is hypothesized advantage $B$ necessary for acquiring linguistic fact $G$?

1. Train BERT without innate structural bias.
2. Check if BERT can acquire fact $F$.
3. If BERT succeeds, and doesn’t have any additional advantage over humans, then innate structural bias is not necessary to explain human acquisition of $F$. 
Is hypothesized advantage $B$ necessary for acquiring linguistic fact $G$?

1. Train BERT without innate structural bias.
2. Check if BERT can acquire subject aux inversion.
3. If BERT succeeds, and doesn’t have any additional advantage over humans, then innate structural bias is not necessary to explain human acquisition of subject aux inversion.
... if the learner doesn’t have any additional advantage over humans
Advantages ANNs Have

Data quantity

Data domain

Orthography
Advantages Humans Have

Multimodal input

Interactive learning
Resources

1. miniBERTas [link]
2. MSGS data/code [link]
3. Probing code [link]
Questions?
Bonus slides
Conclusions
Main Findings

Support for two different stages of learning as data quantity grows:
Main Findings

Support for two different stages of learning as data quantity grows:

1. Linguistic feature learning needs 1M-100M words of data.
Main Findings

Support for two different stages of learning as data quantity grows:

1. Linguistic feature learning needs 1M-100M words of data.
2. Linguistic bias and strong generalization on NLU tasks requires >1B words.
Lessons for Pretraining

...So an LM trained on trillions of words will be better at linguistic generalization?!
Lessons for Pretraining

...So an LM trained on trillions of words will be better at linguistic generalization?!

More important: If we want to improve pretraining, we should make feature preference learning more efficient.
2. Information theoretic MDL probing

Source: Voita & Titov (2020)
2. Information theoretic MDL probing

- Part-of-Speech
- Dependencies
- Constituents
- Relations (SemEval)
- SRL
- Sem. Proto Role 1
- Sem. Proto Role 2
- OntoNotes coref.
- Entities
- Winograd coref.
- Overall

<table>
<thead>
<tr>
<th>Minimum Description Length (kbits)</th>
<th>Model Codelength</th>
<th>Data Codelength</th>
<th>Overall Codelength</th>
<th>Learning Curve</th>
<th>Results</th>
</tr>
</thead>
<tbody>
<tr>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
<td>None</td>
</tr>
<tr>
<td>1M</td>
<td>1M</td>
<td>1M</td>
<td>1M</td>
<td>1M</td>
<td>1M</td>
</tr>
<tr>
<td>10M</td>
<td>10M</td>
<td>10M</td>
<td>10M</td>
<td>10M</td>
<td>10M</td>
</tr>
<tr>
<td>1B</td>
<td>1B</td>
<td>1B</td>
<td>1B</td>
<td>1B</td>
<td>1B</td>
</tr>
<tr>
<td>30B</td>
<td>30B</td>
<td>30B</td>
<td>30B</td>
<td>30B</td>
<td>30B</td>
</tr>
</tbody>
</table>
4. Unsupervised Commonsense Knowledge
5. SuperGLUE: Downstream NLU Tasks
Learning which feature matter

New work in probing emphasizes feature accessibility:

- Minimum description length probing (Voita & Titov, 2020)
- Amnesic probing (Elazar et al., 2020)
- The classic probing paradigm is trivial when taken to the extreme (Pimentel et al., 2020)

We probe feature preference explicitly.
Data Generation

- The MSGS data is generated from templates.
Data Generation

● The MSGS data is generated from templates.

● We always test classifiers’ ability to generalize out-of-domain.
Example: In-domain vs. Out-of-Domain

In domain: *The big dog is yawning.*

Out of domain: *The dog in the dark forest yawned.*
Results:
Ambiguous Experiment
(Fine-grained)
Results: Ambiguous Experiment (Fine-grained)

The bias in favor of absolute position and orthography (surface features) is very strong.
Results: Ambiguous Experiment (Fine-grained)

The bias in favor of sentence length (surface feature) is fairly weak.
Part I: Features/Data/Methods
Feature Learning Experiments

Does model X represent linguistic/surface feature Y?
Feature Learning Experiments

Does model X represent linguistic/surface feature Y?

Two motivations:

1. Feature preferences only make sense for features that are represented.
Feature Learning Experiments

Does model X represent linguistic/surface feature Y?

Two motivations:

1. Feature preferences only make sense for features that are represented.
2. We can compare the difficulty of feature learning and preference learning.
## Surface vs. Linguistic Features

<table>
<thead>
<tr>
<th>Feature type</th>
<th>Feature description</th>
<th>Positive example</th>
<th>Negative example</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Surface</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Absolute position</td>
<td>Is the first token of S “the”?</td>
<td>The cat chased a mouse.</td>
<td>A cat chased a mouse.</td>
</tr>
<tr>
<td>Length</td>
<td>Is S longer than ( n ) (e.g., 3) words?</td>
<td>The cat chased a mouse.</td>
<td>The cat meowed.</td>
</tr>
<tr>
<td>Lexical content</td>
<td>Does S contain “the”?</td>
<td>That cat chased the mouse.</td>
<td>That cat chased a mouse.</td>
</tr>
<tr>
<td>Relative position</td>
<td>Does “the” precede “a”?</td>
<td>The cat chased a mouse.</td>
<td>A cat chased the mouse.</td>
</tr>
<tr>
<td>Orthography</td>
<td>Does S appear in title case?</td>
<td>The Cat Chased a Mouse.</td>
<td>The cat chased a mouse.</td>
</tr>
<tr>
<td><strong>Linguistic</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Morphology</td>
<td>Does S have an irregular past verb?</td>
<td>The cats slept.</td>
<td>The cats meow.</td>
</tr>
<tr>
<td>Syn. category</td>
<td>Does S have an adjective?</td>
<td>Lincoln was tall.</td>
<td>Lincoln was president.</td>
</tr>
<tr>
<td>Syn. construction</td>
<td>Is S the control construction?</td>
<td>Sue is eager to sleep.</td>
<td>Sue is likely to sleep.</td>
</tr>
<tr>
<td>Syn. position</td>
<td>Is the main verb in “ing” form?</td>
<td>Cats who eat mice are purring.</td>
<td>Cats who are eating mice purr.</td>
</tr>
</tbody>
</table>
Fine-tuning

- 9 tasks (4 linguistic + 5 surface)
- 12 miniBERTas + original RoBERTa$_{\text{BASE}}$ (~30B words)
- The training sets are 10k sentences each
Results: Feature Learning Experiments

Performance is at ceiling.

Surface features: Pretraining dataset size
Results: Feature Learning Experiments

Surface features:
Performance is at ceiling.

Linguistic features:
Performance is near ceiling for morphology & syntactic position >1M words.
Results: Feature Learning Experiments

Surface features:
Performance is at ceiling.

Linguistic features:
Performance is near ceiling for morphology & syntactic position >1M words.

Performance for syntactic category & construction is high for >100M words.
Results: Feature Learning Experiments

For subsequent experiments, we’ll exclude any models where feature learning performance < 0.7 (gray points).
Lessons for Language Acquisition

- The very idea that linguistic bias is learnable is controversial.
- We have earlier findings that BERT prefers linguistic generalizations in key empirical domains in this debate (in CogSci; Warstadt & Bowman, 2020)
- Focusing on data quantity is important: Humans are more efficient learners than Transformers.
1. “Standard” classifier probing

Source: Tenney et al. (2019)
### 1. “Standard” classifier probing

<table>
<thead>
<tr>
<th>POS</th>
<th>The important thing about Disney is that it is a global [brand]₁. → NN (Noun)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Constit.</td>
<td>The important thing about Disney is that it [is a global brand]₁. → VP (Verb Phrase)</td>
</tr>
<tr>
<td>Depend.</td>
<td>[Atmosphere]₁ is always [fun]₂ → nsubj (nominal subject)</td>
</tr>
<tr>
<td>Entities</td>
<td>The important thing about [Disney]₁ is that it is a global brand. → Organization</td>
</tr>
<tr>
<td>SRL</td>
<td>[The important thing about Disney]₂ [is]₁ that it is a global brand. → Arg₁ (Agent)</td>
</tr>
<tr>
<td>SPR</td>
<td>[It]₁ [endorsed]₂ the White House strategy... → {awareness, existed_after, ...}</td>
</tr>
<tr>
<td>Coref.₀</td>
<td>The important thing about [Disney]₁ is that [it]₂ is a global brand. → True</td>
</tr>
<tr>
<td>Coref.₁</td>
<td>[Characters]₂ entertain audiences because [they]₁ want people to be happy. → True</td>
</tr>
<tr>
<td></td>
<td>Characters entertain [audiences]₂ because [they]₁ want people to be happy. → False</td>
</tr>
<tr>
<td>Rel.</td>
<td>The [burst]₁ has been caused by water hammer [pressure]₂. → Cause-Effect(e₂, e₁)</td>
</tr>
</tbody>
</table>

Source: Tenney et al. (2019)
1. “Standard” classifier probing

![Graph showing Part-of-Speech performance across different model sizes](image)
1. “Standard” classifier probing

![Graph showing performance across different tasks and conditions.](image-url)
1. “Standard” classifier probing
1. “Standard” classifier probing
1. “Standard” classifier probing

Syntactic feature learning converges \( \sim 10M \) words.
1. “Standard” classifier probing

- Syntactic feature learning converges ~10M words.
- Semantic feature learning converges ~100M words.
1. “Standard” classifier probing

- **Syntactic feature learning** converges ~10M words.
- **Semantic feature learning** converges ~100M words.
- **Winograd coref** requires billions of words.
3. BLiMP: Unsupervised Acceptability Judgments

The Benchmark of Linguistic Minimal Pairs for English

- A collection of thousands of minimal pairs
- 67 types of contrasts, 1000 examples each
- 12 major phenomena in English morphology, syntax, and semantics.

Warstadt et al. (2020)
3. BLiMP: Unsupervised Acceptability Judgments

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>N</th>
<th>Acceptable example</th>
<th>Unacceptable example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anaphor agreement</td>
<td>2</td>
<td>Many girls insulted <strong>themselves</strong>.</td>
<td>Many girls insulted <strong>herself</strong>.</td>
</tr>
<tr>
<td>Argument structure</td>
<td>9</td>
<td>Rose wasn’t <strong>disturbing</strong> Mark.</td>
<td>Rose wasn’t <strong>boasting</strong> Mark.</td>
</tr>
<tr>
<td>Binding</td>
<td>7</td>
<td>It’s himself who <strong>Robert</strong> attacked.</td>
<td>It’s himself who <strong>attacked</strong> Robert.</td>
</tr>
<tr>
<td>Control/Raising</td>
<td>5</td>
<td>Kevin isn’t <strong>irritating</strong> to work with.</td>
<td>Kevin isn’t <strong>bound</strong> to work with.</td>
</tr>
<tr>
<td>Determiner-N agr.</td>
<td>8</td>
<td>Rachelle had bought that <strong>chair</strong>.</td>
<td>Rachelle had bought that <strong>chairs</strong>.</td>
</tr>
<tr>
<td>Ellipsis</td>
<td>2</td>
<td>Anne’s doctor cleans one <strong>important</strong> book and Stacey cleans a few.</td>
<td>Anne’s doctor cleans one book and Stacey cleans a few <strong>important</strong>.</td>
</tr>
<tr>
<td>Filler-gap</td>
<td>7</td>
<td>Brett knew <strong>what</strong> many waiters find.</td>
<td>Brett knew <strong>that</strong> many waiters find.</td>
</tr>
<tr>
<td>Irregular forms</td>
<td>2</td>
<td>Aaron <strong>broke</strong> the unicycle.</td>
<td>Aaron <strong>broken</strong> the unicycle.</td>
</tr>
<tr>
<td>Island effects</td>
<td>8</td>
<td>Which <strong>bikes</strong> is John fixing?</td>
<td>Which is John fixing <strong>bikes</strong>?</td>
</tr>
<tr>
<td>NPI licensing</td>
<td>7</td>
<td>The truck has <strong>clearly</strong> tipped over.</td>
<td>The truck has <strong>ever</strong> tipped over.</td>
</tr>
<tr>
<td>Quantifiers</td>
<td>4</td>
<td>There was a <strong>cat</strong> annoying Alice.</td>
<td>There was each <strong>cat</strong> annoying Alice.</td>
</tr>
<tr>
<td>Subject-Verb agr.</td>
<td>6</td>
<td>These casseroles <strong>disgust</strong> Kayla.</td>
<td>These casseroles <strong>disgusts</strong> Kayla.</td>
</tr>
</tbody>
</table>
3. BLiMP: Unsupervised Acceptability Judgments

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>N</th>
<th>Acceptable example</th>
<th>Unacceptable example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anaphor agreement</td>
<td>2</td>
<td>Many girls insulted themselves.</td>
<td>Many girls insulted herself.</td>
</tr>
<tr>
<td>Argument structure</td>
<td>9</td>
<td>Rose wasn’t disturbing Mark.</td>
<td>Rose wasn’t boasting Mark.</td>
</tr>
<tr>
<td>Binding</td>
<td>7</td>
<td>It’s himself who Robert attacked.</td>
<td>It’s himself who attacked Mark.</td>
</tr>
<tr>
<td>Control/Raising</td>
<td>7</td>
<td>Kevin isn’t irritating to work with.</td>
<td>Kevin isn’t bound to work with.</td>
</tr>
<tr>
<td>Determiner-N agr.</td>
<td></td>
<td></td>
<td>Rachelle had bought that chairs.</td>
</tr>
<tr>
<td>Ellipsis</td>
<td></td>
<td></td>
<td>Anne’s doctor cleans one book and Stacey cleans a few important.</td>
</tr>
<tr>
<td>Filler-gap</td>
<td></td>
<td></td>
<td>Brett knew that many waiters find.</td>
</tr>
<tr>
<td>Irregular forms</td>
<td></td>
<td></td>
<td>Aaron broken the unicycle.</td>
</tr>
<tr>
<td>Island effects</td>
<td></td>
<td></td>
<td>Which is John fixing bikes?</td>
</tr>
<tr>
<td>NPI licensing</td>
<td></td>
<td></td>
<td>The truck has ever tipped over.</td>
</tr>
<tr>
<td>Quantifiers</td>
<td></td>
<td></td>
<td>There was each cat annoying Alice.</td>
</tr>
<tr>
<td>Subject-Verb agr.</td>
<td></td>
<td></td>
<td>These casseroles disgusts Kayla.</td>
</tr>
</tbody>
</table>

\[ P_{IM}(S_{✓}) > P_{LM}(S_{✗}) \]