Machine Translation Using Grammar Materials for LLM Post-Correction

Jonathan Hus[1], Nathaniel Krasner[1], Antonios Anastasopoulos[1,2]
[1]George Mason University, [2] Archimedes, Athena Research Center
jhus@gmu.edu, nkrasner@gmu.edu, antonis@gmu.edu

Abstract

This paper describes George Mason University's submission to the AmericasNLP 2025 Shared Task on Machine Translation into Indigenous Languages. We prompt a large language model (LLM) with grammar reference materials to correct the translations produced by a finetuned Encoder-Decoder machine translation system. This hybrid approach leads to improvements when translating from the indigenous languages into Spanish, indicating that LLMs are capable of using grammar materials to better handle a previously unseen-during-pretraining language.[1]

1 Introduction

Machine translation (MT) systems typically require massive parallel corpora to achieve state-of-the-art results. However, this magnitude of data is not available for low resource languages. To address this dearth of data, we propose a prompt-based approach that incorporates linguistic reference material including grammar books, dictionaries, and a limited number of parallel sentences. This approach was originally proposed in Machine Translation from One Book (MTOB; Tanzer et al., 2023) for a single language (Kalamang) and Hus and Anastasopoulos (2024) expanded to a more large-scale investigation to include 15 additional low resource languages.

In order to improve performance, we have augmented the prompt to include a translation from a dedicated MT system, which has been finetuned on the 13 Latin American indigenous languages using the available parallel sentences from the AmericasNLP 2025 training set. Thus, the large language model (LLM) is provided with a potential translation that can be utilized in conjunction with the reference linguistic material. The reference material consists of the following items:

Dictionaries We obtain dictionaries from PanLex[2] for all our languages. Note that in cases where the number of words in the dictionary was less than 100 we do not include them in the prompt. The size of each dictionary is included in Appendix A

Parallel Sentences Parallel sentences are included in the prompts as translation examples for in-context learning. We use the training set as provided by AmericasNLP 2025 Shared Task on Machine Translation.

Grammar Books The DReaM corpus (Virk et al., 2020) contains digitized versions of thousands of linguistic documents, including grammar books and sketches, for many languages. The source of these documents is often in paper format, and due to the scanning/OCR quality, the digitized versions often contain scanning artifacts. We select one grammar document for each of our languages. We perform slight manual cleanup to remove some items (e.g., scanning artifacts, table of contents) and to ensure that the grammar would fit in the LLM's context size.

2 Methodology

We use the GPT-4o-mini model for our experiments. Its context size of 128k tokens allows large grammar books to be included in the prompt. Additionally, we finetune separate NLLB 3.3B models (Costa-jussà et al., 2022) for each translation direction (xx→es and es→xx) using the provided training data. These NLLB models are then used to provide preliminary "suggested" translations for the LLM to edit.

Prompt Format Our prompts are formatted to contain the following information:

- Prefix - Contains the task description, including the source and target languages

[1] Code and data to reproduce our experiments are here: https://github.com/jonathanhus/americasnlp.

[2] https://panlex.org

Proceedings of the Fifth Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP), pages 92–99
May 4, 2025 ©2025 Association for Computational Linguistics

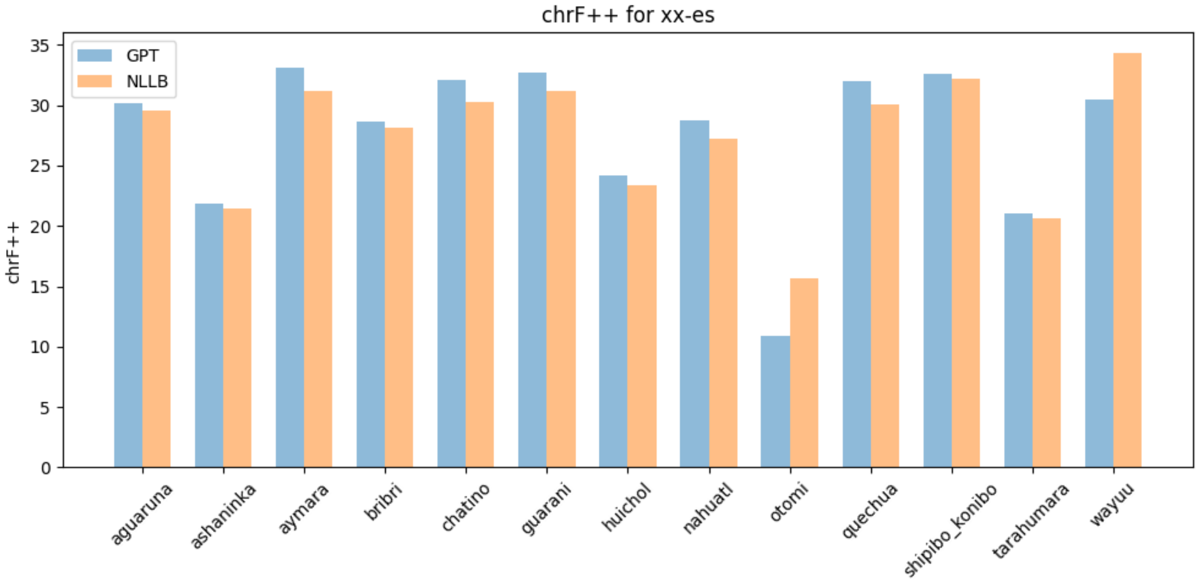

Figure 1: X-to-Spanish Performance on the Dev Dataset

- Dictionary Entries - For each word in the sentence, an entry from the bilingual dictionary is retrieved that closely matches the word. In cases where there is not a direct match of the source word, a selection is made using longest common subsequence (LCS) matching with the available words in the dictionary. The number of dictionary entries to be retrieved is configurable, but for our experiments we chose two, which was the parameter value chosen for evaluation in previous studies.

- Parallel Sentences - For each word in the sentence, a pair of parallel sentences is selected that has a similar word in it. The number of parallel sentences to be retrieved is configurable, but for our experiments we chose two, which was the parameter value chosen for evaluation in previous studies.

- Grammar Book - The full length grammar book for the indigenous language is included in the prompt

- Suggested Translation - Using our finetuned NLLB models, we provide a possible translation, and inform the LLM that it can use that to modify or improve upon it

- Suffix - Finally, we reiterate that the LLM should provide the translation and coax it to attempt the translation even if it does not "speak" the indigenous language

An example prompt is illustrated in Appendix B.

3 Results

We consider two systems when running our tests. The first is the finetuned NLLB system by itself. The second is the prompt-based LLM approach, which uses the finetuned NLLB system as one of its inputs in order to generate a translation. We evaluate both of these systems on the dev dataset and the test dataset.

Using a small sample of 100 sentences in each language from the dev dataset, we compare the chrF++ scores between the NLLB "suggestions" and the final LLM translations. It is clear from Figures 1 and 2 that, in the case of these languages, our grammar-based LLM post-correction is primarily useful for translation into Spanish rather than into languages that the LLM is unfamiliar with. This indicates that the LLM can use the grammar information to better understand the indigenous languages, but it is not enough to produce them, at least under the current prompt format and generation paradigm.

The systems are also evaluated using the test dataset, with results shown in Tables 1 and 2. Similar performance characteristics are observed, with translation into Spanish better performed by the LLM system and translation from Spanish better performed by the NLLB system.

In the previous studies that utilized the prompt-based LLM approach, ablations were performed to assess the performance of the model when given

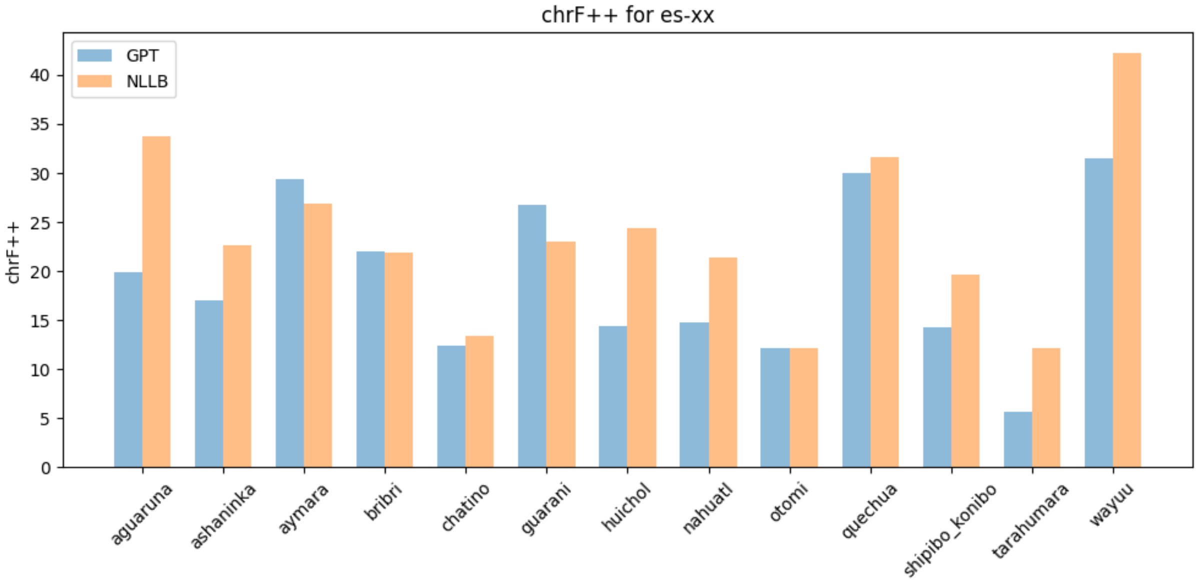

Figure 2: Spanish-to-X Performance on the Dev Dataset

various combinations of reference material input (e.g., providing only parallel sentences or providing only the grammar book.) In addition, a baseline assessment was determined for each language, where the model was provided no reference material. Due to time and cost constraints, that assessment was not performed for the set of languages in this paper. We leave that as a future research activity. A novelty in this paper is that the common language for all of the parallel sentences is Spanish, whereas previous efforts used English as the common language. However, the prompt templates and some of the grammar books are in English. The effect of having English, Spanish, and the indigenous language all represented in the prompt is unknown and this warrants further investigation.

4 Conclusion

We propose two systems to perform machine translation for indigenous languages. The first is an NLLB-based system. The second system utilizes the outputs of the NLLB-based system in addition to linguistic reference material to formulate prompts for LLMs in order to perform translation. We evaluated both our systems on the dev set of 13 different languages, translating into and out of Spanish. We note that the NLLB has superior performance in the es→xx translation direction, while the LLM-based system performs better in the xx→es direction. Both systems show a promising path forward for translation of low resource languages. Since both systems produce similar results, the more computationally efficient NLLB system would appear to be the favored choice, especially for communities lacking the resources necessary for the additional computation. However, additional techniques like Retrieval-Augmented Generation (RAG) could make more efficient use of the model and could provide improved results. Therefore, both NLLB and LLM methods deserve further research.

5 Limitations

Full-length grammar books are provided in the input prompt in order to "teach" a model how to translate into a given language. However, there are some limitations with this approach. First, high quality grammar books are difficult to obtain for many languages. The DReaM corpus does an admirable job of curating and digitizing many linguistic references, but the output is not perfect. Multi-column text documents and tables lose information that is conveyed by the location of text relative to other text on the page. The LLMs, therefore, are most likely not taking full advantage of that information. Additionally, scanning artifacts like headers and page numbers add unnecessary clutter to the reference material.

We used an OpenAI model (gpt-4o-mini) similar to what was used in Back to School (Hus and Anastasopoulos, 2024). While these models are quite performant, there are some drawbacks. First, these are truly closed models, with only an API available. The architecture, weights, and training

	GPT			NLLB			NLLB Baseline
Language	BLEU	ChrF	ChrF++	BLEU	ChrF	ChrF++	ChrF++
agr-es	16.81	38.73	36.59	15.17	38.73	36.52	38.39
aym-es	6.51	27.5	26.09	5.17	26.49	25.23	35.6
bzd-es	6.98	29.14	27.86	6.11	28.77	27.41	30.14
cni-es	5.32	23.72	22.44	4	22.94	21.57	24.86
ctp-es	3.76	15.6	14.47	11.74	28.04	26.16	35.84
gn-es	13.81	34.93	33.84	11.23	33.57	32.31	35.91
guc-es	2.92	25.06	23.1	4.2	26	23.93	24.74
hch-es	5.46	25.91	24.37	4.69	25.53	24.04	26.33
nah-es	7.22	27.14	25.58	5.08	26.18	24.31	26.36
oto-es	2.25	19.69	18.24	1.36	17.76	15.99	20.81
quy-es	12.27	34.64	33.02	10.38	33.5	31.77	37.18
shp-es	13.83	39.93	38.01	12.55	39.4	37.43	47.81
tar-es	2.07	21.53	19.72	1.75	21.23	19.39	18.75

Table 1: System Performance on Test Dataset (XX→ES)

	GPT			NLLB			NLLB Baseline
Language	BLEU	ChrF	ChrF++	BLEU	ChrF	ChrF++	ChrF++
es-agr	1.3	19.16	16.67	8.64	39.75	35.09	36.76
es-aym	0.88	23.12	20.45	1.14	26.26	22.91	31.21
es-bzd	3.85	19.42	20.61	4.41	21.56	22.51	25.52
es-cni	3.63	24.62	21.77	2.47	25.6	22.22	24.39
es-ctp	1.64	15.04	13.33	1.27	15.31	12.25	36.53
es-gn	5.47	32.5	29.95	4.04	27.23	25	35.68
es-guc	0.2	10.94	9.12	1.48	27.42	22.93	24.18
es-hch	5.98	27	23.59	10.04	29.59	26.14	28.26
es-nah	0.64	18.76	15.98	2.02	23.82	20.33	22.42
es-oto	0.98	11.55	10.03	1.33	13.23	11.31	12.78
es-quy	3.8	36.3	31.68	3.7	38.02	32.7	31.88
es-shp	2.68	19.39	17.49	2.79	21.99	19.46	25.76
es-tar	0.77	15.45	13.89	0.39	14.35	12.53	15.96

Table 2: System Performance on Test Dataset (ES→XX)

scheme are not available to researchers. Second, since the model is closed, we do not know whether the linguistic reference material is responsible for improved translation performance or whether the models themselves have this inherent ability.

The sizes of the bilingual dictionaries were inconsistent, with a handful having less than 20 words. We removed these low-volume dictionaries from our experiments. However, larger dictionaries of similar magnitudes would most likely improve the translations and would allow translation performance across the various languages to be better compared.

Acknowledgements

This work is supported by the National Science Foundation under Awards 2327143 and 2346334. This work was partially supported by resources provided by the Office of Research Computing at George Mason University (URL: https://orc.gmu.edu) and funded in part by grants from the National Science Foundation (Award Number 2018631).

References

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 (Paco) Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. URL:https://research.facebook.com/publications/no-language-left-behind/.

Jonathan Hus and Antonios Anastasopoulos. 2024. Back to school: Translation using grammar books. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 20207–20219, Miami, Florida, USA. Association for Computational Linguistics.

David Kamholz, Jonathan Pool, and Susan Colowick. 2014. PanLex: Building a resource for panlingual lexical translation. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3145–3150, Reykjavik, Iceland. European Language Resources Association (ELRA).

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2023. A benchmark for learning to translate a new language from one grammar book. In *Arxiv*.

Shafqat Mumtaz Virk, Harald Hammarström, Markus Forsberg, and Søren Wichmann. 2020. The DReaM corpus: A multilingual annotated corpus of grammars for the world's languages. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 878–884, Marseille, France. European Language Resources Association.

A Resources

For our experiments, we gathered dictionaries, parallel sentences, and grammar books to use in the prompts. Dictionaries were obtained from PanLex (Kamholz et al., 2014) and converted into the format required by the code. The sizes of the dictionaries are shown in Table 3.

Language	ISO 639-3	Dictionary Words	
		es → X	X → es
Aguaruna	agr	2242	2496
Aymara	aym	1827	1555
Bribri	bzd	11	11
Ashaninka	cni	12	10
Chatino	ctp	N/A	N/A
Guarani	gn	3354	3465
Wayuu	guc	2304	2497
Huichol	hch	12	11
Nahuatl	nah	N/A	N/A
Otomi	oto	4416	3439
Quechua	quy	20203	18589
Shipibo-Konibo	shp	1157	1129
Tarahumara	tar	1039	812

Table 3: Number of words in the dictionaries. Note the Chatino and Nahuatl were not found in the PanLex database. Therefore, translations for those words were not included in the prompt.

Language	Grammar Book	Number of Tokens
Aguaruna	Overall, Simon. (2007) A Grammar of Aguaruna. LaTrobe University doctoral dissertation.	109115
Aymara	Hardman, Martha J. (2001) Aymara (LINCOM Studies in Native American Linguistics 35). München: Lincom.	159071
Bribri	Jara Murillo, Carla Victoria. (2018) Gramática de la Lengua Bribri. San José, Costa Rica: E-Digital ED.	130572
Ashaninka	Rojas, Esaú Zumaeta and Gerardo Anton Zerdin. (2018) Ayotero añaane / Guía teórica del idioma asháninka. Nopoki: Universidad Católica Sedes Sapientiae.	164836
Chatino	Pride, Kitty. (1965) Chatino syntax (Summer Institute of Linguistics Publications in Linguistics and Related Fields 12). Norman: Summer Institute of Linguistics of the University of Oklahoma.	44698
Guarani	Gregores, Emma and Jorge A. Suárez. (1967) A Description of Colloquial Guaraní (Janua Linguarum: Series Practica 27). Berlin: Mouton de Gruyter.	
Wayuu	José Álvarez. (2017) Compendio de la gramática de la lengua wayuu. Ms.	114676
Huichol	Iturrioz Leza, José Luis and Paula Gómez López. (2006) Gramática Wixarika I. München: LINCOM.	136345
Nahuatl	Cowan de Beller, Patricia and Richard Beller. (1979) Curso del náhuatl moderno: náhuatl de la Huasteca. Mexico: Instituto Lingüístico de Verano.	57298
Otomi	Priego Montfort de Mostaghimi, Maria Eugenia. (1989) Gramática del otomí (hñähñu) del Mezquital, Mexico. Universität Bielefeld doctoral dissertation.	165311
Quechua	Zariquiey, Roberto and Gavina Córdova. (2008) Qayna, Kunan, Paqarin: Una introducción prática al quechua chanca. Lima: PUCP.	129158
Shipibo-Konibo	Faust, Norma. (1973) Lecciones para el aprendizaje del idioma shipibo-conibo (Documento de Trabajo 1). Yarinacocha: Instituto Lingüístico de Verano.	112794
Tarahumara	Caballero, Gabriela. (2022) A grammar of Choguita Rarámuri: In collaboration with Luz Elena León Ramírez, Sebastián Fuentes Holguín, Bertha Fuentes Loya and other Choguita Rarámuri language experts. Berlin: Language Science Press.	122232

Table 4: Grammar Books and Size

B Prompt Format

Each sentence to be translated is formatted into a prompt for GPT-4. The prompt has six components: prefix, words, sentences, grammar book, suggestion, and suffix. The experiment configuration determines whether words (W), sentences (S), or grammar books (G) are included in the prompt. The prefix and suffix are always included in the prompt. In the following sections, we show the format of the prompt by example, using an Aguaruna-to-Spanish translation task. We heavily used the code provided by the authors of "Machine Translation from One Book" to generate the prompts.

B.1 Prefix

The prefix provides the task to perform (translation), the source and target languages, and the sentence to translate.

> You are an expert translator. Translate the following sentence from Aguaruna to Spanish: Nunik nagkamawaju Timanmi jeen, takai takainakua jimaituk wenak yawejaju.

B.2 Words

For words, we attempt to retrieve the item from the bilingual dictionary. For each word in the source sentence, the top two matching words from the dictionary, as measured by LCS, are included in the prompt.

> To help with the translation, here is one of the closest entries to Nunik in the bilingual dictionary:
> Aguaruna word: nuniktatak
> Spanish translation: a veces
>
> To help with the translation, here is one of the closest entries to Nunik in the bilingual dictionary:
> Aguaruna word: nunik-bau ah-amu
> Spanish translation: causar
>
> *Additional word-level translations are provided for the remaining words of the source sentence.*

B.3 Sentences

For sentences, we attempt to retrieve similar samples from our small corpus of parallel sentences. For each word in the source sentence, we find sentences that contain that word, as measured by LCS, and include the top two matches in the prompt.

> To help with the translation, here is a translated sentence with words similar to N̈unik¨n a list of translated reference sentences:
> Aguaruna sentence: Aatus gobernador aidau chichaman umikag, apu Daríojai chichastatus shiyakajui. Nunik jegajuawag chichajuinak: "¡Apuh, kuashat mijan pujustin ata!
> Spanish translation: Entonces estos jefes principales y los capitanes vinieron al rey y le dijeron: ¡Oh, rey Darío! Ten vida para siempre.
>
> To help with the translation, here is a translated sentence with words similar to N̈unik¨n a list of translated reference sentences:
> Aguaruna sentence: Aatus David tupikaki uwemjauwai. Nunik Samueljai chichastatus yaakat Ramá weuwai. Nuwi jegaa Saúl niina maatag tibaun ashí Samuelan ujakui. Tusa ujaka Samueljai yaakat Naiot Ramá awa nuwi pujustatus weuwai.
> Spanish translation: Entonces David salió en vuelo, se escapó y fue a Ramá, a Samuel, y le contó todo lo que Saúl le había hecho. Y él y Samuel fueron y vivían en Naiot.
>
> *Additional sentence-level translations are provided for the remaining words of the source sentence.*

B.4 Grammar Book

We include the full grammar book in the prompt.

> To help with the translation, here is the full text of a bilingual grammar book:
> —
> ## FULL BOOK INSERTED HERE ##
> This is the end of the bilingual grammar book.
> —

B.5 Hypothesis

The output of our finetuned NLLB system is provided as a hypothesis or suggestion in the prompt.

> Here is a potential translation of the sentence provided by another system that you can modify or improve upon. Only use the suggestion if it improves your response.
> Y los criados de Saúl llegaron a la casa de Timni, y la mitad de su jornada fue en ayunas.

B.6 Suffix

The suffix reiterates the task and prompts for the appropriate translation.

> Now perform the translation. If you are not sure what the translation should be, then give your best guess. Do not say that you do not speak Aguaruna. If your translation is wrong, that is fine, but you have to provide a translation. Provide only the translation as output.
> Aguaruna: Nunik nagkamawaju Timanmi jeen, takai takainakua jimaituk wenak yawejaju.
> Spanish translation: