Findings of the AmericasNLP 2025 Shared Tasks on Machine Translation, Creation of Educational Material, and Translation Metrics for Indigenous Languages of the Americas

Ona de Gibert[~*] **Robert Pugh**[♣*] **Ali Marashian**[♯*] **Raúl Vázquez**[~] **Abteen Ebrahimi**[♯]
Pavel Denisov[♠] **Enora Rice**[♯] **Edward Gow-Smith**[γ] **Juan C. Prieto**[β] **Melissa Robles**[β]
Rubén Manrique[β] **Oscar Moreno Veliz**[⋈] **Ángel Lino Campos**[⋈] **Rolando Coto-Solano**[♡]
Aldo Alvarez[Ω] **Marvin Agüero-Torales**[△▽] **John E. Ortega**[α] **Luis Chiruzzo**[♢]
Arturo Oncevay[⋈] **Shruti Rijhwani**[℧] **Katharina von der Wense**[♯†] **Manuel Mager**[‡†]

[~]University of Helsinki [♣]Indiana University, Bloomington [♯]University of Colorado Boulder
[♠]Fraunhofer IAIS [γ]University of Sheffield [β]Universidad de Los Andes
[⋈]Pontificia Universidad Católica del Perú [♡]Dartmouth College [△]Universidad de Granada, Spain
[Ω]Universidad Nacional de Itapua, Paraguay [▽]Global CoE of Data Intelligence, Fujitsu
[♢]Universidad de la República, Uruguay [α]Northeastern University [℧]Google DeepMind
[†]Johannes Gutenberg University Mainz [‡]Amazon

Abstract

This paper presents the findings of the AmericasNLP 2025 Shared Tasks: (1) machine translation for truly low-resource languages, (2) morphological adaptation for generating educational examples, and (3) developing metrics for machine translation in Indigenous languages. The shared tasks cover 14 diverse Indigenous languages of the Americas. A total of 12 teams participated, submitting 27 systems across all tasks, languages, and models. We describe the shared tasks, introduce the datasets and evaluation metrics used, summarize the baselines and submitted systems, and report our findings.

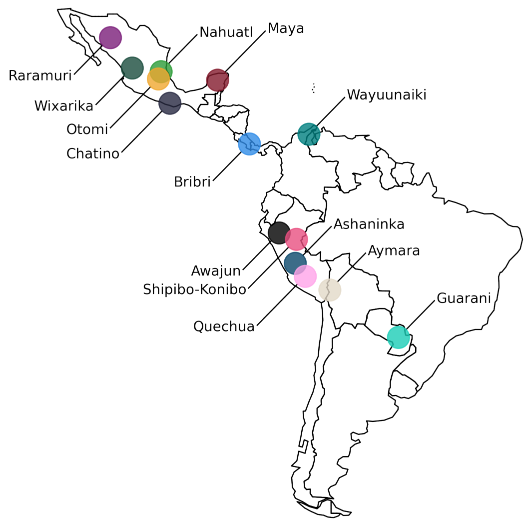

Figure 1: Map of Central and South America presenting an approximate distribution of where each Indigenous language covered by the three Shared Tasks is spoken.

1 Introduction

The recent rapid progress in Natural Language Processing (NLP), significantly accelerated by the improved architectures, training methods, and the rise of Large Language Models (LLMs), has primarily benefited *high-resource languages*, languages that have large amounts of digital text available such as English or French. In contrast, languages with low amounts of data, known as *low-resource languages*, still face considerable challenges in terms of both data availability and the development of appropriate models (e.g., Ignat et al., 2024). Low-resource languages that are native to a specific region, or *Indigenous languages*, remain challenging for even the most novel NLP techniques (Mager et al., 2024; Weerasinghe et al., 2025; Hettiarachchi et al., 2025).

To address these disparities, the Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP) was established with the goal of advancing NLP research for Indigenous languages from the American continent.

Building on the success of last year's Shared Tasks (ST) (Ebrahimi et al., 2024; Chiruzzo et al., 2024), the 2025 edition expands its scope with three STs designed to address critical challenges in working with Indigenous languages. Many of the languages included in the STs are polysynthetic, agglutinative or tonal languages, features which are not mutually exclusive. In addition, they often lack a standardized orthography, exhibit dialectal vari-

[*]In order, the main organizers for shared tasks 1, 2, and 3.
[†] Irrespective of Manuel Mager's listed affiliation, this work is independent of his employment at Amazon.

Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 134–152
May 4, 2025 ©2025 Association for Computational Linguistics

ation and frequent code-switching with dominant regional languages (Mager et al., 2019).

The goal of this effort is not only to advance methodologies for low-resource settings but also to support the development of tools for language learning, preservation, and revitalization. Moreover, we expect to develop technologies that can include the native speakers of these languages in the recent developments in our field. This year's STs include:

- **ST1: Machine Translation (MT) for low-resource languages**, translating between Spanish and 13 Indigenous languages with limited parallel data. This year, it features two new languages (Awajun and Wayuunaiki), and a new translation direction (into Spanish).

- **ST2: Morphological adaptation to generate educational examples** transforming sentences to create grammar exercises for language learners. This year, we include Nahuatl as an additional language.

- **ST3: Developing metrics for MT in Indigenous languages** designing evaluation metrics suited to the linguistic properties of low-resource languages. The first edition of its kind.

Across all tasks, languages, and models, a total of 12 teams participated, submitting 27 systems. The consistent interest from the community highlights the continued interest in developing NLP tools for Indigenous languages.

We publicly release the training and development data through our GitHub repository.[1]

2 Languages

The STs feature 14 Indigenous languages spoken across North, Central, and South America, listed in Table 1. These languages differ in language family, number of speakers, geographical distribution, and resource availability; reflecting their diversity. They vary in their levels of official recognition, and in many cases, speaker population data is based on outdated census information. Figure 1 shows the approximate geographical distribution of the languages included in the tasks. Below, we briefly introduce each of the languages.

[1]https://github.com/AmericasNLP/americasnlp2025/

Language	Family	ISO 639-3	Glottolog	ST
Asháninka	Arawak	cni	asha1243	1
Awajun	Chicham	agr	agua1253	1
Aymara	Aymaran	aym	nucl1667	1
Bribri	Chibchan	bzd	brib1243	1,2,3
Chatino	Oto-Manguean	ctp	chat1268	1
Guarani	Tupi-Guarani	grn	para1311	1,2,3
Maya	Mayan	yua	yuca1254	2
Nahuatl	Uto-Aztecan	nah	azte1234	1,2,3
Otomí	Oto-Manguean	oto	otom1300	1
Quechua	Quechuan	quy	ayac1238	1
Rarámuri	Uto-Aztecan	tar	tara1321	1
Shipibo-Konibo	Panoan	shp	ship1253	1
Wayuunaiki	Arawak	guc	wayuu1243	1
Wixarika	Uto-Aztecan	hch	huic1243	1

Table 1: Languages of the Shared Tasks, their language families, ISO 639-3 and Glottolog codes, and Shared Tasks were they are included.

Asháninka (aka *Campa*) is an Arawakan language spoken primarily in Peru and Brazil by approximately 74,500 speakers. It is agglutinative and polysynthetic and has a Verb-Subject-Object (VSO) word order.

Awajun (aka *Aguaruna*) is a Chicham language spoken in northern Peru, by around 53,400 speakers. It follows a Subject-Object-Verb (SOV) and has rich morphology that consists of agglutinative suffixes. We use the Marañón variant.

Aymara is an Aymaran language spoken in the Andean regions of Bolivia and Peru, with approximately 1.7 million speakers. It is recognized for its agglutinative morphology and polysynthetic nature, typically following a SOV word order. We use Central Aymara variant, spoken in Aymara La Paz.

Bribri is a Chibchan language spoken in southern Costa Rica, by an estimated 7,000 people. The language exhibits morphological ergativity and is tonal, with SOV word order. We use the Amburi variant.

Chatino refers to a group of indigenous Mesoamerican languages within the Zapotecan branch of the Oto-Manguean family, spoken in Oaxaca, Mexico. These languages are tonal and have complex systems of verbal inflection. We use the San Juan Quiahije variant, spoken by about 5,000 people.

Guarani is a Tupi–Guarani language spoken mainly in Paraguay, where it is one of the official languages, as well as in parts of Bolivia, Argentina, and Brazil. It has approximately 6.5 million speakers. It is an agglutinative language. We use the Paraguayan variant, except the training data for ST1, which consists of a mix of dialects.

Maya is a Mayan language spoken on the Yucatán Peninsula of Mexico, northern Belize, and parts of Guatemala, with approximately 800,000 speakers. It is characterized by its use of glottalized consonants and a Verb-Subject-Object (VSO) word order. We use the Yucatec Maya variant.

Nahuatl Nahuatl is a group of related Uto-Aztecan languages spoken throughout Mexico and in parts of Central America, with approximately 1.6 million speakers in total. There are over 30 variants of the language. It is polysynthetic and agglutinative.

For ST1, we use a diverse set of variants, including colonial-era written Nahuatl, for training (from the Axolotl corpus (Gutierrez-Vasques et al., 2016)) and Huasteca Nahuatl for ST1 evaluation as well as for ST3. ST2 focuses on Western Sierra Puebla Nahuatl, a relatively understudied Nahuatl variety.

Otomí (aka *Hñähñu*[2]) is an Oto-Manguean language spoken in central Mexico by about 300,000 people. It has nine variants. Otomí languages are tonal and exhibit a complex system of verb inflection, typically following SVO word order. We focus on the Ixtenco Otomí (OTX), a variant with less than 460 speakers, in the Mexican state of Tlaxcala.

Quechua is a family of languages spoken across the Andean regions of Argentina, Bolivia, Chile, Colombia, Ecuador, and Peru, with approximately 7.2 million speakers. It is recognized as an official language in Peru and Bolivia and is known for its agglutinative structure and SOV word order. We use the Quechua Ayacucho variant, although the training data also includes text in Quechua Cuzco.

Rarámuri (aka *Tarahumara*) is a Uto-Aztecan language spoken in northern Mexico, by around 70,000 speakers. It is polysynthetic and agglutinative. We use the highlands variant.

Shipibo-Konibo is a Panoan language spoken in Peru by approximately 26,000 people. It is characterized by its agglutinative morphology and predominantly SOV word order and uses postpositions.

[2]Other names for the language are used, depending on the language variant.

Wayuunaiki is an Arawakan language spoken in northern Colombia and Venezuela, primarily by the Wayuu community, with about 420,000 speakers. It is an agglutinative language with a predominant SOV word order.

Wixarika (aka *Huichol*) is a Uto-Aztecan language spoken in Mexico, by approximately 35,000 speakers. It is official in Mexico with four variants. It is an agglutinative morphology with strong polysynthetic characteristics and follows the SOV word order. We use the Nayarit version, spoken in Zoquipan.

3 ST1: A ST on Machine Translation on Truly Low-resource Languages

Description Low-resource MT (Haddow et al., 2022) is mainly characterized by the limited availability of parallel corpora, but it also faces additional challenges, such as the scarcity of monolingual data and issues related to data quality.

This task focuses on translation between Spanish and 13 indigenous languages. Now in its fourth iteration (Mager et al., 2021; Ebrahimi et al., 2023, 2024), it continues to push the boundaries of MT for these languages, emphasizing generalization strategies for low-resource MT and the creation of new linguistic resources to support these efforts.

For this year's edition, we introduce two new languages (Awajun, Wayuunaiki) for the ST1 task and expand the ST to cover both translation into an Indigenous language from Spanish (Track 1), as well as translation from an Indigenous language into Spanish (Track 2). These two translation directions are organized as separate tracks within the ST. Furthermore, following the spirit of open science, this year we only take into account submissions which rely solely on open-source weights for the final ranking.

Data Table 7 in the Appendix shows our data statistics. We use the same training data as in previous editions for the repeating languages. This consists of the organizers' collection of parallel sentences, and the data collected by Vázquez et al. (2021) and De Gibert et al. (2023), a combination of scraped sources, and synthetically generated data, obtained through back-translation.

For Wayuunaiki, the train dataset was derived from the work of Prieto et al. (2024), with a thorough curation and selection of the data. It was compiled from grammar books, the Bible, short stories, a dictionary and the Colombian constitution, with a total of 59,715 sentences. To process this data, different extraction techniques were applied based on the structure of each source. Web scraping was used for highly structured texts like the Bible, ensuring precise verse alignment. For more complex sources, such as grammar books and linguistic studies, GPT-4 was used to identify sections of the text containing translated sentences, extracting and tabulating them into a standardized format. In cases were texts were available only as scanned documents or unstructured PDFs, OCR combined with GPT-4 processing enabled the retrieval of bilingual content. Finally, a manual review process was conducted across all sources to filter incomplete translations and correct formatting inconsistencies.

For Awajun, the main part of the training data was extracted from various web sources such as poems, stories, laws, protocols, guidelines, handbooks, the Bible, and news published by Ojo Público,[3] a news media organization that supported the first iteration of the dataset (Moreno et al., 2024). An official translator validated all sources for the corpora to ensure the same dialect is used. Only a few of the sources were aligned automatically, using line breaks and sentence length heuristics as reference, while most of the sources were aligned manually to retain the quality of the translations.

For development and evaluation, we use the AmericasNLP 2021 data (Mager et al., 2021), a multi-way parallel dataset of the XNLI (Conneau et al., 2018) test set into 10 languages of the Americas (Asháninka, Aymara, Bribri, Guarani, Nahuatl, Otomí, Quechua, Rarámuri, Shipibo-Konibo, and Wixarika). The Chatino data comes from Mexican court proceedings. For an in-depth review of the development and evaluation data, please refer to Ebrahimi et al. (2022, 2024) and Mager et al. (2021).

For the new languages, the Wayuunaiki development set is sourced from the work of Prieto et al. (2024), while the test set is created by translating the first 95 pages of the book *Journey to the Center of the Earth* by Verne (1874), with an average of 150 words per page. To uphold high ethical standards, we ensured that translators received fair compensation. The test set also includes the translation of the short story *Benjamin Bunny* by Potter (1904). In the case of Awajun, the development set was split from the available training data. We compile a small test set that contains translations provided by a professional translator in texts extracted from news within the Territorio Amazonas domain, and another portion of the test set are examples extracted from a dictionary by Espejo Apikai et al. (2021) not processed for the train or development set.

Metrics We use ChrF++ (Popović, 2017) as the main metric of the task, although we also report BLEU (Papineni et al., 2002).

ChrF++ is an overlap-based metric at the character-level, which is more suitable than BLEU for our task since most languages are morphologically rich, and BLEU often penalizes morphological variants (Chauhan et al., 2023). The final score for each submission (ChrF++ column in Table 8) is calculated by taking an average over all thirteen languages; if there is no model output for a given language, the score is taken as 0.

Baselines For our baseline, we follow the training set-up of "Submission 3" to the 2023 edition of the ST by Gow-Smith and Sánchez Villegas (2023). We extend the embedding matrix of NLLB-200-distilled-1.3B[4] with language tags for the languages not already covered, and finetune on the task data as well as additional training sources. We finetune two separate models for Track 1 and 2. See the original paper for further training details, our only modification for this year is the addition of the two new languages. We choose the best checkpoint based on the highest average ChrF++ across all languages.

Aiming to assess the current performance of LLMs on the task languages, we also implemented a fine-tuned a LLaMA3.2 model (Dubey et al., 2024) [5] using Low-Rank Adaptation (LoRA) adapters (Hu et al., 2022). This baseline performed poorly, only managing to copy the source sentence; however, we do not rule out the possibility of bugs in our implementation.

Submitted Systems For this year's ST1 we recieved a total of 5 submissions by 3 different teams. Below, we briefly describe each team's participation:

- **George Mason University (GMU)** (Hus et al., 2025): this team submits two systems

[3] https://ojo-publico.com/

[4] facebook/nllb-200-distilled-1.3B

[5] meta-llama/Llama-3.2-3B-Instruct

TEAM	AGR	AYM	BZD	CNI	CTP	GRN	GUC	HCH	NAH	OTO	QUY	SHP	TAR
						TRACK 1: SPA-XXX							
Baseline	**36.76**	**31.21**	**25.52**	**24.39**	**36.53**	**35.68**	**24.18**	**28.26**	**22.42**	**12.78**	31.88	**25.76**	**15.96**
GMU	35.09	22.91	22.51	22.22	13.33	29.95	22.93	26.14	20.33	11.31	**32.70**	19.46	13.89
Syntax Squad	35.16	27.72	22.77	23.17	-	16.21	12.83	26.77	12.64	12.02	31.01	12.76	-
UCSP	-	-	-	-	-	-	-	-	-	-	16.75	-	-
						TRACK 2: XXX-SPA							
Baseline	**38.39**	**35.60**	**30.14**	**24.86**	**35.84**	**35.91**	**24.74**	**26.33**	**26.36**	**20.81**	**37.18**	**47.81**	18.75
GMU	36.59	26.09	27.86	22.44	26.16	33.84	23.93	24.37	25.58	18.24	33.02	38.01	**19.72**
Syntax Squad	33.70	25.78	26.22	20.13	-	24.70	14.40	22.02	13.88	17.80	31.71	30.83	-
UCSP	-	-	-	-	-	-	-	-	-	-	17.87	-	-

Table 2: The best CHRF++ scores for ST1 for each team (across all submitted systems) across all languages. Bold values represent the best performing system overall, while underlined values are the best performing submission to this year's shared task.

for all language pairs in both tracks. First, they finetune NLLB-200-3.3B with the provided data for each language pair separately. Then, they prompt GPT-4o-mini model with external knowledge coming from bilingual dictionaries (a translation word is provided for each word of the sentence), two sample parallel sentences (few-shot approach), a full grammar book on the Indigenous language and a suggested translation, which is the generated hypothesis of the first NLLB-based system. Since GPT4-0-mini is a closed-source model, we only use their NLLB-based approach for the ranking. GMU is the only team to submit entries for all language pairs.

- **Syntax Squad** (Yahan and Amanul Islam, 2025): this team submits one system for 11 language pairs in both tracks and one extra system for translation from Spanish into Aymara. They perform data normalization and then finetune NLLB-200-600M, LLaMA 3.1 8B Instruct, XGLM 1.7B (Lin et al., 2021). They submitt their NLLB-based model, which outperforms the other two in the develoment set.

- **Universidad Católica San Pablo (UCSP)** (Congora et al., 2025): this team participates in the task for Quecha translation from/into Spanish. They dedicate efforts to data collection and data cleaning. Furthermore, they expand their datasets by generating synthetic sentences via the replacement of subjects and verbs in the sentences. They use two methods: Wordnet, which is deemed unsatisfactory, and an LLM (Phi3-mini for English and Phi3.5 for Spanish). Then, they train two different architectures on the augmented dataset: transformer-base (Vaswani et al., 2017) and mT5-small (Xue et al., 2021).

Results The best performance per language for each team is shown in Table 2. In the Appendix, Table 8 provides the official ranking of the ST, which excludes closed-source models, and Table 9 reports the complete results for all submissions and teams. The baseline is hard to beat in both tracks. In both tracks, GMU is the only team to beat it for any language. The strong performance of the baseline indicates the importance of multilingual training, as NLLB is finetuned across all language pairs simultaneously, unlike GMU's NLLB-based submission, which is finetuned on each language individually.

In Track 1 (SPA→XXX), GMU's NLLB-based submission achieves the highest average performance, with a ChrF++ score of 21.95, closely followed by Syntax Squad (17.93) and GMU's GPT-based system (18.81). GMU surpasses the baseline only for Quechua, achieving a +0.82 gain in ChrF++. While Syntax Squad performs well overall, its results are notably weaker for Guarani, Wayuunaiki, Nahuatl, and Shipibo-Konibo.

In Track 2, the best-performing model is also GMU's NLLB-based submission, with an average ChrF++ score of 26.62, slightly ahead of their own GPT-based system (26.41), which performs significantly worse for Chatino. They surpass the baseline for Rarámuri, achieving a +0.97 gain in ChrF++. Overall, GPT-based models appear effective at post-grammar correction for Spanish, but show weaker performance for the Indigenous language targets.

Submissions for Quechua from UCSP underper-

Language	Num. Sentences	Textual features			Grammatical changes	
	(train-dev-test)	Words/Sent	Chars/Word	TTR	Changes/Sent	Num Changes
Nahuatl	391-176-120	3.05	7.69 (20)	0.06	3.5	47
Maya	584-149-310	5.48	4.66 (14)	0.03	1.1	34
Bribri	309-212-480	3.75	3.39 (8)	0.02	2.8	28
Guarani	178-79-364	3.92	6.17 (14)	0.07	1.0	19

Table 3: A comparison of descriptive statistics of the corpora for ST2, calculated on the combination of the train and dev sets. Included features about the text are the average sentence length, average word length, the length of the longest word (in parentheses after the average word length), and the type-token ration for the corpus. With respect to the "Grammatical features", we report the average number of requested grammatical changes per sentence, as well as the total number of unique grammatical changes (i.e. feature-value pairs) in the entire corpus.

form when compared to other submissions, suggesting that training models from scratch has stopped being the most effective approach in low-resource settings.

Findings MT where the target is an Indigenous language appears to have reached a performance plateau. Improvements in the AmericasNLP workshop seem to be difficult given current data limitations. While this may not be the case in general, the most effective strategy in the AmericasNLP workshop remains to be the finetuning of a highly multilingual pretrained model (such as NLLB). In contrast, for translations where the target langauge is a high-resource language like Spanish, LLMs can provide a boost in performance. This is likely due to their extensive pretraining and a stronger representation of the higher-resource target language. However, whether the performance gains justify the practical costs of running these models remains an open question.

4 ST2: A ST on Morphological Adaptation to Generate Educational Examples

Description Language education initiatives, which are critical to many language revitalization efforts, require educational materials that are costly and time-consuming to create.

This task focuses on generating grammar exercises for learners of four Indigenous languages. In its first edition (Chiruzzo et al., 2024), the task involved automatically transforming a given base sentence by modifying its tense, aspect, or other morphosyntactic features into a target sentence. These sentences can later be used to create educational materials for language learners. This year's edition features the addition of an endangered variety of Nahuatl.

Data Four languages are included in this year's task: Bribri, Guarani, and Maya, which were all included in last year's task, and a new addition, Nahuatl. Since the data for the first three languages is the same as in last year's task, we refer the reader to Chiruzzo et al. (2024) for details.

Mexico's *Instituto Nacional de Lenguas Indígenas* (INALI) recognizes 30 Nahuatl varieties (INALI, 2012). The variant included in ST2 is commonly referred to as Western Sierra Puebla Nahuatl or Zacatlán-Ahuacatlán-Tepetzintla Nahuatl (*Náhuatl de la Sierra Oeste de Puebla*, ISO-639-3: `nhi`), spoken in the northwestern sierra region of the state of Puebla, Mexico by less than 20,000 people. This Nahuatl variety is relatively understudied, with most linguistic work, such as a short unpublished grammar and some examination of morphological and phonological phenomena, focusing on the subvariety spoken in the community of San Miguel Tenango, Zacatlán (Schroeder and Tuggy, 2010; Schroeder, 2014, 2015) or the municipality of Ahuacatlán (Sasaki, 2014).

The sentences used (see. Table 3) for the ST come from the community of Omitlán, Tepetzintla, where the specific Nahuatl communalect has been less studied, though it has been included in some recent computational work for the variety, such as a morphological analyzer (Pugh and Tyers, 2021b) and a Universal Dependencies treebank (Pugh et al., 2022). The base sentences are a part of a currently-unreleased corpus of grammatical example sentences, and the transformed sentences were verified by a native-speaking expert from the community.

The set of features used to annotate the Nahuatl data were:

- **Person and number**: Person/number of the subject, object, and indirect object of the Verb, and the possessor of the Noun in the sentence.

System Name	Bribri	Maya	Guarani	Nahuatl	Avg	Rank
NAIST	**41.25**	42.90	32.69	**17.50**	33.59	1◊
JHU_1	22.71	**63.87**	**43.68**	3.33	33.40	2◊
JHU_4	18.75	60.00	40.93	1.67	30.34	3
JHU_2	20.21	59.35	38.19	3.33	30.27	4
JHU_5	15.83	59.03	41.21	2.50	29.64	5
JHU_3	20.21	56.77	38.74	1.67	29.35	6
Syntax Squad	0.42	13.55	1.92	0.00	3.97	8
JHU_6	5.42	9.68	6.32	0.00	5.35	7
FPUNApy	0.00	0.00	8.52	0.00	2.13	9
IUNLP	0.00	2.26	3.85	0.00	1.53	10
RaaVa	1.25	0.00	2.20	0.00	0.86	11
Vasselli et al. (2024)	54.17	53.55	36.81	-	-	-
Baseline	5.66	26.17	22.78	0.00	13.65	-

Table 4: Final Accuracy results table for ST2. Note that while 6 teams submitted results on the test set, only 2 teams submitted system description papers, therefore we only describe the systems for two of the teams (NAIST and JHU). We also report the results from the previous year's winning system and the edit-tree baseline. The overall accuracy difference between ranks 1 and 2 is not significant (see ◊).

Person and number are represented together: 1_SG, 1_PL, 2_SG, 2_PL, 3_SG, 3_PL.

- **Tense**: Past, Present or Future (PRE_SIM, PAS_SIM, FUT_SIM, respectively).
- **Aspect**: Perfective (PERFV) and Imperfective (IMPFV) aspects occur with the past tense, and the Durative (DUR) aspect can occur with Past, Present, or Future tenses.
- **Mood**: Optative (OPT), Imperative (IMP, Conditional (COND), Interrogative (INT), or Indicative (NA).
- **Transitivity**: Nahuatl uses indefinite object prefixes to reduce the valency of a verb (e.g. *nechinnextiliah* "They show them to me" vs. *tetlanextiliah* "They show things to people"). When the valency is reduced by one of these morphemes, the transformation contains the tag TRANSITIV:ITR.
- **Purposive**: Nahuatl verbs can take a Purposive suffix indicating directionality of motion, e.g. "Go and do VERB". This directionality can be either away from (VET) or toward (VEN) the speaker.
- **Honorific**: Nahuatl varieties have as many as four levels of honorifics (Hill and Hill, 1978), though we only include the first in our dataset since it is the most common.
- **Polarity**: Positive or negative.

Metrics The main metric of this task is accuracy (fraction of times the system output matches the expected output). Systems for every language are evaluated separately, in addition to the overall average score, which is used to determine the shared task's winner.

Baselines This year, the baseline was the same as last year's, namely a simplified adaptation of the Prefer Observed Edit Trees (POET) method, which involves learning the edit operations required to convert a source string into a target string (Kann and Schütze, 2016). Learning is performed by calculating the edit tree for each pair of source and target sentences in the training data, and counting the total number of each edit tree associated with the specific grammatical change. During testing, the edit trees for the given grammatical change are applied to the given source sentence in order of decreasing frequency until the succeeding edit tree is found. If no such tree is found, the source sentence is returned as the output.

Submitted Systems We received 11 submissions from 6 teams for the task, but unfortunately only three teams submitted system description papers. Given the lack of description papers from the other 3 teams, we are unable to discuss their submissions.

- **NAIST**: The NAIST submission (Vasselli et al., 2025) developed three different systems: example-based LLM prompting system with additional synthetic data, a transformation-based prompting system where each token is annotated according to its required operation to achieve the sentence-level transformation, and, for Nahuatl, a purely rule-based system which heuristically assigns part-of-speech tags and uses them to infer grammatical features.

- **JHU**: There were a total of six JHU submissions (Lupicki et al., 2025). The submitted systems include multiple variations of prompt-engineering with LLMs, including experimenting with chain-of-thought, few-shot prompting, using additional linguistic data such as parts of speech and a reference book (for Maya, Bribri, and Guarani), and ensembling multiple LLM-based systems. Additionally, they train a pointer generator LSTM model.

- **Syntax Squad**: This team investigated LoRA fine-tuning of LLMs, namely Llama models and XGLM, for the sentence transformation task. The process also involved some text preprocessing, such as removing punctuation and diacritics, and post-processing of the LLM output. They did not describe results for the Nahuatl data.

Results The results of all submissions are listed in Table 4. Two of the three submitted system descriptions correspond to the two highest-performing submissions. The JHU team achieved the best performance for Maya and Guarani with their ensemble method, surpassing the last year's best-performing system on the same data. NAIST achieved the best score for both Bribri (41.25% acc.) and Nahuatl (17.5%), though their system did not outperform last year's winning system for Bribri, a fact the authors attribute to their application of transformations all at once, instead of incrementally as was done in last year's winning system. On the other hand, JHU system 1 had the best performance for Maya (63.87% acc.) and Guarani (43.67% acc.). The overall difference between NAIST and JHU System 1 is not significant[6] we decided for having both teams as winners of this years edition. It is also important to notice the poor performance of most teams on Nahuatl, with 5 submitted systems achieving 0% accuracy, and all, except for NAIST, achieveing less than 4% acc.

[6] Average sample-wise accuracy values with 95% confidence intervals, calculated with the bootstraping approach (Ferrer and Riera), are 36.97 [34.46, 39.56] for the NAIST system, and 36.89 [34.30, 39.48] for the JHU_1 system

The Syntax Squad submission underperformed the baseline for all languages. While it warrants further investigation, it is likely that the dataset sizes were too small to effectively fine-tune the LLMs for this task. Furthermore, they highlight the potential negative impact of excessive preprocessing of the text. For example, for languages like Maya where changes in tone can indicate a change in Voice (one of the features in the Maya dataset), removing this may introduce unwanted noise and make it more challenging for a model to learn the necessary sentence transformations.

Findings For the three languages represented in last year's shared task, we saw year-over-year improvements in the best-performing system for two (Maya and Guarani). None of the submitted systems improved on last year's best performing system on the Bribri data.

Interestingly, Nahuatl proved to be quite challenging, with all teams achieving their lowest score on the Nahuatl data. The best performance on this data was achieved with the purely rules-based system. We suspect that this is due to a combination of lack of representation of the Western Sierra Puebla variety in LLM training data, and a number of language- and dataset-specific features, e.g. longer words, many grammatical transformations per sentence, the largest number of unique grammatical transformations compared to the other languages in the shared task (see Table 3 for details).

While the trend of leveraging pretrained LLMs via prompt engineering and reference data continues to show promise for some languages, the results on the Nahuatl data show that knowledge-based approaches still merit attention, particularly when dealing with complex tasks and data (multiple interacting grammatical transformations, complex morphology with long words) and/or languages with minimal resources (both with respect to LLM training data as well as reference materials and digital dictionaries).

5 ST3: A ST on Creating Metrics for Machine Translation in Indigenous Languages

Description Automatic metrics are a crucial alternative to human evaluation for efficiently evaluating the output of MT systems. However, indigenous languages present unique challenges that standard metrics are not designed to handle. MT evaluation commonly relies on two types of automatic

Language	Num. Sentences	Textual features		
	(dev-test)	Words/Sent	Chars/Word	TTR
Nahuatl	100-200	6.68-6.78	8.27-7.83	0.27-0.23
Bribri	100-200	12.23-11.23	4.78-4.7	0.16-0.14
Guarani	100-200	6.24-6.36	7.94-7.43	0.28-0.24

Table 5: Data statistics for ST3. The textual statistics are for the reference translations, for dev and test sets. We report the average sentence length, average word length, and the type-token ration for the corpus. Overall, 300 sentence pairs were annotated for each language.

metrics: overlap-based and neural. Overlap-based metrics, such as BLEU and ChrF , are less effective for Indigenous languages as these languages often lack standardized orthographies and exhibit polysynthetic structures, making exact word or (to a lesser degree) character overlap unreliable. The limitations of BLEU are well documented (Mathur et al., 2020), and the overreliance of the MT community can potentially negatively affect MT development (Kocmi et al., 2021). Neural metrics, such as COMET (Rei et al., 2020), are also limited because they rely on pretrained models trained on large datasets that rarely include low-resource languages. In the first edition of its kind, this task consists in building metrics to evaluate the quality of translations from Spanish into three Indigenous languages: Guarani, Bribri, and Nahuatl.

Data For each language, a set of 100 sentence pairs are selected from the submissions to AmericasNLP 2024 MT ST, from multiple systems. Although the initial pool of sentences are selected randomly, it is important to select pairs of varying quality to ensure that the metrics can effectively distinguish these differences in quality. We use ChrF++ as a proxy of the quality of submissions, and for a portion of sentences we also include the gold translations [7]. The same set of Spanish sentences were used for all the languages. For the test data, we repeated this process. These sentences were then given to annotators for the human judgment. The annotators are asked to rate each translation on a 5-point scale on two axes: semantics and fluency (Koehn and Monz, 2006). As bilingual speakers, the annotators have access to the source sentence in Spanish, and a candidate translation in the target Indigenous language. Table 5 reports the textual statistics for dev and test sets.

[7]Note that using ChrF++ as a metric could introduce bias. We use ChrF++ mainly to detect the "best" and "worst" translations, but for the majority of Spanish sentences we include random translations. Also, since most of the systems are of lower quality, we expect the introduced bias to be negligible.

Metrics The winning submission will be the one with the highest correlation with the ratings on a held-out test set of size 200. We employ Pearson correlation coefficient as the main evaluation metric, but also report Spearman correlation values. We choose Pearson over Spearman as it measures the linearity of the relationship. Linear metrics are preferred since they offer greater interpretability.

Baselines We use BLEU and Chrf++ as our automatic baselines. ChrF++ is character-based and is shown to correlate better than BLEU with morphologically-rich languages. ChrF outperforms BLEU on non-standardized orthographies as well (Aepli et al., 2023). Therefore, we consider it as the main baseline to beat.

Submitted Systems This ST got a total of 11 submissions by 3 different teams. We only have the descriptions of two of these teams. Below is a concise overview of each team's contribution.

- **Tekio**: The submission of R. Krasner et al. (2025) relies mainly on finetuning Language-agnostic BERT Sentence Encoder (LaBSE; (Feng et al., 2022)) to develop better semantic representations for Indigenous languages. They use the data for the MT ST for contrastive alignment in the finetuning. This finetuned LaBSE is the backbone of four metrics: 1) YiSi-1 (Lo, 2019, 2020) is an MT quality metric that needs representations to evaluate semantic similarity. In the first submission, for each language, they chose the top three intermediate layers based on the performance on the development set and averaged their token embeddings. 2) The same as #1, but they use the three layers that that did best on average for all the languages to avoid overfitting. 3) COMET Estimator Model (Rei et al., 2020) with the finetuned LaBSE as the pretrained model and mean absolute error (MAE) as the loss function. 5-fold cross-validation is

	Guarani		Bribri		Nahuatl		Average	
Method	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson	Spearman	Pearson
ChrF++	<u>0.6725</u>	0.6263	0.4517	0.3823	**0.6783**	0.5549	0.6008	0.5212
BLEU	0.4676	0.4056	0.4518	0.3456	0.3541	0.4061	0.4245	0.3857
Tekio_1	0.6611	0.7196	<u>0.5622</u>	0.6244	0.668	0.6115	**0.6304**	0.6518
Tekio_2	0.6611	0.7196	0.5569	0.63	0.6132	0.5845	0.6104	0.6447
Tekio_3	0.5597	<u>0.7209</u>	0.4892	0.6261	0.4963	0.529	0.5151	0.6254
Tekio_4	0.5605	**0.7234**	0.4909	0.6268	0.5036	0.5351	0.5183	0.6285
RaaVa_1	0.6723	0.6249	0.5356	0.4223	<u>0.6766</u>	0.5657	<u>0.6282</u>	0.5377
RaaVa_2	0.6516	0.6776	**0.5755**	0.5662	0.6145	0.5921	0.6139	0.612
RaaVa_3	0.656	0.7038	0.4829	0.5931	0.6364	0.6263	0.5918	0.6411
RaaVa_4	0.656	0.7038	0.4829	0.5931	0.6364	0.6263	0.5918	0.6411
RaaVa_5	0.6526	<u>0.7209</u>	0.5379	**0.654**	0.6195	**0.6362**	0.6033	**0.6704**
RaaVa_6	0.6429	0.6964	0.5332	<u>0.6523</u>	0.6132	<u>0.6351</u>	0.5965	<u>0.6613</u>
LexiLogic	**0.6811**	0.6529	0.5021	0.3763	0.6717	0.5504	0.6183	0.5265

Table 6: Final results for ST3. The best score for each column is bolded, while the second best score is underlined. The difference between RaaVa_3 and RaaVa_4 is minuscule and can only be seen in the later decimals.

used on all the available annotated scores. 4) The same as #3, but with mean squared error (MSE) as the loss function.

- **RaaVa**: The submission of Raja and Vats (2025) combines various linguistic and computational features, including lexical similarity via Levenshtein distance (Levenshtein et al., 1966), phonetic similarity using Metaphone (Philips, 1990) and Soundex encoding (Russell, 1918), semantic similarity through LaBSE sentence embeddings, and fuzzy token matching to account for morphological variations (Kondrak, 2005). They submit 6 systems: 1) this system integrates character-level lexical overlap via Jaccard similarity with phonetic similarity from Metaphone encodings. 2) Lexical (Damerau-Levenshtein edit distance), phonetic (Metaphone encodings), and semantic similarity (LaBSE sentence encoding) are linearly combined with fixed weights. 3) This system incorporates four similarity metrics, adding fuzzy similarity to the lexical, phonetics, and semantic similarities. Again, the final metric is a weighted average of the individual metrics. 4) Two separate linear regression models are trained for semantic and fluency, based on the four similarity metrics of #3. The regression models are trained on the development sets. 5) Same as #4 but a Ridge regression is used for semantic similarity estimation, while Random Forest regression is used to model fluency. 6) Same as #5, but a Gradient Boosting Regressor (GBR, (Zemel and Pitassi, 2000)) is trained to model fluency.

Results Table 6 shows the final correlation scores for the submitted systems. Overall, RaaVa_5 has the best Pearson performance and is the winner of the shared task, while RaaVa_6 follows closely as the second best system. Tekio_1 has the best Spearman correlation on average, and the third best according to Pearson. None of the systems beat ChrF++ on Spearman for Nahuatl.

Findings In our schema, we weigh fluency and adequacy the same, which could partially explain the superior performance of RaaVa_5 and RaaVa_6 that model those two aspects separately. RaaVa_5 increases the Pearson correlation by 0.149 on average. It must be noted that this framework of human judgment for MT has drawn criticisms (Graham et al., 2013). We adopt this schema for its simplicity for annotators and consistency with previous iterations of MT shared task, but this could potentially change in future iterations.
Table 10 demonstrates the correlation scores of each submitted system with semantics and fluency. Tekio_1 has the highest overall correlation with semantics at 0.6446, while RaaVa_5 is a close second at 0.6432. However, RaaVa_5 has a much higher correlation with fluency than Tekio_1.
The baseline performance on Bribri is relatively poor, hinting that string-based methods are particularly lacking for this language. However, it is important to note that Bribri has much longer sentences in terms of number of words in our study (Table 5). It sees the biggest boost in performance (+0.27) among the three languages. In contrast, Guarani and Nahuatl exhibit more modest gains (+0.1 and +0.08, respectively) but have stronger

baseline results. The agglutinating morphology of Nahuatl could in part explain the strong performance of ChrF++ (Pugh and Tyers, 2021a), whereas Bribri is a fusional language. Taken together, the results suggest that neural approaches hold significant potential for Indigenous languages. This corroborates the findings of Aepli et al. (2023) where neural models based on COMET far outperformed string-based baselines for language variations with non-standardized orthographies.

6 Conclusions

We have introduced the three STs held this year at the AmericasNLP workshop: (1) MT for truly low-resource Languages, (2) morphological adaptation for generating educational examples, and (3) metric development for MT in Indigenous languages. Overall, 12 teams participated across a total of 27 submissions.

In the MT task, the baseline (a 1.3B encoder-decoder model) proves hard to beat for translation from Spanish. The new translation direction into Spanish benefits from the use of GPT-based models. This highlights both the limitations imposed by the current available data and the strength of well-adapted, smaller-scale approaches. For the task on generating examples for educational material, while the use of LLMs through prompt engineering and reference-based approaches proves effective for certain languages, our results suggest that knowledge-based methods still hold value, especially for morphologically complex, low-resource languages and tasks involving multiple interacting grammatical phenomena. In the metrics ST, we find that neural methods far outperform the string-based baselines; in spite of the amount of available data that limits the performance of neural models.

These shared tasks contribute to the broader NLP community by advancing methods specific to highly diverse, underrepresented languages. They also provide publicly available datasets, tools, and benchmarks that serve both academic research and community-driven language technology efforts.

Acknowledgements

We would like to thank all teams for their participation, to all reviewers, and all the community that has supported us for this shared task. We want to specially thank to our annotators Wayuunaiki Language Translation Services S.A.S, María Ximena Juárez Huerta, Ángeles Márquez Hernandez, Luis Samuel Santiago Melchor, Blanca Estela Huerta Acosta, Heber Jabdiel Almaraz Miranda, Yanua Liseth Atamain Uwarai, Amanda Domínguez Galeano, Alex Agustin González Diaz, and Rosana Andrea Alvarez Meza.

This project has benefited from financial support to Ona de Gibert by the European Union's Horizon Europe research and innovation programme under Grant agreement No 101070350 and from UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 10052546].

Ethical statement

All AmericasNLP shared tasks are community-based efforts, and therefore they have a close relationship with the native speakers of all communities. We follow the consensus principles in the NLP field when working with indigenous communities (Bird, 2020; Mager et al., 2023): performing consultation with native speakers and communities for each of the languages; we aim to respect the local culture; we also involve native speakers in the scientific work; and we share and distribute the data and research openly. We also want to emphasize that the systems in this exercise are scientific experiments, are not production-ready, and should not be used to solve real-world problems. We also encourage all participating teams to share their systems, model weights, and additional data, so that the advances can be used at the discretion of each community. For ST1, in some languages, the Bible is used as part of the training data. However, we tried to reduce its usage to a minimum, and never used it for testing, as we aim to have as unbiased a benchmarking set as possible (Hutchinson, 2024). Finally, all translators and manual annotators were paid above the average teacher's salary, depending on their country of origin.

References

Noëmi Aepli, Chantal Amrhein, Florian Schottmann, and Rico Sennrich. 2023. A benchmark for evaluating machine translation metrics on dialects without standard orthography. In *Proceedings of the Eighth Conference on Machine Translation*, pages 1045–1065, Singapore. Association for Computational Linguistics.

Željko Agić and Ivan Vulić. 2019. JW300: A wide-coverage parallel corpus for low-resource languages. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3204–3210, Florence, Italy. Association for Computational Linguistics.

Steven Bird. 2020. Decolonising speech and language technology. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3504–3519, Barcelona, Spain (Online). International Committee on Computational Linguistics.

David Brambila. 1976. *Diccionario rarámuri-castellano (tarahumar)*. Obra Nacional de la buena Prensa.

Shweta Chauhan, Philemon Daniel, Archita Mishra, and Abhay Kumar. 2023. Adableu: A modified bleu score for morphologically rich languages. *IETE Journal of Research*, 69(8):5112–5123.

Luis Chiruzzo, Pedro Amarilla, Adolfo Ríos, and Gustavo Giménez Lugo. 2020. Development of a Guarani - Spanish parallel corpus. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2629–2633, Marseille, France. European Language Resources Association.

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. 2024. 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 2024)*, pages 224–235, Mexico City, Mexico. Association for Computational Linguistics.

Jorge Asillo Congora, Julio Santisteban, and Ricardo Lazo Vasquez. 2025. Ucsp submission to the americasnlp 2025 shared task. 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.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, page 2475. Association for Computational Linguistics.

Ona De Gibert, Raúl Vázquez, Mikko Aulamo, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2023. Four approaches to low-resource multilingual nmt: The helsinki submission to the americasnlp 2023 shared task. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 177–191.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The llama 3 herd of models. *arXiv e-prints*, pages arXiv–2407.

Abteen Ebrahimi, Ona de Gibert, Raul Vazquez, Rolando Coto-Solano, Pavel Denisov, Robert Pugh, Manuel Mager, Arturo Oncevay, Luis Chiruzzo, Katharina von der Wense, and Shruti Rijhwani. 2024. Findings of the AmericasNLP 2024 shared task on machine translation into indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 236–246, Mexico City, Mexico. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios Gonzales, Ivan Meza-Ruiz, and 1 others. 2022. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219, Toronto, Canada. Association for Computational Linguistics.

Hermenegildo Espejo Apikai, Ketty Betsamar García Ruiz, and 1 others. 2021. Awajún chicham jintiatin etejamu. vocabulario pedagógico awajún.

Isaac Feldman and Rolando Coto-Solano. 2020. Neural machine translation models with back-translation for the extremely low-resource indigenous language Bribri. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3965–3976, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic

BERT sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891, Dublin, Ireland. Association for Computational Linguistics.

Luciana Ferrer and Pablo Riera. Confidence intervals for evaluation in machine learning.

Ana-Paula Galarreta, Andrés Melgar, and Arturo Oncevay. 2017. Corpus creation and initial smt experiments between spanish and shipibo-konibo. In *Proceedings of the International Conference Recent Advances in Natural Language Processing, RANLP 2017*, pages 238–244.

Edward Gow-Smith and Danae Sánchez Villegas. 2023. Sheffield's submission to the AmericasNLP shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 192–199, Toronto, Canada. Association for Computational Linguistics.

Yvette Graham, Timothy Baldwin, Alistair Moffat, and Justin Zobel. 2013. Continuous measurement scales in human evaluation of machine translation. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 33–41, Sofia, Bulgaria. Association for Computational Linguistics.

Ximena Gutierrez-Vasques, Gerardo Sierra, and Isaac Hernandez Pompa. 2016. Axolotl: a web accessible parallel corpus for Spanish-Nahuatl. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4210–4214, Portorož, Slovenia. European Language Resources Association (ELRA).

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. *Computational Linguistics*, 48(3):673–732.

Hansi Hettiarachchi, Tharindu Ranasinghe, Paul Rayson, Ruslan Mitkov, Mohamed Gaber, Damith Premasiri, Fiona Anting Tan, and Lasitha Uyangodage, editors. 2025. *Proceedings of the First Workshop on Language Models for Low-Resource Languages*. Association for Computational Linguistics, Abu Dhabi, United Arab Emirates.

JANE H. Hill and Kenneth C. Hill. 1978. Honorific usage in modern nahuatl: The expression of social distance and respect in the nahuatl of the malinche volcano area. *Language*, 54:123.

Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*.

Diego Huarcaya Taquiri. 2020. Traducción automática neuronal para lengua nativa peruana. *Bachelor's thesis, Universidad Peruana Unión*.

Jonathan Hus, Antonios Anastasopoulos, and Nathaniel R. Krasner. 2025. Machine translation using grammar materials for llm post-correction. 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.

Ben Hutchinson. 2024. Modeling the sacred: Considerations when using religious texts in natural language processing. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 1029–1043, Mexico City, Mexico. Association for Computational Linguistics.

Oana Ignat, Zhijing Jin, Artem Abzaliev, Laura Biester, Santiago Castro, Naihao Deng, Xinyi Gao, Aylin Ece Gunal, Jacky He, Ashkan Kazemi, Muhammad Khalifa, Namho Koh, Andrew Lee, Siyang Liu, Do June Min, Shinka Mori, Joan C. Nwatu, Veronica Perez-Rosas, Siqi Shen, and 3 others. 2024. Has it all been solved? open NLP research questions not solved by large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8050–8094, Torino, Italia. ELRA and ICCL.

INALI. 2012. Catálogo de las lenguas indígenas nacionales en riesgo de desaparición. `https://www.cdi.gob.mx/dmdocuments/lenguas_indigenas_nacionales_en_riesgo_de_desaparicion_inali.pdf/`.

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.

Tom Kocmi, Christian Federmann, Roman Grundkiewicz, Marcin Junczys-Dowmunt, Hitokazu Matsushita, and Arul Menezes. 2021. To ship or not to ship: An extensive evaluation of automatic metrics for machine translation. In *Proceedings of the Sixth Conference on Machine Translation*, pages 478–494, Online. Association for Computational Linguistics.

Philipp Koehn and Christof Monz. 2006. Manual and automatic evaluation of machine translation between European languages. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 102–121, New York City. Association for Computational Linguistics.

Grzegorz Kondrak. 2005. N-gram similarity and distance. In *International symposium on string processing and information retrieval*, pages 115–126. Springer.

Vladimir I Levenshtein and 1 others. 1966. Binary codes capable of correcting deletions, insertions, and reversals. In *Soviet physics doklady*, volume 10, pages 707–710. Soviet Union.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, and 1 others. 2021. Few-shot learning with multilingual language models. *arXiv preprint arXiv:2112.10668*.

Chi-kiu Lo. 2019. YiSi - a unified semantic MT quality evaluation and estimation metric for languages with different levels of available resources. In *Proceedings of the Fourth Conference on Machine Translation (Volume 2: Shared Task Papers, Day 1)*, pages 507–513, Florence, Italy. Association for Computational Linguistics.

Chi-kiu Lo. 2020. Extended study on using pretrained language models and YiSi-1 for machine translation evaluation. In *Proceedings of the Fifth Conference on Machine Translation*, pages 895–902, Online. Association for Computational Linguistics.

James Loriot, Erwin Lauriault, and Dwight Day. 1993. *Diccionario Shipibo-Castellano*. Ministerio de Educación del Perú, Perú.

Tom Lupicki, Lavanya Shankar, Kaavya Chaparala, and David Yarowsky. 2025. JHU's submission to the AmericasNLP 2025 Shared Task on the Creation of Educational Materials 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.

Manuel Mager, Diónico Carrillo, and Ivan Meza. 2018. Probabilistic finite-state morphological segmenter for wixarika (huichol) language1. *Journal of Intelligent & Fuzzy Systems*, 34(5):3081–3087.

Manuel Mager, Özlem Çetinoğlu, and Katharina Kann. 2019. Subword-level language identification for intra-word code-switching. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2005–2011, Minneapolis, Minnesota. Association for Computational Linguistics.

Manuel Mager, Abteen Ebrahimi, Shruti Rijhwani, Arturo Oncevay, Luis Chiruzzo, Robert Pugh, and Katharina von der Wense, editors. 2024. *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*. Association for Computational Linguistics, Mexico City, Mexico.

Manuel Mager, Elisabeth Mager, Katharina Kann, and Ngoc Thang Vu. 2023. Ethical considerations for machine translation of indigenous languages: Giving a voice to the speakers. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4871–4897, Toronto, Canada. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 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, Online. Association for Computational Linguistics.

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

Nitika Mathur, Timothy Baldwin, and Trevor Cohn. 2020. Tangled up in BLEU: Reevaluating the evaluation of automatic machine translation evaluation metrics. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4984–4997, Online. Association for Computational Linguistics.

Elena Mihas. 2011. Añaani katonkosatzi parenini, el idioma del alto perené. *Milwaukee, WI: Clarks Graphics*.

Héctor Erasmo Gómez Montoya, Kervy Dante Rivas Rojas, and Arturo Oncevay. 2019. A continuous improvement framework of machine translation for Shipibo-konibo. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 17–23, Dublin, Ireland. European Association for Machine Translation.

Oscar Moreno, Yanua Atamain, and Arturo Oncevay. 2024. Awajun-OP: Multi-domain dataset for Spanish–awajun machine translation. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 112–120, Mexico City, Mexico. Association for Computational Linguistics.

Carla Victoria Jara Murillo. 2018a. *Gramática de la lengua bribri*. éditeur non identifié.

Carla Victoria Jara Murillo. 2018b. *I ttè: historias bribrís*. ,second edition. Editorial de la Universidad de Costa Rica.

Carla Victoria Jara Murillo and Alí García Segura. 2013. *Se'ttö bribri ie: Hablemos en bribri*. Programa de Regionalización Interuniversitaria CONARE.

John E Ortega, Richard Alexander Castro-Mamani, and Jaime Rafael Montoya Samame. 2020. Overcoming resistance: The normalization of an Amazonian tribal language. In *Proceedings of the 3rd Workshop on Technologies for MT of Low Resource Languages*, pages 1–13, Suzhou, China. Association for Computational Linguistics.

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, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Lawrence Philips. 1990. Hanging on the metaphone. *Computer Language*, 7(12):39–43.

Maja Popović. 2017. chrF++: words helping character n-grams. In *Proceedings of the Second Conference on Machine Translation*, pages 612–618, Copenhagen, Denmark. Association for Computational Linguistics.

Beatrix Potter. 1904. *Benjamin Bunny*. Frederick Warne Co., Inc., New York.

Juan Prieto, Cristian Martinez, Melissa Robles, Alberto Moreno, Sara Palacios, and Rubén Manrique. 2024. Translation systems for low-resource colombian indigenous languages, a first step towards cultural preservation. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP 2024)*, pages 7–14, Mexico City, Mexico. Association for Computational Linguistics.

Prokopis Prokopidis, Vassilis Papavassiliou, and Stelios Piperidis. 2016. Parallel Global Voices: a collection of multilingual corpora with citizen media stories. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 900–905, Portorož, Slovenia. European Language Resources Association (ELRA).

Robert Pugh, Marivel Huerta Mendez, Mitsuya Sasaki, and Francis Tyers. 2022. Universal Dependencies for western sierra Puebla Nahuatl. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5011–5020, Marseille, France. European Language Resources Association.

Robert Pugh and Francis Tyers. 2021a. Investigating variation in written forms of Nahuatl using character-based language models. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 21–27, Online. Association for Computational Linguistics.

Robert Pugh and Francis Tyers. 2021b. Towards an open source finite-state morphological analyzer for zacatlán-ahuacatlán-tepetzintla Nahuatl. In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 80–85, Online. Association for Computational Linguistics.

Nathaniel R. Krasner, Justin Vasselli, Belu Ticona, Antonios Anastasopoulos, and Chi-kiu Lo. 2025. Machine translation metrics for indigenous languages using fine-tuned semantic embeddings. 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.

Rahul Raja and Arpita Vats. 2025. Fuse : A ridge and random forest-based metric for evaluating mt in 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.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

Rubén Romano and Sebastián Richer. 2008. Ñaantsipeta asháninkaki birakochaki. http://www.lengamer.org/publicaciones/diccionarios/.

Robert C. Russell. 1918. Soundex system of phonetic indexing.

Mitsuya Sasaki. 2014. A dialectological sketch of Ixquihuacan Nahuatl. , 35(TULIP):139–170.

Petra Schroeder. 2014. *Gramática del Náhuatl de San Miguel Tenango, Zacatlán, Puebla*. Summer Institute of Linguistics. [Draft publication].

Petra Schroeder. 2015. *Phonology of Nahuatl de San Miguel Tenango, Zacatlán, Puebla*. Summer Institute of Linguistics. [Draft publication].

Petra Schroeder and David H. Tuggy. 2010. The consonantal prefixes of San Miguel Tenango Nahuatl, Zacatlán. *Etnografía del estado de Puebla, zona norte*, pages 112–117.

Jorg Tiedemann. 2012. Parallel Data, Tools and Interfaces in OPUS.

Adolfo Constenla Umaña, Feliciano Elizondo Figueroa, and Francisco Pereira Mora. 2004. *Curso básico de bribri*. Editorial de la Universidad de Costa Rica.

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.

Justin Vasselli, Haruki Sakajo, Arturo Martínez Peguero, Frederikus Hudi, and Taro Watanabe. 2025. Leveraging Dictionaries and Grammar Rules for the Creation of Educational Materials 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.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Raúl Vázquez, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2021. The Helsinki submission to the AmericasNLP shared task. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 255–264, Online. Association for Computational Linguistics.

Jules Verne. 1874. *A Journey to the Centre of the Earth*. Scribner, Armstrong & Co., New York.

Ruvan Weerasinghe, Isuri Anuradha, and Deshan Sumanathilaka, editors. 2025. *Proceedings of the First Workshop on Natural Language Processing for Indo-Aryan and Dravidian Languages*. Association for Computational Linguistics, Abu Dhabi.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mt5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498.

Mahshar Yahan and Mohammad Amanul Islam. 2025. Leveraging large language models for spanish-indigenous language machine translation at americasnlp 2025. 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.

Richard Zemel and Toniann Pitassi. 2000. A gradient-based boosting algorithm for regression problems. *Advances in neural information processing systems*, 13.

A Dataset Statistics for ST1

Table 7 shows the number of sentences for each language in the dataset.

Language	Train Source	Train
Chatino (ctp)	(Ebrahimi et al., 2023)	357
Asháninka (cni)	(Ortega et al., 2020; Romano and Richer, 2008; Mihas, 2011)	3,883
Otomí (oto)	(Mager et al., 2021)	4,889
Aymara (aym)	(Prokopidis et al., 2016; Tiedemann, 2012)	6,531
Bribri (bzd)	(Feldman and Coto-Solano, 2020; Margery Peña, 2005; Murillo, 2018a; Umaña et al., 2004; Murillo and Segura, 2013; Murillo, 2018b)	7,508
Wixarika (hch)	(Mager et al., 2018)	8,966
Shipibo-Konibo (shp)	(Montoya et al., 2019; Galarreta et al., 2017; Loriot et al., 1993)	14,592
Rarámuri (tar)	(Brambila, 1976)	14,720
Nahuatl (nah)	(Gutierrez-Vasques et al., 2016)	16,145
Awajun (agr)	(Moreno et al., 2024)	21,964
Guarani (grn)	(Chiruzzo et al., 2020)	26,032
Wayuunaiki (guc)	(Prieto et al., 2024)	59,715
Quechua (quy)	(Agić and Vulić, 2019; Huarcaya Taquiri, 2020)	125,008

Table 7: Dataset statistics for ST1, together with the sources for the tr
Languages are listed in increasing order of available training data. Americ
indigenous languages from a set of different sources (please see the corre

B ST1 Ranking

Table 8 shows the main ranking of all submitted systems for ST1.

Rank	Team	Ver.	Count	Tot. BLEU	Tot. ChrF	Tot. ChrF++
			Track 1: SPA-XXX			
1	GMU	2	13	43.72	324.12	285.37
2	Syntax Squad	1	11	36.24	265.50	233.07
3	Syntax Squad	2	1	2.02	30.13	26.31
4	UCSP	1	1	0.07	21.73	16.75
-	GMU	1	13	31.83	273.23	244.56
			Track 2: XXX-SPA			
1	GMU	2	13	93.44	368.14	346.06
2	Syntax Squad	1	11	75.31	279.68	261.19
3	UCSP	1	1	1.52	20.70	17.87
-	GMU	1	13	99.19	363.52	343.34

Table 8: Main ranking of all submitted systems for ST1. Ver denotes the t
denotes the number of languages a particular system was submitted for, w
total sum of the metric score across submissions. The final three column
languages of the shared task, with ChrF++ being used to calculate the c

C ST1 Full Results

Table C shows the full results of ST1.

LANG.	TEAM	VER.	BLEU	CHRF	CHRF++
TRACK 1: SPA-XXX					
agr-spa	GMU	0	16,81	38,73	36,59
agr-spa	GMU	1	15,17	38,73	36,52
agr-spa	Syntax Squad	0	13,21	36,11	33,70
aym-spa	GMU	0	6,51	27,50	26,09
aym-spa	Syntax Squad	0	5,89	27,53	25,78
aym-spa	GMU	1	5,17	26,49	25,23
bzd-spa	GMU	0	6,98	29,14	27,86
bzd-spa	GMU	1	6,11	28,77	27,41
bzd-spa	Syntax Squad	0	5,87	27,53	26,22
cni-spa	GMU	0	5,32	23,72	22,44
cni-spa	GMU	1	4,00	22,94	21,57
cni-spa	Syntax Squad	0	3,06	21,34	20,13
ctp-spa	GMU	1	11,74	28,04	26,16
ctp-spa	GMU	0	3,76	15,60	14,47
grn-spa	GMU	0	13,81	34,93	33,84
grn-spa	GMU	1	11,23	33,57	32,31
grn-spa	Syntax Squad	0	15,14	26,15	24,70
guc-spa	GMU	1	4,20	26,00	23,93
guc-spa	GMU	0	2,92	25,06	23,10
guc-spa	Syntax Squad	0	3,14	16,19	14,40
hch-spa	GMU	0	5,46	25,91	24,37
hch-spa	GMU	1	4,69	25,53	24,04
hch-spa	Syntax Squad	0	3,98	23,69	22,02
nah-spa	GMU	0	7,22	27,14	25,58
nah-spa	GMU	1	5,08	26,18	24,31
nah-spa	Syntax Squad	0	4,00	15,40	13,88
oto-spa	GMU	0	2,25	19,69	18,24
oto-spa	Syntax Squad	0	1,50	19,91	17,80
oto-spa	GMU	1	1,36	17,76	15,99
quy-spa	GMU	0	12,27	34,64	33,02
quy-spa	GMU	1	10,38	33,50	31,77
quy-spa	Syntax Squad	0	10,60	33,26	31,71
quy-spa	UCSP	0	1,52	20,70	17,87
shp-spa	GMU	0	13,83	39,93	38,01
shp-spa	GMU	1	12,55	39,40	37,43
shp-spa	Syntax Squad	0	8,94	32,58	30,83
tar-spa	GMU	0	2,07	21,53	19,72
tar-spa	GMU	1	1,75	21,23	19,39

LANG.	TEAM	VER.	BLEU	CHRF	CHRF++
		TRACK 2: XXX-SPA			
spa-agr	Syntax Squad	0	7,82	40,10	35,16
spa-agr	GMU	1	8,64	39,75	35,09
spa-agr	GMU	0	1,30	19,16	16,67
spa-aym	Syntax Squad	0	1,96	31,61	27,72
spa-aym	Syntax Squad	1	2,02	30,13	26,31
spa-aym	GMU	1	1,14	26,26	22,91
spa-aym	GMU	0	0,88	23,12	20,45
spa-bzd	Syntax Squad	0	4,55	21,68	22,77
spa-bzd	GMU	1	4,41	21,56	22,51
spa-bzd	GMU	0	3,85	19,42	20,61
spa-cni	Syntax Squad	0	2,43	26,96	23,17
spa-cni	GMU	1	2,47	25,60	22,22
spa-cni	GMU	0	3,63	24,62	21,77
spa-ctp	GMU	0	1,64	15,04	13,33
spa-ctp	GMU	1	1,27	15,31	12,25
spa-grn	GMU	0	5,47	32,50	29,95
spa-grn	GMU	1	4,04	27,23	25,00
spa-grn	Syntax Squad	0	3,46	17,84	16,21
spa-guc	GMU	1	1,48	27,42	22,93
spa-guc	Syntax Squad	0	0,11	15,86	12,83
spa-guc	GMU	0	0,20	10,94	9,12
spa-hch	Syntax Squad	0	11,07	30,47	26,77
spa-hch	GMU	1	10,04	29,59	26,14
spa-hch	GMU	0	5,98	27,00	23,59
spa-nah	GMU	1	2,02	23,82	20,33
spa-nah	GMU	0	0,64	18,76	15,98
spa-nah	Syntax Squad	0	0,65	15,73	12,64
spa-oto	Syntax Squad	0	0,76	14,16	12,02
spa-oto	GMU	1	1,33	13,23	11,31
spa-oto	GMU	0	0,98	11,55	10,03
spa-quy	GMU	1	3,70	38,02	32,70
spa-quy	GMU	0	3,80	36,30	31,68
spa-quy	Syntax Squad	0	3,07	36,14	31,01
spa-quy	UCSP	0	0,07	21,73	16,75
spa-shp	GMU	1	2,79	21,99	19,46
spa-shp	GMU	0	2,68	19,39	17,49
spa-shp	Syntax Squad	0	0,37	14,94	12,76
spa-tar	GMU	0	0,77	15,45	13,89
spa-tar	GMU	1	0,39	14,35	12,53

Table 9: Full results of ST1.

D ST3 Results

Table 10 shows the results for ST3 broken down between semantics and fluency scores.

	Guarani		Bribri		Nahuatl		Average	
Method	Semantics	Fluency	Semantics	Fluency	Semantics	Fluency	Semantics	Fluency
ChrF++	0.63	0.5323	0.4078	0.3018	0.5681	0.4929	0.5353	0.4424
BLEU	0.4207	0.3314	0.3515	0.2908	0.4257	0.351	0.3993	0.3244
Tekio_1	**0.6899**	0.6474	0.6369	0.5236	0.6069	0.5618	**0.6446**	0.5776
Tekio_2	**0.6899**	0.6474	**0.6404**	0.5307	0.5789	0.5381	0.6364	0.5721
Tekio_3	0.603	<u>0.7411</u>	0.6002	0.5657	0.49	0.5203	0.5644	0.609
Tekio_4	0.6054	**0.7433**	0.6036	0.5634	0.4972	0.5248	0.5687	<u>0.6105</u>
RaaVa_1	0.6367	0.5227	0.4644	0.3187	0.5818	0.5	0.561	0.4471
RaaVa_2	0.6518	0.6073	0.5852	0.4667	0.5896	0.5423	0.6089	0.5388
RaaVa_3	0.6793	0.6284	0.5689	0.5355	**0.625**	0.5722	0.6244	0.5787
RaaVa_4	0.6793	0.6284	0.5689	0.5355	**0.625**	0.5722	0.6244	0.5787
RaaVa_5	<u>0.6816</u>	0.6584	0.6314	**0.5862**	0.6165	**0.5991**	<u>0.6432</u>	**0.6146**
RaaVa_6	0.6661	0.628	<u>0.6372</u>	<u>0.5768</u>	<u>0.621</u>	<u>0.5927</u>	0.6414	0.5992
LexiLogic	0.6512	0.5608	0.4233	0.274	0.5645	0.488	0.5463	0.4409

Table 10: Pearson correlation scores of each submitted system with adequacy (semantics) and fluency of the annotated instances in the test dataset for ST3. The best score(s) for each column is bolded, while the second best score is underlined.