Incorporating Rule-Based Methods with Prompt-Based Techniques for Indigenous Language Generation

Justin Vasselli Arturo Martínez Peguero Taro Watanabe
Nara Institute of Science and Technology
vasselli.justin_ray.vk4@is.naist.jp

Abstract

In this work, we explore how to leverage the meta-linguistic knowledge of large language models (LLMs) by combining rule-based techniques with few-shot prompting to produce new sentences in Indigenous languages, despite the LLMs having little to no prior knowledge of the language. Integrating rule-based preprocessing for Bribri significantly improves accuracy—over six times the edit-tree baseline and twice that of few-shot prompting—while a simplified version enhances performance for Maya and Guarani. This research provides a generalizable solution for addressing linguistic challenges in low-resource settings through combining structured linguistic resources with LLM meta-linguistic capabilities to support language revitalization and preservation.[1)]

1 Introduction

The disappearance of a language represents a loss of cultural and historical knowledge, but advances in technology offer tools to prevent this extinction. Fostering new speakers through education is essential for the survival of a language. The AmericasNLP 2024 Shared Task [1] addressed this need by focusing on creating educational materials for Indigenous languages, including Maya, Bribri, and Guarani, contributing to efforts in language revitalization.

Each of the languages of the shared task had their own unique challenges, and we focused primarily on Bribri, which features complex verb morphology. The challenges were further compounded by a lack of overlap between verbs in the training and testing data making straighforward few-shot prompting less effective.

To address these challenges, we propose a hybrid methodology that combines rule-based methods with the generative capabilities of LLM. Rule-based methods leverage grammatical frameworks and expert-curated lexicons to address complex morphology and syntax. The metalinguistic capabilities of LLMs allow them to apply linguistic patterns from limited data, effectively generating example sentences or translations when guided by structured input.

This work advances low-resource NLP by designing a pipeline that integrates rule-based preprocessing with LLM prompting for educational material generation, demonstrating the role of rule-based methods in improving linguistic accuracy for languages with unique features, and providing insights into how structured linguistic resources can be combined with LLM capabilities to support underrepresented languages.

Experimental results demonstrate the effectiveness of integrating rule-based methods with LLM prompting. On Bribri, the baseline edit-tree approach achieved an accuracy of only 8.75% on the shared task test set. By delegating the complex verb morphology to a rule-based conjugator, our method achieved a sixfold improvement, with accuracy over 53%. We further find that a simpler version of this technique improves performance on other low resource languages such as Maya and Guarani.

2 Prior Work

Our approach builds on prior research like Rosetta Stone puzzles [2], which simulate low-resource NLP scenarios requiring grammatical inference and two-way translation. LLMs such as GPT4 have been shown to do well on this task that requires high level of metalinguistic reasoning ability [3, 4]. Rule-based methods have been combined with LLM prompting to address machine translation for no-resource languages, showing how structured frameworks enhance LLM flexibility for complex tasks [5].

Our hybrid approach combining rule-based techniques

1) https://github.com/JVasselli/JAJ-Americas2024

with LLM prompting excelled in the AmericasNLP Shared Task, particularly for Bribri, where rule-based systems improved accuracy beyond few-shot prompting [6].

3 Methodology

3.1 Data and Task Description

The dataset from the AmericasNLP 2024 Shared Task 2 (Americas2024ST2) is a parallel set of source sentences and target sentences both in the same indigenous language. Each entry also has one or more grammatical changes that were used to transform the source sentence into the target sentence. These transformations included morphosyntactic changes such as tense, aspect, and negation. For example:

Source sentence: *Ye' shka'* ("I walked")
Expected change: Polarity: Negative
Target sentence: *Ye' k'ë shkànwē* ("I didn't walk")

The Bribri data for Americas2024ST2 was constructed using examples from textbooks, grammar books, and a treebank. The focus was on Bribri's verbal morphology, particularly its tense-aspect-mood suffixes. A total of 64 original sentences were selected and conjugated into all possible forms based on linguistic resources, resulting in 1,001 example sentences. These included a mix of transitive, intransitive, locative, and copular sentences. Furthermore, the dataset incorporated irregular verbs due to their high frequency in the language (e.g., tso for 'is' vs. bák for 'was'). Sentences were categorized by features such as polarity, mood, tense, aspect, voice, number of arguments, and type of pronoun. These transformations formed the clusters of sentences used for training, development, and testing.

As this was a low-resource language task, the size of the Americas2024ST2 data was quite small: 309 training instances, 212 in the development split, and 480 in the test data.

3.2 The Prompt

Our system leverages the capabilities of large language models by prompting them with relevant example cases tailored to the target language. The base prompt was adapted from one used for the Rosetta Stone Puzzles [3]. The simplest prompt we test includes only relevant examples as in the following example:

This is a linguistic puzzle. Below are example sentences in a foreign language and sets of changes to apply to them. The examples are followed by the problem sentence and desired change. Your task is to look closely at the example sentences and to change the sentence correctly.

Example 1:
Sentence: Ye' shka'
Change(s): TYPE:NEG
Answer: Kë ye' shkàne

(more examples)

Here is the problem. Answer first, then explain your reasoning.
Sentence: Pûs kapë'wa
Change(s): TYPE:NEG

To select examples for Bribri, we focused on aligning test cases with relevant training examples by grouping similar grammatical changes. For compound changes, our system decomposed them into smaller, sequential steps processed independently. For instance, changes such as "ABSNUM:PL" and "PERSON:3_PL" were combined where possible to streamline processing. This ensured examples were representative and directly applicable to the grammatical transformations required for each test case.

3.3 POS Tags

A key component of our system is the application of custom, simplified part of speech (POS) taggers tailored to each target language. These taggers are primarily dictionary-based and are used to supplement the example sentences being passed to the LLM by explaining better the grammatical role of the words of the provided examples. Our tagger was built on Professor Haakon S. Krohn's online Bribri dictionary[2] [7].

With POS tags, the examples and problem text of the prompt are altered to include the additional information. The above example would become:

Example:
Sentence: Ye' shka'
((Ye', PRON:1_SI) (shka', VERB))
Change(s): TYPE:NEG
Answer: Kë ye' shkàne
((Kë,NEG) (ye', PRON:1_SI) (shkàne, VERB))

Problem:
Sentence: Pûs kapë'wa ((Pûs, NOUN) (kapë'wa, VERB))
Change(s): TYPE:NEG

2) https://www.haakonkrohn.com/bribri/index.html

3.4 Verb Conjugation

The complexity of Bribri verb conjugation, particularly for irregular verbs, required targeted strategies to improve translation accuracy. To evaluate potential performance enhancements, we conducted an experiment using oracle verb conjugation "hints" to provide the correct verb forms directly to the LLM in the prompt. We tested this oracle verb conjugation hint on the development set by manually annotating the verb in the target sentence and providing it to the prompt. Our initial experiment showed an increase in accuracy from 15% to 65%.

Motivated by the success of the oracle hint, we developed a rule-based verb conjugation tool, built on a database of verb conjugations from [8][3)]. In our system, the verbs identified by our POS tagger are retrieved from the database and the correct form is produced from a series of conjugation rules. For example, in the sentence *Ye' tö i k'ötwa* with changes `TYPE:NEG, TENSE:FUT_CER, ASPECT:IPFV`, the verb *k'ötwwa* is located by the POS tagger and looked up in the verb conjugation database. It is found to be the perfect remote form of *ujt'ökwwa*. The conjugator transforms the verb into *ujtèpawa* for the negative certain future tense. This transformation is then included as a hint at the end of the prompt: "The correct form of *k'ötwa* is likely *ujtèpawa*."

4 Edit-Tree Baseline

To contextualize the performance of our system, we compare it against the edit-tree baseline implemented by the shared task organizers. The baseline system was based on a simplified adaptation of the Prefer Observed Edit Trees (POET) method. An edit tree is a hierarchical structure representing a sequence of edit operations needed to transform a source sentence into a target sentence. Nodes in the tree either perform substitutions or match substrings, recursively applying these operations to produce the desired transformation. During training, the system built edit trees for all source-target sentence pairs in the dataset and counted their frequency for each morphosyntactic change. At inference, the most frequent edit tree for a given change was applied to the input sentence. If the transformation failed, the system attempted less frequent trees. If no successful transformation occurred, the input sentence was returned unchanged.

3) https://www.lenguabribri.com/gramtica-de-la-lengua-bribri

5 Experimental Results

We tested generation on Americas2024ST2-dev using `gpt-3.5-turbo-0125`, `gpt-4-0125-preview`, and `Mixtral-8x7B-Instruct-v0.1` [9]. For the GPT models, we used temperature of 0. For Mixtral we used a greedy search. Table 1 details the performance across different LLMs. As `gpt-4-0125-preview` had the highest score on the development set, we used that on the test set.

Table 1 The results on the development set prompting different LLMs. The best result in each column is bolded.

System	Acc.	BLEU	ChrF
Mixtral	34.43	42.86	72.06
GPT-3.5	40.57	61.15	77.04
GPT-4	**47.17**	**67.01**	**80.75**

Our system improved accuracy by over six times higher than the edit-tree baseline, and more than twice that of few-shot prompting alone. This is likely due to the challenges of complex verb conjugation using pattern matching approach. The complete results can be seen in Table 2.

Table 2 The results of the test set of AmericasNLP2025.

System	Accuracy	BLEU	ChrF
Edit-tree	8.75	22.11	52.73
Few-shot prompt	17.71	39.48	69.28
+ Hints (ours)	**53.55**	**78.41**	**91.53**

5.1 Error Analysis

Despite the rule-based conjugation, verb conjugation errors still posed the most significant challenge, comprising 57% of total errors. These ranged from minor accent issues (e.g., *sur* instead of *sùr*) to completely incorrect verb forms (e.g., *k'ötwwa* instead of *ujtèkèulur*). The inclusion of numerous irregular verbs in the Bribri data, as noted by [10], compounded these challenges, especially given the lack of overlap between training and testing verbs. Omissions made up 19% of errors, where words were missing as in *Pp'ö* instead of *I pp'ö*. Extraneous words, such as *Ye' wa stsa'* instead of *Ye' stsa',* accounted for 9%, while pronoun mismatches caused 8% of errors. The final 6% of errors involved incorrect word order (e.g., *K'e ie' stsö* instead of *Ie' k'e stsö*).

The high proportion of errors related to verb conjugation suggests that while the rule-based conjugator contributed significantly to system performance, there remains room for improvement in handling irregular forms and less frequent patterns. Streamlining these enhancements could address gaps in linguistic coverage, which would result in better generalization across unseen data.

5.2 Expanding to Maya and Guarani

As rule-based methods require significant time and linguistic expertise to develop, we aimed to test a more minimal-effort version of our hybrid approach on Maya and Guarani. This simplified approach utilized only POS tagging, which can be quickly constructed using a list of words and their corresponding parts of speech. See Appendix A for details on the datasets for these languages. See Table 3 for the size of the dataset for each language.

Table 3 The number of instances in the training, development, and test splits for each language.

Lang	Train	Dev	Test
Bribri	309	212	480
Maya	594	149	310
Guarani	178	79	364

We made language-specific alterations as follows:

Maya The POS tagger for Maya focuses predominantly on function words, as these play a crucial role in understanding the grammatical structure of sentences. Although we did not create a full dictionary for Maya, we ensured coverage of key aspect markers such as *t'a'an* and pronouns like *in* or *teen* [11]. Additionally, the tagger is designed to recognize and handle common suffixes such as *e'ex*.

Guarani The POS tagger for Guarani locates prefixes that indicate the person performing the action, pronouns, and determinants. It tags verbs based on conjugations and guesses at nouns using sentence structure. All other parts of speech remain untagged.

We did not build verb conjugators for these languages, focusing instead on testing the feasibility of our hybrid method with only the minimally developed POS tagging system.

Our experiments demonstrated that the hybrid approach improved performance for Maya and Guarani compared to baseline approaches, though the gains were less pronounced than for Bribri. For Guarani, 75% of errors involved incorrect verb forms, highlighting the potential benefit of a rule-based verb conjugator. In contrast, Maya's strong baseline performance, likely due to its larger training dataset, minimized the impact of verb conjugation on overall accuracy. Errors in Maya were primarily linked to inconsistencies in training data and syntactic complexities, such as the placement of wáaj in interrogatives (25% of errors).

Table 4 The results of our hybrid method on Maya and Guarani

Language	Data	Accuracy	BLEU	ChrF
Maya	Edit-tree	25.81	53.69	80.23
	Our system	**54.17**	**71.72**	**82.78**
Guarani	Edit-tree	14.84	25.03	76.10
	Our system	**36.81**	**48.29**	**84.12**

6 Conclusion

This study demonstrates that combining rule-based methods with LLM prompting provides a viable framework for generating educational materials in low-resource settings. The integration of a tailored rule-based verb conjugator significantly improved accuracy on the Bribri data of Americas2024ST2, demonstrating the importance of addressing linguistic complexity in low-resource settings. Experiments on Maya and Guarani, using a minimal-effort adaptation focused solely on partial POS tagging, also showed improvements over the edit-tree baseline. Future work should explore scalable methods to expand rule-based frameworks more efficiently while maintaining high accuracy, as well as integrating advanced prompting techniques like chain-of-thought reasoning or Retrieval-Augmented Generation (RAG) to enrich contextual understanding. Our proposed method of integrating rule-based techniques into LLM prompts offers a practical and scalable approach for revitalizing underrepresented languages, as the LLM does not have to be trained on a language directly to be able to complete tasks in it effectively.

Acknowledgments

We would like to express our gratitude to Junehwan Sung for his invaluable advice and support throughout the AmericasNLP 2024 Shared Task. We are also deeply thankful to Professor Carla Victoria Jara Murillo and Professor Haakon S. Krohn for granting us permission to use and repackage their Bribri dictionary and textbook

References

[1] Luis Chiruzzo, Pavel Denisov, Samuel Canul Yah, Lorena Hau Ucán, Marvin Agüero-Torales, Aldo Alvarez, Silvia Fernandez Sabido, Alejandro Molina Villegas, Abteen Ebrahimi, Robert Pugh, Arturo Oncevay, Shruti Rijhwani, Rolando Coto-Solano, Katharina von der Wense, and Manuel Mager. Findings of the AmericasNLP 2024 shared task on the creation of educational materials for indigenous languages. In **Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)**. Association for Computational Linguistics, June 2024.

[2] Bozhidar Bozhanov and Ivan Derzhanski. Rosetta stone linguistic problems. In Ivan Derzhanski and Dragomir Radev, editors, **Proceedings of the Fourth Workshop on Teaching NLP and CL**, pp. 1–8, Sofia, Bulgaria, August 2013. Association for Computational Linguistics.

[3] Jannis Vamvas. Translation puzzles are in-context learning tasks, 2022.

[4] Nathan Chi, Teodor Malchev, Riley Kong, Ryan Chi, Lucas Huang, Ethan Chi, R. McCoy, and Dragomir Radev. ModeLing: A novel dataset for testing linguistic reasoning in language models. In Michael Hahn, Alexey Sorokin, Ritesh Kumar, Andreas Shcherbakov, Yulia Otmakhova, Jinrui Yang, Oleg Serikov, Priya Rani, Edoardo M. Ponti, Saliha Muradoğlu, Rena Gao, Ryan Cotterell, and Ekaterina Vylomova, editors, **Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP**, pp. 113–119, St. Julian's, Malta, March 2024. Association for Computational Linguistics.

[5] Jared Coleman, Bhaskar Krishnamachari, Ruben Rosales, and Khalil Iskarous. LLM-assisted rule based machine translation for low/no-resource languages. In Manuel Mager, Abteen Ebrahimi, Shruti Rijhwani, Arturo Oncevay, Luis Chiruzzo, Robert Pugh, and Katharina von der Wense, editors, **Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)**, pp. 67–87, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[6] Justin Vasselli, Arturo Martínez Peguero, Junehwan Sung, and Taro Watanabe. Applying linguistic expertise to LLMs for educational material development in indigenous languages. In Manuel Mager, Abteen Ebrahimi, Shruti Rijhwani, Arturo Oncevay, Luis Chiruzzo, Robert Pugh, and Katharina von der Wense, editors, **Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)**, pp. 201–208, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[7] Haakon S Krohn. Diccionario bribri–español español–bribri, 2023.

[8] Carla Victoria Jara Murillo. **Gramática de la lengua bribri**, Vol. 1. EDigital, San José, 2018.

[9] Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mixtral of experts, 2024.

[10] Luis Chiruzzo, Pavel Denisov, Alejandro Molina-Villegas, Silvia Fernandez-Sabido, Rolando Coto-Solano, Marvin Agüero-Torales, Aldo Alvarez, Samuel Canul-Yah, Lorena Hau-Ucán, Abteen Ebrahimi, Robert Pugh, Arturo Oncevay, Shruti Rijhwani, Katharina von der Wense, and Manuel Mager. Findings of the AmericasNLP 2024 shared task on the creation of educational materials for indigenous languages. In Manuel Mager, Abteen Ebrahimi, Shruti Rijhwani, Arturo Oncevay, Luis Chiruzzo, Robert Pugh, and Katharina von der Wense, editors, **Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)**, pp. 224–235, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[11] David Bolles and Alejandra Bolles. **A Grammar of the Yucatecan Mayan Language**. 1996.

A Maya and Guarani

A.1 Data

Americas2024ST2 includes data for Maya and Guarani as well as Bribri. The data collected is as follows:

Maya The Maya dataset focused on Yucatec Maya, a language with complex grammatical features distinct from European languages. The data originated from a collaborative effort between SEDECULTA (the Secretariat of Culture and the Arts of Yucatán) and CentroGeo for the development of a machine translation system. This initial data included 13,873 Maya-Spanish parallel sentences, which were later refined and annotated for the shared task.

The shared task data consisted of 1,400 phrases derived from this corpus, annotated with 12 grammatical tags, including predicate type, statement type, mood, aspect, and transitivity. Sentences were grouped into clusters, where each cluster contained a base sentence and several variations with minor grammatical modifications. These clusters aimed to reflect diverse linguistic features, including affirmatives, negatives, interrogatives, and different tenses.

Guarani The Guarani dataset focused on the Paraguayan variety, a language spoken by approximately six million people across South America. Guarani' s morphology is highly complex, with verbs inflected for person, number, tense, aspect, and mood, and often involving circumfixes for negation. This dataset aimed to challenge models with these intricate linguistic features.

The data was sourced from three main contributors: the Jojajovai parallel corpus, Mozilla Common Voice transcriptions, and a grammar-based generator for Guarani-Spanish sentence pairs. The generator provided around 80% of the training and development clusters, while the Common Voice data accounted for 33% of the test set. Sentences were manually reviewed and annotated by three linguists, including two native speakers. To increase difficulty, verbs seen in the training data were excluded from the test set, requiring systems to generalize across unseen examples. Annotation features included person, number, polarity, aspect, and verb nasal/oral categorization, the latter influencing affix compatibility.

A.2 Results

We conducted experiments with multiple LLMs on all three languages. Table 5 details the performance across different LLMs, noting that while Mixtral scored more competitively with GPT 3.5 for Maya, it was very ineffective for Guarani. GPT-4 resulted in the highest accuracy for all three languages.

Table 5 The results on the development set for the different LLMs for Maya and Guarani.

Lang	System	Acc.	BLEU	ChrF
Maya	Mixtral	44.97	69.19	83.52
	GPT-3.5	42.28	67.84	86.04
	GPT-4	**56.38**	**78.26**	**91.33**
Guarani	Mixtral	12.66	20.95	69.84
	GPT-3.5	36.71	51.38	83.35
	GPT-4	**41.77**	**55.81**	**86.12**