Language Models and Human Language Acquisition

Alex Warstadt
ETH Zürich
For most of history, humans were the only thing in the known universe that could learn language.

In the last few years, remarkable improvements in neural language models (LMs) make us seem a little less unique.

Timeline: Things that can “learn language” *(not to scale)*
Carried out language deprivation experiments in learning environment
Roadmap

1. BACKGROUND
2. INDUCTIVE BIAS
3. INDIRECT EVIDENCE
4. FUTURE DIRECTIONS
...but first, what is a language model?

\[ p(x_1, \ldots, x_T) \]
Language Modeling as Pretraining

predict the next word from context.

repeat billions of times

Text generation.

Fine-tuning.

In-context learning.

...
Minimal Pairs

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

✓ Betsy is *eager* to sleep.
✗ Betsy is *easy* to sleep.

1. Targeted
2. Reproducible
3. Unsupervised

\[ P_{LM}(S_{✓}) > P_{LM}(S_{✗}) \]
The Benchmark of Linguistic Minimal Pairs (BLiMP)  
(Warstadt et al., 2020)

<table>
<thead>
<tr>
<th>Phenomenon</th>
<th>N</th>
<th>Acceptable Example</th>
<th>Unacceptable Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Anaphor agr.</td>
<td>2</td>
<td>Many girls insulted themselves.</td>
<td>Many girls insulted herself.</td>
</tr>
<tr>
<td>Arg. 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>Carlos said that Lori helped him.</td>
<td>Carlos said that Lori helped himself.</td>
</tr>
<tr>
<td>Control/raising</td>
<td>5</td>
<td>There was bound to be a fish escaping.</td>
<td>There was unable to be a fish escaping.</td>
</tr>
<tr>
<td>Det.-noun 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>Whose hat should Tonya wear?</td>
<td>Whose should Tonya wear hat?</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>No boy knew fewer than six guys.</td>
<td>No boy knew at most six guys.</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>

- 67 different minimal pair contrasts
- 1000 sentences each
- 12 broad categories
The Benchmark of Linguistic Minimal Pairs (BLiMP)  
(Warstadt et al., 2020)
The MiniBERTas

RoBERTa Base

30B words

1M words

10M words

100M words

1B words
The MiniBERTas on BLiMP

Long-distance wh-dependencies are still improving with >1B words.
The Data Efficiency Gap
Acquiring Inductive Bias

Inductive biases determine how a learner generalizes given ambiguity in the input.

Language model pretraining is thought to work because it “induces a hypothesis space $H$ that should be useful for many other NLP tasks” (Howard & Ruder, 2018)
Linguistic vs. Surface Bias

Can a preference for linguistic generalizations over surface generalizations be acquired with sufficient exposure to language? It is impossible to formulate as a transformation such a simple operation as the insertion of a symbol in the middle of a string of even length.

(Chomsky, 1965)
Poverty of the Stimulus Design

Wilson, 2006 (see also McCoy et al. 2018, 2020; Warstadt & Bowman, 2020; Kim & Linzen, 2020; Hupkes et al., 2022; and others)
Poverty of the Stimulus Design + Inoculation

Inoculation data: 0.1% | 0.3% | 1%
# Mixed Signals Generalization dataSet (MSGS)

<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>
Results on MSGS

Linguistic bias is acquired, and increases smoothly as a function of the amount of linguistic exposure.

Linguistic bias score (LBS) = \begin{cases} 
1, & \text{if fully linguistic} \\
-1, & \text{if fully surface} 
\end{cases}
At EMNLP, 2020.

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

Alex Warstadt,1 Yan Zhang,2 Haau-Sing Li,1 Haokun Liu,2 Samuel R. Bowman1,2,3

1Dept. of Linguistics, 2Dept. of Computer Science, 3Center for Data Science
New York University
Why should we conduct language acquisition experiments on language models?

How does the distribution of syntactic phenomena in the input affect grammatical generalization?
Subject Auxiliary Inversion

Adults always acquire the linguistic generalization... Children never even entertain the surface generalization.

(Crain and Nakayama, 1987)

Example: McCoy et al. (2020)
“Surely, if children hear enough [disambiguating examples], then they could reject the [linear] hypothesis. But if such evidence is virtually absent from the linguistic data, one cannot but conclude that children do not entertain the [linear] hypothesis, because the knowledge of structure dependency is innate.”

(Legate & Yang, 2001)
The man who **has** gone **has** seen the cat.

**Surface Generalization:**
Move the *first* auxiliary to the front.

**Linguistic Generalization:**
Move the *structurally highest* auxiliary to the front.

Has the man who ___ gone **has** seen the cat?
The Indirect Evidence Hypothesis

While a child may not receive direct evidence about the correctness of a particular hierarchical phrase structure rule..., there is vast indirect evidence for the general superiority of syntax with that structure throughout language. A learner who adopts a hierarchical phrase structure framework for describing the syntax of English will arrive at a much simpler, more explanatory account of her observations than a learner who adopts a linear framework.

(Perfors, Tenenbaum, Regier, 2011)
LMs and Subject Auxiliary Inversion

Earlier findings:

- LMs trained from scratch on ambiguous data usually adopt the surface generalization. (McCoy, Frank, and Linzen, 2018, 2020; Petty and Frank, 2022)

- Pretrained LMs fine-tuned on ambiguous data usually adopt the linguistic generalization. (Warstadt and Bowman, 2020; Mueller et al. 2020)

Confound: Pretraining data contains some direct evidence.
Language Deprivation Experiment

Questions:

1. Does direct evidence have a causal impact on generalization?
2. Is indirect evidence sufficient to learn the linguistic generalization?
Models

48 RoBERTa models pretrained from scratch

- 2 main conditions
- 4 sizes
- 3 runs (failed runs discarded)
- 2 domains (written, spoken)
Results: General acceptability judgments on BLiMP

Question: Did the removal of direct evidence have effects on unrelated phenomena?

Answer: No
Results: Subject Aux Inversion

Question: Did the removal of direct evidence affect generalization on subject auxiliary inversion?

Answer: Slightly, only in the written domain.
Question: Is indirect evidence sufficient to acquire the linguistic generalization?

Answer: Yes, but only in the best case.
Roadmap

Summary

CHAPTER 6

The Role of Indirect Evidence in Grammar Learning:
Investigations with Causal Manipulations of the
Learning Environment


Can neural networks acquire a structural bias from raw linguistic data?

Alex Warstadt (warstadt@nyu.edu)
Department of Linguistics, New York University
New York, NY 10003 USA

Samuel R. Bowman (bowman@nyu.edu)
Department of Linguistics & Center for Data Science & Department of Computer Science, New York University
New York, NY 10003 USA

Advantages & Disadvantages: The data efficiency gap
UPCOMING Shared task @ CMCL/CoNLL 2023

Objectives:

1. Data efficient pretraining
2. Plausible cognitive models
3. Democratization of pretraining research

Challenge

- 100 million words
- Mostly transcribed speech
- Test on acceptability and downstream tasks

Track 1: Strict

Track 2: Strict-small

Track 3: Loose

- Unlimited non-linguistic data
- Unlimited model-generated data

Plausible corpus
Is a Picture Really Worth a Thousand Words 
(with Theodor Amariucai & Ryan Cotterell)

Big Question: How much can we close data-efficiency gap using multimodal input?

Prior work:

● Multimodal vision + text models are becoming ubiquitous.
● Models are typically pretrained LMs, fine-tuned on captions data.
● Models are rarely tested in a language-only setting.

Our approach: Multitask multimodal learning on complex and abstract texts.
Is a Picture Really Worth a Thousand Words
(with Theodor Amariucai & Ryan Cotterell)

Some Probing Task

Image quantity

Text quantity

Performance

100%

Schematic

100%

chance
Interactive Language Mode (With Lennart Stoehler, Mitja Nikolaus, and Ryan Cotterell)

Big Question: How much can we close data-efficiency using inter-agent interaction

Prior work:

Jenny is wearing a crown.

Lazaridou et al., 2020; Nikolaus & Fourtassi, 2021
Interactive Language Mode
(With Lennart Stoeppler, Mitja Nikolaus, and Ryan Cotterell)

Our approach:
Prosody and LMs

(With Lukas Wolf, Tamar Regev, Eghbal Hosseini Ethan Wilcox, & Ev Fedorenko)

Question 1: How much information does prosody encode that isn’t in the text?

An utterance can be decomposed into two variables:

- $T =$ the text (i.e., a string of words)
- $P =$ the prosody (i.e., pitch + loudness + duration)

What is $\text{MI}(T; P)$?

Method: Train the most powerful possible probe to predict prosodic features from text. (Pimentel et al., 2020)
Question 2: How much can we close the data-efficiency gap by adding prosodic information to LM training data.

Methods:

1. Extract text & prosody from audio corpus.
2. Predict prosody from our probe for a text-only corpus, and give those representations to the LM during training.
Roadmap

Summary

What Artificial Neural Networks Can Tell Us About Human Language Acquisition

Alex Warstadt, Samuel R. Bowman

In Algebraic Structures in Natural Language, 2022.

Call for Papers - The BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus

https://babylm.github.io/

Alex Warstadt
ETH Zürich

Leshem Choshen
IBM Research

Aaron Mueller
Johns Hopkins University

Ethan Wilcox
ETH Zürich

Adina Williams
Meta AI

Chengxu Zhuang
MIT

At CoNLL and CMCL, forthcoming in 2023.
Conclusions

Figure 1: The Transformer - model architecture.

Figure 2: Human baby
Thank you!

Collaborators: Sam Bowman, Amanpreet Singh, Alicia Parrish, Yian Zhang, Haokun Liu, Haau-Sing Li, Sheng-Fu Wang, Anhad Mohananey, Wei Peng, Theodor Amariucai, Lennart Stoepler, Ryan Cotterell
Bonus Slides
The Recipe for Model Learners
As with any scientific model, there are obvious limitations with LMs.
Relevance to humans

Differences from humans
Debates in language acquisition often center around the necessary and sufficient conditions for human-learnability.
Suppose the model **SUCCEEDS** given some experimental manipulation. How likely are humans also to succeed?
Suppose the model **FAILS** given some experimental manipulation. How likely are humans also to succeed?

Model is at a great advantage  

Likelihood that humans show same result  

Human is at a great advantage
Positive results will generally be more relevant than negative results.

Model is at a great advantage

Likelihood that humans show same result

Model failure

Model success

Model is at a great advantage

Human is at a great advantage
Positive results will generally be more relevant than negative results.

Model is at a great advantage

Likelihood that humans show same result

Model failure

Model success

Model is at a great advantage

Human is at a great advantage
A recipe for relevant model learners:

- Maximize relevance of positive results by minimize advantages that models have over humans.
- Maximize chances of positive results by minimizing advantages that humans have over models.
Environmental vs. Innate advantages

- It’s relatively obvious how to apply this recipe to environmental advantages.
- But how do we apply this recipe to innate properties of the learner?

Typical ANNs appear to have weak language-specific advantages. But measuring and manipulating inductive bias is a serious problem where we don’t have great solutions.
Indirect evidence
Distribution of direct evidence (by domain)
Discussion: What does indirect evidence for hierarchical structure look like?

1. Classic constituency tests

**Fragment answers**

**Who** has seen the cat? [The man who was here this afternoon]

**Coordination**

John and [the man who was here this afternoon] are friends.

**Pronominalization**

[The man who was here this afternoon] left. **He** saw the cat.
Discussion: What does indirect evidence for hierarchical structure look like?

2. Other hierarchical rules

Subject Verb Agreement
[The man who saw the cats] is here.

Passivization
I greeted [the man who saw the cat.] → [The man who saw the cat] was greeted by me.
Intro stuff
The Mystery of Human Language Acquisition

Thousands of linguists have spent decades trying to describe the grammar of human language (and only partly succeeding).

How does a single child acquire the grammar of their native language in a matter of years?
Language acquisition is based on the child's discovery of what from a formal point of view is a deep and abstract theory a generative grammar of his language — many of the concepts and principles of which are only remotely related to experience by long and intricate chains of unconscious quasi-inferential steps.

A consideration of the character of the grammar that is acquired, the degenerate quality and narrowly limited extent of the available data, the striking uniformity of the resulting grammars, and their independence of intelligence, motivation, and emotional state, over wide ranges of variation, leave little hope that much of the structure of the language can be learned by an organism initially uninformed as to its general character.

(Chomsky, 1965)
Two Sources of Grammatical Knowledge

- **Innate Bias**
- **The Environment**
My research uses language models to study how the environment affects language learning.
Learning which features matter
An example
Pretraining → Feature Learning

- Dependency structures can be extracted from BERT (Hewitt & Manning, 2019)
- Contextual embeddings contain POS, semantic roles, coreference, etc. (Tenney et al., 2019a/b)

...and many more (see Rogers et al., 2020)
But feature learning isn’t everything.
Representing $F \neq \text{Using } F$

Models that represent linguistic features can still fail to use them during fine-tuning (McCoy et al., 2019).
Representing $F \neq \text{Using } F$

Models that represent linguistic features can still fail to use them during fine-tuning (McCoy et al., 2019).

*Inductive bias* is also crucial to good generalization.
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.*
Results:
Experiment 2
(Ambiguous)
(Fine-grained)
The bias in favor of absolute position and orthography (surface features) is very strong.
The bias in favor of sentence length (surface feature) is fairly weak.
Data Generation

- The MSGS data is generated from templates.
- We always test classifiers’ ability to generalize out-of-domain.

In domain: *The big dog is yawning.*

Out of domain: *The dog in the dark forest yawned.*
Fine-tuning

- 9 tasks (4 linguistic + 5 surface)
- 12 miniBERTas + original RoBERTa\textsubscript{BASE} (~30B words)
- The training sets are 10k sentences each
Results: Experiment 1 (Feature Learning)

Surface features: Performance is at ceiling.
Results: Experiment 1 (Feature Learning)

Surface features:
Performance is at ceiling.

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

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.
For subsequent experiments, we’ll exclude any models where feature learning performance <0.7 (gray points).
Experiment 2: Ambiguous Data

Does model X prefer linguistic feature A or surface feature B?
Experiment 2: Ambiguous Data

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

We fine-tune X on an ambiguous binary classification task.
Experiment 2: Ambiguous Data

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

We fine-tune X on an ambiguous binary classification task.

Poverty of the Stimulus design (Wilson, 2006)

- Also used by McCoy et al. (2018, 2020), Warstadt & Bowman (2020), and others.
NLP

EVALUATION

ENGINEERING

EXPERIMENTATION

Linguistic theory

Dev. psych & Cogsci
The MiniBERTas

- 4 incremental datasets: 1M, 10M, 100M, 1B words
- 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 data size, we train at least 10 models with varying hyperparameters (e.g., # of parameters) & select the best 3.

https://huggingface.co/nyu-mll
Hypothetical Human Inductive Biases

Linguistic features
- Inflectional form
- Syntactic category
- Syntactic position
- Semantic roles

Surface features
- Linear position
- Length
- Lexical content
- Orthography
- Linear precedence
Syntactic filtering

Training data: 1B words from books & Wikipedia

- Percent filtered: 1.7%
- Recall (% of direct evidence removed): 99%
- Precision (% of removed data that is direct evidence): 51%
Evaluation

We do BLiMP-style evaluation on a hand-crafted test suite of subject-auxiliary inversion minimal pairs.

We designed minimal pairs following 8 different templates to probe generalization to different syntactic structures, and compared LM scores for the good and bad sentences.
Results: General acceptability judgments on BLiMP

This result holds across all phenomena in BLiMP.
Takeaways

The results support the indirect evidence hypothesis, but with important caveats.

- How reproducible is the best model’s success?
- How important are small amounts of direct evidence that passed through the filter?
- Can models succeed with the same data-volume limitations as humans?
- Can we identify and quantify indirect evidence?
The Corpus of Linguistic Acceptability (CoLA)

BUT...
Roadmap

1. BACKGROUND
2. INDUCTIVE BIAS
3. INDIRECT EVIDENCE
4. FUTURE DIRECTIONS
Developments in text generation (2015-now)

how it started

===Widely accepted grammars===

There are twelve dialects which concern under the language of which which in sufficient, areas will be surprising before the racial controversy, probably those who in history, and no consensual is sincere.

Karpathy (2015)
http://karpathy.github.io/2015/05/21/rnn-effectiveness/
(h/t Will Merrill)

how it’s going

Generate a wikipedia article titled:

===Widely accepted grammars===

In linguistics, grammar refers to the set of rules that govern the structure of a language…. One of the most well-known grammars is the generative grammar proposed by Noam Chomsky in the 1950s.

GPT-4 (OpenAI, 2023)
Acceptability Judgments

An empirically adequate grammar of a language $L$ generates all and only the grammatical strings of $L$.

Acceptability judgments are the primary behavioral test of grammatical theories in linguistics.

Examples from linguistics publications

- ✓ Mary should know that you must go to the station.
- ✓ I promised that around midnight he would be there.
- ✓ Susan whispered the news to Rachel.
- ✓ When time will you be there?
- ✗ Patrick is likely that left.
- ✗ Harry coughed us into a fit.
The MiniBERTas on BLiMP
Results:

Experiment 1
(Fully Ambiguous)

- 20 tasks * (12 miniBERTas + RoBERTa base)
- Linguistic bias score = 1 if linguistic, -1 if surface.
- <1B words: surface bias
- RoBERTa base: 50/50
Results: Subject Aux Inversion (BEST CASE)
The Recipe for Model Learners

1. Minimize any advantages that language models have over humans learners.

2. Provide language models with more of the advantages that we know humans have.

3. Gather training data from developmentally plausible sources.