Findings of the AmericasNLP 2024 Shared Task on the Creation of Educational Materials for Indigenous Languages

Luis Chiruzzo[◇] **Pavel Denisov**[♠] **Alejandro Molina-Villegas**[§] **Silvia Fernandez-Sabido**[§]
Rolando Coto-Solano[♡] **Marvin Agüero-Torales**[△▽] **Aldo Alvarez**[Ω] **Samuel Canul-Yah**[η]
Lorena Hau-Ucán[ψ] **Abteen Ebrahimi**[♯] **Robert Pugh**[♣] **Arturo Oncevay**[⋈]
Shruti Rijhwani[℧] **Katharina von der Wense**[♯†] **Manuel Mager**[‡]

[◇]Universidad de la República, Uruguay [♠]University of Stuttgart [§]CentroGEO, Mexico
[♡]Dartmouth College [△]Universidad de Granada, Spain [η]Independent linguist, Mexico
[Ω]Universidad Nacional de Itapua, Paraguay [▽]Global CoE of Data Intelligence, Fujitsu
[ψ]SEDECULTA, Mexico [♯]University of Colorado Boulder [℧]Google DeepMind
[⋈]Pontificia Universidad Católica del Perú [♣]Indiana University, Bloomington
[†]Johannes Gutenberg University Mainz [‡]Amazon AWS AI

Abstract

This paper presents the results of AmericasNLP 2024 Shared Task on the Creation of Educational Materials for Indigenous Languages, the first natural language processing (NLP) shared task on automatically creating educational resources for languages Indigenous to the Americas. Teams are tasked with generating variations of sentences according to linguistic features that could be used for grammar exercises. The languages involved in this task are Bribri, Maya, and Guarani. Seven teams took part in the challenge, submitting a total of 22 systems, obtaining very promising results.

1 Introduction

The AmericasNLP 2024 Shared Task on the Creation of Educational Materials for Indigenous Languages is a competition aimed at encouraging the development of Natural Language Processing systems (NLP) to help with the teaching and diffusion of Indigenous languages of the Americas. Many of the Indigenous languages of the Americas are vulnerable or endangered. This means that, depending on the language, no or only a few children are learning them and, generally, they are only spoken by a few small groups of people. Because of this, some of these languages are at a high risk of becoming extinct in the near future. Many communities are carrying out revitalization efforts, including teaching their languages to their community members. Creating materials to teach these languages is an urgent priority, but this process is expensive and time consuming. NLP presents an opportunity to help with these efforts.

In addition to being endangered, the Indigenous languages of the Americas are so-called low-resource languages (Joshi et al., 2020): the data needed to train any NLP systems, let alone deep learning-based systems, is severely limited. This means that many approaches used for high-resource languages, such as English and Chinese, are not directly applicable or perform poorly. On top of this, many Indigenous languages exhibit linguistic properties uncommon among languages frequently studied in NLP. This constitutes an additional difficulty.

In this task, participants built systems for transforming sentences in an Indigenous language according to some linguistic feature (such as negation or tense), in a way that could enable to automatically create grammar exercises. This often implies inflecting the main verb of the sentence, but other types of changes could be necessary as well, such as including different adverbs or particles, or making adaptations according to agreement rules.

We hope that this challenge helps to motivate researchers to develop systems for these Indigenous languages, as well as spark the interest in NLP research for the huge diversity of languages across the American continent, as is the goal of the AmericasNLP workshop since its inception (Mager et al., 2021).

2 Related Work

NLP for Educational Applications Over the last years, NLP has been used more and more in educational contexts. Examples for this are NLP-based tutors (Wollny et al., 2021; Dyke et al., 2013; Macina et al., 2023), feedback systems for teachers (Suresh et al., 2022), or automatic student assessment (Andersen et al., 2013). Closest in spirit to the AmericasNLP Shared Task on the Creation of Educational Materials for Indigenous Languages is work on automatic exercise creation (Hill and Simha, 2016; Perez and Cuadros, 2017): with this shared task, we aim at automatically creating sen-

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 224–235

tence pairs in Indigenous languages, where the first sentence can be given to a learner with the task to correctly produce the second one by applying the indicated change.

Morphological Inflection This task shares similarities with morphological inflection shared tasks such as the SIGMORPHON 2016 Shared Task on Morphological Reinflection (Cotterell et al., 2016), in which the participants were presented with a word and a target morphological feature, and they had to inflect the word into a form corresponding to that feature. This could start with the lemma (subtask 1), with an inflected word with known morphology (subtask 2), or with an inflected word that is not annotated (subtask 3). The present task is more similar to that last subtask in that the participants are presented with an unannotated inflected form and they have to generate another inflected form, but in our case the word is used in the context of a sentence and often other words in the sentence might be affected by the change as well, so it is a case of reinflection in context.

The most important precedent of a task about reinflection in context is the CoNLL–SIGMORPHON 2018 Task 2 (Cotterell et al., 2018), where participants were presented in a cloze test format, with a sentence containing a gap and a lemma, and they had to produce the appropriate inflected form that fits the gap. In our case, we are presenting a whole sentence without gaps, and the participants have to detect the words they have to change in order to adapt it to the expected features.

These previous competitions have featured some Indigenous languages of the Americas in their data: Cotterell et al. (2016) included Navajo, while Cotterell et al. (2018) also added Quechua, Mapundungun and Greenlandic Inuit (alongside 100 more languages). As far as we know, this is the first time the Bribri, Mayan, and Guarani languages are featured in a task of these characteristics.

3 Task Description

The idea of this shared task is to automatically convert sentences in Indigenous languages into small exercises for language learners. In particular, we aim to create grammar exercises in which students must tweak a sentence changing its tense, aspect, or other morphosyntactic features. In order to do this, participants have to create systems that can automatically modify sentences with regard to a given property (e.g., they must create a negated version of a sentence). Those sentences could then be used as exercises by either asking learners to do the same transformation or by masking out all changed words in the sentence and asking learners to fill in the blank.

For instance, if a model can correctly reproduce the linguistic labels, it will also be capable of transforming simple sentences from first-person singular to first-person plural, as in the following example in Maya:

Original Sentence:
`J-jaan en tin najil (1s)`
tr. I ate at my home.

Transformed Sentence:
`J-jaano'ob tu najil (1p)`
tr. We ate at his/her/their house.

Using that pair of sentences, we could come up with the grammar exercise below.

Exercise 1. Transform the following sentence to first-person plural:
`J-jaan ____ najil`
`a) béet u`
`b) o'ob tu`
`c) o tin`
`d) o'ob janal`

Task Format The participants were provided with one data file for each language, containing the following columns:

- ID: unique identifier of the example.
- original sentence: this would be used as the system input.
- change to be conducted: tag indicating the morphosyntactic change to perform.
- target sentence: sentences expected as system output.

Systems were expected to take the original sentence and a morphosyntactic feature marker, and generate the target sentence as output. Internally, the examples were organized in clusters in which, starting from one original sentence, one or more morphosyntactic variations (deltas) were created.

The task was evaluated in terms of exact accuracy (fraction of times the system output matched the expected output), and also two classic metrics for generative tasks: BLEU (Papineni et al., 2002)

and chrF (Popović, 2015). The main metric for the task was exact accuracy.

4 Dataset

Table 1 shows a summary of the data created for this task. In each case we present the number of clusters and the number of total examples provided.

		Train	Dev	Test	Total
Bribri	Examples	309	212	480	1001
	Clusters	15	17	32	64
Maya	Examples	594	149	310	1053
	Clusters	179	53	89	321
Guarani	Examples	178	79	364	621
	Clusters	56	14	34	104

Table 1: Size of the dataset.

4.1 Bribri

Bribri (Glottolog `brib1243`) is a Chibchan language spoken in Southern Costa Rica. It is spoken by approximately 7000 people (INEC, 2011) and it is closely related to other Chibchan languages like Cabécar (Quesada, 2007). Bribri is vulnerable (Sánchez Avendaño, 2013), in that some children are not learning to speak the language from their parents.

Bribri is a morphologically ergative SOV language. Its verbs have fusional morphology, with suffixes to indicate voice, tense, aspect and mood. Bribri is a tonal language with five tones, and these also form minimal pairs in the verbal morphology (e.g., falling tone *ché* 'said' versus high tone *chè* 'is saying'). Most nouns do not have any morpheme that indicates the plural, but some animate plural nouns do trigger morphological changes in the verb, either by the use of a suffix for number agreement (e.g., *I túr* 'he runs' versus *I túndak* 'they are running'), or by changing the verb to a suppletive root for the plural (e.g., *Chìchi dör bêrie* 'The dog is big' versus *Chìchi dör wi̱wi̱* 'The dogs are big').

There are numerous published educational materials for Bribri. These include a grammar book (Jara, 2018), two textbooks (Constenla et al., 2004; Jara Murillo and García Segura, 2013), two dictionaries (Margery, 2005; Krohn, 2021), several books for school children (Sánchez Avendaño et al., 2021a,b) and several books with transcribed oral literature (Jara, 1993; Jara and García Segura, 1997; García Segura, 2016, 2021; Jara Murillo and García Segura, 2022). There is also an oral corpus (Flores-Solórzano, 2017a,b) with audiovisual recordings of oral literature.

The data included in this shared task was constructed by using examples from the textbooks and the grammar cited above, as well as examples from the treebank in Coto-Solano et al. (2021). We focused on the verbal morphology, particularly the tense-aspect-mood suffixes. We selected a total of 64 sentences and then conjugated the verbs in all their possible forms, based on the information in the books and on the conjugations in the morphological analysis of Flores-Solórzano (2017c). We included a number of irregular verbs in the example, given their high frequency in the language (e.g., *tso* 'is' versus *bák* 'was'). The 64 original examples included 33 transitive sentences, as well 17 intransitive, 8 locative intransitive and 6 copular sentences. After the conjugations, we had a total 1,001 example sentences, which were split as shown in table 1. The following are the main categories used to conjugate and derive the examples:

- **Polarity:** Sentences can be `positive` or `negative`.
- **Verbal mood:** Verbs can be conjugated for `indicative`, `imperative`, `adversative`, `exhortative` and `optative` moods. They can also be in the `knowledge` mood, which is used when someone "knows" how to do something, and is similar to the potential mood in languages like Japanese.
- **Tense and aspect:** Past tenses include the `anterior`, `perfect remote`, `perfect continuous` and `perfect recent`. Tenses that cover the present tense include the `imperfect recent`, `imperfect continuous` and `imperfect habitual`. Tenses that cover the future include the `potential future` and the `certain future`.
- **Aspect:** As a complement to the tense-aspect tag, we have a macro-tag to classify the aspect as `imperfect`, `perfect` or `inchoative`.
- **Voice:** Verbs can be in the `active` or `middle` voice.
- **Number of the absolutive:** Verbs do not have conjugations for person. Therefore, we have

included information for whether the absolutive argument is `singular`, `plural` or `zero`. We have done this because verbs can change their conjugation for some plural absolutive arguments.

- **Pronoun type:** Finally, we included information for pronoun subjects, whether they were absolutive or ergatives. Pronominal subjects can be `1SG`, `2SG`, `2PL`, `3SG` and `3PL`. The language also has a clusivity distinction between `1PL.INCL` and `1PL.EXCL` pronouns. Finally, a sentence is tagged as `no pronoun` if the subject is zero or a full nominal.

4.2 Maya

In this task, we focused on the Yucatec Maya variety (Glottolog `yuca1254`). The first version of the data in Maya was created in 2022 at the request of Duolingo[1], an American educational technology company that produces learning apps. The Secretariat of Culture and the Arts of Yucatan (SEDECULTA) served as the starting point for generating the initial Maya-Spanish aligned data. The company requested the translation from Spanish to Maya, as well as the alignment of tokens for some phrases. Although the integration of Maya into Duolingo was discarded (because the Maya did not fit the simplistic scheme that was required), scientists from The Geospatial Information Sciences Research Center (CentroGeo, Mexico) promoted the follow-up and resumed data generation with the aim of creating technologies in Maya.

With the help of several Maya speakers, linguists, NLP practitioners, and volunteers, the process of creating aligned Maya-Spanish phrase corpora continued, part of which is included in the challenge data. These background details are important because they explain why the data has a simple structure, covers everyday topics, and features slight variations in grammatical characteristics. These data were always intended as inputs for educational materials.

As the Maya-Spanish aligned data was created with the aim of generating an automatic translator, it includes themes of everyday contexts: greetings, farewells, park, market, house, cornfield, lot, school, weather, courtesy, family, work, town, location, daily life, physical description, shopping, travel, pets, birds, insects among others. At the end of 2022, from November to December, three native-speaking Mayan scholarship recipients generated the phrases. Each one created 3,200 phrases in Mayan and their corresponding translation into Spanish. 9,600 parallel phrases were achieved, which, added to those that had previously been generated for Duolingo, reached a total of 13,873.

Before starting to create the phrases, the speakers were trained giving them the instruction that, for each assigned topic, they should consider the most commonly used expressions in orality, making a written version that was as natural as possible. In this way they would be useful to learn Mayan as a second language. The initial production went through a testing phase and several revisions. In the final phase, they were instructed to make simple phrases using the demonstrative, phrases with different aspects and people, affirmative, negative, transitive and intransitive phrases, and descriptive, among others. Of the 13,873 phrases, 1,400 were selected to generate the groups with labels for this challenge.

The grammatical annotation of the corpus was done by NLP specialists and a native speaker linguist, whose invaluable help provided insights on how Mayan grammar is very difficult to analyze with a Eurocentric linguistic mindset. We had hundreds of phrases in Mayan with their translation into Spanish and we had to give each one grammatical labels that mainly indicated the type of phrase (affirmative, negative, interrogative,...), person (1st singular, 2nd plural,...), verbal tense (present, past, future), among other categories. We naively thought that it was a tedious but simple task, believing we could rely on the Spanish version to achieve a good classification.

Everything went through a double or triple check, and in case of disagreement a few minutes of discussion were enough to reach a consensus and continue. But it was time for a complex and fascinating situation that had no simple solution: establishing the verb tense of Mayan phrases. This is because the very concept of verbal tense simply does not exist in this language, and this information is conveyed by other means. We noticed that on many examples there was no difference in the time they occur, but rather the degree of completeness of the action (mood) and the intention in carrying it out (aspect) (Briceño Chel, 2021; Yoshida, 2016; Chan Dzul, 2010). The tense of a phrase exists but not as an inflection of the verb, it is introduced with additional particles such as adverbs (Yoshida,

[1]https://duolingo.com/

2016).

Finally, we had a selection of 1,400 annotated phrases with 12 grammatical tags: **predicate_type**, **statement_type**, **statement_subtype**, **mood**, **action_state**, **verbal_aspect**, **adverb_tense**, **tense**, **person**, **voice**, **transitivity**, and **mark**. Additionally, the phrases were classified into clusters with one base and several deltas in each one. Each delta contains one or two grammatical differences from the base. The used split was 50% train, 20% dev and 30% test.

4.3 Guarani

Guarani is a language belonging to the Tupian stock with around 6 million native speakers in several countries of South America, mainly in Paraguay and some regions of Argentina, Bolivia and Brazil. As many Indigenous languages of the Americas, Guarani has a very complex noun and verbal morphology, with words that change their POS according their affixes and the way they are used in the sentence. The verbal category is the most complex one, containing prefixes that encode person and number, many possible suffixes that encode for voice, tense, aspect, mood and grade, and also a circunfix to create negative forms (Academia de la Lengua Guaraní, 2018).

In this task we focused on the Paraguayan variety of the Guarani language (Glottolog `para1311`). Although this variety is not considered immediately endangered, it is considered vulnerable due to the massive borrowing of Spanish terms and idioms (Moseley, 2010) as a result of the contact with European languages since the 16th century (Rodríguez Gutiérrez and Núñez Méndez, 2018).

For this dataset we used three sources of sentences: the blogs subset of the Jojajovai corpus (Chiruzzo et al., 2022); the transcriptions of the Guarani data from Mozilla Common Voice[2], already used in (Ebrahimi et al., 2022); and a simple generator of Guarani-Spanish pairs based on feature grammars and transfer rules (Lucas et al., 2024). We always started with an original sentence in Guarani annotated with their corresponding morphosyntactic features, then selected a few variations in the features to create a cluster of between 5 and 10 examples, finally we wrote the modified sentences manually. The training and development data were collected from the generator (around 80% of the clusters) and the Jojajovai data (around 20% of the clusters), plus a few examples written manually. The test data was collected from the generator (around 67% of the clusters) and the Common Voice dataset (around 33% of the clusters). The Common Voice sentences were the hardest to work with, as they were much more complex than the other sources, and often featured more than one verbal construction.

Three annotators, two of them native speakers of Guarani, took part in this annotation process, and all the final sentences in the dataset were reviewed by the native speakers. In order to make the task more challenging, we tried as much as possible to keep examples that use the same main verb on the same split, so that systems need to generalize the different inflection types to unseen examples.

The set of features used to annotate the Guarani variations is the following:

- **Person and number:** Combinations of first, second and third person, both singular and plural. Also, Guarani distinguishes between forms that include or exclude the interlocutor for the first person plural (`1SG`, `2SG`, `3SG`, `1PL.INCL`, `1PL.EXCL`, `2PL`, `3PL`).
- **Tense:** `Present`, `Simple future`, `Recent past`, `Imperfect past`, `Pluperfect past`.
- **Polarity:** `Affirmative` or `Negative` forms of the verb.
- **Aspect:** Besides the base form, we included the `Imperfective` (progressive or continuous) and `Intermittent` (an action performed occasionally, but not always) aspects.

These features are often marked as affixes of the verb or as accompanying adverbs. Another important feature in Guarani is the categorization of verbs and other words as nasal or oral terms. This categorization is based on the pronunciation of words, and impacts the types of affixes and pronouns that could be used, in a phenomenon called nasal/oral agreement (Academia de la Lengua Guaraní, 2018).

5 Approaches and Results

This section describes the different approaches that the participants used to solve the task, as well as the baseline approach we implemented, and then presents the results obtained by these approaches. Seven teams took part in the shared task, submitting

[2] `https://commonvoice.mozilla.org/`

a total of 22 systems. All seven teams submitted results for the Bribri and Maya languages, while for Guarani only four teams presented results. The methods of both systems by the anonymous submission are not known.

5.1 Baseline

Our baseline system is a simplified adaptation of the Prefer Observed Edit Trees (POET) method (Kann and Schütze, 2016). An edit tree (Chrupała, 2008) is a tree of edit operations which are applied recursively to a source string (source sentence) to obtain a target string (target sentence). There are two types of nodes in edit trees: a substitution node and a match node. A substitution node outputs a fixed target string given a fixed source string. A match node splits a source string to a possibly empty prefix of a fixed length, a fixed matched substring, and a possibly empty suffix of a fixed length. Prefix and suffix point to their own edit trees. An output of a match node is a concatenation of the output of prefix edit tree applied to the prefix, the matched substring, and the output of suffix edit tree applied to the suffix. Given a source and a target strings, an edit tree is built by recursive execution of two steps. The first step is to find the longest common substring (LCS) (Gusfield, 1997) between the source and target strings. If the LCS has a zero length, create a substitution node with the source and target strings. If the LCS length is larger than zero, the second step is to create a match node with the LCS as its match, and lengths of the parts of the source string before and after the LCS as the prefix and suffix lengths of the node. After that, the first step is repeated for the prefix and suffix. Fig. 1 shows an example edit tree for one of the training samples. We utilize the spaCy implementation of the edit trees structures[3].

During the training stage, we build an edit tree for each combination of a source sentence, a change and a target sentence in the training data, and count numbers of occurrences of each tree for each change. During the testing stage, we try to apply the most frequent edit tree for a given change to a given source sentence. If the output is not empty, we return it as a target sentence, otherwise we try to apply the next less frequent edit tree for a given change. If a target sentence is not defined after trying all edit trees observed in the training data for the given change, we return the source sentence without changes.

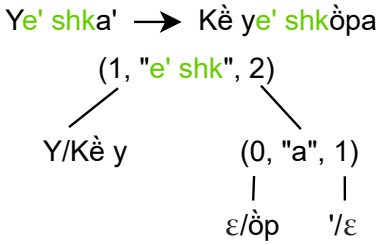

Figure 1: Edit tree for the training sample `Bribri0315`, from *Ye' shka'* 'I walked' to *Kè ye' shkòpa* 'I won't walk'. The root node is a match node with the match "e' shk", prefix length 1 and suffix length 2. Its prefix node is a substitution node that replaces "Y" with "Kè y". Its suffix node is a match node with the match "a", prefix of length 0, and suffix of length 1. Both prefix and suffix trees are substitution nodes replacing an empty string (ϵ) with "òp" for the prefix, and "'" with an empty string for the suffix.

5.2 JAJ (/dʒæz/)

The JAJ team (Vasselli et al., 2024) experimented with several LLMs and submitted predictions for Bribri and Maya languages from the system based on GPT-4 (OpenAI et al., 2024), which performed best on the development set. The LLM was given the prompt adapted from the one (Vamvas, 2022) used for the Rosetta Stone Puzzles. The prompt integrates the examples from the training set, part of speech tags generated with a dictionary based method, and some language specific hints. Language specific hints include short summaries of grammatical rules related to the changes extracted from textbooks, and, for Bribri, possible target verb form generated with a rule-based verb conjugator. Besides that, the team applied such preprocessing steps to the data, as duplicate removal and capitalization normalization, tag collapsing for the changes that mostly appear together, generation of additional training samples by labeling from the target back to source, and decomposition of certain compound changes to simple changes for sequential execution.

5.3 Meenzer Team

The Meenzer team (Bui and von der Wense, 2024) submitted predictions of four different ensembles of models for all three languages. System 1 incorporates the largest combination of mod-

[3] https://github.com/explosion/spaCy/tree/2e2334632beb0e91abc1d7820a0471a10af61489/spacy/pipeline/_edit_tree_internals

els: 10 character-level pointer-generator LSTMs (Bahdanau et al., 2015; See et al., 2017; Vinyals et al., 2015), 12 finetuned Mixtral 8x7B (Instruct) (Jiang et al., 2024) models, and 2 GPT-4 (OpenAI et al., 2024) based systems. System 2 incorporates LSTMs and Mixtral models only, system 3 incorporates LSTMs only, and system 4 incorporates Mixtral models only. The LSTMs are selected from the pool of 100 models trained with various hyperparameters, first on the training data for all three languages combined, and subsequently finetuned for each language separately. The desired set of grammatical changes is encoded as a sequence containing one token per change, combined with a language tag, and is fed to a separate LSTM encoder. The Mixtral models are finetuned using the unsupervised in-context learning (SICL) method (Li et al., 2023) with 5, 10, or 20 examples per prompt for 10 or 20 epochs, resulting in 6 different models. Each Mixtral model and GPT-4 system are used in 2 ways, differing with the order of examples in the prompt. The ensemble output is decided with majority voting.

5.4 Giving it a Shot

The Giving it a Shot team (Haley, 2024) submitted predictions of three systems based on three LLMs, namely Command R+ from Cohere (system 1), and GPT-3.5 Turbo and GPT-4 (OpenAI et al., 2024) from OpenAI (systems 2 and 3 respectively). The prompt simply listed several lines of the training data in CSV format, the instruction to fill in the column, and a line with a test sample having the missing last column. Examples are selected from the training data according to the grammatical change in the test sample. In cases when more than 10 samples are available, examples are selected for the highest sum of BLEU and chrF scores of source sentence with the test sample. In cases when a compound grammatical change does not have any examples in the training data, this change is split and examples are searched for the resulting simple changes.

5.5 LECS Lab

The LECS Lab team (Ginn et al., 2024) submitted predictions of nine systems, one of which does not include Maya, and eight other include all three languages. System 1 is based on GPT-4 (OpenAI et al., 2024), which is prompted with the complete training set and chunks of 20 testing samples. System 8 is based on mBART (Liu et al., 2020). All other systems are based either on a standard encoder-decoder LSTM (Bahdanau et al., 2015) (systems 2, 3, 4, and 9) or pointer-generator LSTM (See et al., 2017) (systems 5, 6, and 7), and utilize different data augmentation methods.

The team develops a variation of the *lemma copying* technique (Liu and Hulden, 2022; Yang et al., 2022), which they name *sentence copying*. The idea is to create additional training samples by copying same sentence as both source and target with an empty change field. All LSTM systems except of system 9 use the external sentence copying for data augmentation, where the copied sentences are taken from external datasets, namely transcriptions from the Yucatec Maya DoReCo dataset (Skopeteas, 2022) for Maya, Guarani portion of the CC-100 dataset (Conneau et al., 2020) for Guarani, and Bribri portion of the Americas-NLP2024 Shared Task 1 data for Bribri. System 3 additionally performs the sentence copying with all sentences from the training data.

Another data augmentation method is called *stem permutation* and it is based on the idea to replace stems with random characters (Silfverberg et al., 2017; Anastasopoulos and Neubig, 2019). Instead of identifying which parts of words are stems, the team randomly changes one or two characters in a source sentence and relies on the edit tree built from the original source and target sentences to see if the change is valid. If the edit tree still applies to the modified source sentence, then this sentence is added to the training data with the original target sentence. The stem permutation method is used to augment data for systems 4 and 7.

Training data for system 7 also uses concatenation data augmentation, which finds pairs of training samples with exactly same grammatical change and creates a new sample by concatenating source sentences and target sentences from such pair.

Training data for system 6 is augmented with *transitive transformations* method. This method utilizes pairs of training samples sharing same source sentence while having grammatical changes with same attributes, but different values. One of the target sentences from such pair can be used as a source sentence to another target sentence in a new training sample, because it can be inferred that these target sentences share all grammatical and lexical content except of the attributes specified in the change.

System 9 works with byte-pair encoding sub-

Team	System	Bribri			Maya			Guarani			Average Acc.
		Acc.	BLEU	chrF	Acc.	BLEU	chrF	Acc.	BLEU	chrF	(Rank)
JAJ (/dʒæz/)	1	**54.17**	**71.72**	**82.78**	53.55	78.41	91.53	0.00	0.00	0.00	35.91 (1)
Meenzer Team	4	19.38	46.93	73.02	**53.87**	77.68	90.94	23.90	36.94	79.48	32.38 (2)
	1	17.50	44.20	70.09	38.39	66.81	83.70	**34.62**	**49.60**	**84.93**	30.17 (5)
	2	17.50	44.20	70.09	38.39	66.81	83.70	23.08	35.95	79.71	26.32 (7)
	3	8.54	32.50	61.24	27.74	58.59	79.29	12.64	20.01	71.61	16.31 (11)
Giving it a Shot	3	17.71	39.48	69.28	**53.87**	**78.54**	**91.66**	25.00	40.55	81.71	32.19 (3)
	2	11.67	33.80	65.51	50.97	75.09	89.76	18.13	31.94	79.36	26.92 (6)
	1	7.08	31.68	62.45	49.03	73.09	88.54	9.34	22.64	73.40	21.82 (8)
LECS Lab	1	12.08	36.95	66.75	51.61	76.82	90.29	30.77	45.18	82.33	31.49 (4)
	7	2.50	14.65	41.51	30.00	65.22	83.28	12.09	22.73	72.11	14.86 (13)
	8	0.83	9.90	36.47	35.16	68.11	86.04	3.30	13.84	61.46	13.10 (14)
	5	0.21	3.34	21.78	24.19	56.05	77.64	7.69	20.53	71.26	10.70 (15)
	3	2.29	10.87	37.35	15.16	50.77	74.38	9.34	13.08	66.93	8.93 (17)
	6	0.21	2.01	18.80	12.90	43.31	69.27	11.81	17.62	68.88	8.31 (18)
	2	1.67	11.49	41.00	15.48	55.22	76.58	7.69	17.80	70.54	8.28 (19)
	4	2.29	11.88	42.76	13.55	52.83	75.94	8.24	15.59	66.90	8.03 (21)
	9	0.83	7.91	47.76	0.00	0.00	0.00	0.55	3.80	56.21	0.46 (22)
UF_NLP	2	26.88	48.71	74.83	33.23	74.36	86.59	0.00	0.00	0.00	20.04 (9)
	1	9.79	37.92	65.33	37.42	69.59	85.77	0.00	0.00	0.00	15.74 (12)
Arizona Linguistics	1	9.38	17.13	55.07	25.81	50.36	79.46	14.84	22.55	73.18	16.67 (10)
Anonymous	1	12.50	31.51	57.20	16.45	54.20	77.87	0.00	0.00	0.00	9.65 (16)
submission	2	9.79	29.91	56.99	14.52	51.28	76.06	0.00	0.00	0.00	8.10 (20)
Baseline (edit trees)		8.75	22.11	52.73	25.81	53.69	80.23	14.84	25.03	76.10	
Max		54.17	71.72	82.78	53.87	78.54	91.66	34.62	49.60	84.93	

Table 2: Results over the test set. The last column shows the average accuracy over the three languages and the rank of each submission. Teams are ordered according to their best performing submissions.

words, unlike the other LSTM models, which work with characters. System 9 uses only one data augmentation that aims to replace frequent non-inflection subwords with their synonyms in both source and target sentences. The synonyms are identified using separate word2vec models, which are trained on external data for Guarani and Bribri languages.

5.6 UF_NLP

The UF_NLP team (Su et al., 2024) submitted predictions of two systems for Bribri and Maya languages. System 1 is NLLB-200-3.3B model (NLLB Team et al., 2022) finetuned separately for each language. Its input is concatenation of a source sentence and a grammatical change tag. System 2 is Claude 3 Opus LLM. Its prompt contains all training samples with sample IDs replaced with row numbers.

5.7 Arizona Linguistics

The Arizona Linguistics team (Hammond, 2024) submitted predictions from one system for all three languages. This system adopts the baseline and relaxes the requirement of strict match of grammatical change for selection of candidate edit trees. More precisely, if none of originally selected edit trees could be applied to the test source sentence, then the system considers the full list of edit trees from the training data and attempts to apply them in the order of similarity of their grammatical changes to the testing grammatical change.

5.8 Task Results

Table 2 shows the results of the different systems for our task. The JAJ team got the first position in the task according to average accuracy, although none of the teams was a clear winner for the three languages: the JAJ team obtained the best performance for Bribri, the Giving it a Shot team for Maya, and the Meenzer Team for Guarani. The JAJ team obtained on average the best accuracy results, even considering they did not submit their results for the Guarani language. This accuracy metric was very strict, and we can see that it was the metric for which the participants got the lowest results.

The results in terms of chrF were very high, but this was expected as the target sentences in general share many words and morphemes with the source sentence, so the character n-gram overlap between them should already be very high. The language that got on average the worst results was Guarani, having only 34% accuracy and 49.6 BLEU score.

It was also the language that was tackled by fewer teams: only four out of seven. One possible explanation for the lower results is the division of clusters with different verbs in the different splits, or the fact that a different (more difficult) combination of sources was used for the test set.

6 Conclusions

We presented the results of the first task on the creation of educational materials for Indigenous languages of the Americas. In this task, the participants had to create systems that could transform a source sentence into a target sentence by changing some linguistic feature, usually associated to the main verb (e.g., negation, aspect or tense). These pairs of sentences can be used to create grammar exercises for students of the Indigenous languages.

The languages targeted in this task were Bribri, Maya and Guarani, three Indigenous languages of the Americas with different characteristics. Seven teams took part in the task, submitting 22 systems. Different teams obtained the best results for each language: JAJ for Bribri, Giving it a Shot for Maya, and Meenzer Team for Guarani. The results in general were very promising, obtaining high scores in terms of the generative task metrics BLEU and chrF, but still with a lot of room for improvement in terms of the main accuracy metric.

Notably, most of the teams used neural methods, in particular LLMs like GPT-4 or Mixtral, often with some strategies for data augmentation. This is interesting because such models have often shown worse performance on lower-resource languages than those with higher resources, but in this case where the systems did not need to generate a full sentence but make some localized changes, they seem to work quite well.

Acknowledgements

We would like to thank all teams for their participation in this shared task. The Maya data provided is a collaborative work between CentroGeo and SEDECULTA, as cited in Agreement SEDECULTA-DASJ-149-04-2024.

References

Academia de la Lengua Guaraní. 2018. *Gramática guaraní*. Editorial: Servilibro.

Antonios Anastasopoulos and Graham Neubig. 2019. Pushing the limits of low-resource morphological inflection. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 984–996, Hong Kong, China. Association for Computational Linguistics.

Øistein E. Andersen, Helen Yannakoudakis, Fiona Barker, and Tim Parish. 2013. Developing and testing a self-assessment and tutoring system. In *Proceedings of the Eighth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 32–41, Atlanta, Georgia. Association for Computational Linguistics.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *Proc. of ICLR*.

Fidencio Briceño Chel. 2021. *Los verbos del maya yucateco actual. Clasificación y conjugación*. Instituto Nacional de Lenguas Indígenas (INALI). Second Edition.

Minh Duc Bui and Katharina von der Wense. 2024. Jgu mainz's submission to 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.

Miguel Óscar Chan Dzul. 2010. Los marcadores de aspecto y modo en el maaya t'aan. Bachelor's thesis in linguistics. Universidad de Oriente de Yucatán.

Luis Chiruzzo, Santiago Góngora, Aldo Alvarez, Gustavo Giménez-Lugo, Marvin Agüero-Torales, and Yliana Rodríguez. 2022. Jojajovai: A parallel guarani-spanish corpus for mt benchmarking. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2098–2107.

Grzegorz Chrupała. 2008. *Towards a machine-learning architecture for lexical functional grammar parsing*. Ph.D. thesis, Dublin City University.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Adolfo Constenla, Feliciano Elizondo, and Francisco Pereira. 2004. *Curso Básico de Bribri*. Editorial de la Universidad de Costa Rica.

Rolando Coto-Solano, Sharid Loáiciga, and Sofía Flores-Solórzano. 2021. Towards Universal Dependencies for Bribri. In *Proceedings of the Fifth Workshop on Universal Dependencies (UDW, SyntaxFest 2021)*, pages 16–29.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Arya D. McCarthy, Katharina Kann, Sabrina J. Mielke, Garrett Nicolai, Miikka Silfverberg, David Yarowsky, Jason Eisner, and Mans Hulden. 2018. The CoNLL–SIGMORPHON 2018 shared task: Universal morphological reinflection. In *Proceedings of the CoNLL–SIGMORPHON 2018 Shared Task: Universal Morphological Reinflection*, pages 1–27, Brussels. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared task—morphological reinflection. In *Proceedings of the 2016 Meeting of SIGMORPHON*, Berlin, Germany. Association for Computational Linguistics.

Gregory Dyke, Iris Howley, David Adamson, Rohit Kumar, and Carolyn Penstein Rosé. 2013. Towards academically productive talk supported by conversational agents. *Productive multivocality in the analysis of group interactions*, pages 459–476.

Abteen Ebrahimi, Manuel Mager, Adam Wiemerslage, Pavel Denisov, Arturo Oncevay, Danni Liu, Sai Koneru, Enes Yavuz Ugan, Zhaolin Li, Jan Niehues, et al. 2022. Findings of the second americasnlp competition on speech-to-text translation. In *NeurIPS 2022 Competition Track*, pages 217–232. PMLR.

Sofía Flores-Solórzano. 2017a. Corpus oral pandialectal de la lengua bribri. `http://bribri.net`.

Sofía Flores-Solórzano. 2017b. *Un primer corpus pandialectal oral de la lengua bribri y su anotación morfológica con base en el modelo de estados finitos*. Ph.D. thesis, Universidad Autónoma de Madrid.

Sofía Margarita Flores-Solórzano. 2017c. *Un primer corpus pandialectal oral de la lengua bribri y su anotación morfológica con base en el modelo de estados finitos*. Ph.D. thesis, Universidad Autónoma de Madrid.

Alí García Segura. 2016. *Ditsö rukuö - Identity of the seeds: Learning from Nature*. IUCN.

Alí García Segura. 2021. *Se' dör stè - We are art: The Teaching of Awá*. International Tree Fund.

Michael Ginn, Ali Marashian, Bhargav Shandilya, Claire Benet Post, Enora Rice, Juan Vásquez, Marie C. McGregor, Matthew J. Buchholz, Mans Hulden, and Alexis Palmer. 2024. On the robustness of neural models for full sentence transformation. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*. Association for Computational Linguistics.

Dan Gusfield. 1997. *Algorithms on Strings, Trees, and Sequences: Computer Science and Computational Biology*. Cambridge University Press.

Coleman Haley. 2024. The unreasonable effectiveness of large language models for low-resource clause-level morphology: In-context generalization or prior exposure? In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*. Association for Computational Linguistics.

Michael Hammond. 2024. The role of morphosyntactic similarity in generating related sentences. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*. Association for Computational Linguistics.

Jennifer Hill and Rahul Simha. 2016. Automatic generation of context-based fill-in-the-blank exercises using co-occurrence likelihoods and Google n-grams. In *Proceedings of the 11th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 23–30, San Diego, CA. Association for Computational Linguistics.

INEC. 2011. X Censo Nacional de Población y VI de Vivienda 2011 - Territorios Indígenas - Principales Indicadores Demográficos y Socioeconómicos.

Carla Victoria Jara. 1993. *I Ttè - Historias Bribris*. Editorial de la Universidad de Costa Rica.

Carla Victoria Jara. 2018. *Gramática de la lengua bribri*. E-Digital ED.

Carla Victoria Jara and Alí García Segura. 1997. *Kó Késka - El Lugar del Tiempo*. Editorial de la Universidad de Costa Rica.

Carla Jara Murillo and Alí García Segura. 2022. Sẽbliwak Francisco García ttö̀. `https://www.lenguabribri.com/las-palabras-de-francisco`.

Carla Victoria Jara Murillo and Alí García Segura. 2013. *Se' ttő bribri ie Hablemos en bribri*. E Digital.

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. 2024. Mixtral of Experts.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Katharina Kann and Hinrich Schütze. 2016. Single-model encoder-decoder with explicit morphological representation for reinflection. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 555–560, Berlin, Germany. Association for Computational Linguistics.

Haakon S. Krohn. 2021. Diccionario digital bilingüe bribri. http://www.haakonkrohn.com/bribri.

Chengzu Li, Han Zhou, Goran Glavaš, Anna Korhonen, and Ivan Vulić. 2023. On Task Performance and Model Calibration with Supervised and Self-Ensembled In-Context Learning.

Ling Liu and Mans Hulden. 2022. Can a transformer pass the wug test? tuning copying bias in neural morphological inflection models. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 739–749, Dublin, Ireland. Association for Computational Linguistics.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Agustín Lucas, Alexis Baladón, Victoria Pardiñas, Marvin Agüero-Torales, Santiago Góngora, and Luis Chiruzzo. 2024. Grammar-based data augmentation for low-resource languages: The case of guarani-spanish neural machine translation. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics.

Jakub Macina, Nico Daheim, Sankalan Chowdhury, Tanmay Sinha, Manu Kapur, Iryna Gurevych, and Mrinmaya Sachan. 2023. MathDial: A dialogue tutoring dataset with rich pedagogical properties grounded in math reasoning problems. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5602–5621, Singapore. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios Gonzales, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, et al. 2021. Findings of the americasnlp 2021 shared task on open machine translation for indigenous languages of the americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217.

Enrique Margery. 2005. *Diccionario fraseológico bribri-español español-bribri*, second edition. Editorial de la Universidad de Costa Rica.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. Unesco.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, et al. 2024. GPT-4 Technical Report.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Naiara Perez and Montse Cuadros. 2017. Multilingual CALL framework for automatic language exercise generation from free text. In *Proceedings of the Software Demonstrations of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, pages 49–52, Valencia, Spain. Association for Computational Linguistics.

Maja Popović. 2015. chrf: character n-gram f-score for automatic mt evaluation. In *Proceedings of the tenth workshop on statistical machine translation*, pages 392–395.

Juan Diego Quesada. 2007. *The Chibchan Languages*. Editorial Tecnológica de Costa Rica.

YV Rodríguez Gutiérrez and E Núñez Méndez. 2018. Language contact and the indigenous languages of Uruguay. *Biculturalism and Spanish in contact: sociolinguistic case studies*, pages 217–238.

Carlos Sánchez Avendaño. 2013. Lenguas en peligro en Costa Rica: vitalidad, documentación y descripción. *Revista Káñina*, 37(1):219–250.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational*

Linguistics (Volume 1: Long Papers), pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Miikka Silfverberg, Adam Wiemerslage, Ling Liu, and Lingshuang Jack Mao. 2017. Data augmentation for morphological reinflection. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 90–99, Vancouver. Association for Computational Linguistics.

Stavros Skopeteas. 2022. Yucatec Maya DoReCo dataset.

Jim Su, Justin Minh Ho, George Aaron Broadwell, Sarah Moeller, and Bonnie J. Dorr. 2024. A comparison of fine-tuning and in-context learning for morphological inflection. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*. Association for Computational Linguistics.

Abhijit Suresh, Jennifer Jacobs, Margaret Perkoff, James H. Martin, and Tamara Sumner. 2022. Fine-tuning transformers with additional context to classify discursive moves in mathematics classrooms. In *Proceedings of the 17th Workshop on Innovative Use of NLP for Building Educational Applications (BEA 2022)*, pages 71–81, Seattle, Washington. Association for Computational Linguistics.

Carlos Sánchez Avendaño, Alí García Segura, et al. 2021a. *Se' Dalì Diccionario y Enciclopedia de la Agricultura Tradicional Bribri*. Editorial de la Universidad de Costa Rica.

Carlos Sánchez Avendaño, Alí García Segura, et al. 2021b. *Se' Má Diccionario-Recetario de la Alimentación Tradicional Bribri*. Editorial de la Universidad de Costa Rica.

Jannis Vamvas. 2022. Translation Puzzles are In-context Learning Tasks.

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)*. Association for Computational Linguistics.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. *Advances in neural information processing systems*, 28.

Sebastian Wollny, Jan Schneider, Daniele Di Mitri, Joshua Weidlich, Marc Rittberger, and Hendrik Drachsler. 2021. Are we there yet?-a systematic literature review on chatbots in education. *Frontiers in artificial intelligence*, 4:654924.

Changbing Yang, Ruixin (Ray) Yang, Garrett Nicolai, and Miikka Silfverberg. 2022. Generalizing morphological inflection systems to unseen lemmas. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 226–235, Seattle, Washington. Association for Computational Linguistics.

Shigeto Yoshida. 2016. Un análisis morfosintáctico sobre el tiempo y el aspecto en la lengua maya yucateca. *Latin American and Caribbean Studies*, 23(1):39–51.