Learning What Models Can't: A Case Study on Rare Linguistic Phenomena

Ling 575 Project Proposal
Iskar Deng, Nathalia Xu

1. Introduction

In recent years, researchers have explored whether artificial neural networks can offer insights into human language acquisition. Warstadt and Bowman (2024)[1] propose that if a model weaker than a human learner can acquire a linguistic phenomenon, humans likely can too. Because models are disadvantaged in perceptual input, cognitive biases, and environmental interaction, a model's successful learning outcome is highly generalizable, while its failure to learn does not necessarily imply human incapacity, given that humans possess richer learning mechanisms. This suggests that there may indeed exist languages that are impossible to learn for models, but possible for humans.

Recent studies have further highlighted models' inductive limitations. Kallini et al. (2024)[2] systematically constructed impossible languages and showed that even large pretrained models (e.g., GPT-2) exhibit significantly reduced learning ability when confronted with artificial languages that violate the statistical properties of natural languages. This provides evidence that languages difficult for models to acquire can indeed be constructed. Building on this, this project asks: **Are there real but rare linguistic phenomena in natural languages that models find difficult to acquire from pure textual input, but humans are nevertheless able to learn?** Focusing on counterintuitive structures with low information locality—such as case-marking based on relational animacy—the project will systematically rewrite controlled English corpora and conduct modeling experiments to test the learning abilities of small-scale models under different conditions of linguistic relatedness and information locality, aiming to better understand the broader relationship between human language acquisition and model inductive capacity.

2. Methods

The overall training process follows the framework established in *Mission: Impossible Language Models* (Kallini et al., 2024)[2], with particular emphasis on controlling input size, the scope of data modifications, and model initialization conditions.

2.1 Model and Training Strategy

The primary model used in this study is GPT-2 small (Radford et al., 2019)[3]. The training strategy adopts random initialization from scratch, without loading any pretrained weights, in order to avoid the influence of prior learning biases.

2.2 Data

The corpus used in this project is based on the BabyLM Challenge Dataset (Warstadt et al., 2023)[4], which contains approximately 100 million tokens sourced from child-directed speech, simulating the environment of human language acquisition.

To prepare the data, sentences from subsets of the BabyLM corpus are systematically rewritten to introduce the target grammatical phenomena. Each sentence is first annotated using *Stanza* (Qi et al., 2020)[5] before rule-based modifications are applied. To preserve the lexical naturalness throughout this process, artificial markers are only introduced when necessary.

Training and validation sets are separately constructed to ensure that the evaluation of inductive learning capabilities can be conducted independently and reliably.

2.3 Language Phenomena

This project initially selects relational animacy-based case marking systems as the target phenomenon for preliminary experiments, and takes the Naxi language as a starting point. In Naxi, transitive sentences where the subject has a higher animacy than the object, neither argument requires an explicit case marker; however, when the subject's animacy is not greater than that of the object, both the subject and object must be marked with a case particle. This rule depends on cross-argument comparison between the subject and object rather than properties of a single argument.

Several factors motivate the selection of relational animacy-based case marking in Naxi:

1. **High adaptability:** As an analytic language with little verbal inflection, Naxi allows straightforward rewrites from English without major morphological adjustments.
2. **Low information locality:** Adding a case marker requires comparing subject and object animacy, which violates the principle of information locality that natural languages typically rely on (Kallini et al., 2024)[2], thereby posing a challenge to model generalization.
3. **Rarity:** Direct case marking based on subject-object relationships is rare (Comrie, 1989)[6], as most systems rely on single-argument features like nominative-accusative or ergative-absolutive alignments.

Previous works show that neural models like Transformers strongly favor information locality, struggling with long-distance and hierarchical dependencies (Kallini et al., 2024; Yedetore et al., 2023)[2] [7]. Thus, selecting a real phenomenon that violates locality offers an effective test of a model's inductive capacity with atypical input. Depending on time and progress, the project also plans to extend the study to other rare case-marking systems, such as direct-inverse systems, ergative-absolutive alignment, and animacy-based case systems. Through these extensions, the project aims to systematically examine the effects of two key factors on model learning outcomes: (1) linguistic relatedness — the genealogical distance between the source language of the phenomenon and English, and (2) information locality strength — the degree to which local information is relied upon in inductive judgments.

2.4 Evaluation

The evaluation will focus on three primary metrics.

1. **Training set perplexity** will be tracked to measure changes in the model's confidence in the input distribution. The rate of reduction in perplexity can reflect the model's learning speed and stability.

2. **Few-shot prediction accuracy** will be assessed for case-marking classification or generation tasks, testing whether the model can correctly apply the target grammatical rules when given only limited support examples.
3. **Rule induction success** will be evaluated to determine whether the model can infer the underlying grammatical rules without direct prompts or explicit supervision.

In addition to the primary metrics, differential analysis will be conducted as an auxiliary evaluation. Specifically, the project will compare the model's success rates in acquiring phenomena drawn from different source languages, categorized by linguistic relatedness and information locality strength. This analysis aims to explore which types of linguistic structures are more easily or more challengingly induced by the model.

3. Possible Results

3.1 Phenomena Acquisition Outcomes

Successful acquisition cases are expected primarily for adapted corpus phenomena that rely on local and explicit cues, such as the presence of clear case particles near the verb. In particular, when the target phenomena originate from languages that are syntactically or morphologically close to English (e.g., Indo-European languages, African American Vernacular English [AAVE]), the models are likely to show faster learning curves (as reflected in a quicker decline in perplexity) and higher few-shot prediction accuracy.

Failure cases or induction failures are anticipated for phenomena that require cross-constituent reasoning (such as relational animacy comparisons between subject and object), involve non-explicit marking, or derive from languages whose overall structure differs substantially from English (e.g., the case-marking system of Naxi). In these cases, the models may exhibit slower perplexity reduction, lower prediction accuracy, and higher induction error rates under few-shot conditions.

3.2 Linguistic Relatedness

For phenomena originating from languages closely related to English, transfer is expected to be relatively smooth. Models are likely to adapt more quickly to new grammatical patterns, even when working with adapted corpora, if the source languages (e.g., German, French, AAVE) share many syntactic features with English.

In contrast, for phenomena from distantly related languages, transfer is expected to be much more limited. When the phenomena stem from structurally divergent languages such as Naxi or Basque, the model's induction success rate is expected to drop significantly, even when surface lexical items are adjusted to English.

A potential inference from these findings is that linguistic relatedness may affect not only surface-level lexical or syntactic structures but also reflect deeper differences in syntactic organization (e.g., case systems, ontological assumptions about argument structure). Successful cross-linguistic adaptation may

require a deeper alignment of core syntactic principles, whether in constructed experiments such as this project or in real-world cases of language contact and fusion.

3.3 Modulatory Effects of Information Locality

Phenomena with strong information locality—where determining case marking requires reference only to nearby elements (e.g., case particles near the verb)—are expected to be learned more easily by the models.

Phenomena with weak information locality—where successful case marking requires global reasoning across multiple constituents (such as comparing the animacy of the subject and object)—are expected to present significant learning challenges. Even with few-shot examples, models are likely to struggle with successful induction.

As suggested by Kallini et al. (2024)[2], models may sometimes develop shallow heuristics that simulate correct behavior without truly mastering the underlying rules. For example, a model might default to leaving animate subjects unmarked, rather than genuinely learning the animacy comparison mechanism.

3.4 Expected Theoretical Implications

The results are expected to support Warstadt and Bowman's (2024)[1] framework, particularly that successful model learning is broadly generalizable, while failure does not imply human failure. Findings will also align with Kallini et al. (2024)[2], showing that neural models strongly favor typical statistical properties like local dependencies and low structural complexity. Finally, the project aims to raise new hypotheses, possibly suggesting that information locality may better predict model inductive success than linguistic relatedness, and that deep structural features, such as argument role encoding, may significantly influence cross-linguistic transfer difficulty.

4. Division of Labor + Timeline

Timeframe	Task	Responsible
Early May	Select initial language phenomena (starting with relational animacy comparison) and formulate rewriting guidelines	Both members
Mid-May	Rewrite a subset of BabyLM corpus to construct the adapted dataset	Both members
Late May	Annotate and organize adapted corpus with phenomena labels and documentation	Both members
Early June	(If time permits) Conduct small-scale model training runs and record initial learning curves	Both members
By June 12	Compile final project report, including dataset construction, task design, and expected analyses	Both members

References

[1] A. Warstadt and S. R. Bowman, “What Artificial Neural Networks Can Tell Us About Human Language Acquisition,” Feb. 11, 2024, *arXiv*: arXiv:2208.07998. doi: 10.48550/arXiv.2208.07998.

[2] J. Kallini, I. Papadimitriou, R. Futrell, K. Mahowald, and C. Potts, “Mission: Impossible Language Models,” Aug. 02, 2024, *arXiv*: arXiv:2401.06416. doi: 10.48550/arXiv.2401.06416.

[3] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, and I. Sutskever, “Language Models are Unsupervised Multitask Learners”.

[4] A. Warstadt *et al.*, “Findings of the BabyLM Challenge: Sample-Efficient Pretraining on Developmentally Plausible Corpora,” in *Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning*, A. Warstadt, A. Mueller, L. Choshen, E. Wilcox, C. Zhuang, J. Ciro, R. Mosquera, B. Paranjabe, A. Williams, T. Linzen, and R. Cotterell, Eds., Singapore: Association for Computational Linguistics, Dec. 2023, pp. 1–34. doi: 10.18653/v1/2023.conll-babylm.1.

[5] P. Qi, Y. Zhang, Y. Zhang, J. Bolton, and C. D. Manning, “Stanza: A Python Natural Language Processing Toolkit for Many Human Languages,” Apr. 23, 2020, *arXiv*: arXiv:2003.07082. doi: 10.48550/arXiv.2003.07082.

[6] B. Comrie, *Language Universals and Linguistic Typology: Syntax and Morphology*. University of Chicago Press, 1989.

[7] A. Yedetore, T. Linzen, R. Frank, and R. T. McCoy, “How poor is the stimulus? Evaluating hierarchical generalization in neural networks trained on child-directed speech,” Jun. 06, 2023, *arXiv*: arXiv:2301.11462. doi: 10.48550/arXiv.2301.11462.