Leveraging Dictionaries and Grammar Rules for the Creation of Educational Materials for Indigenous Languages

Justin Vasselli, Haruki Sakajo
Arturo Martínez Peguero, **Frederikus Hudi**, **Taro Watanabe**
Nara Institute of Science and Technology
{vasselli.justin_ray.vk4, sakajo.haruki.sd9,
martinez_peguero.arturo.ma3, frederikus.hudi.fe7, taro}@is.naist.jp

Abstract

This paper describes the NAIST submission to the AmericasNLP 2025 shared task on the creation of educational materials for Indigenous languages. We implement three systems to tackle the unique challenges of each language. The first system, used for Maya and Guarani, employs a straightforward GPT-4o few-shot prompting technique, enhanced by synthetically generated examples to ensure coverage of all grammatical variations encountered. The second system, used for Bribri, integrates dictionary-based alignment and linguistic rules to systematically manage linguistic and lexical transformations. Finally, we developed a specialized rule-based system for Nahuatl that systematically reduces sentences to their base form, simplifying the generation of correct morphology variants.

1 Introduction

The development of educational materials for Indigenous languages presents unique challenges due to their low-resource nature, limited digital representation, and morphological complexity. The AmericasNLP 2025 Shared Task (de Gibert et al., 2025) addresses these challenges by focusing on the creation of accurate grammatical modifications in sentences across several Indigenous languages: Bribri, Maya, Guarani, and Nahuatl. The goal of the shared task was to apply specified grammatical transformations to source sentences in order to generate appropriate new sentences that could be used in educational content for language learning and preservation.

Historically, language processing tasks such as grammatical transformations, have relied on extensive corpora. However, such resources are scarce or entirely unavailable for many Indigenous languages. Building on our successful approach from the AmericasNLP 2024 Shared Task, we again leverage dictionaries and linguistic rules combined with the generative capabilities of GPT-4o (Achiam et al., 2023). This year we try a new technique which proved to be less effective than our technique from 2024, but still resulted in strong scores for Bribri. We also tested an entirely rule-based system for Nahuatl, which while still in early stages, nevertheless achieves significant improvements over LLM prompting.

Our submission comprises three distinct translation systems. The first system, submitted for Maya and Guarani, employs a straightforward GPT-4o few-shot prompting technique, enhanced by synthetically generated examples to ensure coverage of all grammatical variations encountered. The second system, used for Bribri, integrates dictionary-based alignment with GPT-4o, inspired by the edit-tag method used in the Grammatical Error Correction Tagged with Edits (GECTOR) system (Omelianchuk et al., 2020), to manage lexical and morphological transformations systematically. Finally, recognizing the specific complexities of Nahuatl, we developed a specialized rule-based system that classifies grammatical features, reduces sentences to a base form, and generates the target sentence from that base form.

2 Task and Data Description

In this shared task, the provided dataset includes original sentences along with the grammatical transformations to be applied to these sentences. The goal is to develop systems capable of applying these transformations accurately to the base sentences, producing grammatically modified versions suitable for educational use.

While many instances in the data consisted of a single change, there were many compound changes as well, where multiple types of transformations were combined, especially for Nahuatl and Bribri (See Appendix A). For example, a negative type alteration (TYPE:NEG) may be combined with a

Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 112–118

Language	Original	With Synthetic
Bribri	309	533
Maya	594	615
Guarani	178	186
Nahuatl	391	391

Table 1: Number of example sentences initially provided versus the number actually utilized after adding synthetic examples. We did not create synthetic examples for Nahuatl.

Source	Ie' dúwa̲
Change	TYPE:NEG, TENSE:PRF_PROG
Target	Ie kɛ̀ ku̲'bak dawőkwa̲
KEEP:	ie'
ADD:	kɛ̀ (negation particle)
ADD:	ku̲'bak (NEG PRF_PROG marker)
CHANGE: form dawőkwa̲	base form dúwa̲ -> PRF_PROG

Table 2: Example with change description

change to an interrogative (SUBTYPE:INT). This would have the effect of going from "I walked" to "Didn't you walk?" in English. This may be further combined with transformations to subject, such as to 3rd person plural (PERSON:3_PL): "Didn't they walk?"

We synthetically enhanced the training set by expanding changes into component substeps, combining alterations to make more compound changes. The number of sentences before and after expansion are listed in Table 1.

Sub-step Expansion We decomposed complex grammatical transformations into simpler, sequential sub-steps. For example, a change labeled `TYPE:NEG, SUBTYPE:INT` was expanded into two distinct steps: initially applying `TYPE:NEG` to reach an intermediate form, followed by `SUBTYPE:INT` to attain the final sentence.

Change Combination Additionally, we introduced new examples by combined changes. For example, a change in tense or mood would be combined with a person's changes. We aimed to have comprehensive coverage of all grammatical transformation combinations.

3 System Description

We implemented three systems, varying in their dependence on prompting versus rule-based processing. For each language, we selected the system that performed best on the dev set.

3.1 Example-Based Prompt

The first system leverages GPT-4o exclusively through few-shot prompting, relying on synthetic examples to maximize its coverage of grammatical variations. In this approach, we choose examples from the training data with the exact same change, from which the LLM can hopefully learn to generalize and perform similar modifications on new sentences. As mentioned in Section 2, there was not always an exact match for the change in the training data. This approach differed from the submission last year, JAJ (Vasselli et al., 2024), which addressed the lack of comprehensive coverage of change combinations by iteratively processing the test cases, applying sub-changes in a different order for each language. We also experimented with translating the prompt into Spanish, which improved scores for Bribri, but did not help Maya, Guarani, or Nahuatl.

3.2 Transformation-Based Prompt

The second system is based on the intuition that grammatical changes typically require only a small number of edits to the source sentence. Inspired by GECTOR (Omelianchuk et al., 2020), we annotate each training example with an explicit transformation sequence. Each transformation is framed in terms of token-level operations:

- **KEEP** for words that remain unchanged
- **ADD** for newly inserted words
- **REMOVE** for words that are removed.
- **REPLACE** for words that are replaced with different word types.
- **CHANGE** for cases where the word form changes, but the base word type is preserved (e.g., tense/person inflection).

This format allows GPT-4o to operate more conservatively by avoiding unnecessary rewrites, leading to more interpretable predictions and improved generalization. In addition, it facilitates automatic double-checking of each transformation using dictionary lookups or morphological rules, further enhancing the reliability of the output. An example can be seen in Table 2.

Using this method greatly improved performance on Bribri. Even moreso when the tagged change output was postprocessed. See Table 3 for ablation results on the development set.

System	Acc.	BLEU	ChrF
Examples	4.25	9.77	35.21
+ Description	15.09	40.94	58.24
+ Postprocessing	36.79	60.83	70.80

Table 3: Ablation experiments on the Bribri development set using examples only, with change descriptions, and postprocessing the change description output.

3.3 Pure Rule Based Transformation

The third system is a fully rule-based approach developed specifically for Nahuatl. Unlike Bribri, we lacked a digitized dictionary, preventing us from applying the transformation-based method described in Section 3.2. Nahuatl also presents more grammatical changes per sentence than Maya or Guarani, making the example-based approach less effective.

To address this, we created rules to heuristically assign part-of-speech tags using word position and known affixes. These tags were then used to infer grammatical features of each sentence—such as subject, object, and indirect object person markers, honorific status, type, and purposive direction.

Grammatical Feature Identification Evaluation We used the training data to infer grammatical features by identifying sentences that appeared in multiple transformation pairs. Table 4 shows two such examples. [1]

From the first pair, we infer that the target sentence is honorific (HON:1), has a 2nd person plural subject, a 3rd person plural object, a 3rd person singular indirect object, and is not purposive. This implies that the source sentence differs in those respects, but the only meaningful thing we learn about the source is that it is not honorific.

However, the same source sentence appears as the target in the second pair. From that example, we infer that "tehuatl amo otinechnextilito nin tlatzotzonal" has a 2nd person singular subject, is negative, and expresses purposive intent toward the speaker. Since these features were not listed as changed in the first pair, we can propagate them to the first target as well, inferring that the target of the first pair is also negative. We also infer that the second source sentence is not honorific.

[1]There is an error in this sentence which affects five other examples in the provided data: "otinechnextilito" should be "otinechnoxtilico" for PURPOSIVE:VEN. This error, in an already infrequent change category, may have contributed to the challenge of learning the PURPOSIVE feature.

Source	tehuatl amo otinechnextilito nin tlatzotzonal
Change	HON:1, PERSON[IOBJ]:3_SI, PERSON[OBJ]:3_PL, PERSON[SUBJ]:2_PL, PURPOSIVE:NA
Target	nimehuantzitzin amo onoconnextilihqueh nin tlatzotzonal
Source	yehuatl onechnextileh nin tlatzotzonal
Change	PERSON[SUBJ]:2_SI, PURPOSIVE:VEN, TYPE:NEG
Target	tehuatl amo otinechnextilito nin tlatzotzonal

Table 4: Examples from the Nahuatl training set

Quality	Training	Development
Honorific	93.7	100.0
Subject	59.0	88.6
Possessor	69.0	100.0
Object	31.0	-
Ind. Object	0.0	-
Tense	64.8	82.1
Mood	75.9	83.3
Aspect	58.6	88.9
Purposive	0.0	-
Type	100.0	100.0
Transitivity	0.0	-

Table 5: Results of rule-based classification. "-" indicates there was not enough information in the set to generate test cases for this quality.

By iterating over the dataset in this way, we assembled a more complete set of grammatical features for each sentence. These annotations allowed us to evaluate our rule-based system by assigning source and target features, applying transformations, and comparing the result.

Table 5 shows classification results on training and dev sets. While our system performs well on simpler features like type (positive or negative), it struggles with indirect object, transitivity, and purposive features, indicating areas for future improvement.

Inference Time At inference time, we used the classifier to predict the grammatical features of a new source sentence. These predicted features were then modified according to the specified changes to derive the expected target sentence features. We decomposed the source sentence into a normalized default form—non-honorific, 3rd person singular subject, no possessor, present simple tense, no mood, and positive type—by systematically stripping or converting known morphological indicators. From this base form, we then generated the target sentence by applying all grammatical features required by the target configuration.

This rule-based generation pipeline still requires

System	Bribri			Guarani			Maya			Nahuatl		
	Acc.	BLEU	ChrF	Acc.	BLEU	ChrF	Acc.	BLEU	ChrF	Acc.	BLEU	ChrF
Edit-tree baseline	5.66	20.35	45.56	22.78	34.99	77.14	26.17	52.38	78.72	0.00	1.38	34.32
Example-based Prompt	4.25	9.77	35.21	**45.57**	**55.53**	**86.77**	**45.64**	**71.21**	**87.28**	0.57	3.40	34.76
+ Spanish prompt	8.49	31.32	55.90	37.97	51.68	84.14	42.28	70.18	86.28	0.57	1.64	31.54
Transformation-based Prompt	15.09	40.94	58.24	39.24	50.58	85.59	42.95	69.13	84.62	-	-	-
+ Postprocessing	**36.79**	**60.83**	**70.80**	15.19	42.31	77.18	40.94	70.22	84.77	-	-	-
Rule-based Transformation	-	-	-	-	-	-	-	-	-	**26.14**	**26.64**	**52.19**

Table 6: Results on the development set. "-" indicates the system does not currently support that language.

System	Bribri			Guarani			Maya			Nahuatl		
	Acc.	BLEU	ChrF	Acc.	BLEU	ChrF	Acc.	BLEU	ChrF	Acc.	BLEU	ChrF
Edit-tree baseline	8.75	22.11	52.73	14.84	25.03	76.10	25.81	53.69	80.23	-	-	-
JAJ (Vasselli et al., 2024)	**54.17**	**71.72**	**82.78**	**36.81**	**48.29**	**84.12**	**53.55**	**78.41**	**91.53**	-	-	-
Ours	41.25	62.57	74.99	32.69	49.21	84.98	42.90	71.81	88.97	**17.5**	**40.50**	**65.40**

Table 7: Results on the test set. Ours was the best performing system for each language on the development set: Postprocessed transformation-based prompt for Bribri, English language Example-based Prompt for Maya and Guarani, and Rule-based Transformation for Nahuatl.

further refinement, particularly for accurate reconstruction of morphologically complex forms. However, the system proved to be more effective than the example-based prompting approach when evaluated on the Nahuatl development set.

4 Results

As seen in Table 7, across all four languages, our systems outperformed the edit-tree baseline provided in the shared task in terms of accuracy, BLEU, and ChrF scores. However, our results did not reach the performance levels of the JAJ system from last year.

For Maya and Guarani, our approach this year applied all changes at once using synthetically constructed examples, whereas the JAJ system applied transformations incrementally. The iterative strategy appears to reducing the complexity at each transformation step, improving accuracy.

In Bribri, two factors probably contributed to our lower scores. First, as with Maya and Guarani, we did not apply changes iteratively. Second, we omitted explicit conjugation hints from the prompt, which were included in the JAJ system and likely contributed to the improved performance. Although our post-processing step was designed to enforce correct conjugation, it is unknown whether it is less effective than targeted prompting. A combination of the edit-tag prompting method with conjugation hints and iterative change application is a promising direction for future experiments.

Nahuatl was introduced to the task for the first time this year and was the most challenging for our system. Although our rule-based system performed better than the example-based prompting baseline, it still falls short of ideal performance. The lack of a digitized dictionary and the large number of interacting grammatical features per sentence continue to pose significant challenges.

5 Related Work

Rosetta Stone Puzzles In Rosetta Stone puzzles (Bozhanov and Derzhanski, 2013), solvers are given a limited set of bilingual sentence pairs and asked to translate sentences into the other language. These puzzles contain machine translation and grammatical transformation. Şahin et al. (2020) tested several algorithms for those problems, including statistical algorithms and Transformer-based language models (Vaswani et al., 2017). Sung et al. (2024) explored the metalinguistic awareness of pre-trained language models. Chi et al. (2024) and Bean et al. (2024) developed benchmarks in the same format as Rosetta Stone puzzles and tested several LLMs. The results demonstrate that LLMs potentially have the capabilities to apply linguistic knowledge and extract linguistic features from limited data.

LLM-Assisted Rule-Based Approach An LLM-assisted rule-based approach demonstrates promising performance, particularly for low-resource languages. Low-resource languages have limited linguistic resources, resulting in the challenging performance of LLMs. To address this issue, several studies have leveraged existing linguistic knowledge to develop pipeline systems that apply rule-based processing to input in low-resource

languages before passing it to LLMs. Coleman et al. (2024) introduced a new methodology, LLM-Assisted Rule-Based Machine Translation, and explored the performance and advantages. Zhang et al. (2024) proposed a method that decomposes inputs into morphemes with morphological analyzers, assigns glosses to each morpheme with dictionaries, and uses them for translation. Both methods leverage rule-based approaches to narrow the candidates or add rich information to the original input, guiding LLMs to the correct output.

6 Conclusion

We presented three systems for generating educational sentence transformations in Indigenous languages, varying in their use of prompting and linguistic rules. Our systems consistently outperformed the baseline across all four languages, but results suggest several areas for refinement.

For Maya and Guarani, applying all changes at once proved less effective than the iterative approach used in previous work. For Bribri, the absence of conjugation cues in the prompt may have hindered performance, even with post-processing. For Nahuatl, our rule-based system offered improvements over prompting alone, but remains limited by the lack of digitized lexical resources.

Future work will focus on refining the rule-based system, incorporating a Nahuatl dictionary to support edit-tag prompting, and adopting iterative application of changesa strategy that yielded strong results in prior shared tasks.

The interplay between LLM-based reasoning and structured linguistic knowledge emerged as a key factor in producing reliable transformations—especially when creating educational tools for under-resourced Indigenous languages.

Limitations

The purely prompt-based approach is highly sensitive to the quality and coverage of examples. When faced with compound grammatical transformations, our system often failed to generalize.

The transformation-based system relies on accurate alignments, which in turn relies on complete dictionaries. While effective for Bribri, incomplete dictionaries may lead to missing or incorrect transformation annotations, which in turn affect the system's outputs.

For Nahuatl, the rule-based system is based on hand-crafted heuristics and POS inference rules. These rules are not always accurate and can misclassify grammatical qualities. Additional work must be done to make this system more accurate.

Acknowledgments

We extend our gratitude to the authors of the initial effort for Guarani Wordnet (Chiruzzo et al., 2023) for access to their data. Special thanks to Professor Carla Victoria Jara Murillo and Professor Haakon S. Krohn who generously allowed us to use and repackage their Bribri data (Jara Murillo, 2018; Krohn, 2023) for use in this project. Also, a big thank you to our reviewers who pointed out the error in the Nahuatl example and helped us improve this paper.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, and 1 others. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Andrew Michael Bean, Simeon Hellsten, Harry Mayne, Jabez Magomere, Ethan A Chi, Ryan Andrew Chi, Scott A. Hale, and Hannah Rose Kirk. 2024. LINGOLY: A benchmark of olympiad-level linguistic reasoning puzzles in low resource and extinct languages. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Bozhidar Bozhanov and Ivan Derzhanski. 2013. Rosetta stone linguistic problems. In *Proceedings of the Fourth Workshop on Teaching NLP and CL*, pages 1–8, Sofia, Bulgaria. Association for Computational Linguistics.

Nathan Chi, Teodor Malchev, Riley Kong, Ryan Chi, Lucas Huang, Ethan Chi, R. McCoy, and Dragomir Radev. 2024. ModeLing: A novel dataset for testing linguistic reasoning in language models. In *Proceedings of the 6th Workshop on Research in Computational Linguistic Typology and Multilingual NLP*, pages 113–119, St. Julian's, Malta. Association for Computational Linguistics.

Luis Chiruzzo, Marvin Agüero-Torales, Aldo Alvarez, and Yliana Rodríguez. 2023. Initial experiments for building a Guarani WordNet. In *Proceedings of the 12th Global Wordnet Conference*, pages 197–204, University of the Basque Country, Donostia - San Sebastian, Basque Country. Global Wordnet Association.

Jared Coleman, Bhaskar Krishnamachari, Ruben Rosales, and Khalil Iskarous. 2024. LLM-assisted rule based machine translation for low/no-resource languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages*

of the Americas (AmericasNLP 2024), pages 67–87, Mexico City, Mexico. Association for Computational Linguistics.

Ona de Gibert, Raul Vazquez, Robert Pugh, Abteen Ebrahimi, Pavel Denisov, Ali Marashian, Enora Rice, Edward Gow-Smith, Juan C. Prieto, Melissa Robles, Rubén Manrique, Oscar Moreno Veliz, Ángel Lino Campos, Rolando Coto-Solano, Aldo Alvarez, Marvin Agüero-Torales, John E. Ortega, Luis Chiruzzo, Arturo Oncevay, and 3 others. 2025. Findings of the AmericasNLP shared tasks on machine translation, creation of educational material, and translation metrics for indigenous languages. In *Proceedings of the 5th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2025)*, Albuquerque, New Mexico. Association for Computational Linguistics.

Carla Victoria Jara Murillo. 2018. *Gramática de la lengua bribri*, volume 1. EDigital, San José.

Haakon S Krohn. 2023. Diccionario bribri–español español–bribri.

Kostiantyn Omelianchuk, Vitaliy Atrasevych, Artem Chernodub, and Oleksandr Skurzhanskyi. 2020. GECToR – grammatical error correction: Tag, not rewrite. In *Proceedings of the Fifteenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 163–170, Seattle, WA, USA → Online. Association for Computational Linguistics.

Gözde Gül Şahin, Yova Kementchedjhieva, Phillip Rust, and Iryna Gurevych. 2020. PuzzLing Machines: A Challenge on Learning From Small Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1241–1254, Online. Association for Computational Linguistics.

Junehwan Sung, Hidetaka Kamigaito, and Taro Watanabe. 2024. Exploring metalinguistic awareness in pretrained language models through the international linguistics olympiad challenges. In *Proceedings of the Thirtieth Annual Meeting of the Association for Natural Language Processing*, Kobe, Japan. Association for Natural Language Processing.

Justin Vasselli, Arturo Martínez Peguero, Junehwan Sung, and Taro Watanabe. 2024. Applying linguistic expertise to LLMs for educational material development in indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 201–208, Mexico City, Mexico. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, NIPS'17, pages 6000–6010, Red Hook, NY, USA. Curran Associates Inc.

Kexun Zhang, Yee Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. Hire a linguist!: Learning endangered languages in LLMs with in-context linguistic descriptions. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 15654–15669, Bangkok, Thailand. Association for Computational Linguistics.

A Data Distribution

We observed that Maya and Guarani examples typically involved only one or two grammatical changes per instance, whereas Bribri and Nahuatl frequently included compound transformations affecting multiple features simultaneously. This discrepancy is illustrated in Table 8 and Figure 1. We hypothesize that this difference in complexity contributed to the weaker performance of purely prompt-based systems on Bribri and Nahuatl, as those systems may struggle to generalize when required to model multiple interacting changes at once as illustrated in Figure 2.

		1	2	3	4	5	6	7	8	Total
bribri	train	51 (16.5%)	89 (28.8%)	75 (24.3%)	60 (19.4%)	26 (8.4%)	7 (2.3%)	1 (0.3%)	-	309
	dev	46 (21.7%)	62 (29.2%)	51 (24.1%)	33 (15.6%)	16 (7.5%)	3 (1.4%)	1 (0.5%)	-	212
	test	83 (17.3%)	141 (29.4%)	125 (26.0%)	85 (17.7%)	41 (8.5%)	5 (1.0%)	-	-	480
guarani	train	175 (98.3%)	3 (1.7%)	-	-	-	-	-	-	178
	dev	79 (100.0%)	-	-	-	-	-	-	-	79
	test	361 (99.2%)	3 (0.8%)	-	-	-	-	-	-	364
maya	train	538 (90.6%)	47 (7.9%)	6 (1.0%)	1 (0.2%)	2 (0.3%)	-	-	-	594
	dev	138 (92.6%)	8 (5.4%)	1 (0.7%)	1 (0.7%)	1 (0.7%)	-	-	-	149
	test	222 (71.6%)	83 (26.8%)	5 (1.6%)	-	-	-	-	-	310
nahuatl	train	17 (4.3%)	69 (17.6%)	98 (25.1%)	90 (23.0%)	72 (18.4%)	28 (7.2%)	14 (3.6%)	3 (0.8%)	391
	dev	17 (9.7%)	49 (27.8%)	52 (29.5%)	38 (21.6%)	19 (10.8%)	-	1 (0.6%)	-	176
	test	16 (13.3%)	36 (30.0%)	41 (34.2%)	15 (12.5%)	7 (5.8%)	3 (2.5%)	2 (1.7%)	-	120

Table 8: Number of changes.

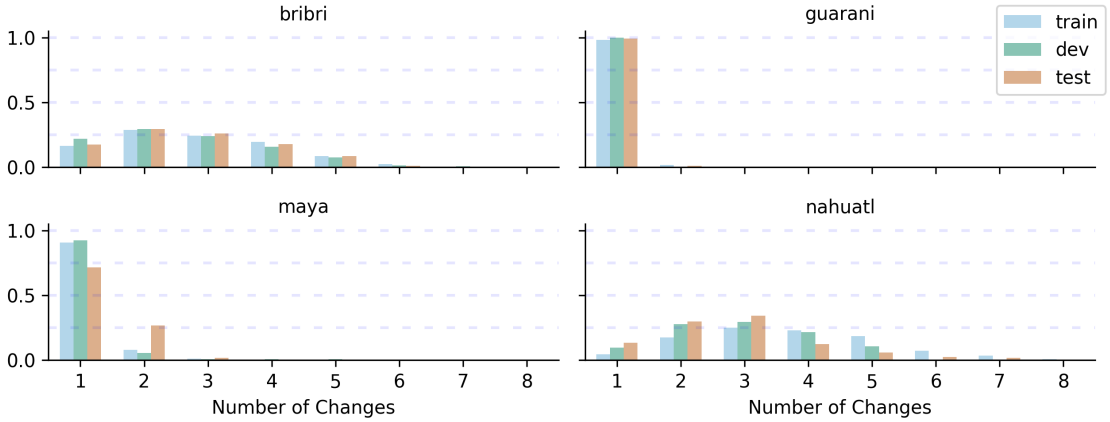

Figure 1: Ratio of number of changes across datasets.

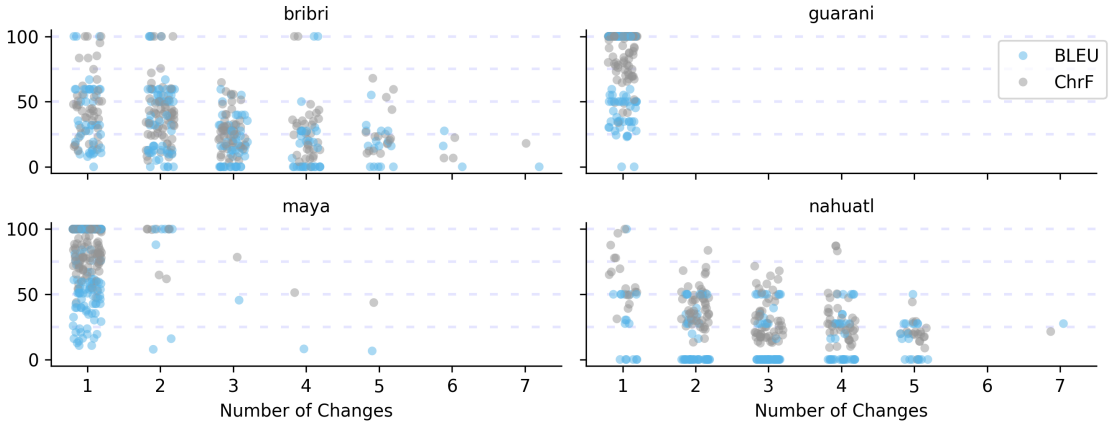

Figure 2: Performance w.r.t. number of changes in devset.