Minding the LLM Gap: MultiBERT's Performances on Text from Different Languages Replication

Jaanhvi Agarwal
Colgate University
jagarwal@colgate.edu

Sarah Cryan
Colgate University
scryan@colgate.edu

Abstract

This paper replicates a multilingual syntactic evaluation using monolingual and multilingual BERT models, applying them to the Cross-Linguistic Assessment of Models on Syntax (CLAMS) five-language dataset (Mueller et al., 2020). Our evaluation focuses on both simple and complex syntactic structures. The results show that monolingual BERT consistently outperforms multilingual BERT, particularly on more complex syntactic tasks, such as relative clauses and prepositional phrases. Yet, both models perform similarly on simpler tasks. Notably, our replication successfully generated results the original study could not produce, indicating that improvements in dataset preparation enhanced model evaluations. These findings encourage further research on that addressing language-specific syntactic nuances to help improve multilingual model performance, especially for non-Latin scripts (Arivazhagan et al., 2019) and more complex syntactic dependencies. Future work could explore fine-tuning architectures or adapting models for diverse script systems (Clark et al., 2022) to enhance cross-linguistic generalization.

1 Introduction

Current language modeling demonstrates greater success with monolingual word prediction than in the past, particularly seen in a model's ability to identify sentence structures. However, the training datasets used for these models may not present sufficient complexity to further enhance performance. To address this gap, this paper, along with the study it replicates, incorporates additional languages—specifically French, Hebrew, German, and Russian—to provide more challenging syntactic structures. Although these languages differ in vocabulary and script, they share some syntactic rules, offering more diverse examples of complex sentence structures. This diversity allows models to learn new linguistic patterns and improves their ability to generalize across languages.

The original study developed the Cross-Linguistic Assessment of Models on Syntax (CLAMS) dataset (Mueller et al., 2020), which expanded on (Marvin and Linzen, 2018)'s dataset of minimally different English sentences (ungrammatical vs. grammatical). By introducing minimally different French, German, Hebrew, and Russian sentences, CLAMS provides a more comprehensive multilingual dataset. This allows for a deeper evaluation of language models' abilities to handle complex grammatical constructions across different languages. Examples of increasingly complex sentence structures can be seen below:

Structure Type	Example
Simple Agreement	The cat chases the mouse.
VP Coordination (Short)	The cat chased the mouse and caught it.
VP Coordination (Long)	The cat chased the mouse, climbed the tree, and jumped across the fence.
Across Subject Relative Clause	The cat that chased the mouse is sitting on the couch.
Within Object Relative Clause	The cat chased the mouse that was hiding in the hole.
Across Object Relative Clause	The cat that the dog chased is now hiding under the couch.
Across Prepositional Phrase	The cat on the mat near the window chased the mouse.

Table 1: Types of Sentence Structures with Examples

The original research evaluated both LSTM and transformer architectures, including monolingual BERT and multilingual BERT (mBERT). The results showed that LSTM models accurately distinguished correct from incorrect subject-verb agreement but struggled with more complex structures, such as center-embedded clauses. Surprisingly, training LSTM models on multiple languages introduced language interference, reducing overall performance. The researchers suggested that different architectures might better handle this issue.

In contrast, transformer models like BERT

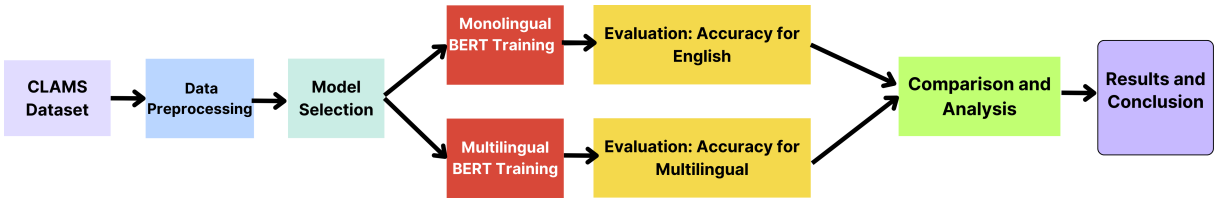

Figure 1: Pipeline for the NLP Scholar Experiment

showed a higher level of generalization, with monolingual BERT outperforming mBERT on complex syntactic structures. mBERT did perform well on certain grammatical tasks, but monolingual BERT was consistently more sensitive to syntactic nuances, particularly in languages with more grammatical complexity.

In this replication study, we recreated the multilingual CLAMS dataset and focused on evaluating monolingual BERT and mBERT models. While the original study also included LSTM models, we chose to concentrate on BERT due to our prior experience with transformer models. By making some necessary formatting adjustments to the dataset, we successfully replicated the evaluation process and observed similar trends. Our results confirmed that monolingual BERT consistently achieved higher accuracy, particularly with complex syntactic structures, than mBERT, and both models performed comparably well on simpler grammatical tasks. While there was some differences between the accuracy scores of the original paper and our replication paper, there were similar trends in which grammar, languages, and models had higher or lower accuracy scores.

To illustrate the steps of our replication experiment, Figure 1 provides an overview of the pipeline we followed. This pipeline outlines the stages from dataset preparation, model selection, and evaluation, through to the final analysis of accuracy scores across different grammatical categories and languages. While these methods produced minor differences in accuracy between our replication and the original study, the general trends and findings remained consistent and can be seen in Table 2 and Table 3.

2 Background

The original paper and our replication paper rely on understanding what a well-performing model means and how we are able to analyze a grammatical rule across multiple languages.

A well-performing language model accurately distinguishes between grammatical and ungrammatical sentences, often tested with example pairs such as "the cat eats" (correct) versus "the cat eat" (incorrect). The model predicts the next word based on prior context, so a well-performing model would recognize that "eat" is an incorrect prediction after "cat," demonstrating an understanding of subject-verb-agreement. A singular subject, like "cat," must be followed by a singular verb form, such as "eats." Evaluating whether the model made the correct choice mirrors human native-speaker judgments, where a good model can intuitively differentiate between correct and incorrect language constructions, much like a native speaker would.

In the original study, two types of language models were used: LSTM (a type of sequential neural network) and BERT (a transformer-based model). LSTMs store and recall information across sequences of words, relying on previous context to predict the next word. In contrast, BERT uses a bidirectional approach, where it analyzes the entire sentence context to predict words. This allows BERT to capture more complex relationships between words and handle syntactic dependencies better than LSTMs in certain scenarios.

Evaluating a model's ability to generalize across multiple languages adds complexity to the analysis. Each language has unique syntactic structures and rules that models must learn. To test how well these models handle cross-linguistic syntactic variation, the original study used Attribute-Varying Grammars (AVGs). AVGs generate sentence pairs—one grammatical, one ungrammatical—by modifying specific features such as verb conjugations or word order. For example, in French, the AVG framework might produce "je pense" (correct) and "je penses" (incorrect) to test the model's understanding of verb agreement. This method allows for consistent syntactic evaluation across multiple languages, enabling researchers to assess how well a model can generalize grammar rules beyond a single language.

While multilingual models like mBERT can gen-

eralize across different languages, they often struggle with language-specific syntactic rules due to potential language interference. For instance, a model trained on multiple languages may learn general patterns that do not fully capture the intricacies of any single language's syntax. In contrast, monolingual models, such as BERT trained on English, tend to perform better on complex syntactic tasks within their target language. Understanding how these models handle grammatical variations across languages is crucial for improving their robustness and applicability in multilingual NLP tasks.

3 Methods

Our replication study evaluated language models with the same goal as the original study, to determine their ability to distinguish grammatical from ungrammatical sentences using new, larger, challenging datasets. We compared monolingual and multilingual model performance across different languages and syntactic constructions, using datasets generated from attribute-varying grammars (AVGs).

3.1 Models

We used BERT for our models, testing monolingual BERT for evaluating English and using multilingual BERT for evaluating all 5 languages (English, French, German, Hebrew, and Russian). Using these models allowed us to test and observe how well multiple languages improve the models' performance and allowed us to work with a familiar model architecture.

3.2 Datasets

We relied on the same CLAMS dataset generated in the original study, but we modified the data to fit our specific format for running NLPScholar, a toolkit for running natural language processing experiments (Prasad and Davis, 2024). We created a script that changed the original format [true/-false] [sentence] to follow the NLPScholar format: [sent id] [condition] [context id] [pair id] [comparison] [lemma] [sentence] [Region Of Interest]. These necessary changes allowed us to run evaluate and analyze on the sentences from CLAMS. Our datasets can be found here.

This modification was necessary to ensure compatibility with the MinimalPair analysis, which performs binary comparisons (e.g., grammatical vs. ungrammatical) by measuring the predictability or surprisal of words in sentences. The toolkit requires two types of TSV files:

1. **Predictability TSV**: Contains sub-word token information, probabilities, and surprisal values.

2. **Data TSV**: Contains sentence-level information, including IDs for sentences, pairs, and contexts, along with the comparison (expected/unexpected).

These files are generated by running the NLPScholar toolkit using a configuration file that specifies the dataset paths, models to be used, and evaluation measures. The configuration file, written in YAML format, defines the details of the experiment, including the type of model used (e.g., multilingual BERT), the location of the input datasets, and the method for computing predictability (e.g., surprisal). A sample configuration is shown below:

Listing 1: NLPScholar Configuration Example

```yaml
exp: MinimalPair

mode:
    - evaluate
    - analyze

models:
  hf_masked_model:
    - google-bert/bert-base-multilingual
        -cased
    - google-bert/bert-base-uncased

datafpath: /path/to/eval_dataset.tsv
predfpath: /path/to/predictions.tsv
resultsfpath: /path/to/summary.tsv

pred_measure: surp
```

These changes allowed us to run more detailed evaluations, where the predictability of specific words could be compared across different syntactic structures. By restructuring the data, we were able to replicate the experiments from the original paper, but with the enhanced flexibility and functionality provided by the *NLPScholar* toolkit. The scripts and config files used to transform the dataset can be found in our GitHub repository at https://github.com/Jaanhvi18/clams---Midterm-Replication-NLP.git.

3.3 Evaluation

For the evaluation of our models, we utilized the *NLPScholar* toolkit, which allowed us to compare

Table 2: Results from the replication using NLPScholar: Multilingual BERT accuracies across different languages and constructions.

	English	French	German	Hebrew	Russian
Simple agreement	**0.78**	**0.76**	**0.84**	**0.72**	0.78
VP coordination (short)	**0.82**	**0.78**	0.90	0.76	0.79
VP coordination (long)	0.81	0.86	**0.93**	0.83	0.90
Across subject relative clause	0.55	0.56	0.61	0.64	0.75
Within object relative clause	0.58	0.75	0.50	0.52	0.67
Across object relative clause	0.56	0.57	**0.57**	0.53	0.76
Across prepositional phrase	0.57	0.53	**0.75**	0.55	0.68

the predictability and surprisal of words across different syntactic structures and languages. The evaluation followed a binary comparison method between grammatical and ungrammatical sentence pairs, derived from the *Cross-Linguistic Assessment of Models on Syntax (CLAMS) dataset*.

Our analysis focused on comparing the performances of *monolingual BERT* (trained on English) and *multilingual BERT* (trained on multiple languages) across a diverse set of syntactic constructions such as subject-verb agreement, VP coordination, and relative clauses. We structured the evaluation based on the following steps:

- **Predictability TSV and Data TSV Generation**: Using the NLPScholar toolkit, we generated these two types of TSV files:
- **Evaluation Metrics**: The primary metric used to assess model performance was *accuracy*, which measured how effectively the models distinguished between grammatical and ungrammatical sentences. Additionally, we assessed *surprisal values* to quantify how unexpected a word is in its syntactic context based on the model's training data. A lower surprisal score indicates a more predictable word within the given context, suggesting better syntactic understanding by the model. We used surprisal differences between grammatical and ungrammatical sentence pairs to determine the model's ability to capture syntactic nuances.
- **Syntactic Complexity Focus**: Our evaluation focused on both simpler constructions, such as *subject-verb agreement* and *VP coordination*, and more complex structures like *relative clauses* and *prepositional phrases*. This range enabled us to assess how the models handle basic grammar rules as well as more intricate syntactic dependencies.
- **Cross-Linguistic Comparison**: We evaluated multilingual BERT on five languages: English, French, German, Hebrew, and Russian. By comparing the model's performance across these structurally diverse languages, we aimed to assess how well it generalized syntactic rules across languages, each with varying degrees of similarity to English.

This combination of steps allowed us to best evaluate the performance of our models across multiple languages and sentence structures.

	Mono	Multi
Subject-Verb Agreement		
Simple	0.79	0.78
VP coordination (short)	0.89	0.81
VP coordination (long)	**0.98**	0.87
Across subject rel. clause	0.76	**0.62**
Within object rel. clause	**0.83**	0.60
Across object rel. clause	**0.74**	0.60
Across prepositional phrase	0.72	**0.62**
Average accuracy	**0.82**	**0.70**

Table 3: Results from the replication using NLPScholar: Monolingual BERT accuracies compared to Multilingual BERT accuracies averaged across all 5 languages.

4 Results

Our replication study yielded results that are largely in line with the findings from the original study, with some notable differences. We evaluated the performance of both monolingual BERT (trained on English) and multilingual BERT (trained on multiple languages), focusing on their ability to

distinguish between grammatical and ungrammatical sentences across different languages and syntactic structures. Below, we present our findings in comparison to the original study's results.

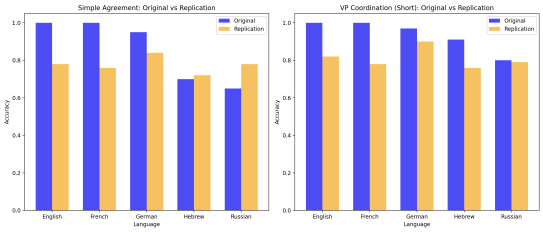

Figure 2: Comparison of Simple Agreement and VP Coordination (Short) - Original vs Replication Across Languages

Simple Agreement and VP Coordination In our replication, all five languages achieved high accuracy (> 0.70) in simpler syntactic constructions such as *simple agreement*, *VP coordination (short)*, and *VP coordination (long)*, as shown in Table 2. While slightly lower than the original paper's results, our findings using *NLPScholar* remain strong. For *VP coordination (long)*, the accuracies were 0.81 for English, 0.86 for French, 0.93 for German, 0.83 for Hebrew, and 0.90 for Russian, all within 0.12 of the original study's results, demonstrating similar performance between our results and the original paper, as seen in Figure 2. The results also indicate multilingual BERT's relatively strong ability to handle these structures across diverse languages.

Across subject relative clause and Across prepositional phrase There were also lower accuracies in certain complex syntactic constructions, similar to the original paper's results, such as *across subject relative clauses* and *across prepositional phrases* Table 2. For *across subject relative clauses*, the replication accuracies were 0.55 for English, 0.56 for French, 0.61 for German, 0.64 for Hebrew, and 0.75 for Russian. These results align with the lower accuracies reported in the original paper as can be seen in Figure 3, further highlighting the challenges faced by the models in handling more complex structures. Similarly, for across prepositional phrases, the replication accuracies were 0.57 for English, 0.53 for French, 0.75 for German, 0.55 for Hebrew, and 0.68 for Russian, demonstrating a mix of lower and higher scores across the two papers.

When analyzing complex syntactic structures,

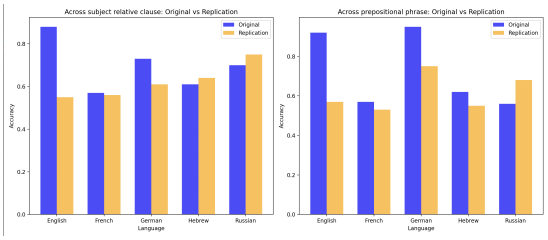

Figure 3: Accuracy Comparison Across Subject Relative Clauses and Prepositional Phrases: Original vs Replication Across Languages

performance varied significantly across different languages. In the case of *across subject relative clauses*, Hebrew and Russian exhibited better performance (0.64 and 0.75, respectively) compared to English (0.55) and French (0.56). The differences suggest that the model struggles more with these constructions in languages that share less structural similarity with English, potentially affecting the model's ability to generalize across languages.

A similar pattern was observed for *across prepositional phrases*, where German achieved the highest accuracy (0.75), followed by Russian (0.68). This may be due to the syntactic characteristics of German and Russian, where prepositional phrases are structured differently compared to English. These results further emphasize the challenges faced by multilingual models when dealing with language-specific grammatical rules, especially when those rules diverge significantly from the English-based training data used for BERT.

Mono vs Multi BERT Performance Both monolingual and multilingual BERT showed strong performance in simpler structures such as *VP coordination (long)*, achieving accuracies of 0.98 and 0.87, respectively Table 3. However, as the syntactic complexity increased, accuracy noticeably declined for both models. For example, in *across-subject relative clauses*, monolingual BERT outperformed multilingual BERT with an accuracy of 0.76 compared to 0.62 Table 3. This drop indicates that more complex dependencies between sentence elements, such as those found in relative clauses, are more challenging for both models, but especially for multilingual BERT.

Similarly, in *within-object relative clauses*, monolingual BERT maintained a higher accuracy of 0.83, while multilingual BERT achieved only 0.60 Table 3. This consistent performance gap between the two models across different sentence structures suggests that monolingual BERT is bet-

ter at capturing and generalizing syntactic rules within a single language, likely because it is trained on more language-specific data. In contrast, multilingual BERT, while versatile across multiple languages, may struggle with deep syntactic nuances due to the trade-offs made during multilingual training.

Interestingly, unlike the original study, where hyphens were used to indicate out-of-vocabulary cases for certain constructions (e.g., *Within object relative clause* in Hebrew and *VP coordination* (long)" in Russian, our replication successfully computed results for these constructions. This suggests that modifications in our dataset preparation or model training—such as enhanced preprocessing for better word coverage or fine-tuning—may have improved the models' ability to handle these previously problematic cases. By generating results for these constructions, our replication offers a more comprehensive evaluation of syntactic structures across multiple languages.

These results highlight that although multilingual models are useful for handling multiple languages, they still face challenges in mastering the intricacies of complex syntactic structures. This is likely because the training data for multilingual BERT must generalize across languages, which may dilute its ability to focus on language-specific syntactic rules. Monolingual BERT, on the other hand, benefits from language-specific data, allowing it to perform more accurately in these complex syntactic scenarios.

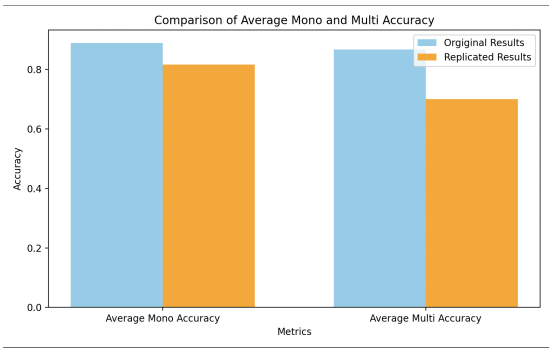

Figure 4: Comparison of Mono vs Multi BERT Performance - Replication vs Original

Overall the average results of our models, monolingual BERT and mBERT,were lower but still close to the orginial study's results, as seen in Figure 4.

5 Discussion

The main motivation behind our paper was to replicate a cross-linguistic syntactic evaluation, using the same dataset, investigating how word prediction models perform across multiple languages. We hypothesized that our more expansive dataset with matching syntax in different languages would provide more data for these models to learn from.

In our results as can be seen in Table 3, we observed better performance for monolingual BERT and multilingual BERT on simpler structures like *Simple* (mono: 0.79, multi: 0.78), *VP (short)* (mono: 0.89, multi: 0.81), and *VP (long)* (mono: 0.98, multi: 0.87), similar to the original study Figure 4. For more complex structures, though, both models had lower accuracy, aligning with the results from the original paper and supporting their conclusions that BERT models can learn from multiple languages for simpler sentence structures. However, we did not have as strong of evidence to conclude that some languages perform far better than others.

One notable difference in our replication is that we were able to compute results for constructions where the original study used hyphens to indicate out-of-vocabulary cases, such as *Within object relative clause* in Hebrew and *VP coordination (long)* in Russian. This improvement is likely due to modifications we made during dataset preprocessing to fit the *NLPScholar* format. The changes we made ensured better compatibility and may have enhanced word coverage, allowing the models to handle previously out-of-vocabulary cases. Additionally, *NLPScholar*'s evaluation processes may have contributed to these improvements by facilitating more robust comparisons across sentence structures.

Both the original study and our replication focus on only five languages and on a specific set of subject-verb sentence structures. While the limited number of languages and sentence structures made it easier to compare performance across languages, it also restricted the variety of sentences tested. In other words, there are other sentence structures that may have yielded useful insights that we could not test. Also, using *NLPScholar* in our replication may have impacted how the models processed and evaluated certain sentences, explaining some of the discrepancies we observed compared to the original study.

Our replication results support the original

study's conclusions about monolingual and multilingual BERT. The multilingual models can apply syntax rules across languages, but they struggle with more complex sentence structure, and monolingual BERT consistently outperformed the multilingual BERT. Following the recommendations from the original paper, we agree that future work could explore architectural changes or fine-tuning on non-Latin script languages to enhance performance further.

Acknowledgements

Figure 5: https://xkcd.com/262/

References

N. Arivazhagan, Ankur Bapna, Orhan Firat, Dmitry Lepikhin, Melvin Johnson, Maxim Krikun, Mia Xu Chen, Yuan Cao, George F. Foster, Colin Cherry, Wolfgang Macherey, Z. Chen, and Yonghui Wu. 2019. Massively multilingual neural machine translation in the wild: Findings and challenges. *ArXiv*, abs/1907.05019.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language understanding. *Transactions of the Association for Computational Linguistics*.

Rebecca Marvin and Tal Linzen. 2018. Targeted syntactic evaluation of language models. In *Conference on Empirical Methods in Natural Language Processing*.

Aaron Mueller, Garrett Nicolai, Panayiota Petrou-Zeniou, Natalia Talmina, and Tal Linzen. 2020. Cross-linguistic syntactic evaluation of word prediction models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5523–5539, Online. Association for Computational Linguistics.

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*, Bangkok, Thailand. Association for Computational Linguistics.