SYNTACTIC ISLANDS AND LLMS

ATAKAN INCE
Mercor

1. Introduction[1]

The aim of this study is to find out whether Large Language Models (llms) can classify sentences with phrasal extraction out of syntactic islands as ungrammatical, in contrast to sentences with syntactic islands but with phrasal movement originating from a position outside the island. Being a classification study, it is also supervised, i.e. the training dataset is labeled. Below, both 1a & 1b have the same adjunct island. However, the first sentence is grammatical because movement of *the blogger* originates from outside the island, whereas 1b is ungrammatical because the DP *the client* moves out of the adjunct island:[2]

1. a. I like [the blogger]$_1$ who _____$_1$ worries **ADJUNCT ISLAND** if the manager praises the client.

   b.* I like [the client]$_2$ who the blogger worries **ADJUNCT ISLAND** if the manager praises _____$_2$.

2. Data

In this study, as an exemplary case, we look at data that includes certain ungrammatical vs. grammatical constructions of (a) wh-questions and (b) relative clauses:

2. a. [Which student]$_1$ _____$_1$ wonders **WH-ISLAND** whether John bought a car?

   b. *[Which car]$_2$ do you wonder **WH-ISLAND** whether John bought _____$_2$?

[1] Thanks to Alex Krauska for editorial help, Niels Dickson, Nihan Ketrez, Ivan Ortega-Santos, Asad Sayeed, Ashwini Tayade, Juan Uriagereka and one anonymous reviewer for helpful feedback. I am indebted to Howard Lasnik for his guidance in linguistics, linguistic thinking and ping pong over the years.
[2] With 'island', we are referring to specific constructions whether extraction occurs out of them or not, such as *wh-islands*, *subject islands*, *complex NP phrases*, *adjunct islands*, etc. Islands are highlighted.

3. a. I called [the secretary]$_1$ who _____$_1$ worries if the lawyer insults the client.

ADJUNCT ISLAND

b. *I called [the client]$_2$ who the secretary worries if the lawyer insults _____$_2$.

ADJUNCT ISLAND

In the ungrammatical cases, a phrase ('which car' in 2b & 'the secretary' in 3b) is an argument of a verb inside a so-called 'syntactic island' ('bought' and 'insults', respectively) (Ross 1967) but occurs outside the island. In 2a&2b, a phrase ('which student' in 2a & 'the secretary' in 3a) is an argument of a verb outside a syntactic island ('wonders' & 'worries', respectively), hence no ungrammaticality.

We limited the study to four types of syntactic island:

I. Wh-island[3]

*[Which car]$_1$ do you wonder whether Mary sold ____$_1$?

II. Complex NP island

*[Which tart]$_2$ did you hear the statement that Jeff baked ____$_2$?

III. Subject island

*[Which lobbyist]$_3$ do you think the letter from ____$_3$ prompted the rumor about the Senator?

IV. Adjunct island

*[Which file]$_4$ do you worry if the inspector leaves ____$_4$ at the office?

We used two constructions to form the training dataset:

4. a. Wh-questions

a'. Which paper did you read?

[3] We include *whether*-islands in *wh*-islands.

b. Relative Clauses

b'. The paper I read yesterday was great.

The other factors we included are whether a phrase is extracted out of a matrix clause or embedded clause (matrix vs embedded) and whether the sentence includes an island or not (island vs non-island), whether a phrase is extracted out of the island or not. Therefore, when we include the construction types, we have a 2 x 2 x 2 factorial design for the training dataset:[4]

5. (dependency) x (matrix vs embedded) x (island vs non-island)

6. a. (wh-dependency, matrix, non-island)
   Who thinks that John bought a car?
   b. (wh-dependency, embedded, non-island)
   What do you think that John bought?
   c. (wh-dependency, matrix, island)
   Who wonders whether John bought a car?
   d. (wh-dependency, embedded, island)
   *What do you wonder whether John bought?

7. a. (rc-dependency, matrix, non-island)
   I know the manager who heard that Laura is dating the CEO.
   b. (rc-dependency, embedded, non-island)
   I know the CEO who the manager heard that Laura is dating.
   c. (rc-dependency, matrix, island)
   I know the manager who heard the rumor that Laura is dating the CEO.
   d. (rc-dependency, embedded, island)
   *I know the CEO who the manager heard the rumor that Laura is dating.

[4] The reviewer suggests including data related to "*that*-trace effects" and "island repair". However, as stated above, this study focuses on certain island cases, without any theoretical basis. Including data related to "*that*-trace effects" would require a totally new factorial design, which is beyond the scope of this study. We assume "island repair" cases would not be relevant to this specific study.

In sum, 25% of the training dataset includes extraction out of syntactic islands, hence such sentences are ungrammatical.[5] Since the train dataset is labeled as 'grammatical' vs. 'ungrammatical', this is a **supervised learning** experiment.

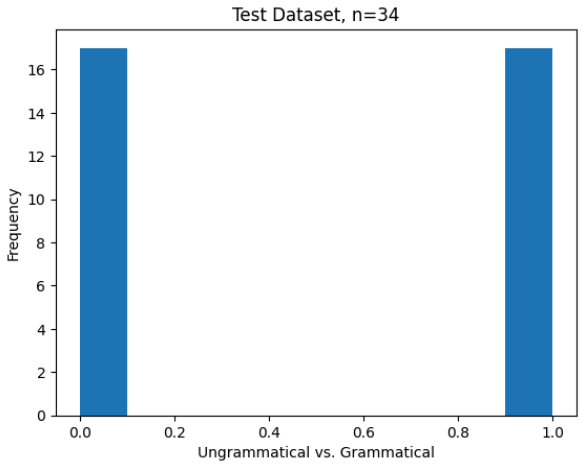

Figure 1. Test Dataset Used in all the experiments

As to the test dataset, we included minimal pairs. In each minimal pair, one sentence is grammatical and the other is ungrammatical. All sentences include a Relative Clause with a syntactic island inside it.The test dataset has 34 sentences, with 17 minimal pairs.[6]

[5] The fact that the train dataset is imbalanced is a natural consequence of the factorial design. The factorial design requires that only 25% of the data be ungrammatical. Again, when a model is fine-tuned, the distribution of data is kept identical to the real world data, and rarely is an equal distribution observed in the real world. Therefore, equal distribution is not a requirement.

[6] The reviewer is worried about the test size, pointing out that "[i]t is standard practice that the test set be at least 1/4 the size of the training data to make sure the results are significant." However, there is no such requirement. There are many studies with very small test sets. Besides, by designing the test dataset as minimal pairs, we believe the test dataset is challenging enough, quality being over quantity. We also believe that if a fine-tuned model can discriminate between such minimal pairs, the results are significant.

3. Fine-Tuning

For this study, we chose DistilBERT (Sanh et. al. 2019) because it is smaller and faster than BERT. We used Google Cloud Platform and a GPU.[7] We used 20% of the training data for validation and ran three epochs in each experiment.

3.1. EXPERIMENT 1

In Experiment 1, we use 750 sentences of the (labeled) training dataset. It includes both wh-questions and relative clauses of the four patterns in (6) & (7) above.

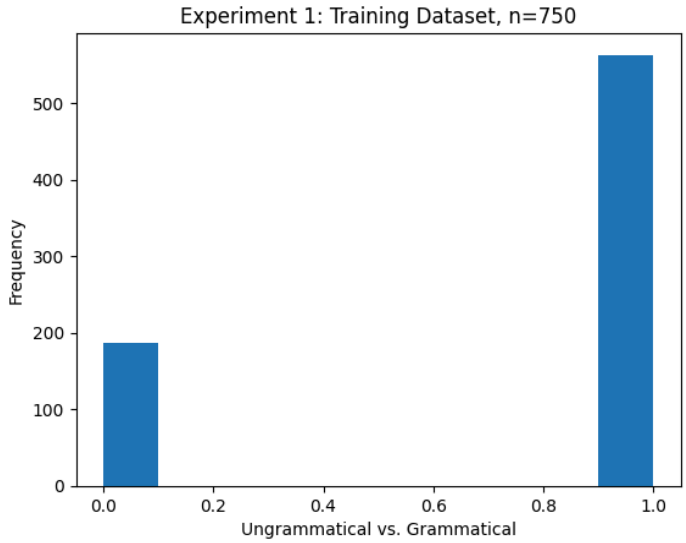

Figure 2: Experiment 1: both wh-questions and relative clauses in the training dataset (n=750)

[7] The reason we choose this model is that it is smaller than the standard BERT models. This way, we aim to consume less energy during fine-tuning and see how well a smaller model can do. Generally, the larger models perform better than smaller models. Therefore, we do not see a reason to run the same experiment with larger BERT models. Neither do we aim to compare the performance of different models (Asad Sayeed, pers. comm.).

Results of Experiment 1

The confusion matrix below summarizes the predictions of the model on the test dataset.[8] The model predicts grammatical sentences correctly (50%) but also predicts 14 (41%) out 17 ungrammatical cases as grammatical as well, and only 3 ungrammatical sentences are predicted correctly (8.9%).

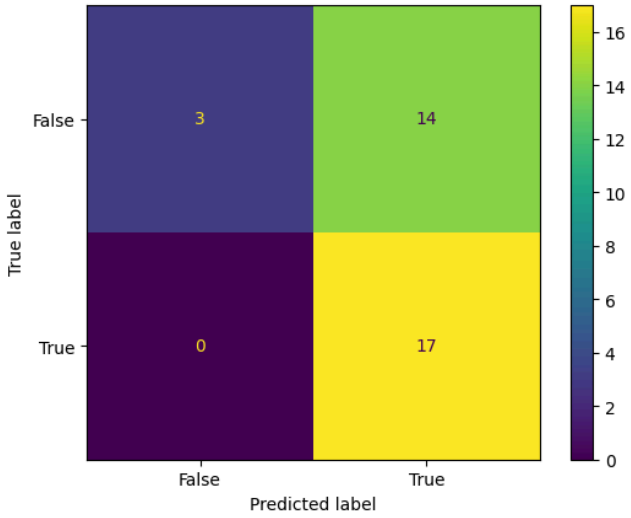

Figure 3: Confusion Matrix for the Test Dataset Predictions (Experiment 1)

	precision	recall	f1-score	support
Ungrammatical	**1.00**	**0.18**	**0.30**	**17**
Grammatical	**0.55**	**1.00**	**0.71**	**17**
accuracy			**0.59**	**34**
macro avg	**0.77**	**0.59**	**0.50**	**34**
weighted avg	**0.77**	**0.59**	**0.50**	**34**

Table 1: Classification Report (Experiment 1)

In the first experiment, we see that F1 score is 0.50 (both macro and weighted). The model predicts by chance. In other words, it has not learned classifying the grammatical vs. ungrammatical cases in the test dataset. Precision and Recall are high, but Specificity (out of the

[8] The reviewer also recommends using ROC AUC curves/scores. However, ROC AUC curves/scores are used to compare the effect of different thresholds. Since we are not manipulating any parameter such as thresholds, there is no need for ROC AUC curves/scores.

ungrammatical sentences, how many of them were correctly classified as ‘ungrammatical’?) is low (0.18). In sum, the model overgeneralizes.

3.2. EXPERIMENT 2

In the second experiment, we increase the training data to n=1500 (labeled sentences), keeping data structure as in Experiment 1, and see if increasing the data size would increase the precision, recall, F1 and specificity scores.

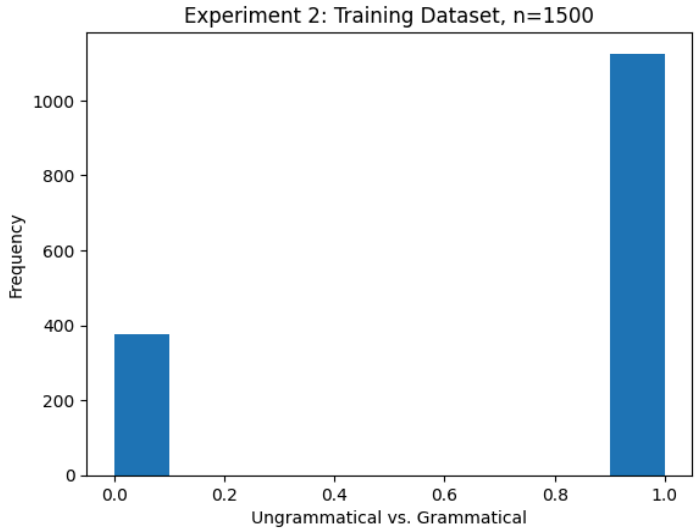

Figure 4: Experiment 2: both wh-questions and relative clauses in the training dataset (n=1500)

Results of Experiment 2

As seen in the confusion matrix below, the model predicts both grammatical and ungrammatical sentences correctly.

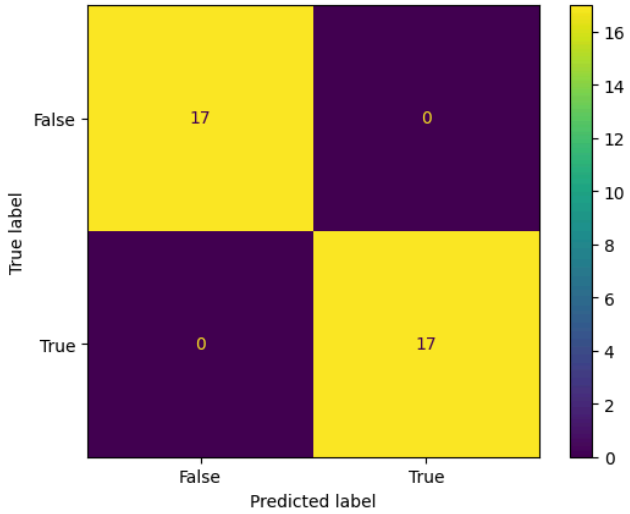

Figure 5: Confusion Matrix for the Test Dataset Predictions (Experiment 2)

	precision	recall	f1-score	support
Ungrammatical	**1.00**	**1.00**	**1.00**	**17**
Grammatical	**1.00**	**1.00**	**1.00**	**17**
accuracy			**1.00**	**34**
macro avg	**1.00**	**1.00**	**1.00**	**34**
weighted avg	**1.00**	**1.00**	**1.00**	**34**

Table 2: Classification Report (Experiment 2)

In the second experiment we see perfect scores: 1 for precision, recall, F1 and specificity. However, this could be a result of **rote learning** by the llm (Ashwini Tayade, pers. comm.). In other words, the model could have memorized the patterns, since sentences in the training data designed by "factorial design" are very similar. Since factorial design is used in forming the training dataset, words are distributed equally across sentences. Therefore, word distribution cannot be a factor in overfitting. Also, the position of functional heads such as *if*, *who* etc. are the same between grammatical and ungrammatical sentences. Neither can they be a factor in overfitting. In most ungrammatical cases, it's the Direct Object of the verb inside the island that is extracted. In the grammatical versions of the same sentence, the Direct Object of the verb inside the island is not extracted, the rest of the sentence being the same in terms of token similarity and sentence structure similarity. This could be a major factor in **rote learning**.[9]

[9] I am indebted to the reviewer for suggesting possibilities for **rote learning** in the study.

3.3. Interim Conclusion

In the two experiments above, we see that the fine-tuned model can learn to classify the sentences with movement out of a syntactic island as ungrammatical. As the training data size increased from Experiment 1 to Experiment 2, so did the precision, recall, precision and F1 scores. However, in Experiment 2, we see a perfect model, with 100% accuracy, which could be a result of **rote learning**.

To block possible rote learning, we will run two more experiments. We will use only sentences with wh-questions and exclude the ones with relative clauses. The test dataset will remain the same, consisting only of sentences with relative clauses. Thus, we will minimize any possibility of **rote learning**. We will keep the training data size the same as in the first 2 experiments: Experiment 3 (n=750), and Experiment 4 (n=1500), to see if accuracy drops between Experiment 1 vs Experiment 3, and Experiment 2 vs Experiment 4. We will also see whether accuracy increases as the training data size increases from Experiment 3 to Experiment 4.

3.4. EXPERIMENT 3

In the experiment 3, we will include only sentences with wh-questions and exclude the ones with relative clauses, keeping the number of (labeled) sentences at 750 for the training dataset.

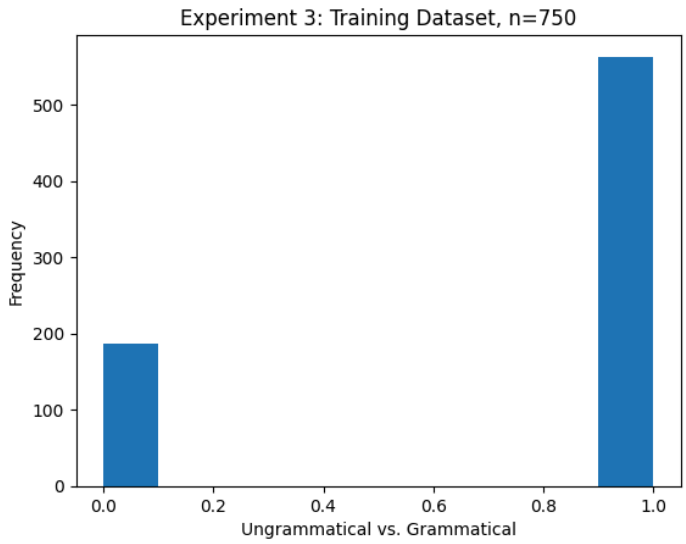

Figure 6: Experiment 3: only wh-questions in the training dataset (n=750)

Results of Experiment 3

In the confusion matrix below, we see that the model predicts every sentence in the test dataset as grammatical.

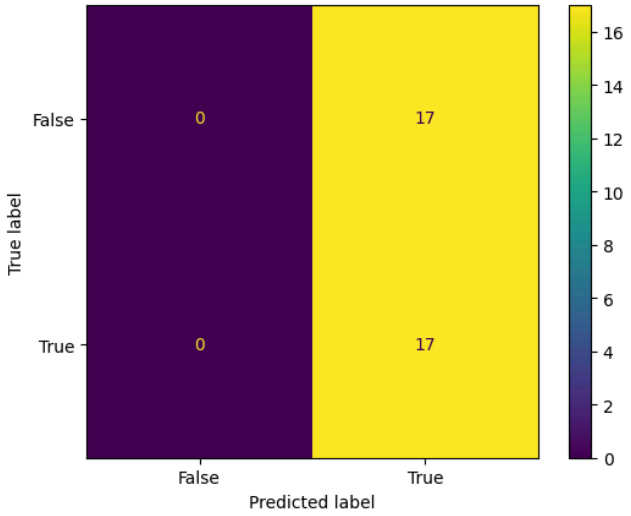

Figure 7: Confusion Matrix for the Test Dataset Predictions (Experiment 3)

	precision	recall	f1-score	support
Ungrammatical	**0.00**	**0.00**	**0.00**	**17**
Grammatical	**0.50**	**1.00**	**0.67**	**17**
accuracy			**0.50**	**34**
macro avg	**0.25**	**0.50**	**0.33**	**34**
weighted avg	**0.25**	**0.50**	**0.33**	**34**

Table 3: Classification Report (Experiment 3)

In this experiment, we see precision, recall and F1 are low. Specificity is 0%. In other words, the model has not learned to classify sentences with syntactic islands correctly. Compared to Experiment 1, specificity is much worse, while precision, recall and F1 are lower, due to the fact that the training dataset does not include any sentences with relative clauses.

3.5. EXPERIMENT 4

In Experiment 4, the last experiment, we will increase the training dataset to n=1500.

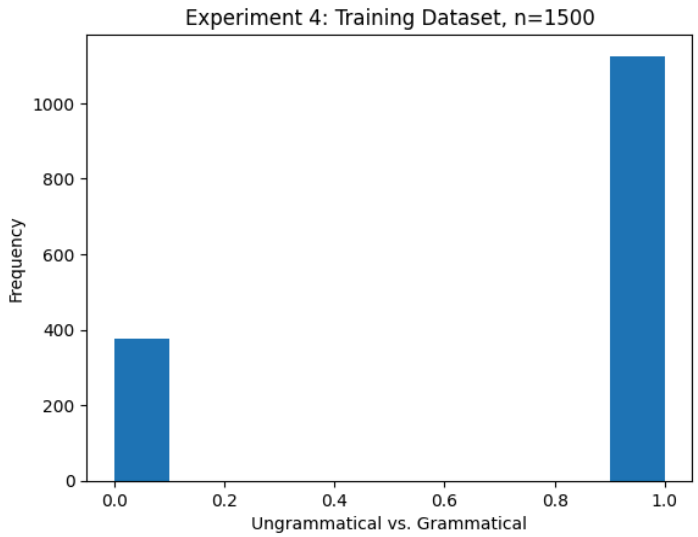

Figure 8: Experiment 4: only wh-questions in the training dataset (n=1500)

Results of Experiment 4

The confusion matrix below shows that 11 ungrammatical sentences were predicted as grammatical (32%), only 6 ungrammatical sentences were predicted correctly (18%), 2 grammatical sentences were predicted as ungrammatical (6%), and 15 sentences were predicted correctly (44%).

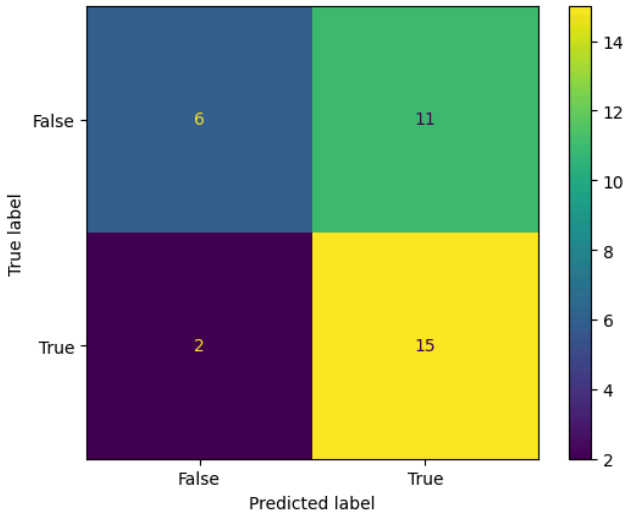

Figure 9: Confusion Matrix for the Test Dataset Predictions (Experiment 4)

	precision	recall	f1-score	support
Ungrammatical	**0.75**	**0.35**	**0.48**	**17**
Grammatical	**0.58**	**0.88**	**0.70**	**17**
accuracy			**0.62**	**34**
macro avg	**0.66**	**0.62**	**0.59**	**34**
weighted avg	**0.66**	**0.62**	**0.59**	**34**

Table 4: Classification Report (Experiment 4)

In this experiment, we see an increase in the scores, compared to the results in the previous experiment. In Particular, precision jumps from 0% to 30%. Although precision is still low (below chance level), the increase demonstrates that the model is learning. With more training data, precision could increase. Also, compared to Experiment 2, all over scores are lower, which means that **rote learning** has been minimized.

3.6. LEARNABILITY

In this section, we would like to compare llms and children as to language learning. First, llms are trained on four/five orders of magnitude more language data than children (Frank 2023: 990). This makes children very efficient learners (named as ‘Poverty of the Stimulus’, ‘Plato’s Problem’, ‘the logical problem of language acquisition’) (Lasnik & Lidz 2017). Second, Primary Linguistic Data (PLD), which children are exposed to, does not include negative data, whereas the data used for fine-tuning llms (training data) includes negative data as well (as in this study). However, children can still make judgments for negative data. Third, the data llms are exposed to is text data, whereas children are exposed to multimodal data in the real world, mostly through structured social interactions. Since the input data and its amount are different for children and llms, it is hard to compare their performance with the same tools and metrics. Children’s language comprehension and production is generally measured with multimodal data (pictures, videos, etc.), and llms are tested on benchmarks based on text data.

4. Conclusion

In this paper, we tackled the question whether llms can learn to classify cases of extraction out of syntactic islands as ungrammatical. For this purpose, a supervised classification study was designed. We designed a test dataset that consists of minimal pairs, which could be challenging because the sentences in each pair are very similar structure-wise and token-wise. To minimize rote learning, we also used training datasets consisting of (a) wh-questions + relative clauses (Experiments 1&2) and (b) only wh-questions (Experiments 3&4), whereas the test dataset consists of only relative clauses. In Experiments 3&4, we see that as the size of the training dataset increases, the model performs better. This would mean that the model can learn how to classify the ungrammatical sentences. In conclusion, this study shows that with carefully designed training and test datasets, we can test whether llms can learn certain syntactic constructions, minimizing any confounding factors such as rote learning.

References

Balloccu, S., P. Schmidtová, M. Lango, and O. Dusek. 2024. Leak, Cheat, Repeat: Data Contamination and Evaluation Malpractices in Closed-Source LLMs. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 67–93, St. Julian's, Malta. Association for Computational Linguistics.

Frank, Michael C. 2023. "Bridging the data gap between children and large language models", *Trends in Cognitive Sciences* 27/11: 990-992.

Lasnik, Howard. 2011. Another look at Island Repair by Deletion. Islands in Contemporary Linguistic Theory, University of the Basque Country.

Lasnik H, Lidz J. 2017. The argument from the poverty of the stimulus. In *Oxford Handbook of Universal Grammar*, ed. I Roberts, pp. 221–48. Oxford, UK: Oxford Univ. Press

Ross, John Robert. 1967. *Constraints on variables in syntax*. Doctoral dissertation, MIT, Cambridge, Mass. Published as *Infinite syntax!* Norwood, N.J.: Ablex (1986).

Sanh, V., I. Debut. J. Chaumond, T. Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108.

Schwarzschild, A., Z. Weng, P. Maini, Z. C. Lipton, J. Z. Colter. 2024. Rethinking LLM Memorization through the Lens of Adversarial Compression. arXiv preprint arXiv:2404.15146

Sprouse, J., I. Caponigro, C. Greco, & C. Cecchetto. 2016. Experimental syntax and the variation of island effects in English and Italian. Natural Language and Linguistic Theory 34: 307-344.