Paper Replication: Investigating Syntax Evaluation Across Languages

Lily Gamburg, William Johnson, Edgar Suritis

Natural Language Processing
Department of Computer Science
Colgate University

1 Introduction

Recent advances in neural language models have led to major improvements in natural language processing, but there is still debate about what these language models actually learn. One key question is whether language models capture the underlying grammatical rules of language or simply memorize the surface-level patterns from their training data. Much of the previous research on this topic has focused exclusively on English, where models with LSTM and BERT architectures have been shown to correctly distinguish between grammatical and ungrammatical sentences. However, it remains unclear whether this syntactic knowledge extends to other languages, especially those that differ significantly in structure and morphological complexity. Investigating this question is essential to reveal whether these models truly generalize linguistic knowledge across languages.

The paper *Cross-Linguistic Syntactic Evaluation of Word Prediction Models* (Mueller et al., 2020) addresses this issue by evaluating how well neural language models capture syntax across multiple languages. The authors developed the CLAMS (Cross-Linguistic Assessment of Models on Syntax) dataset (Mueller, 2020), which includes pairs of grammatical and ungrammatical sentences for a range of syntactic constructions such as subject-verb agreement and reflexive anaphora. Their study compared monolingual and multilingual language models to see whether syntactic knowledge learned in one language could be transferred to others. Mueller et al. found that monolingual models demonstrated some understanding of syntax, but their accuracy declined for complex structures, and multilingual models generally performed worse, suggesting that exposure to multiple languages may lead to interference rather than improvement.

In this paper, we recreate Mueller et al.'s results and examine two related questions. Firstly, how well do BERT-based models trained for next-word prediction capture subject-verb agreement and reflexive anaphora? Secondly, is the accuracy of a model related to the morphological complexity of the language being tested? To address these questions, we used both the English-trained BERT-base-cased and the multilingual BERT-base-multilingual-cased models. English sentences were evaluated on both models to compare monolingual performance, while the multilingual model was also tested on additional languages to evaluate cross-linguistic patterns. All sentences were taken directly from Mueller et al.'s CLAMS dataset and processed through the NLPScholar toolkit for evaluation.

Our results were largely consistent with the original study. The monolingual BERT model performed better than the multilingual BERT model on nearly all English sentence constructions, with the largest differences observed in negative polarity item constructions. Multilingual BERT's accuracy decreased for languages with higher morphological complexity, showing a negative correlation between morphological complexity and model accuracy similar to that found by Mueller et al. Despite some differences in implementation, such as the inclusion of all test sentences regardless of vocabulary coverage, the overall trends support the original conclusion. While multilingual models capture some syntax-like patterns, their syntactic understanding remains limited and highly influenced by language structure, vocabulary coverage, and training composition.

2 Background

Understanding whether neural language models truly learn grammar or just memorize word patterns has been a central question in computational linguistics. Earlier work, including (Linzen et al., 2016) and (Marvin and Linzen, 2018), explored this question using acceptability judgment tasks. In

these experiments, models are tested on pairs of sentences that differ only in grammaticality. For example, "The dogs run" versus "The dogs runs." If a model correctly identifies the grammatical sentence, it is taken as evidence that it has learned some aspect of syntactic structure rather than just reflecting co-occurrence statistics.

This approach assumes that human acceptability judgments, our intuitive sense of whether a sentence "sounds right", can serve as a valid benchmark for measuring a model's grammatical understanding. It also draws from theories in linguistics, especially Chomsky's generative grammar framework (Chomsky, 1965), which proposes that syntax is rule-based and universal across languages. From this perspective, if a model truly learns syntax, that knowledge should generalize across languages rather than being limited to English word order or morphology.

The paper we are replicating extended this line of research by introducing the CLAMS dataset, which applies the minimal-pair testing method to multiple languages. By comparing models trained on single languages (monolingual) and those trained on many (multilingual), the paper tests whether exposure to multiple linguistic systems strengthens or interferes with syntactic learning.

To explore whether differences in performance might depend on the complexity of a language, Mueller et al. incorporated a measure called the CWALZ score (Bentz et al., 2016). Based on the World Atlas of Language Structures (WALS) (Dryer and Haspelmath, 2013), this score quantifies morphological complexity by counting the presence of certain grammatical features such as inflection, case marking, or agreement. The more such features a language has, the higher its CWALZ score, under the assumption that these features reflect structural complexity. Comparing model performance against CWALZ scores helps test whether languages with richer morphology pose greater challenges for syntactic generalization.

Together, these ideas form the conceptual foundation for our replication. They clarify what it means to measure a model's syntactic understanding and why evaluating that understanding across multiple languages provides a stronger test of true grammatical learning.

3 Methods

Our experiment aims to replicate part of Mueller et al.'s *Cross-Linguistic Syntactic Evaluation of Word Prediction Models* (Mueller et al., 2020) study to test how well BERT-based models capture syntactic agreement across languages. We first compared English and multilingual BERT's accuracy in predicting grammatical English sentences across different construction types. We then explored whether a relationship exists between a language's morphological complexity and the multilingual model's performance on the same syntactic constructions in different languages.

To investigate these topics we utilized NLP-Scholar (Prasad and Davis, 2024), a toolkit that integrates with HuggingFace (Wolf et al., 2020; Lhoest et al., 2021) models to run controlled NLP experiments. We fed sets of minimal pair sentences from the CLAMS dataset into NLPScholar and evaluated model predictions.

3.1 Models

Following the original study's design, we evaluated two pretrained BERT models:

- BERT-base-cased (BERT) (monolingual, trained only on English)
- BERT-base-multilingual-cased (mBERT) (multilingual, trained on 104 languages)

Both models share the same transformer architecture which allows for a direct comparison of how multilingual training affects syntactic learning. English sentences were evaluated on both models, while the non-English sentences were only evaluated with mBERT.

The original paper included a study of a custom-trained LSTM model (Marvin, 2025), but this model was unavailable on HuggingFace and NLP-Scholar. Since both BERT models are transformer-based and widely used for syntactic evaluation, we used them as comparable replacements for testing our hypotheses.

3.2 Datasets

We used the CLAMS dataset as it provides minimal pairs for several syntactic constructions across multiple languages. Each pair consists of one grammatical and one ungrammatical version of the same sentence.

We formatted the data for each language into a TSV file compatible with NLPScholar's evaluate mode. Each file included columns for:

- Sentence ID
- Sentence pair ID
- Construction type
- Language
- Grammaticality label (expected / unexpected)
- Sentence text

Some of the NPI constructions in the LMsyneval (Marvin, 2025) dataset had three sentences instead of pairs. Instead, we utilized the CLAMS dataset, which has the same sentences but in pairs because NLPScholar cannot directly evaluate triplets and only works with minimal pairs.

3.3 Evaluation

Each dataset was evaluated using configuration files specifying the model, data paths, and experiment setup within NLPScholar. The evaluate mode generated token-level predictability scores for each sentence, while the analyze mode was used to compute the difference in predicted probabilities between grammatical and ungrammatical sentences within each pair.

Results were automatically grouped by languages and syntactic construction, allowing comparison across both dimensions. For English, accuracies were compared between BERT and mBERT to assess the effect of multilingual training. For other languages, accuracy was calculated as the proportion of minimal pairs correctly predicted by mBERT (the grammatical sentence receiving a higher probability than the ungrammatical one).

Finally, to explore cross-linguistic patterns, we compared overall model accuracy for each language with its CWALZ morphological complexity score from (Bentz et al., 2016). This analysis tested whether greater morphological complexity corresponded to lower syntactic prediction accuracy.

4 Results[1]

The original paper excluded negative polarity items (NPIs) from their tables since they differed in more than one word position. Since we are not using the Region of Interest (ROI) column in the NLPScholar experiment, we decided to include this data and found that its performance mirrored that of other constructions. These NPI constructions, along with all other English constructions, are shown in Table 1.

	Mono	**Multi**
SUBJECT VERB AGREEMENT		
Simple	**0.90**	0.77
In a Sentential Complement	**0.82**	0.80
VP Coordination (Short)	**0.93**	0.80
VP Coordination (Long)	**0.96**	0.89
Across Subject rel. Clause	**0.82**	0.54
Within Object rel. Clause	**0.92**	0.55
Within Object rel. Clause (no *that*)	**0.76**	0.55
Across Object rel. Clause	**0.90**	0.70
Across Object rel. Clause (no *that*)	**0.86**	0.53
Across Prepositional Phrase	**0.90**	0.74
Average Accuracy	**0.86**	0.62
REFLEXIVE ANAPHORA		
Simple	**0.99**	0.95
In a Sentential Complement	0.90	**0.91**
Across a relative clause	**0.93**	0.67
Average Accuracy	**0.93**	0.71
NEGATIVE POLARITY ITEM		
Simple	**1.00**	0.46
Across Object rel. clause	**0.92**	0.66
Average Accuracy	**0.93**	0.65

Table 1: BERT (Mono) vs mBERT (Multi) accuracy by construction type

The monolingual BERT model outperformed the multilingual mBERT model on English sentences for all constructions except reflexive anaphora in sentential complement. An example of this construction is "The mechanics said the author hurt (himself/themselves)." In this case, mBERT performed slightly better, likely because its multilingual training includes exposure to languages with richer reflexive structures, allowing it to better track antecedents. Apart from this construction, BERT was more accurate for all other English constructions.

The difference was most pronounced with NPI structures: BERT achieved above 90% accuracy for both constructions, while mBERT remained below 70%. English NPIs often depend on specific trigger words like "ever", which may strongly signal grammaticality in monolingual training data but appear less consistently in multilingual corpora.

[1]GitHub with our configuration files and results: https://github.com/lilygamburg/midterm-lily-will-edgar

	English	French	German	Hebrew	Russian
Simple Agreement	0.77	0.79	**0.96**	0.77	0.81
VP Coordination (Short)	0.80	0.85	**0.91**	0.89	0.84
VP Coordination (Long)	0.89	0.95	**0.96**	0.82	0.94
Across Subject rel. Clause	0.54	0.56	**0.84**	0.67	0.75
Within Object rel. Clause	0.55	0.81	0.74	0.63	**0.85**
Across Object rel. Clause	0.70	0.71	**0.88**	0.56	0.81
Across Prepositional Phrase	0.74	0.65	**0.91**	0.57	0.81
Average Accuracy	0.65	0.69	**0.85**	0.62	0.81

Table 2: Accuracy of mBERT by language and construction type

Overall, mBERT's shared vocabulary of 110,000 tokens across all languages may weaken its English performance, while BERT's English-only training provides stronger syntactic and lexical associations.

The original paper found that mBERT outperformed BERT in certain constructions (for example sentential complements, long VP coordination, subject relative clauses, and preposition phrases). However, Mueller et al. excluded sentences whose focus verbs were not present in mBERT's vocabulary. Since NLPScholar lacks the ability to filter such sentences, all test items were included in this experiment, likely contributing to mBERT's lower English performance.

Across languages, high morphological complexity was associated with lower mBERt accuracy. As shown in Figure 1, mBERT's accuracy decreases as CWALZ complexity scores increase. English had a way lower than expected accuracy, likely due to vocabulary mismatches and its inconsistent morphological marking (for example eat/eats/ate). Russian performed exceptionally well, potentially because its case system marks subjects and objects with distinct suffixes, offering strong grammatical cues for mBERT.

Despite these surprising results, the same overall trend reported by Mueller et al. was observed: languages with greater morphological complexity generally showed lower mBERTaccuracy. The correlation here was weaker, possibly due to the differences in tokenization and vocabulary coverage between NLPScholar and the original paper's custom tokenizer.

Finally, Table 2 shows accuracies by language and construction. German and Russian consistently outperformed the other languages across nearly all constructions. German exceeded Russian in every construction except within-object relative clauses. English was the second least accurate language,

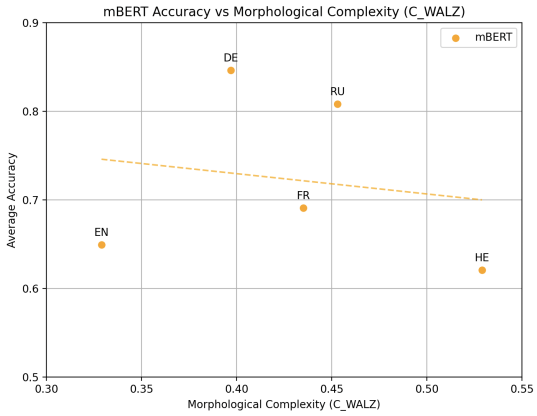

Figure 1

despite its similarity to German, likely due to its limited morphological markers.

Simple constructions and VP coordination yielded the highest accuracies across languages, with long VP coordination outperforming short VP coordination due to increased contextual information. In contrast, sentences with relative clauses performed worse overall, likely because longer dependencies and additional noise make grammatical agreement harder for the model to capture. Language-specific grammatical markers further influenced these results, improving performance in morphologically rich languages and weakening it in languages with fewer overt clues.

5 Discussion

Previous research has shown that neural language models can learn syntactic and grammatical structures in English. Mueller et al investigates whether such grammatical understanding generalizes across languages. This is an important area of study, as most NLP research remains concentrated on English as well as a few other high resource languages.

Their results indicated that monolingual models tend to perform better on English grammatical construction, though multilingual models can show improvements for specific constructions. They also found an inconclusive relationship between morphological complexity and model accuracy; LSTM models showed a positive correlation while mBERT showed a negative one.

This replication more clearly supports the hypothesis that monolingual models trained on English outperform multilingual models on English grammatical acceptability tasks. Except for one reflexive anaphora construction, BERT consistently outperformed mBERT, echoing and strengthening the original finding. Like Mueller et al., this experiment found a negative correlation between morphological complexity and mBERT accuracy, though with weaker correlation strength. English's poor performance and Russian's strong performance highlight how vocabulary coverage and language structure can shape outcomes.

Model accuracy appears highly dependent on vocabulary composition and coverage. Because NLPScholar does not allow the removal of test sentences with unseen focus verbs, mBERT's results were inferior when compared to the original paper. When using mBERT, the amount of English in the vocabulary is diluted by the other languages, sharply decreasing its overall accuracy compared to BERT.

Limitations Both the original study and this replication are constrained by vocabulary coverage. Models like BERT and mBERT have finite vocabularies, and unseen tokens can substantially affect sentence probability calculations and grammaticality judgements. Mueller et al. mitigated this by filtering out-of-vocabulary items, while this experiment included all sentences, leading to lower scores overall.

The datasets also differ in structure. For instance, LMSyneval (Marvin, 2025) includes triplet minimal pairs for NPI constructions, while CLAMS and NLPScholar can only support pairs, limiting direct comparison. Additionally, NLPScholar does not provide access to the LSTM model used in the original study, restricting replication to transformer-based models.

6 Conclusion

More research in this field is needed as results are not entirely conclusive and are dependent on particular variables such as model vocabulary and evaluation sentences. Overall this experiment reinforces that monolingual models are more reliable for predicting grammatical acceptability in English, while multilingual models' performance depends heavily on linguistic features and vocabulary overlap. Our results further confirm a negative relationship between morphological complexity and model accuracy, though sensitive to implementation details. Future research should focus on refining multilingual vocabularies and evaluation datasets. Underrepresented languages like Hebrew may benefit from dedicated monolingual models, while others with complex morphology may require specialized architectures. Vocabulary design remains a critical factor in achieving consistent cross-linguistic generalization in neural language models. While monolingual models may be more consistent at predicting grammatical sentences for particular languages, multilingual models serve an important purpose of identifying particular patterns that are more clear in some languages than others that can permit better understanding of certain elaborate structures

References

Christian Bentz, Tatyana Ruzsics, Alexander Koplenig, and Tanja Samardži´samardži´c. 2016. A comparison between morphological complexity measures: Typological data vs. language corpora.

Noam Chomsky. 1965. *Aspects of the Theory of Syntax*. MIT Press, Cambridge, MA.

Matthew S. Dryer and Martin Haspelmath, editors. 2013. *WALS Online (v2020.4)*. Zenodo.

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, Angelina McMillan-Major, Philipp Schmid, Sylvain Gugger, Clément Delangue, Théo Matussière, Lysandre Debut, Stas Bekman, Pierric Cistac, Thibault Goehringer, Victor Mustar, François Lagunas, Alexander Rush, and Thomas Wolf. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. 2016. Assessing the ability of LSTMs to learn syntax-

sensitive dependencies. *Transactions of the Association for Computational Linguistics*, 4:521–535.

Becky Marvin. 2025. Lm_syneval: A toolkit / evaluation framework. GitHub repository.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202, Brussels, Belgium. Association for Computational Linguistics.

Aaron Mueller. 2020. Syntactic evaluation sets, attribute-varying grammars, and code for replicating the clams paper. `https://github.com/aaronmueller/clams`.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models.

Grusha Prasad and Forrest Davis. 2024. Training an NLP scholar at a small liberal arts college: A backwards designed course proposal. In *Proceedings of the Sixth Workshop on Teaching NLP*, pages 105–118, Bangkok, Thailand. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.