AmericasNLP 2024

4th Workshop on Natural Language Processing for Indigenous Languages of the Americas

Proceedings of the Workshop

June 21, 2024

The AmericasNLP organizers gratefully acknowledge the support from the following sponsors.

Platinum

amazon

Bronze

ADITU

Order copies of this and other ACL proceedings from:

Association for Computational Linguistics (ACL)
317 Sidney Baker St. S
Suite 400 - 134
Kerrville, TX 78028
USA
Tel: +1-855-225-1962
acl@aclweb.org

ISBN 979-8-89176-108-7

Introduction

We would like to welcome you to AmericasNLP 2024, the Fourth Workshop on Natural Language Processing for Indigenous Languages of the Americas!

The main goals of the workshop are to:

- encourage research on NLP, computational linguistics, corpus linguistics, and speech around the globe to work on Indigenous American languages.
- promote research on both neural and non-neural machine learning approaches suitable for low-resource languages.
- connect researchers and professionals from underrepresented communities and native speakers of endangered languages with the machine learning and NLP communities.

In 2024, AmericasNLP is being held in Mexico City, Mexico, on June 21. Prior to the workshop two shared tasks were hosted: (1) the Shared Task on Machine Translation into Indigenous Languages and (2) the Shared Task on the Creation of Educational Materials for Indigenous Languages. During the workshop, there will be 2 invited talks, a poster session, and multiple paper and shared task submission presentations.

We received a total of 37 submissions: 21 research papers, 6 previously published papers, and 10 shared task system description papers (across both shared tasks). 16 archival papers were accepted (acceptance rate: 76%) – in addition to the previously published and system description papers.

AmericasNLP would not have been possible without the help of the following: first, we thank our sponsors, Amazon and Aditu, and second, we would like to acknowledge all the time and effort put into the reviewing process, and thank for program committee members for helping us create a high-quality program. Finally, we also thank all the authors who submitted their work to the workshop, the participants of both shared tasks, and everyone who will be at the workshop, both in-person and remote, to exchange and discuss their ideas for improving natural language technologies for Indigenous languages of the Americas!

Manuel Mager, Abteen Ebrahimi, Shruti Rijhwani, Arturo Oncevay, Luis Chiruzzo, Robert Pugh, and Katharina von der Wense

AmericasNLP 2024 Organizing Committee

Organizing Committee

Organizing Committe

Manuel Mager, Amazon AWS AI, USA
Abteen Ebrahimi, University of Colorado Boulder, USA
Shruti Rijhwani, Google Deepmind, USA
Arturo Oncevay, Pontificia Universidad Católica del Perú, Peru
Luis Chiruzzo, Universidad de la República, Uruguay
Robert Pugh, Indiana University, Bloomington, USA
Katharina von der Wense, University of Colorado Boulder, USA; Johannes Gutenberg University Mainz, Germany

Program Committee

Program Committee

Eduardo Blanco, University of Arizona
Paulo Cavalin, IBM Research - Brazil
Raghuveer Chanda, Google
Rolando Coto-Solano, Dartmouth College
Ruixiang Cui, University of Copenhagen
Shridhar Devamane, Global Academy of Technology, Bangalore
C.M. Downey, University of Washington
Cristina España-Bonet, DFKI GmbH
Luke Gessler, University of Colorado, Boulder
Kartik Kannapur, Amazon Web Services
Satyapriya Krishna, Harvard University
Éric Le Ferrand, Boston College
Yukyung Lee, korea university
Zoey Liu, Department of Linguistics, University of Florida
Daniela Moctezuma, Centrogeo
Sarah Moeller, University of Florida
Remo Nitschke, University of Arizona
John E. Ortega, Northeastern University
Tanmay Parekh, University of California Los Angeles
Angeles Belem Priego Sanchez, Universidad Autónoma Metropolitana
Nathaniel Robinson, Johns Hopkins University
Amit Sah, Department of Computer Science, South Asian University
Shabnam Tafreshi, UMD:ARLIS
Atnafu Lambebo Tonja, Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC)
Daan Van Esch, Google Research
Aniket Vashishtha, Microsoft Research India
Raul Vazquez, University of Helsinki
Ekaterina Vylomova, University of Melbourne
Koichiro Watanabe, pluszero, inc
Adam Wiemerslage, University of Colorado Boulder
Si Wu, Northeastern University

Keynote Talk

Graham Neubig
Carnegie Mellon University

Bio: My research is concerned with language and its role in human communication. In particular, my long-term research goal is to break down barriers in human-human or human-machine communication through the development of natural language processing (NLP) technologies. This includes the development of technology for machine translation, which helps break down barriers in communication for people who speak different languages, and natural language understanding, which helps computers understand and respond to human language. Within this overall goal of breaking down barriers to human communication, I have focused on several aspects of language that both make it interesting as a scientific subject, and hold potential for the construction of practical systems.

Keynote Talk

Jaime Pérez González
University of California, Santa Barbara

Bio: As a morpho-syntactician, I study word formation and I am especially interested in the interface between morphology, syntax and semantics in human languages. I look at functional factors that shape the use of certain morphological constructions in agglutinative languages. Speakers activate these different domains when using their language, so there must be abstract and formal principles that determine how these patterns are represented in their minds. I investigate these topics in lesser-studied languages, with focus on Mayan languages and Miskitu (Misumalpan language spoken in Nicaragua and Honduras).

Table of Contents

NLP for Language Documentation: Two Reasons for the Gap between Theory and Practice

Luke Gessler **Katharina von der Wense**
University of Colorado Boulder
{luke.gessler,katharina.kann}@colorado.edu

Abstract

Both NLP researchers and linguists have expressed a desire to use language technologies in language documentation, but most documentary work still proceeds without them, presenting a lost opportunity to hasten the preservation of the world's endangered languages, such as those spoken in Latin America. In this work, we empirically measure two factors that have previously been identified as explanations of this low utilization: curricular offerings in graduate programs, and rates of interdisciplinary collaboration in publications related to NLP in language documentation. Our findings verify the claim that interdisciplinary training and collaborations are scarce and support the view that interdisciplinary curricular offerings facilitate interdisciplinary collaborations.

1 Introduction

In 2019, 5 out of 68 indigenous languages from Colombia were about to become extinct: one of them, Tinigua, had only a single speaker left;[1] for the others, the situation looked only marginally better. Globally, approximately half of humanity's roughly 7,000 languages are considered endangered (Bromham et al., 2022). While many people in Latin America and other places around the world want their languages to be preserved, language documentation – the process of producing grammars and texts to record a language – is very labor-intensive. Demand for individuals who can perform language documentation far outstrips supply worldwide, and there is little reason to expect this will change any time soon.

In the past 20 years,[2] the computational linguistics (CL) and natural language processing (NLP) communities have responded with systems which can automate some of the labor required in language documentation (LD). For example, ELPIS (Foley et al., 2018) can transcribe audio into text even under the challenging conditions endemic to the LD process, such as low data volumes. Despite the considerable number of computational systems which have been proposed and described over this period, they have seen little practical use (see, e.g., Good et al. 2014; Flavelle and Lachler 2023).

It is puzzling, *prima facie*, that systems with proven potential to facilitate LD have not been integrated into LD projects, and several explanations of this have been offered: Gessler (2022) cites lack of interoperability between NLP systems and LD apps. Flavelle and Lachler (2023) cite an array of organizational barriers that linguists, NLP researchers, and community members face in their collaborations, including conflicting professional incentives and a lack of understanding of the other party's conceptual frameworks. They further observe that "coursework in computational linguistics is rarely required (or even available) to students training to be documentary linguists, and vice-versa", with the consequence that they "miss out on the opportunity to learn even the basic concepts of each other's fields, they also miss out on the opportunity to build connections with others who may go on to specialize in those areas". We expect that there is plenty of room for many explanations to be correct, as this issue is multifaceted.

In this work, we aim to quantify two potential reasons for the lack of usage of NLP systems in real-world LD projects and to compare situations across countries. Specifically, we ask the following research questions (RQs): (1a) How many top-25 universities offer graduate programs in which students can learn about both NLP and LD? (1b) How does the answer to the aforementioned question differ across countries? (2a) What percentage of papers on NLP for LD are the result of truly interdisciplinary collaborations between NLP re-

[1] https://www.eafit.edu.co/noticias/eleafitense/113/universo-linguistico-colombiano-68-idiomas-propios

[2] For an early example, see Kuhn and Mateo-Toledo (2004).

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 1–6
June 21, 2024 ©2024 Association for Computational Linguistics

searchers and documentary linguists? (2b) How does the answer to the aforementioned question differ across countries? (3) Finally, is there a connection between the answers to (1a) and (2a)?

To answer our RQs, we use publicly available data from two sources: graduate program curricula and academic publications. We treat each *country* as as an individual unit: quantities we gather are aggregated per country before we proceed with analysis. Aggregation at any smaller unit (e.g., at the university or individual level) would make data collection impractical, and while it is true that countries are not monolithic with respect to curricular offerings or publishing cultures, we observe that these differences are in sum much more pronounced between rather than within countries.

2 University Curricula

We examine five countries: the United States, Germany, Brazil, Mexico, and Colombia. We choose the United States and Germany because of their prevalence in AI publication venues and because their academic cultures are quite distinct. We additionally choose Brazil, Mexico, and Colombia, as these countries, like much of Latin America, have many indigenous languages.

For each country, we consider the 25 top-ranked universities according to QS World University Rankings 2024.[3] For each university, we then determine whether it offers a graduate program in computer science (CS) or linguistics (Ling). We define a "graduate program" as anything that is at least partially beyond the scope of a United States bachelor's degree: any MS or PhD program would qualify, though some degree programs such as *licenciaturas* vary in whether they include graduate-level training, and we examine their curricula on an individual basis.

As for whether a program qualifies as "computer science" or "linguistics", we would like to capture the programs that have the highest densities of NLP researchers and documentary linguists. To this end, we define a "computer science" program as any program that has "natural language processing" or "computational linguistics" in its name, or has a graduate course in algorithms; and we define a "linguistics" program as any program that offers at least one graduate course in theoretical linguistics.[4]

[3] https://www.topuniversities.com/world-university-rankings

[4] We do not consider whether the program contains the word "linguistics", as this would include programs that are primarily focused on language teaching or learning and would be unlikely to host a documentary linguist student.

Country	CS Dept	Ling Dept	Both Depts	NLP in CS, LD in Ling	LD in CS	NLP in Ling	NLP & LD in Ling
USA	25	17	17	12	0	12	9
Germany	25	20	20	1	0	6	0
Mexico	12	3	2	0	0	1	0
Brazil	23	12	12	1	0	2	1
Colombia	8	4	3	0	0	1	0

Table 1: University curriculum data by country. Among the 25 top-ranked universities in each country, the columns display the number of universities which have a qualifying computer science program; have a qualifying linguistics program; have both programs; have both programs *and* offer both an NLP course in the computer science program and a LD course in the linguistics program; have an LD course in the computer science program; have an NLP/CL course in the linguistics program; and have both an NLP/CL and an LD course in the linguistics program.

For each eligible program, we determine whether it offers coursework in NLP/CL or LD. For a CS department, an NLP course must be dedicated to just NLP (an introduction to AI with a couple of weeks introducing NLP does not qualify) and a LD course must either include real LD or study systems which are explicitly intended for use in LD settings. For a linguistics department, an NLP/CL course should cover the use and/or development of NLP systems which can automatically perform linguistic analysis (such as finite-state automata, PoS taggers, or parsers) or modern NLP, and a LD course should be structured as a typical field methods course where students document a language through the full term of the course.

2.1 Results

We give a summary of our findings in Table 1. See [REDACTED] for full data.

RQ1a First, for CS departments, we can see that none of the 93 departments offered an LD-related course. Of the 56 linguistics departments, 22 offered an NLP/CL course, and 10 offered both an NLP/CL course and an LD course. This indicates low overall availability of interdisciplinary training to both populations, as it is overall not common for graduate students to take courses outside of their departments.

RQ1b Considering now the differences between countries, we can see that 9 out of 10 of the universities offering both an LD and NLP/CL course are in the US, as are 12 out of the 22 offering an NLP/CL course. We also see that 12 out of the 14 universities which have both a CS program with an NLP course and a linguistics program with an LD course are in the US, which we view as potentially facilitative of interdisciplinary collaborations. Thus while rates of interdisciplinary training and contact are low overall, they are comparatively higher for the US, and if the view that this ought to encourage collaboration is correct, then we should expect to see higher rates of interdisciplinary publications among works from the US (cf. RQ3).

3 Publications

We collect a large number of publications, each with the following annotations:

1. Relevance – whether the work's core contribution is a resource or system that *could* directly aid the efforts of documentation projects. We operationalize this requirement in two ways. First, a relevant paper must use at least one dataset that is an order of magnitude smaller (by tokens/hours) or more than typical high-resource datasets for the task, and the language of this dataset must be unrepresented among these high-resource datasets. (For example, a work on Universal Dependencies parsing that used the Thai treebank would count, because at 22K tokens, the Thai treebank is over an order magnitude smaller than a typical English treebank, EWT, which has 250K tokens.) We further require that a relevant paper's task be one that has direct relevance to LD activities, such as morphological parsing or machine translation.

2. Country – the country with the most representation among the authors' organizational affiliations. (We do not consider authors' nationalities, whatever they may be—only their institutional affiliations at the time of the work's publication.) If there is a tie, we take the country of the first author.

3. Documentation as Purpose (DaP) – whether LD was explicitly mentioned in the paper as a motivation for the work.

4. Performance of Documentation (PoD) – whether the collection of novel documentary data was a part of the work, where "collection" means the creation of digital primary language data that did not exist before the work.

5. Interdisciplinarity (Int) – whether the author list contains at least one NLP researcher and one documentary linguist. Any individual researcher may belong to at most one of these groups. Authors are assessed on the basis of what venues they have published in: typical NLP venues include ACL conferences, and typical documentary linguistics venues include LD&C and ICLDC.

The population of relevant (as defined above) papers is diverse and distributed throughout many publication venues, which makes it non-trivial to sample from it. We employ two resources for gathering data which have complementary strengths: the ACL Anthology,[5] a machine-readable repository of publications from venues associated with the Association for Computational Linguistics, and Semantic Scholar,[6] an academic publication aggregator with advanced querying capabilities.

AmericasNLP & ComputEL The first part of our data comes from two venues contained in the ACL Anthology which we identify as having the highest potential density of relevant papers of any publication outlet we are aware of. These are the AmericasNLP workshop[7] and the ComputEL workshop.[8] All documents which belong to one of our target countries are annotated. This dataset is useful because of its concentration of highly relevant papers, but its weakness is that it is biased heavily towards the relevant papers that are most concerned with LD as a primary goal.

Semantic Scholar The second part of our data comes from Semantic Scholar's bulk search feature, which we use to find documents which contain at least one keyword related to LD and at least one keyword related to NLP.[9] Results are shuffled, and the first 50 relevant papers for each of our target countries are annotated. This dataset is useful because it ought to offer a wider view of relevant papers, but its weakness is that its keyword-based approach likely excludes many relevant papers.

3.1 Results

For any one of the Latin American countries we consider in the previous section, we are unable to find more than 5 relevant papers despite an exhaustive review of the over 3,000 publications that were

[5] https://aclanthology.org/
[6] https://www.semanticscholar.org/
[7] https://aclanthology.org/venues/americasnlp/
[8] https://aclanthology.org/venues/computel/
[9] See Appendix A for details.

Country	DaP	PoD	Int	Total
ComputEL & AmericasNLP				
USA	37	11	22	48
Germany	3	1	1	5
Total	*40*	*12*	*23*	*53*
Semantic Scholar				
USA	18	6	8	50
Germany	4	3	5	50
Total	*22*	*9*	*13*	*100*

Table 2: The number of relevant papers for each country (Total) which respectively had documentation as an explicit purpose (DaP), actually performed documentation (PoD), and had an interdisciplinary authorship (Int).

returned by our query, and we therefore consider only Germany and the United States in this section. We give a summary in Table 2, and all data is publicly available at [REDACTED].

RQ2a Looking at the *Total* rows in Table 2, we see that less than half of all relevant papers are the results of truly interdisciplinary collaborations – for Semantic Scholar, as little as 13%. While not all papers that could be relevant for LD necessarily benefit from being interdisciplinary, we claim that this is desirable at least for papers that cite LD as their main motivation. As those papers number 62 overall, while only 36 are interdisciplinary, we find the latter number to be unfortunately small. This shows that there is much room for growth in the formation of interdisciplinary collaborations.

RQ2b For the ACL Anthology data, the United States has significantly more representation than Germany, and around 80% of works name LD as an explicit goal. Curiously, a large but smaller number of works are interdisciplinary, which could be interpreted as evidence of a degree of awareness within the NLP community in the United States of the need for NLP in LD.

A different but consistent picture emerges in the S2 data. While many American publications still cite documentation as a motivation, the proportion is smaller, and the number of interdisciplinary authorships is also smaller. This corroborates our initial conjecture that ComputEL and AmericasNLP papers would be disproportionately focused on documentation relative to the population of relevant papers as a whole. Fewer than 10% of German papers cite documentation as a purpose or have an interdisciplinary team.

RQ3 Unfortunately, the amount of papers we are able to find, especially for Latin American countries, is too small to give a definite answer to RQ3. However, the fact that more US-based than Germany-based researchers motivate their work with LD and the larger number of interdisciplinary paper collaborations could be interpreted as evidence of a higher degree of awareness of LD challenges in the NLP community as well as a larger number of LD researchers who are aware of NLP. This, in turn, could potentially stem from more readily accessible education on LD as well as from programs that offer courses in both LD and NLP.

4 Conclusion

We have presented what is to our knowledge the first evidence that provides an empirical understanding of two factors in the adoption of language technologies in LD: university curricula and collaboration trends between NLP researchers and documentary linguists. Our data confirms previous claims that rates of interdisciplinary training and collaboration are low, even for work that cites application in language documentation as a motivation.

Moreover, while the scale of our data precludes a firm conclusion, it is consistent with the claim that interdisciplinary coursework is a partial determinant of collaboration rates, as the higher rates of interdisciplinary course offerings in the United States (relative to Germany) are mirrored by higher rates of interdisciplinary publishing by authors working in the United States. This broadly supports the view that interdisciplinary graduate coursework is important for supporting the incorporation of human language technologies into LD practice. More evidence is needed, however, in order to investigate other possible factors: perhaps other influences, such as nation-level grant programs or academic cultures, are directly affecting both curricula and rates of interdisciplinary collaborations.

We therefore join Flavelle and Lachler (2023) in identifying interdisciplinary curricular offerings as an important way for the NLP and linguistics communities to work towards the ultimate goal of aiding LD with language technologies. Additionally, we observe that many of the same benefits could be gained from interdisciplinary workshops such as the LTLDR workshop (Neubig et al., 2020), which gathered documentary linguists, NLP researchers, and community members for the explicit purpose of fostering interdisciplinary collaborations.

Limitations

Our findings are limited by the quantity of data that we have collected and the methods we used to sample the data points that we have. For universities, this comes out in our selection of 5 particular countries and our consideration of 25 universities from each, as we were unable to include more countries and universities given the high time cost of annotating a single university. For publication data, this is instantiated in our two methods for collecting papers which, as we described, we expect introduced sampling bias, though these two methods seemed that they would introduce the least sampling bias of any of the other methods we considered while still remaining practical to perform.

References

Lindell Bromham, Russell Dinnage, Hedvig Skirgård, Andrew Ritchie, Marcel Cardillo, Felicity Meakins, Simon Greenhill, and Xia Hua. 2022. Global predictors of language endangerment and the future of linguistic diversity. *Nature ecology & evolution*, 6(2):163–173.

Darren Flavelle and Jordan Lachler. 2023. Strengthening relationships between indigenous communities, documentary linguists, and computational linguists in the era of NLP-assisted language revitalization. In *Proceedings of the First Workshop on Cross-Cultural Considerations in NLP (C3NLP)*, pages 25–34, Dubrovnik, Croatia. Association for Computational Linguistics.

Ben Foley, Josh Arnold, Rolando Coto-Solano, Gautier Durantin, T. Mark Ellison, Daan van Esch, Scott Heath, František Kratochvíl, Zara Maxwell-Smith, David Nash, Ola Olsson, Mark Richards, Nay San, Hywel Stoakes, Nick Thieberger, and Janet Wiles. 2018. Building Speech Recognition Systems for Language Documentation: The CoEDL Endangered Language Pipeline and Inference System (ELPIS). In *6th Workshop on Spoken Language Technologies for Under-Resourced Languages (SLTU 2018)*, pages 205–209. ISCA.

Luke Gessler. 2022. Closing the NLP gap: Documentary linguistics and NLP need a shared software infrastructure. In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 119–126, Dublin, Ireland. Association for Computational Linguistics.

Jeff Good, Julia Hirschberg, and Owen Rambow, editors. 2014. *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*. Association for Computational Linguistics, Baltimore, Maryland, USA.

Jonas Kuhn and B'alam Mateo-Toledo. 2004. Applying computational linguistic techniques in a documentary project for Q'anjob'al (Mayan, Guatemala). In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Graham Neubig, Shruti Rijhwani, Alexis Palmer, Jordan MacKenzie, Hilaria Cruz, Xinjian Li, Matthew Lee, Aditi Chaudhary, Luke Gessler, Steven Abney, Shirley Anugrah Hayati, Antonios Anastasopoulos, Olga Zamaraeva, Emily Prud'hommeaux, Jennette Child, Sara Child, Rebecca Knowles, Sarah Moeller, Jeffrey Micher, Yiyuan Li, Sydney Zink, Mengzhou Xia, Roshan S Sharma, and Patrick Littell. 2020. A summary of the first workshop on language technology for language documentation and revitalization. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 342–351, Marseille, France. European Language Resources association.

A Search Criteria

We use Semantic Scholar's bulk search API[10], which accepts queries in a rich structured format which features several operators which form trees over keyword arguments. Our query is provided below in an abstract syntax tree. The two main parts of it contain keywords related to language documentation and NLP models, respectively. The tilde operator `x~n` specifies that up to `n` words may intervene between the words in `x`. Both keyword lists are joined with the logical or operator `|` which is satisfied if any one of the keyword expressions are satisfied, and both keyword lists are finally joined with the logical and operator `+` which is satisfied only if both subexpressions are satisfied.

```
[
    "+",
    [
        '|',
        '"low-resource"',
        '"low resource"~1',
        '"less-resourced"',
        '"less resourced"~1',
        '"under-resourced"',
        '"under resourced"~1',
        '"under-studied"',
        '"under studied"~1',
        '"less-studied"',
```

[10] https://api.semanticscholar.org/api-docs/graph#tag/Paper-Data/operation/get_graph_paper_bulk_search

```
        '"less studied"~1',
        '"endangered language"~1',
        '"indigenous language"~1',
        '"language documentation"',
        '"document language"',
        '"language revitalization"',
        '"revitalize language"',
        '"language maintenance"',
        '"maintain language"',
        '"language revival"',
        '"revive language"',
        '"ELAN"',
        '"FLEx"',
        '"FieldWorks Language Explorer"',
        '"LingSync"',
        'typological',
    ],
    [
        '|',
        'model',
        'resource',
        'lexicon',
        'parser',
        'corpus',
        'dataset',
        'document',
        'dictionary',
        'grammar',
        'segmentation',
        'orthographic',
        'normalization',
        'evaluation',
        'experiments',
        '"machine translation"',
        '"automatic translation"',
        'predict',
        'neural'
    ]
]
```

Translation systems for low-resource Colombian Indigenous languages, a first step towards cultural preservation

Juan C. Prieto , Cristian A. Martinez, Melissa Robles, Alberto Moreno, Sara Palacios, Rubén Manrique

The Department of Systems and Computing Engineering
Universidad de Los Andes
{jc.prietoa,ca.martinez2,mv.robles,a.morenoc23,s.palaciosc,rf.manrique}@uniandes.edu.co

Abstract

The use of machine learning and Natural Language Processing (NLP) technologies can assist in the preservation and revitalization of indigenous languages, particularly those classified as "low-resource". Given the increasing digitization of information, the development of translation tools for these languages is of significant importance. These tools not only facilitate better access to digital resources for indigenous communities but also stimulate language preservation efforts and potentially foster more inclusive, equitable societies, as demonstrated by the AmericasNLP workshop since 2021. The focus of this paper is Colombia, a country home to 65 distinct indigenous languages, presenting a vast spectrum of linguistic characteristics. This cultural and linguistic diversity is an inherent pillar of the nation's identity, and safeguarding it has been increasingly challenging given the dwindling number of native speakers and the communities' inclination towards oral traditions. Considering this context, scattered initiatives exist to develop translation systems for these languages. However, these endeavors suffer from a lack of consolidated, comparable data. This paper consolidates a dataset of parallel data in four Colombian indigenous languages - Wayuunaiki, Arhuaco, Inga, and Nasa - gathered from existing digital resources. It also presents the creation of baseline models for future translation and comparison, ultimately serving as a catalyst for incorporating more digital resources progressively.

1 Introduction

In the field of natural language processing (NLP), low-resource languages are characterized by limited written or spoken digital material. Consequently, applying translation models rooted in advanced neural architectures to these languages is challenging due to the models' high dependence on substantial data volumes (Wang et al., 2021). However, recent years have seen a growing trend towards working with low-resource languages for machine translation based on transformers. Notably, efforts to preserve indigenous languages have significantly contributed to advancements in this area (Mager et al., 2018; Ortega et al., 2020a,b; Chen and Fazio, 2021). The Americas, in particular, host numerous endangered indigenous languages spoken by small populations. In response, several researchers have devoted their work towards developing translation models for some of these languages (Ngoc Le and Sadat, 2020). Highlighting this effort, AmericasNLP was convened in 2021 as the first global workshop dedicated to the application of NLP to American indigenous languages (Mager et al., 2021).

The development of translation tools catering specifically to indigenous languages has the potential to confer numerous advantages, including enriched access to digital resources and the promotion of language preservation efforts. As the availability of digital resources proliferates, it becomes increasingly imperative for indigenous communities to have access to information in their respective languages, in order to safeguard their unique cultures and traditional ways of living. Regrettably, the limited resources available in indigenous languages restrict their accessibility to crucial digital materials. The creation of translation tools could serve as a resolution to this issue, enabling communities to translate digital resources into their languages. Beyond simply providing improved access to information, these tools could stimulate language preservation by offering a medium for language revitalization. Furthermore, translation tools could provide a platform for indigenous communities to engage with the global community, contributing to their economic and social development. By fostering cross-cultural understanding and mutual respect, translation tools could play a critical role in constructing more inclusive and equitable societies.

As acknowledged by the National Indigenous

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 7–14
June 21, 2024

Organization of Colombia (ONIC) (de Gobierno Indígena – ONIC, 2015), Colombia's linguistic diversity is marked by the existence of 65 indigenous languages, alongside Spanish and two Creole languages. Among the 32 different departments in the territory, the regions of Amazonas and Vaupés, located in the southern sector of Colombia, stand out for their significant diversity of indigenous languages. This assortment of indigenous languages in Colombia is highly distinctive due to their different characteristics. For instance, Colombia accommodates tonal languages similar to those found in Southeast Asia and Central Africa, along with languages representing all four classic morphological types. These include inflectional (as exemplified by Kogui and Arhuaco), agglutinative (like Achagua, Andoque, and Páez), isolating akin to the Malayo-Polynesian languages (such as Embera and Creole), and polysynthetic (such as Kamsá) (de Estudios de Lenguas Aborígenes , C.C.E.L.A).

According to ONIC, the task of preserving these indigenous voices is increasingly formidable, primarily due to their endangerment amid the dwindling number of native speakers. Over half of these languages, alarmingly, have fewer than a thousand active speakers, thus exacerbating their preservation and conservation challenges (de Gobierno Indígena – ONIC, 2015). Further complexities arise from the proclivity of certain communities to uphold their oral traditions over written modes. Take, for example, the Inga community, wherein cultural identity preservation is embedded in the prioritization of oral tradition as the primary vehicle for knowledge transmission. Such communities contend that the detachment from oral traditions can incite a loss of numerous practices that necessitate face-to-face interaction and stimulate dialogue (Rodríguez and Narváez, 2022).

Despite the existence of scattered efforts aimed at creating translation systems for some indigenous Colombian languages, we found no consolidated data that allows for reuse and/or comparison (Sierra et al., 2015; Sierra Martínez et al., 2016, 2018; Fernandez et al., 2013). Only recently have initiatives emerged that have promoted data replicability and openness in Colombian Indigenous languages (Graichen et al., 2023). The primary objective of this work is to consolidate a dataset of parallel data in four Colombian indigenous languages: Wayuunaiki, Arhuaco, Inga, and Nasa. This dataset is a product of the compilation, processing, and alignment of already existing digital resources in these languages. It aims to serve as a starting point to encourage the incorporation of new digital resources progressively. A second objective is to create a set of baseline models for translation that can be used for comparison in future research.

2 Related Work

Low-resource languages, which lack the digital or written material needed to build a corpus or a linguistic collection, include indigenous or endangered languages, region-specific dialects, or languages without substantial digital resources despite the existence of millions of speakers. The shortage of available data, which often results from limited access to technology, opens up opportunities to apply various techniques like Data Augmentation, Back-Translation, and Transfer Learning to mitigate this scarcity.

Studies on diverse languages have explored Neural Machine Translation, Transfer Learning, and advanced model architectures. For instance, an investigation of the Tigrinya Ethiopian language yielded positive results using Transfer Learning (Öktem et al., 2020). In another study comparing three different models, the Transformer Network performed the best with a parallel corpus of Yoruba and English (Adebara et al., 2021). Also, a modification of the Transformer architecture led to improved results for South African languages (van Biljon et al., 2020). Numerous studies demonstrate the effectiveness of Transfer Learning and advanced modeling techniques across global languages. For instance, Finnish and Czech were used as parent languages to assist low-resource Estonian and Slovak languages via Transfer Learning, leading to improvements over the baseline for almost all pre-trained models (Kocmi and Bojar, 2018). The combination of attention layers and byte-pair encoding in Transfer Learning also notably enhanced translation capabilities for Turkic languages (Nguyen and Chiang, 2017). Notably, the pairing of the Transformer architecture with the Back-Translation technique resulted in improved translation quality for several language pairs (Przystupa and Abdul-Mageed, 2019).

American Indigenous Languages, historically deficient in written records, pose unique challenges for language preservation. The AmericasNLP workshop provided a platform to unite global research groups to address these challenges, with a focus on machine translation across various indigenous languages. Numerous techniques and strategies were employed across the participating teams in the different versions, yielding promising results. While some teams achieved success with multilingual neural networks (Vázquez et al., 2021; Knowles et al., 2021; Moreno, 2021), others found Statistical Machine Translation more effective (Parida et al., 2021). Additionally, using unique sources such as Wikipedia and biblical texts to build the corpus yielded significant results, enhancing progress beyond baseline starts (Billah-Nagoudi et al., 2021). In the recent 2023 edition of the Workshop, the winning team's shared-task strategy comprised extending and finetuning several variants of the NLLB-200 (NLLBTeam, 2022). This cutting-edge machine translation model is specifically tailored for environments with scarce resources. Their submission surpassed the baseline by an average chrF of 11% across all languages, yielding especially considerable enhancements for Aymara, Guarani, and Quechua (Gow-Smith and Sánchez Villegas, 2023).

As for the indigenous Colombian languages' translations, Graichen et al. (2023) and Robles et al. (2024) studies are the only works we discovered that present a translation system from Wayuunaiki to Spanish and Ika (Arhuaco) to Spanish. Graichen et al. (2023) applied various unsupervised and semisupervised subword segmentation methods to enrich the data used to train a transformer-based model with linguistic information. According to the results, the incorporation of linguistic knowledge helps the system to generate improved translation. Nonetheless, these methodologies introduced substantial noise into the process.

3 Data

3.1 Langauges

The linguistic diversity of Colombia is characterized by a variety of indigenous languages, including Wayuunaiki, Nasa Yuwe, Arhuaco (Ika), and Inga. Wayuunaiki, predominantly spoken by the Wayuú community in the La Guajira region, is the most widely spoken indigenous language in Colombia. The 2005 DANE census report (de Gobierno Indígena – ONIC, 2015) indicates a population of 270.413 Wayuú individuals, making it the largest indigenous demographic. Wayuunaiki is an agglutinative language, characterized by the combination of independent morphemes to form words.

On the other hand, Nasa Yuwe, primarily spoken by the Nasa people in the Cauca department and smaller regions such as Valle del Cauca, Tolima, and Huila, is the second most spoken indigenous language. Although traditionally classified as part of the Chibchan language family, it is now largely considered an isolated language.

Likewise, the Arhuacos, who inhabit the western and southeastern regions of the Sierra Nevada de Santa Marta, speak the Ika language. Ika, a member of the Chibchan language family, is distinguished by its sentence structure, which involves the addition of various morphemes to a root or lexeme.

Lastly, the Inga community, descendants of the Inca civilization, primarily inhabit the Sibundoy Valley within the Putumayo region, with additional settlements in Nariño and Cauca. Their language, Inga, belongs to the Quechuan family. The linguistic diversity of these communities contributes to the rich cultural tapestry of Colombia.

3.2 Data Collection

Locating documents written in both indigenous languages and Spanish is challenging due to a lack of translated resources, as shown by limited translations of the constitution. The Colombian Center for Studies in Aboriginal Languages (de Estudios de Lenguas Aborígenes, 1994) has only translated the constitution into seven indigenous languages: Inga, Guambiano, Arhuaco, Kamentsa, Kubeo, Nasa Yuwe, and Wayuunaiki. This lack of translated resources extends to religious texts as well. For instance, complete translations of the Bible are only available in a handful of indigenous languages. Specifically, complete or partial translations of the Bible exist in Wayuunaiki, Arhuaco, and Nasa Yuwe.

For the Wayuunaiki language, in addition to the Bible and the constitution, there are various documents that delve into the characteristics of the language and provide sections with translated excerpts. An example is the document "La conjugación del verbo en la lengua Wayuu" (Álvarez, 2016) which offers a comparative perspective on verbal conjugation in Wayuunaiki and Spanish, addressing semantic, morphological, and syntactic aspects. Similar to this, the book "Compendio de la Gramática de la Lengua Wayuu" (Álvarez, 2017) and the article "Panorámica de la fonología y morfología de la lengua Wayuu" (Álvarez González, 2021) detail the important features of the morphology, phonology, and syntax of the language, presenting comparative examples between the two languages. The book "Vamos a hablar nuestra lengua" (Flórez et al., 2020) provides accurate information about writing, grammar, and cultural aspects implicit in everyday expressions and words. The consolidation of this language dataset was ultimately achieved by utilizing a short story (Cue, 2012) and a dictionary (Amaya, 2021). This last resource, compiled in 2021 by Rafael Jose Negrette Amaya, encompasses a total of 74.583 translated phrases and words in both languages.

To consolidate the Wayunnaiki dataset, sentences were extracted from PDFs or web pages. The dataset for the New Testament of the Bible (YouVersion, 2023) was constructed using a web scraping process. This entailed systematically pairing sentences by verse and chapter of the book. In some of the other sources, we relied on additional pre-processing steps via Large Language Models (LLMs). Since the texts did not follow a defined format, GPT-4 (OpenAI, 2023) was employed to extract candidate texts from the selected documents. A prompt template was utilized to identify and tabulate sentences in both Wayuu and Spanish languages, with this process being conducted at three-page intervals. An illustrative example of the prompts used is, "`Identify sections in the text where Wayuu and Spanish sentences co-occur and create a tabulated representation...`". Then a manual review process was carried out, filtering incomplete translations, as the Spanish sentence contained blank spaces or non-alphabetic characters. The combined use of web scraping and GPT-4 in this manner allowed for the creation of a comprehensive and well-structured dataset, thereby enhancing the overall readability and coherence of the information.

The primary data source for Nasa Yuwe was the constitution (de Estudios de Lenguas Aborígenes, 1994). This document is partitioned into sections: introductory letters, articles, and a dictionary that is a compendium of frequently translated words and phrases from Nasa Yuwe to Spanish. An Optical Character Recognition (OCR) (Smith et al., 2009) process was employed to extract the 23 translated articles, along with introductory letters and acknowledgments. Discrepancies were observed in the introductory letters as the Nasa Yuwe translation occasionally contained more content than the Spanish version. A manual review of the letter text was necessitated to pinpoint precise word translations and sentence terminations and to eliminate any additional Nasa Yuwe content not found in the Spanish text. Additionally, a dictionary (originarios. Lenguas de América) containing 3.729 words and brief phrases in both Spanish and Nasa Yuwe was included. This dictionary is presented in HTML format and was processed using the Beautiful Soup tool (Richardson, 2007), a web scraping library. This was followed by a manual error correction procedure to guarantee the precision of the extraction process further.

For the Arhuaco our biggest data source was the Bible. We again used web scraping and the BeautifulSoup library on selected chapters of the Old and New Testaments (para el Desarrollo de Pueblos Marginados). Our second source for Arhuaco was the constitution (de Estudios de Lenguas Aborígenes, 1994). Given the low quality of the online document, it was processed through a text identification procedure with the assistance of Google's DocumentAI OCR. This particular API employs a neural network designed to enhance the recognition of text within PDF documents, which facilitates the conversion of visual document data into text, organizing the content into distinct paragraphs. This segmentation significantly simplifies the subsequent concatenation of sentences, thereby streamlining the text analysis and processing tasks. The initial phase of the processing involved the analysis of introductory letters, which were messages in either Spanish or Arhuaco, expressing gratitude, detailing efforts, and explaining the reasons behind the creation of the book. The document was processed in blocks of text, and sentences were consolidated by matching paragraphs and sentences separated by period. After the letters, we moved to process the constitution articles, which involved creating a correspondence between titles in Spanish and Arhuaco and identifying the pairs. Finally, a manual pairing process was carried out for the dictionary section of the constitution. The last source used for Arhuaco was a book titled "Cantando desde la Sierra" (de la comunidad arhuaca de Jewrwa, 2014), which contains various short stories and poems in Arhuaco, accompanied by their Spanish translations. To utilize this, each poem was detected and subsequently matched with its corresponding translation in the other language.

The Inga language was undoubtedly the most challenging. We employ as a primary source a comprehensive dictionary of words and phrases (de Educación Inga de la Organización “Musu Runakuna”, 1997). This dictionary was processed using a methodology similar to the one used for the Wayuunaiki language. This involved the use of the GPT-4 (OpenAI, 2023) for identifying candidate pairs translations within the dictionary. Subsequently, a further step of data cleaning was undertaken to minimize the occurrence of false positives. The second source was the constitution (de Estudios de Lenguas Aborígenes, 1994), which was processed using the same procedure used for Arhuaco, using an Google OCR and a manual cleaning step. Table 1 shows the size of the training data for each language.

4 Baseline construction

Following the recent results obtained by the work of (Robles et al., 2024) and (Gow-Smith and Sánchez Villegas, 2023) we use NLLB-200, a state-of-the-art machine translation model specifically designed for low-resource settings. We experiment with different distilled versions of NLLB-200 with 600M and 1.3B parameters. Each dataset was randomly divided into training, validation, and testing sets, comprising 80%, 10%, and 10% of the total number of sentences, respectively. The resulting partition, along with the models and code, can be accessed at `https://github.com/juanks235/MT-Colombian-Indigenous-Languages`.

4.1 Experimental Setup

In our approach, we execute model training distinctively for each language pair available in our dataset. We refine the embedding matrix to encompass tags for newly added languages, scrutinizing for not recognized tokens and employing text normalization to reduce potential problems related to unrecognized punctuation. The application of normalization ensures the accurate processing of the text, obviating any unknown tokens and providing the promise that the vocabulary of the tokenizer doesn’t require an update for the target language. Nevertheless, if it becomes necessary for the tokenizer’s vocabulary to be updated, we implement an update to include any new or unrecognized tokens previously overlooked. Our experimental framework operates on four A40-48GB, with a batch size of 16, 1000 warmup steps, 57000 training steps, featuring a learning rate of $1e-4$ and a weight decay of $1e-3$. For automated evaluations, we leverage SacreBLEU (Post, 2018) for computing BLEU scores and chrF2++ (Popović, 2017) to measure chrF2.

4.2 Results

Table 2 presents the results obtained with different trained models. As expected, the 1.3B model performed better in translations from Spanish to the target language.

The influence of the inclusion of dictionary data was evaluated for the Arahauco language due to the difficulty of its extraction. The results of the models using all data and excluding the dictionary generally show a low contribution from the dictionary. For Arhuaco, the best model achieved a BLEU of 7.72 and a chrF2 of 24.17 (spanish to target).

A similar evaluation was carried out for NASA. However, in this case, the more challenging extraction process was with letters from the constitution. Therefore, the model trained with all data and the model trained without constitutional letters were evaluated. The results in terms of BLEU suggest a little contribution from the letters, likely due to difficulties encountered in the alignment process due to enriched translations for the indigenous community.

In Wayuu, we evaluated the use of all data and without the dictionary, which corresponds to the largest fragment of the dataset. Contrary to what was expected, the inclusion of the dictionary did not have a positive impact on the results. In fact, the highest scores for the BLEU metric were achieved without the use of this data source, while the ChrF scores did not show a significant difference. Therefore, the most effective model was the one trained without the dictionary using the 1.3B model, which achieved a BLEU score of 15.37 and a ChrF2 score of 32.06.

Finally, for Inga, we found it to be the most challenging language, both for data collection and for the translation model. Our most successful model yielded a BLEU score of 1.71 and a ChrF2 score of 18.40, achieved without the utilization of the dictionary. However, the dictionary represented the largest dataset, and this experiment only considered 212 sentences. The built models can be downloaded and accessed from the repository, hoping to constitute a baseline for future efforts in

Language	Description	Sentences
Wayuunaiki	Dictionary (Amaya, 2021)	74583
Wayuunaiki	Bible (YouVersion, 2023)	6220
Wayuunaiki	Book (Álvarez, 2017)	534
Wayuunaiki	Book (Flórez et al., 2020)	467
Wayuunaiki	Book (Álvarez González, 2021)	229
Wayuunaiki	Book (Álvarez, 2016)	109
Wayuunaiki	Short story (Cue, 2012)	39
Wayuunaiki	Constitution (de Estudios de Lenguas Aborígenes, 1994)	37
Nasa Yuwe	Dictionary (originarios. Lenguas de América)	3729
Nasa Yuwe	Letters (Constitution) (de Estudios de Lenguas Aborígenes, 1994)	57
Nasa Yuwe	Common words (Constitution) (de Estudios de Lenguas Aborígenes, 1994)	53
Nasa Yuwe	Articles (Constitution) (de Estudios de Lenguas Aborígenes, 1994)	23
Arhuaco	Bible (para el Desarrollo de Pueblos Marginados)	5542
Arhuaco	Articles (Constitution) (de Estudios de Lenguas Aborígenes, 1994)	88
Arhuaco	Letters (Constitution) (de Estudios de Lenguas Aborígenes, 1994)	67
Arhuaco	Dictionary (Constitution) (de Estudios de Lenguas Aborígenes, 1994)	46
Arhuaco	Short stories (de la comunidad arhuaca de Jewrwa, 2014)	42
Inga	Dictionary (de Educación Inga de la Organización "Musu Runakuna", 1997)	3048
Inga	Constitution (de Estudios de Lenguas Aborígenes, 1994)	212

Table 1: Parallel data collected for each language

these languages.

5 Conclusion and Future Work

The preservation of indigenous languages, encompassing their stories, wisdom, and traditions is instrumental in fostering cross-cultural understanding. However, working with low-resource languages such as these often presents unique challenges, particularly in regions like Colombia, which are teeming with linguistic diversity. We constructed a dataset of parallel data in four indigenous Colombian languages, and the resulting dataset is freely accessible and usable for future research projects. Additionally, we developed baseline translation models for each language pair. Our findings demonstrated that the NLLB 1.3B model excelled overall in comparison to the 600M model as expected. Also, a contrast emerged in the range of the BLEU score: from as low as 1.71 (Inga) to as high as 15.37 (Wayuu). Such a significant difference is attributable to the disparities in data volume, with Inga being the most challenging language. We also tested the influence of the inclusion in the training data of some of the sources in particular those that were challenging in the extraction phase. Although this project did not involve direct engagement with community members, future work will prioritize establishing connections with these communities to expand the dataset and evaluate translation systems more thoroughly. Our focus will be on incorporating additional Indigenous Colombian languages and exploring alternative models or architectures to potentially enhance translation outcomes.

6 Acknowledgements

Melissa Robles acknowledges the support of a UniAndes-GoogleDeepMind Scholarship 2023.

References

2012. *Putunkaa Serruma: Duérmete, pajarito blanco. Arrullos y relatos indígenas de cinco etnias colombianas*. Instituto Colombiano de Bienestar Familiar ICBF.

Ife Adebara, Muhammad Abdul-Mageed, and Miikka Silfverberg. 2021. Translating the unseen? yoruba-english mt in low-resource, morphologically-unmarked settings.

Rafael Jose Negrette Amaya. 2021. Osf spanish-wayuunaki.

El Moatez Billah-Nagoudi, Wei-Rui Chen, Muhammad Abdul-Mageed, and Hasan Cavusoglu. 2021. IndT5: A Text-to-Text Transformer for 10 Indigenous Languages. In *Proceedings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas*. Association for Computational Linguistics.

William Chen and Brett Fazio. 2021. Morphologically-guided segmentation for translation of agglutinative low-resource languages.

Comité de Educación Inga de la Organización "Musu Runakuna". 1997. *Diccionario Inga*.

Centro Colombiano de Estudios de Lenguas Aborígenes. 1994. *Constitución Política de 1991 traducida a Lenguas Indígenas*.

Target Language	Data	NLLB - Model	Spanish - Target		Target - Spanish	
			BLEU	chrF2	BLEU	chrF2
Arhuaco	All	600M	6.00	23.04	8.29	33.28
Arhuaco	All	1.3B	7.72	**24.17**	8.29	32.98
Arhuaco	Without Dict	600M	7.28	23.06	8.29	32.95
Arhuaco	Without Dict	1.3B	**8.19**	23.08	**8.75**	**33.38**
Nasa	All	600M	2.65	18.18	4.02	**18.70**
Nasa	All	1.3B	2.70	**18.98**	2.22	17.92
Nasa	Without Letters	600M	3.02	15.99	**4.78**	18.21
Nasa	Without Letters	1.3B	**3.87**	15.68	3.19	16.60
Wayuu	All	600M	11.89	30.94	12.50	39.10
Wayuu	All	1.3B	13.81	**32.62**	14.19	40.51
Wayuu	Without Dict	600M	14.38	31.01	17.30	41.26
Wayuu	Without Dict	1.3B	**15.37**	32.06	**18.93**	**42.79**
Inga	All	600M	1.27	20.43	**3.08**	27.21
Inga	All	1.3B	0.78	**21.15**	1.08	27.38
Inga	Without Dict	600M	0.74	19.21	1.86	**30.00**
Inga	Without Dict	1.3B	**1.71**	18.40	1.10	29.30

Table 2: Scores (BLUE, chrF) on test partitions for all languages pairs per NLLB model

El Centro Colombiano de Estudios de Lenguas Aborígenes (C.C.E.L.A). 1994. *Estructuras sintácticas de la predicación: lenguas amerindias de Colombia*.

Autoridad Nacional de Gobierno Indígena – ONIC. 2015. *65 Lenguas Nativas de las 69 en Colombia son Indígenas*.

Alumnos de la comunidad arhuaca de Jewrwa. 2014. *Niwi úmuke pari ayunnuga, Cantando desde la Sierra*.

Dayana Fernandez, Jose Atencia, Ornela Gamboa, and Oscar Bedoya. 2013. Design and implementation of an web api for the automatic translation colombia's language pairs: Spanish-wayuunaiki case. In *Communications and Computing (COLCOM), 2013 IEEE Colombian Conference on*, pages 1–9.

Yasir Bustos Flórez, Norys Jiménez Pitre, and Delio Pontilus. 2020. *Joo'uya waashajaaiwa Wanüiki. Vamos a hablar nuestra lengua*. Institución Educativa Indígena No 4 de Maicao - sede Majayutpana.

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.

Nora Graichen, Josef Van Genabith, and Cristina España-bonet. 2023. Enriching Wayúunaiki-Spanish neural machine translation with linguistic information. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 67–83, Toronto, Canada. Association for Computational Linguistics.

Rebecca Knowles, Darlene Stewart, Samuel Larkin, and Patrick Littell. 2021. NRC-CNRC Machine Translation Systems for the 2021 AmericasNLP Shared Task. In *Proceedings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas*. Association for Computational Linguistics.

Tom Kocmi and Ondř ej Bojar. 2018. Trivial transfer learning for low-resource neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*. Association for Computational Linguistics.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Annette Rios, Ivan Vladimir Meza Ruiz, Alexis Palmer, Graham Neubig, and Katharina Kann, editors. 2021. *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*. Association for Computational Linguistics, Online.

Oscar Moreno. 2021. The REPUcs' Spanish–Quechua Submission to the AmericasNLP 2021 Shared Task on Open Machine Translation. In *Proceedings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas*, pages 241–246. Association for Computational Linguistics.

Tan Ngoc Le and Fatiha Sadat. 2020. Revitalization of indigenous languages through pre-processing and neural machine translation: The case of Inuktitut. In *Proceedings of the 28th International Conference* *on Computational Linguistics*, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Toan Q. Nguyen and David Chiang. 2017. Transfer learning across low-resource, related languages for neural machine translation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 296–301, Taipei, Taiwan. Asian Federation of Natural Language Processing.

NLLBTeam. 2022. No language left behind: Scaling human-centered machine translation.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

Pueblos originarios. Lenguas de América. Diccionario páez-español.

John E. Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020a. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

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

Fundación para el Desarrollo de Pueblos Marginados. Visor biblia iku.

Shantipriya Parida, Subhadarshi Panda, Amulya Dash, Esau Villatoro-Tello, A. Seza Doğruöz, Rosa M. Ortega-Mendoza, Amadeo Hernández, Yashvardhan Sharma, and Petr Motlicek. 2021. Open Machine Translation for Low Resource South American Languages. In *Proceedings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas*. Association for Computational Linguistics.

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

Matt Post. 2018. A call for clarity in reporting bleu scores.

Michael Przystupa and Muhammad Abdul-Mageed. 2019. Neural machine translation of low-resource and similar languages with backtranslation. In *Proceedings of the Fourth Conference on Machine Translation*, Florence, Italy. Association for Computational Linguistics.

Leonard Richardson. 2007. Beautiful soup documentation. *April*.

Melissa Robles, Cristian A. Martínez, Juan C. Prieto, Sara Palacios, and Rubén Manrique. 2024. Preserving heritage: Developing a translation tool for indigenous dialects. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining*, WSDM '24. Association for Computing Machinery.

Geraldyn Otavo Rodríguez and Melissa Lizette Portilla Narváez. 2022. *Relatos ancestrales: una alternativa para la preservación de la identidad cultural oral del territorio Inga*.

Luz Marina Sierra, Carlos Alberto Cobos, Juan Carlos Corrales, and Tulio Rojas Curieux. 2015. Building a nasa yuwe language test collection. In *Computational Linguistics and Intelligent Text Processing*, pages 112–123, Cham. Springer International Publishing.

Luz Marina Sierra Martínez, Carlos Cobos, and Juan Corrales. 2016. Tokenizer adapted for the nasa yuwe language. *Computacion y Sistemas*, 20:355–364.

Luz Marina Sierra Martínez, Carlos Alberto Cobos, Juan Carlos Corrales Muñoz, Tulio Rojas Curieux, Enrique Herrera-Viedma, and Diego Hernán Peluffo-Ordóñez. 2018. Building a Nasa Yuwe Language Corpus and Tagging with a Metaheuristic Approach. *Computación y Sistemas*, 22:881 – 894.

Ray Smith, Daria Antonova, and Dar-Shyang Lee. 2009. Adapting the tesseract open source ocr engine for multilingual ocr. In *MOCR '09: Proceedings of the International Workshop on Multilingual OCR*, ACM International Conference Proceeding Series. ACM.

Elan van Biljon, Arnu Pretorius, and Julia Kreutzer. 2020. On optimal transformer depth for low-resource language translation.

Raúl Vázquez, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2021. The Helsinki submission to the AmericasNLP shared task. In *Proceedings of the AmericasNLP 2021 Shared Task on Open Machine Translation for Indigenous Languages of the Americas*. Association for Computational Linguistics.

Rui Wang, Xu Tan, Renqian Luo, Tao Qin, and Tie-Yan Liu. 2021. A survey on low-resource neural machine translation.

YouVersion. 2023. Biblia en wayuu, https://www.bible.com/es/bible/1584/mat.1.guc.

José Álvarez. 2016. La conjugación del verbo en la lengua wayuu.

José Álvarez. 2017. Compendio de la gramática de la lengua wayuu.

José Ramón Álvarez González. 2021. Panorámica de la fonología y morfología de la lengua wayuu. 61.

Alp Öktem, Mirko Plitt, and Grace Tang. 2020. Tigrinya neural machine translation with transfer learning for humanitarian response.

Word-level prediction in Plains Cree: First steps

Olga Kriukova
Department of Linguistics,
University of Saskatchewan
Saskatoon, SK, Canada
olga.kriukova@usask.ca

Antti Arppe
Department of Linguistics,
University of Alberta
Edmonton, AB, Canada
arppe@ualberta.ca

Abstract

Plains Cree (nêhiyawêwin) is a morphologically complex and predominantly prefixing language. The combinatory potential of inflectional and derivational/lexical prefixes and verb stems in Plains Cree makes it challenging for traditional auto-completion (or word suggestion) approaches to handle. The lack of a large corpus of Plains Cree also complicates the situation. This study attempts to investigate how well a BiLSTM model trained on a small Cree corpus can handle a word suggestion task. Moreover, this study evaluates whether the use of semantically and morphosyntactically refined Word2Vec embeddings can improve the overall accuracy and quality of BiLSTM suggestions. The results show that some of the models trained with the refined vectors provide semantically and morphosyntactically better suggestions. They are also more accurate in predictions of content words. The model trained with the non-refined vectors, in contrast, was better at predicting conjunctions, particles, and other non-inflecting words. The models trained with different refined vector combinations provide the expected next word among top-10 predictions in 36.32 to 37.34% of cases (depending on the model).

1 Introduction

Auto-complete systems and predictive text input have become integral components of our daily interactions with our devices and digital platforms. These applications heavily rely on robust language models capable of accurately predicting the next word in a given sequence of text. While substantial progress has been made in developing efficient language models for major languages, the challenges persist for low-resource languages where scarcity of training data poses a significant obstacle. This challenge is especially found for Indigenous languages that are often also morphologically rich.

With advances in the NLP and machine learning fields, small training datasets have become less of a problem; however, the handling of the morphological complexity still presents a challenge. Lane and Bird (2020) approached this problem with the development of an interactive word-completion system for Kunwingku (an Indigenous language spoken in Northern Australia) based on a finite state recognizer which included most morphology for some 500 verbs. Their tool suggests a completion up to the next morpheme boundary and helps to avoid the so-called "combinatorial explosion of possible words" typical for the prefixing polysynthetic languages.

Lane et al. (2022) further successfully extend this method to Plains Cree, with a full-fledged model including all parts of speech, covering most inflectional morphology for the inflecting verbs and nouns, and based on a lexicon of well over 20k lexemes. The tool is based on a finite state morphosyntactic analyzer of Plains Cree (nêhiyawêwin, an Indigenous language spoken mainly in on the Western Canadian Plains) (Snoek et al., 2014; Harrigan et al., 2017). It uses corpus-based information about Cree prefixes to predict the most probable and common next morpheme in a word (based on a small corpus of some 150k Cree words). While the results were perceived as surprisingly good, given the small size of the corpus, there remained yet quite many valid optional completions, since their tool did not make use of preceding lexical or morphosyntactic context. Similarly, with Lane and Bird (2020), Plains Cree interactive word completion could be used by non-fluent Cree speakers and learners who may struggle to build word forms. Nevertheless, as the successful use of the model requires a broad knowledge of the language and its word formation, in order to be able to choose the completion appropriate to the context, they considered that the compilation model might be most useful for fluent speakers. Additionally, the model is helpful to fluent speakers who have difficulties with diacritics for vowel length and other aspects

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 15–23
June 21, 2024

of spelling, in support of which they also included a spelling correction component.

The present research draws inspiration from these pioneering works of Lane et al. (2022) and Lane and Bird (2020), and seeks to continue the experiments in the field of word completion for Plains Cree. The present study, however, aims to investigate the feasibility of a complete word prediction and seeks to provide fluent Cree speakers with morphosyntactically and contextually appropriate, if not accurate, word suggestions which can potentially speed up the typing process.

To achieve this, we train a Bidirectional Long Short-Term Memory (BiLSTM) model to predict the next word in a sequence. LSTMs, a type of recurrent neural network (RNN), have demonstrated remarkable success in capturing contextual dependencies in sequential data, making them a compelling choice for natural language generation tasks (Hochreiter and Schmidhuber, 1997; Sundermeyer et al., 2012). Moreover, the fact that several studies working with low-resource agglutinative and polysynthetic languages used LSTMs for word prediction task (Kosyak and Tyers, 2022), makes it a compelling choice for the Plains Cree case.

To improve the model performance, we also train Word2Vec embeddings (Mikolov et al., 2013) for words in the Plains Cree corpus (see Section 2.2). Additionally, in this study, we explore the effect of vector augmentation—based on the words' morphosyntactic analyses, WordNet semantic classes (if applicable) (Miller, 1995), and lemmas—on the overall model accuracy and quality of word suggestions.

The paper is structured as follows. The data used in this paper and data preprocessing are described in Section 2. The Word2Vec vectors training and refinement, and BiLSTM model training are described in Section 3. Section 4 presents the results, and Section 5 presents their discussion. Possible directions for further research are examined in Section 6.

2 Data

2.1 Plains Cree

Plains Cree (endonimically – nêhiyawêwin, ISO 639-3: crk) is an Algonquian language spoken in Alberta, Saskatchewan, and Northwest Territories in Canada, and in the northern part of Montana in the US. This is the most widely spoken dialect of Cree. Cree is an agglutinative and polysynthetic language of predominantly prefixing nature. Although Cree is among the most spoken Indigenous languages of Canada, only a small corpus of Plains Cree is available currently.

2.2 Plains Cree corpus

The training data for this study comes from a combination of the Ahenakew-Wolfart Corpus (Arppe et al., 2020), the Bloomfield Corpus (Schmirler, 2023), and the Corpus of Miscellaneous Plains Cree Texts (*misi-mîkiwâhp pêsêkinosa ohci*) (Dacanay and Arppe, 2024), which has been morphosyntactically analyzed and lemmatized with the finite-state model mentioned before (Snoek et al., 2014; Harrigan et al., 2017), morphosyntactically disambiguated with a CG parser (Schmirler et al., 2018; Schmirler, 2023), and then annotated for WordNet semantic class (for nouns and verbs, when available). The WordNet classes are based on the classification by Dacanay (2022) of over 20k Cree entries in the lexical database underlying the *Cree Words/nêhiyawêwin: itwêwina*, a bilingual English-to-Cree dictionary by Wolvengrey (2001). All this information about each word was organized in a .tsv file, where each row included word form, analysis, and WordNet class as shown below:

awâsisak awâsis+N+A+Pl (n) child#1

2.3 Preprocessing

Before the corpus could be used for training it required significant preprocessing. First, the standard corpus normalization steps were made: 1) punctuation signs were removed, 2) words were converted to lowercase, and 3) Arabic and Roman numerals were removed.

Secondly, some special notes from the corpus were taken out. They include speakers' initials, new segment markers (e.g., 'Part'), web URLs for the texts taken from the Internet pages, and transcribers' notes (e.g., 'laughs', 'gesture').

Thirdly, English pieces were removed from the corpus along with the personal names. Lastly, the word 'and', connecting multiple WordNet classes, were removed, while all the spaces were replaced with underscore signs, as exemplified below:

(n) bannock#1 and (n) bread#1 and (n) flour#1

-> (n)_bannock#1_(n)_bread#1_(n)_flour#1

For words that lack WordNet class (e.g., conjunctions, pronouns) the UNK (unknown) code was added.

Lastly, the morphosyntactic analyses were also preprocessed. Originally, each analysis contained

	W/o preverbs	With preverbs
Word tokens	224,440	281,269
Word types	50,313	40,404
Lemmas	28,200	28,250
WordNet types	5,540	5,545
Analysis types	4,595	4,637

Table 1: Training dataset features.

a word lemma, for instance:

kâwiy+N+A+Px1Sg+Sg *or*

PV/ki+***itêw***+V+TA+Ind+X+3SgO

Lemmas were extracted from the morphosyntactic analyses, to provide an additional source of information about each word, resulting in the following representations:

+N+A+Px1Sg+Sg *or*

PV/ki+V+TA+Ind+X+3SgO

All the manipulations with the corpus were done with regular expressions.

Next, we separated verbs from preverbs for easier processing. In Plains Cree, preverbs are a broad category that includes both grammatical and derivational/lexical morphemes. As their name suggests, they appear before the verb stem. Plains Cree verbs can have multiple preverbs attached. Preverbs are usually separated from each other and the verb stems by hyphens (Okimâsis, 2004, 17). For instance, *nikakwê-nêhiyawân* 'I try to speak Cree', where the preverb *kakwê-* means 'try to, attempt to'. The number of combinations that preverbs can form is enormous, as shown by the Lane et al. (2022). Therefore, we decided to separate preverbs from their stems and treat them as separate entries in the training dataset for the purposes of this study. By doing so, we expect to improve our model prediction accuracy, because it will be able to learn preverbs combinations and their relations with different verb stems.

2.4 Training dataset

After all the aforementioned preprocessing steps, the dataset presented in Table 1 was obtained. The left column shows the size of the corpus before preverb separation and the right column - after separation.

3 Language modelling

3.1 Word2Vec pre-training

To improve the performance of the LSTM model, we decided to pre-train Word2Vec vectors using the CBOW approach. We experimented with different window sizes and settled with window size 5, because it was giving the best results. Considering the amount of information about each word available in the dataset, we decided to make the most of it during the Word2Vec pre-training. In order to do so, separate vectors were trained on the sequences of words[1], their WordNet semantic classes, their lemmas, and their morphosyntactic analyses, giving us four sets of vectors. The average of four vectors was calculated, and the original vectors for words were updated with the refined ones. Thus, the final word vectors are based not only on the neighbouring words but also on the semantic classes, lemmas, and morphosyntactic features of these neighbouring words.

Some adjustments had been made, however, to address the case of the words without WordNet class (such as non-inflecting particles) and/or morphosyntactic analysis. The vector refinement was organized in such a way that vectors for the 'Unknown' WordNet class or analysis were not included in the vectors' averaging. So, for example, the refined vector for the word *mêskanaw* (mêskanaw; N+I+Sg; (n)_road#1_(n)_trail#2) was the average of the vectors of *mêskanaw* and its lemma, analysis, and semantic class. However, for the particle *iyikohk* ('to such a degree; to such an extent'), the WordNet class is unavailable, so its averaged vector is based on the word, analysis, and lemma vectors only.

3.2 Model training

We trained a BiLSTM language model using the pre-defined word vectors as embeddings. As was mentioned earlier, the Bidirectional Long Short-Term Memory model was chosen for the purposes of this study. The LSTM model was chosen for our experiments because it was previously successfully used for predictive text tasks for polysynthetic low-resource languages (Kosyak and Tyers, 2022). At the beginning of the study, both unidirectional and bidirectional models were tested, and BiLSTM

[1]For convenience, we will be using the term 'word' to refer to all the Cree tokens in the training dataset, which also include preverbs.

showed better results. Thus, we decided to proceed with BiLSTM.

The data for language modelling was split into modelling and testing subsets at a ratio of 90% to 10%. The proportionately larger training set is due to the relatively small amount of data. The models' evaluation was based on the 3-fold cross-validation results.

The Word2Vec training parameters are provided in Appendix A. The model training hyperparameters are provided in Appendix B. All the BiLSTM models were trained on the university's high-performance computing cluster. The average training time for each model was 2-3 hours.

Vectors	Correct prediction		
	Top 1	Top 5	Top 10
Non-refined	15.06	31.15	**38.55**
Word+Analysis	14.19	29.92	**37.34**
Word+WordNet	13.92	29.59	37
Word+Analysis +Lemma	13.8	29.57	37.14
Word+Analysis +WordNet	13.67	29.17	36.54
Word+WordNet +Analysis+Lemma	13.46	28.86	36.32

Table 2: Prediction accuracy for the models trained with different Word2Vec sets (mean scores of 3-fold cross-validation).

4 Results

4.1 Refined BiLSTM model

To evaluate our model we looked at its top 10 predictions. The number 10 was chosen because it covers the number of suggestions provided by all commonly used tools, such as a smartphone keyboard (3-9 suggestions), search engine (9 suggestions), and text editors (5-9 suggestions). The resulting model predicted the correct word in the top 10 predictions in 36.3% of cases. Mostly, the model handled the prediction of nouns, adverbs, and preverbs better than of verbs. In comparison to the model trained with non-refined vectors, the overall quality of predictions slightly improved, and suggestions became more contextually and grammatically suitable (see Section 4.2 for comparison).

In cases when the model could not provide the correct completion among the first 10 predictions, the first letter of a word was provided to the model, simulating the beginning of user input. As a result, 41% of previously non-predicted words were eventually suggested as a possible completion in the top 10 predictions. The majority of the words predicted with the first letter input were verbs.

To sum up, the model predicts the next word by preceding context in 36.3% of cases and predicts the next word by context and the first letter in 28% of cases. However, as will be shown in the next section, this version of the model was not the most accurate.

4.2 Model performance with different Word2Vecs

To evaluate how the refined word vectors affected the overall model performance, we conducted several experiments with different combinations of refined vectors. It was mainly done to scrutinize how information about semantic, morphosyntactic, and lemma sequences contributes to the accuracy of the BiLSTM model. Table 2 shows the results of these experiments with a percentage of correct predictions in the top one, top 5, and top 10 predictions. It should be noted that these numbers do not evaluate the semantic and morphosyntactic appropriateness of the suggestions.

As can be seen from Table 2, the best accuracy of predictions was shown by the model that was trained with non-refined word vectors. However, the results of the models trained with different sets of refined vectors are not dramatically different as well. Nevertheless, it is interesting to compare the performance of the refined vectors' models. First, the averaging with lemma vectors does not seem to provide better prediction results. In both cases when they were used the overall accuracy dropped in comparison to the same vectors' combinations without lemma. The morphosyntactic analysis, on the contrary, seems to provide valuable information about the word's neighbours. The Word+Analysis model provides the best results (37.34%) among the models with refined vectors. The next best result is shown by the Word+Analysis+Lemma model (37.14%). The model trained with the full Word2Vec set shows the lowest accuracy results.

To better understand how the suggestions changed with the refined vectors, we compared the models' results. We began with a comparison of the predictions produced by the BiLSTM trained with the Word+WordNet+Analysis+Lemma refined vectors (hereafter full BiLSTM) and the one trained with the simple word vectors (hereafter BiLSTM).

This comparison revealed that the full BiLSTM had better accuracy in predicting nouns and verbs, in contrast to other parts of speech that do not have WordNet class. The overall top-10 suggestions became more semantically and grammatically suitable to the context than those predicted by BiLSTM. In some cases, predictions of full BiLSTM were not equal to the originally occurring word, but they all were in the correct morphosyntactic form. For example, in one case, the word *okimâwa* 'another chief' was expected, and the full BLSTM had it as a top-10 prediction, but it also offered *iskwêwa* 'another woman', *oskinîkiwa* 'another young man', *nâpêwa* 'another man', *mostoswa* 'another cow'. Most of them (except the 'cow') are semantically close to the expected word and denote humans, and all of them are in the expected obviative form. The top-10 BiLSTM's predictions for this case there were also some relevant suggestions in the correct morphosyntactic form (e.g., *oskinîkiwa*, *iskwêwa*, *mostoswa*); however, it had more semantically unsuitable suggestions than full BiLSTM like *mostoswa* 'another cow', *wâkayôsa* 'another black bear', *êskana* 'another antler', *misatimwa* 'another horse', *ôhi* 'this one'. Another example like this is illustrated in Table 3. Both models predicted the correct word in the top 3 suggestions, but the overall prediction quality is better in the full BiLSTM case. Interestingly, both Word+Analysis and Word+WordNet models predicted this word as top-1 and did not consider the preverb 'kâ-' in top-5 suggestions. Word+Analysis model predictions were all in the correct morphosyntactic form. Expectedly, Word+WordNet predictions were better semantically sorted, but not all of them were in the obviative form.

Another example (see Table 4), represents how some of the suggestions of the full BiLSTM, although a bit 'ambitious' do not sound completely absurd as well. All of them fit the overall structure of the sentence, and some words fit the context nicely (e.g., gift, decision). The 'ôma' suggestion after the 'ôma' in the preceding context is, most likely, a result of the word repetitions natural for spoken language presented in the corpus. The Word+Analysis BiLSTM also predicted the expected word, but other suggestions were less satisfactory. The simple BiLSTM did not predict the next word in this sequence.

In the case of verb predictions, we can observe a more or less similar situation. Full BiLSTM provides better suggestions than simple BiLSTM. For instance, in the case of the example provided in Table 5 the full BiLSTM correctly predicts the next verb *âtotamân* 'S/he will tell' in the top 3 predictions. The regular BiLSTM, in contrast, could not provide the expected verb. Interestingly, all the other refined models predicted the correct verb, with Word+Analysis+WordNet and Word+Analysis models providing the best suggestions. The Word+Analysis model's suggestions are also provided in Table 5.

Moreover, the separation of the preverbs from the verb stems allowed all the models to suggest out-of-vocabulary preverb+verb combinations. However, the Word+Analysis refined model offered the best preverbs and verb suggestions. When the first preverb was provided, for example, 'ê-', this model suggested possible next preverbs (e.g., kî-), as well as possible verbs, to follow it. It is also important to note that some of the predictions are only possible, if they are not prefixed, as they incorporate initial change, e.g. êtwêhk, êtwêt, and êtât. Thus, further work is needed to address these cases.

Finally, we observed that the refined models were less effective in the prediction of particles. The simple BiLSTM was on average 10% more successful in predicting them (e.g., *awa* or *ôma* 'this', *êkosi* 'so, thus'). Moreover, refined models often failed to predict low-frequency words that, in addition, did not have a WordNet class and morphosyntactic analysis assigned in our training dataset.

5 Discussion and Conclusions

Although testing of the models shows that the overall accuracy is higher for the simple BiLSTM, we argue that these results need further analysis and discussion before we can come to the final conclusion about vector refinement efficiency for Plains Cree word prediction. In order to quantify the results of our qualitative observations, we did two additional tests on the out-of-fold prediction results. First, we analyzed how semantically close were the predictions to the expected word. Second, we measured the morphosyntactic similarity of the predictions and the expected words.

For the first test, we measured a Wu-Palmer Similarity between the WordNet classes of the predictions and the target words with the NLTK WordNet package. The Wu-Palmer similarity value repre-

Input: *...cêskwa! itwêw awa sihkihp. êkotê isi kapâw. miyosiyiwa ôhi _*		
Eng: Wait! S/he says this is a waterhen. Towards there, s/he goes ashore. Someone is beautiful, this is_'		
No.	**Full BiLSTM predictions**	**BiLSTM predictions**
1	*ka-* Preverb	*kâ-* Preverb
2	***oskinîkiskwêwa*** 'another young woman'	*iskwêwa* 'another woman'
3	*iskwêwa* 'another woman'	***oskinîkiskwêwa*** 'another young woman'
4	*oskinîkiwa* 'another young man'	*ê-* Preverb
5	*wâkayôsa* 'another black bear'	*oskinîkiwa* 'another young man'
6	*nâpêwa* 'another man'	*nâpêwa* 'another man'
7	*kisêyiniwa* 'another old man'	*another wâkayôsa* 'another black bear'
8	*nâpêsisa* 'another boy'	*mostoswa* 'another cow'
9	*okimâwa* 'another chief'	*kisêyiniwa* 'another old man'
10	*nôtokêsiwa* 'another old woman'	*okimâwa* 'another chief'

Table 3: Predictions comparison 1

Input: *...pîhci ôma owiyasiwêwin piko ta-kawotinikêhk ôma _*		
Eng: 'By law, s/he must take back this _'		
No.	**Full BiLSTM predictions**	**BiLSTM predictions**
1	*miyikosiwin* 'gift'	*ôma* 'this'
2	CRTC	*ka-* Preverb
3	***askiy*** 'land'	*owiyasiwêwin* 'law, decision'
4	*wîhtamâkêwin* 'statement, announcement'	*nêhiyaw* 'Cree person'
5	*ôma* 'this'	*mâmiskôcikâtêwin* 'discussion'
6	*pîkiskwêwina* 'words'	*kistêyihtcikâtêwin* 'importance; principle'
7	*tahtoskânêsiwak* 'United Nations'	***askiy*** 'land'
8	*owiyasiwêwin* 'law, decision'	*wîhtamakêwin* 'statement'
9	*mâmawâyâwinihk* 'community, group'	*wiyastêwin* 'context, foundation'
10	*wiyastêwin* 'structure, arrangement, format'	*miyo-âyâwin* 'prosperity, good health'

Table 4: Predictions comparison 2

Input: *...âcimowin ôma k-ôh-nitotamâkawiyân k- _*		
Eng: 'This story, you (pl.) have not told me _'		
No.	**Full BiLSTM predictions**	**Word+Analysis BiLSTM predictions**
1	*ôh-* Preverb	*ôh-* Preverb
2	*ayâyân* 'I will say'	*âti-* Preverb
3	***âtotamân*** 's/he will tell'	***âtotamân*** 's/he will tell'
4	*êsiyîhkâtêk* 'it will be called'	*êtwêt* 's/he says so'
5	*âtotamân* 'you will tell us about it'	*êtwêhk* 'people say'
6	*êtwêhk* 'people say'	*êtât* 'you (sg) say thus to him'
7	*êtwêt* 's/he will say'	*êsiyîhkâtêk* 'it will be called'
8	*ês-âsotamawiyâhk* 'you (sg) promise to us'	*êtwêyân* 'I will say'
9	*ây-* Preverb	*âyâyâhk* 'for us to be there'
10	*êtât* 'you will say to him/her'	*êsiyîhkâsot* 's/he is called so'

Table 5: Predictions comparison 3

Model	Wu-Palmer similarity	MorphSyn similarity
Non-refined	37.9	42.17
Word+Analysis	**38.22**	**42.5**
Word+WordNet	37.86	41.86
Word+Analysis +Lemma	37.93	42.21
Word+Analysis +WordNet	37.48	42.1
Word+WordNet +Analysis+Lemma	37.78	42

Table 6: Average Wu-Palmer similarity and morphosyntactic (MorphSyn) similarity of the out-of-fold predictions to the actual labels

sents the distance between two synsets within the WordNet semantic hierarchy tree. It ranges from 0 to 1; the higher the value the more semantically similar two words are. The second column of Table 6 shows the average semantic similarity of top-10 predictions made by the models to the corresponding target words. The similarity was counted for/with applicable words only, i.e. words marked for the WordNet class.

For the second test, we calculated the Jaccard coefficient index for the morphosyntactic analyses of target words and predicted words. This comparison intended to show how well the suggested words were able to fit the morphosyntactic structure of the sentence. The third column of Table 6 demonstrates the average Jaccard similarity index of all the predictions and target words' pairs. Similarly, with the first test, the similarity index was calculated only for the words with the morphosyntactic analysis in the dataset.

The results of the tests showed that Word+Analysis and Word+Analysis+Lemma refined models provided contextually and grammatically better suggestions in comparison to other models. An average suggestion of the Word+Analysis model was 42.5% grammatically and 38.22% semantically similar to a target word. However, the difference is too small to claim with certainty that the refined vectors significantly improved the quality of predictions.

Although vector refinement does not provide a substantial prediction improvement, the main experiment and the additional tests indicate that the morphosyntactic information about words contributed the most to the refined models' accuracy and quality of suggestions. Probably, this result is due to the lower number of words lacking morphosyntactic analysis (in comparison to words lacking WordNet class). The lemma information does not seem to contribute to the overall accuracy of prediction. Nevertheless, it seems to improve the overall quality of suggestions. The information about neighbouring WordNet classes did not improve the accuracy or quality of word prediction. However, it is most likely related to the high number of words in the dataset that were not assigned a WordNet class yet or do not have a WordNet class.

The analysis of the models trained with refined vectors and the regular model revealed a disadvantage of embeddings refinement in the present settings. Although refined vectors contributed to the slight prediction quality improvement for the words that had all the extra information like morphosyntactic analysis, WordNet class, and lemma, they did not provide an adequate representation for the words lacking some or all of this extra information. The vector refinement function did not update its vectors based on the additional information. Consequently, after refinement, they could appear further from the words they originally co-occurred with, because their 'neighbours' vectors were updated. This highlights the necessity for a better refinement approach in further studies. In this study we used simple vector averaging, but in the follow-up studies the more sophisticated approaches like those proposed by Faruqui et al. (2015) and Mrkšić et al. (2017) should be explored.

To interpret the results of this study, it is also important to keep in mind that the model was tested on a small chunk of the corpus. Our corpus, in general, has a large portion that comes from transcribed Plains Cree narratives and fiction stories. Transcribed narratives often have more filler words, while fictional stories often have rare literary words. Both significantly differ from the writing we use on a day-to-day basis (texting, search queries, etc.). In some cases, the preceding context may have many words without semantic class and that makes predictions of the following words very tricky. Thus, we are sure that under the present circumstances, the experiment with refined BiLSTM models training yielded promising results. In future, we want to experiment with more standardized texts for training and testing and explore the possibility of excluding filler words for the Word2Vec training. Moreover, we believe that the predictions' quality can be also improved by reducing the number of non-analyzed words with unknown WordNet classes in our corpus. Further improvements in the training dataset can allow the model to learn more about contextual neighbours of each word, WordNet class and morphosyntactic analysis.

To conclude, this research lays the groundwork for a future predictive text model for Plains Cree. It shows that full-word prediction is not impossible for Plain Cree, and with certain improvements and modifications, can reach higher accuracy levels. This study also explores the use of augmented word embeddings in data scarcity cases; however, the efficiency of this method requires further analysis with a fuller dataset. Potentially, use of the model in tandem with other rule-based tools and resources developed for Plains Cree, such as morphosyntactic analyzer (Snoek et al., 2014; Harrigan et al., 2017), constraint grammar parser (Schmirler, 2023), or weighted Plains Cree morpheme combinations (Lane et al., 2022), can lead to more accurate results. Naturally, significant improvements are required before speakers and learners can use this tool.

6 Future work

Naturally, this study is only the beginning of the journey to the full-scale tool for the predictive text for Plains Cree. Hence, there are several directions for future research and experiments that we plan to pursue next to address the gaps in the present study.

First, we would like to try using fastText embeddings (Bojanowski et al., 2017) to capture regularities on the sub-word level of Plains Cree beyond preverbs. FastText embeddings were already successfully used for Mi'kmaq (Boudreau et al., 2020), another Algonquian language, and provided substantial improvements to the Mi'kmaq word prediction model. It would be interesting to compare the results of the word prediction model trained with refined word vectors used in this study, and the one trained with fastText embeddings. Hypothetically, the fastText-based language model should handle Plains Cree verbs better, because it will be able to capture other aspects of rich Plains Cree morphology (like verbal suffixes), by learning it on a sub-word level.

Secondly, we would like to implement partial input suggestions beyond preverbs for long morphologically complex words (e.g., verbs with 3+ preverbs). This approach will require a different methodology for dataset preparation and prediction assessment. Moreover, the keystroke saving tests will be required to explore the efficiency of partial input suggestions for keyboard users.

Limitations

Since this study is experimental, many problems have not been addressed here yet. First of all, our model does not have mechanisms to work with OOV (out-of-vocabulary) words. It only knows the words and prefixes it encountered in the corpus. This significantly limits the model at this stage, however, we plan to address this issue during the next development phase. Secondly, the model has difficulties in predicting longer and more morphologically complex words. As mentioned above, we plan to fix this by implementing partial input predictions. Thirdly, the model does not yet have a spelling relaxation function that would allow users to type without diacritics and still get predictions.

Ethics Statement

The tools described in this manuscript have been developed in order to support the explicit objectives of the language communities in question, to support their language instruction, maintenance, and revitalization activities.

Acknowledgements

We are grateful to the Plains Cree (nêhiyawak) communities, in particular those associated with the Maskwacîs Education Schools Commission (MESC) in Maskwacîs, Alberta, Canada, for the opportunity to work with their language. This research was supported in part by a Partnership Grant (#895-2019-1012) from the Social Sciences and Humanities Research Council (SSHRC) of Canada.

We also want to thank our anonymous reviewers for their valuable feedback that helped strengthen the study's quality.

References

Antti Arppe, Katherine Schmirler, Atticus G Harrigan, and Arok Wolvengrey. 2020. A morphosyntactically tagged corpus for plains cree. In *49th Algonquian Conference (PAC49)*, volume 49, pages 1–16.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Jeremie Boudreau, Akankshya Patra, Ashima Suvarna, and Paul Cook. 2020. Evaluating the impact of subword information and cross-lingual word embeddings on Mi'kmaq language modelling. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2736–2745.

Daniel Dacanay. 2022. *A Comparative Analysis of Manual and Vector Semantic Organisation using a Bilingual Dictionary of Plains Cree*. Bachelor's Thesis, University of Alberta, Edmonton, Canada.

Daniel Dacanay and Antti Arppe. 2024. misi-mîkiwâhp pêsêkinosa ohci – A corpus of miscellaneous Plains Cree texts. In *Papers of the 55th Algonquian Conference (PAC55)*.

Manaal Faruqui, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Edward Hovy, and Noah A. Smith. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the North American Chapter of the Association for Computational Linguistics – Human Language Technologies*, pages 1606–1615.

Atticus G Harrigan, Katherine Schmirler, Antti Arppe, Lene Antonsen, Trond Trosterud, and Arok Wolvengrey. 2017. Learning from the computational modelling of plains cree verbs. *Morphology*, 27:565–598.

Sepp Hochreiter and Jurgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735 – 1780.

Sergey Kosyak and Francis M. Tyers. 2022. Predictive text for agglutinative and polysynthetic languages. In *Proceedings of the First Workshop on Field Linguistics*, pages 77–85.

William Lane and Steven Bird. 2020. Interactive word completion for morphologically complex languages. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4600–4611.

William Lane, Atticus Harrigan, and Antti Arppe. 2022. Interactive word completion for Plains Cree. *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 1:3284–3294.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space.

George A. Miller. 1995. WordNet: a lexical database for English. *Communications of the ACM*, 38(11):39–41.

Nikola Mrkšić, Ivan Vulić, Diarmuid Ó Séaghdha, Ira Leviant, Roi Reichart, Milica Gašić, Anna Korhonen, and Steve Young. 2017. Semantic specialization of distributional word vector spaces using monolingual and cross-lingual constraints. *Transactions of the Association for Computational Linguistics Transactions of the Association for Computational Linguistics*, 5:309–324.

Jean L. Okimâsis. 2004. *Cree: Language of the Plains/nêhiyawêwin: paskwâwi-pîkiskwêwin.* University of Regina Press.

Katherine Schmirler. 2023. *Syntactic Features and Text Types in 20th Century Plains Cree: A Constraint Grammar Approach.* PhD Dissertation, University of Alberta, Edmonton, Canada.

Katherine Schmirler, Antti Arppe, Trond Trosterud, and Lene Antonsen. 2018. Building a Constraint Grammar Parser for Plains Cree Verbs and Arguments. In *Proceedings of the 11th Language Resources and Evaluation Conference*, pages 2981–2988.

Conor Snoek, Dorothy Thunder, Kaidi Lõo, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2014. Modeling the Noun Morphology of Plains Cree. In *ComputEL: Workshop on the use of computational methods in the study of endangered languages*.

Martin Sundermeyer, Ralf Schlüter, and Hermann Ney. 2012. LSTM neural networks for language modeling. *Interspeech*.

Arok Wolvengrey. 2001. *nêhiyawêwin: itwêwina / Cree: Words (Bilingual edition).* University of Regina Press.

A Word2Vec hyperparameters

The following hyperparameters were used to train Word2Vec embeddings.

Parameter	**Value**
vector size	300
window	5
min count	1
workers	4

Table 7: Word2Vec hyperparameters

B BiLSTM hyperparameters

The following BiLSTM hyperparameters provided the best training results.

Parameter	**Value**
BiLSTM layers	3
embedding dim	300
layers dropout	0.3
sequence length	21
optimizer	Adam
learning rate	0.001

Table 8: BiLSTM hyperparameters

Mapping ‘when’-clauses in Latin American and Caribbean languages: an experiment in subtoken-based typology

Nilo Pedrazzini
The Alan Turing Institute (London, United Kingdom)
npedrazzini@turing.ac.uk

Abstract

Languages can encode temporal subordination lexically, via subordinating conjunctions, and morphologically, by marking the relation on the predicate. Systematic cross-linguistic variation among the former can be studied using well-established token-based typological approaches to token-aligned parallel corpora. Variation among different morphological means is instead much harder to tackle and therefore more poorly understood, despite being predominant in several language groups. This paper explores variation in the expression of generic temporal subordination (‘when’-clauses) among the languages of Latin America and the Caribbean, where morphological marking is particularly common. It presents probabilistic semantic maps computed on the basis of the languages of the region, thus avoiding bias towards the many world’s languages that exclusively use lexified connectors, incorporating associations between character n-grams and English *when*. The approach allows capturing morphological clause-linkage devices in addition to lexified connectors, paving the way for larger-scale, strategy-agnostic analyses of typological variation in temporal subordination.

1 Introduction

Across the 7000+ world’s languages recorded by the Glottolog database (Nordhoff and Hammarström 2011, Hammarström et al. 2023)[1] there is great variation in how temporal relations between different eventualities can be encoded in a sentence or discourse unit. English has one main generic temporal subordinator, *when*, which is relatively underspecified with respect to the temporal semantic relation between the clause it introduces and its matrix clause, compared to semantically more precise connectors (e.g. *after*, *before*, or *while*). The number and scope of generic temporal subordinators can vary cross-linguistically from one (e.g. Italian *quando*), to two (e.g. German *wenn*/*als*) or several more (e.g. Pular *nde*/*si*/*ɓay*/*fewndo*/*tuma*; Evans 2017; Pedrazzini 2023). Crucially, languages can additionally or exclusively encode WHEN-clauses[2] morphologically on the predicate, rather than using a lexified subordinator (cf. Spanish *viendo* ‘see.GER’[3] as opposed to *cuando vio* ‘when saw.3.SG’; Ukrainian *pobačyvšy* ‘see.GER’ as opposed to *koly vyn pobačyv* ‘when he saw’). Because of the very nature of competition, overarching semantic differences between subordination strategies within individual languages cannot be fully captured in terms of discrete, categorical variables, but they should be modeled as a continuum allowing for a degree of overlap, aiming to reveal broader patterns in a probabilistic, rather than a fully deterministic way. Previous studies (Haug and Pedrazzini 2023) have employed a ‘token-based approach’ (Levshina 2019, 2022) to explore the semantic ground covered by English *when* and *induce* cross-linguistically common semantic dimensions from parallel corpora. In Haug and Pedrazzini (2023), probabilistic semantic maps (Croft and Poole 2008; Wälchli and Cysouw 2012) of WHEN were generated from a massively parallel corpus of 1400+ linguistic varieties (Mayer and Cysouw 2014), to capture systematic variation in the ways languages tend to divide the semantic space of English WHEN by using different lexical items for its different meanings. One of the greatest limitations of a purely token-based typological approach to the study of temporal subordination in the world’s languages is that it does not

[1] https://glottolog.org

[2] Small caps WHEN is used to refer to the semantic concept of ‘generic temporal subordination’, rather than the English lemma *when* (written in italics).

[3] The following abbreviations are used in glosses throughout this paper: GER = gerund, 3 = third person, SG = singular, PL = plural, SBJ = subject, VIS = visible (speaker’s area), SS = same subject, DS = different subject, DISTR = distributive, NARR = narrative, NSBJ = non-subject, LOC = locative, AS2 = secondary assertion, *pro* = prominent, PFV = perfective.

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 24–33

allow to account for variation *within* the semantic space covered by non-lexified WHEN-clauses cross-linguistically. That is, it will merely allow us to observe that particular *subsets* of *when*-occurrences are more likely to lack a parallel token in the target languages, without further identifying typologically widespread constructions (or *gram types*; Dahl and Wälchli 2016) within the semantic subspace of non-lexified WHEN-clauses.

While languages using predominantly or exclusively morphological means to express generic temporal subordination are relatively uncommon among European languages, non-lexified WHEN-clauses are instead particularly frequent among Latin American languages, as evidenced by the plethora of areal studies on converbal, clause-bridging, and, especially, switch-reference morphology in the region (among others, van Gijn et al. 2011; van Gijn 2012, 2016; Overall 2014, 2016).

This paper zooms in on the languages of Latin American and the Caribbean, given the particular computational challenges posed by their common, extensive use of non-lexified WHEN-clauses (exclusively so or in addition to lexified means). As in previous experiments, Mayer and Cysouw's (2014) massively parallel corpus of New Testament translations is used, and probabilistic semantic maps are adopted as a base method to induce typologically relevant dimensions within the semantic space of WHEN, since they allow capturing the gradience and overlap between different means in any given language, as well as the language-internal variation which is inherent to the very concept of competition. The goal of this paper is twofold:

a. incorporate associations between character n-grams and English *when* for capturing differences among WHEN-clauses that are expressed morphologically *as well as* lexically, and generate probabilistic semantic maps based on the parallel dataset thus refined. As detailed in Section 2, this method builds on Asgari and Schütze's (2017) 'SuperPivot' approach, but with substantial changes to their pipeline. Crucially, it gets rid of the assumption that there should be at most one 'pivot' (i.e. a marker in a parallel language) per linguistic feature (e.g. 'past' in Asgari and Schütze's 2017 example), reflecting instead the existing typological knowledge about the nature of generic temporal subordination as a phenomenon with great language-internal variation. The code to achieve this is released alongside this paper as a generalized tool, which starts from one or several lexical items in a source language and can be used to look for systematic cross-linguistic variation in a parallel dataset, both at the lexical and morphological level;

b. generate probabilistic semantic maps that are built exclusively on the basis of the languages of the region, thus avoiding bias towards the many world's languages that exclusively or predominantly use lexified connectors. The resulting maps and parallel data enriched with n-gram annotation are also released to facilitate future computational experiments.[4]

2 Methods

Dataset creation The Latin American and Caribbean parallel language data used in this experiment is a subset of Mayer and Cysouw's (2014) massively parallel corpus. To identify Latin American and Caribbean varieties in the massively parallel corpus, a `GeoJSON` dataset was manually created using `https://geojson.io/` to define the geographical region of interest. The approximate coordinates for each language variety in the dataset were taken from Glottolog and assigned to each New Testament translation based on its associated ISO 639-3 code. All varieties whose approximate coordinates were outside of the polygon defined by the `GeoJSON` dataset were filtered out from the corpus. The resulting data consisted of 335 varieties, representing approximately one-third of all the languages (1,005) recorded for Latin America and the Caribbean by Glottolog.[5] Figure 1 shows the areal distribution of the languages in our dataset among all the languages with an ISO 639-3 code from the region.

Word alignment & semantic mapping SyMGIZA++ (Junczys-Dowmunt and Szał 2012) was used to align the English version of the New Testament to each of the translations in our dataset at the token level, achieving a one-to-one token alignment for each language (i.e. each English token corresponds to at most one token in the target language, in contrast to possible one-to-many or many-to-one

[4]The code, datasets and all the maps, only a very small portion of which is presented in this paper, can be found in the associated repository.

[5]This number excludes sign languages, as we focus on textual data.

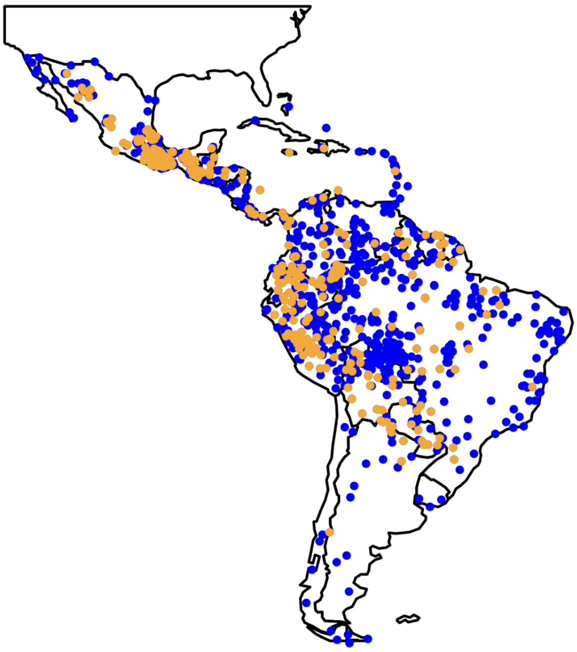


Figure 1: Approximate areal distribution of the languages in the dataset (orange) among the languages listed by Glottolog for the region (blue).

alignments). The occurrences of English *when* and its parallels in all Latin American and Caribbean languages in the dataset were then extracted. The quality of the automatic alignment was evaluated based on a sample of 300 *when*-clauses manually aligned to the Huichol translation, against which automatic alignment achieved a precision of 0.66, recall of 1, and F1-score of 0.79.[6]

Each instance of *when* and its parallel in every target language was treated as one usage point for WHEN. Hamming distance was applied as a measure of dissimilarity between pairs of usage points, by counting the number of languages using two different words, as opposed to the same word, for the two usage points in each pair. Multidimensional scaling (MDS) was then used to reduce the resulting Hamming-distance matrix to two dimensions, which were then treated as coordinates to plot the semantic map of WHEN as shown in Figure 2. Each dot in the semantic map represents a context for WHEN (i.e., a New Testament verse), and the farther apart two dots are, the more different their semantics is assumed to be, and the more likely they are to be encoded by different linguistic means cross-linguistically.

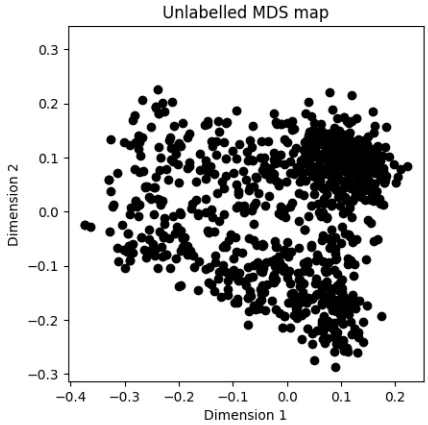


Figure 2: Unlabelled semantic map of WHEN.

Benchmarking As a form of evaluation of the methods and results, this experiment leveraged detailed typological and grammatical descriptions of the morphological system of one particular Latin American language, Huichol (or Wixárika). Huichol is among the several Latin American languages that show a clear division of labor between lexified and non-lexified WHEN-clauses (Pedrazzini 2023). In particular, Huichol uses *switch-reference* marking, a morphological system for tracking referents in an ongoing discourse (Roberts 2017, 538). In a 'canonical' switch-reference system (cf. Haiman and Munro 1983, ix), a clause is marked to signal whether its subject is co-referential or not with the subject of another, usually adjacent, clause, even though switch-reference has now long been shown to serve a much broader purpose than merely signaling referential (non-)identity (cf. Stirling 1993; McKenzie 2012, 2015a,b; Keine 2013). With subject co-reference, a same-subject marker is used (SS), else a different-subject marker is employed (DS). Switch-reference is overwhelmingly present in languages that allow and use *clause chaining*, which is the possibility of asyndetically stacking up several *deranked* verb forms (Stassen

[6] To calculate precision and recall, the presence of an aligned word in the target language was considered a 'positive', whereas the lack of an alignment ('NULL'-alignment) was considered a 'negative'. For an alignment to be considered a 'true negative', English *when* needed to have a NULL-alignment in Huichol in cases where Huichol does not use a subjunction to render the WHEN-clause, but expresses temporal subordination morphologically. Conversely, 'false negatives' corresponded to any NULL-alignment which should have been aligned to a 'when' word in Huichol. 'False positives', then, were considered cases in which *when* was aligned to a token in Huichol, despite the language using a morphological subordination strategy (i.e. switch-reference) or an independent clause, rather than a *quepaucua*-('when')-clause. Finally, 'true positives' corresponded to all 'when' instances correctly aligned to a 'when' word in Huichol.

1985; Croft 2002; Cristofaro 1998, 2019), that is, lacking marking of one or more tense, aspect, or mood distinctions compared to independent clauses in the same language, to signal their status as 'medial' clauses or 'converbs'. Switch-reference marking is well-known to serve that purpose particularly commonly among South American languages (cf. van Gijn 2016). In other words, by capturing switch-reference markers, we also capture the morphological means (i.e. the n-grams, or most common morphemes) that signal subordination, in our case, specifically, temporal clauses. (1) is a Huichol example of canonical switch-reference from our dataset, where switch-reference markers are used on the dependent verb to signal its subordinate status, where the English version has a *when*-clause in both cases.[7]

(1) Huichol/Wixárika (Uto-Aztecan)

a. *Hesüana* ***me-'u'-axüa-cu***
to.him 3.PL.SBJ-VIS-arrive.PL-DS
müpaü
thus
ti-ni-va-ru-ta-hüave
DISTR-NARR-3.PL.NSBJ-PL-SG-say
'When they came to him he said to them' (Acts 20:18)

b. *Hesüana* ***me-'u'-axüa-ca***
to.him 3.PL.SBJ-VIS-arrive.PL-SS
müme, müpaü
men thus
me-te-ni-ta-hüave
3.PL.SBJ-DISTR-NARR-SG-say
'When the men had come to him they said' (Luke 7:20)

Huichol additionally has a lexified 'when' subordinator (*quepaucua*), in which case switch-reference marking is absent, as in (2).

(2) *Mericüsü* ***quepaucua*** *yemuri-sie*
then when mountain-LOC
m-a-ca-ne, *teüteri yumüiretü*
AS2-PRO-down-arrive.PFV, people many
me-ca-n-i-veiya-caitüni
3.PL.SBJ-NARR-NARR-3.SG.OBJ-follow-IPFV
'When he came down from the mountain, great crowds followed him' (Matthew 8:1)

The concurrent presence of both a lexified connector and easily isolable morphemes for morphological subordination makes the language an ideal initial benchmark for experimenting with automatically detecting morphological *and* lexified markers of temporal subordination in the parallel corpus. As a form of evaluation for the character n-gram search system described below, the Huichol translation of the New Testament was enriched with annotation for different switch-reference markers. The markers were identified by using existing descriptions of Huichol switch-reference (i.e. Comrie 1983, 1982; Bierge 2017). The language has easily isolable switch-reference morphemes, namely *-ku* and *-ka* (spelled as *-cu* and *ca* in our dataset), for 'different-subject' and 'same-subject' marker, so the placeholders DS and SS were inserted before any word in the Huichol text ending with the respective forms, thus allowing the alignment model to capture the placeholders as dummy subordinators. Based on the annotated dataset, the location of SS and DS markers in the semantic map (Figure 3) can be compared with the location of morphological markers identified automatically via character n-gram search (Figure 4 in Section 3).

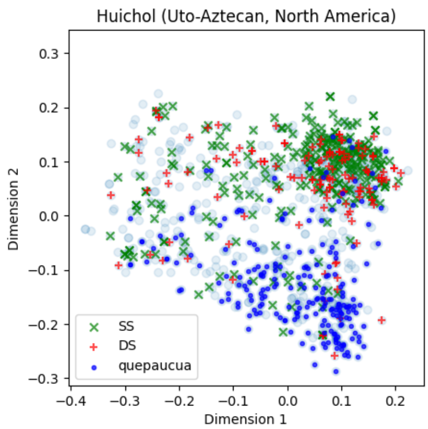


Figure 3: Probabilistic semantic map of WHEN, showing the location of lexified subordinators and switch-reference markers in Huichol after direct annotation (used as benchmark).

N-gram search Character n-grams were leveraged to identify potential morphological markers that are highly correlated with English *when*-clauses in our dataset, *in addition* to lexified means. As mentioned in the Introduction, the identification of potentially meaningful n-grams (i.e. those expressing a particular meaning of WHEN) is based on the approach by Asgari and Schütze (2017), albeit with additional steps and different n-gram ranges. Similarly to Asgari and Schütze (2017), χ^2 is used as a score of association between a 'head pivot'

[7] In the Huichol examples, the spelling of the Bible translation in Mayer and Cysouw's (2014) corpus was kept. Note, however, that this is not the most common orthography found in most studies on Huichol today.

(in our case always *when*) and a character n-gram, and it is calculated based on how many times *when* is aligned to a word containing that n-gram, how many times it is aligned to other n-grams and the frequency of both *when* and the n-gram. The raw alignments by SymGIZA++ were used as a starting point to identify tokens on which the n-gram search should be carried out. The following steps were followed to subsequently refine the parallel dataset with potentially meaningful n-grams:

1. a bespoke list of stopwords in English was established, based on their being either extremely frequent (*Jesus*, *Herod*, *Peter*, *Paul*) or very likely to introduce noise in a study on temporal subordination because of their distributional overlap with subordinators in terms of absolute position in a sentence (*and*, *behold*, *then*). χ^2 was used to find highly associated forms and parallel forms with an associated p-value of 0 were removed from the target language;

2. associations were identified between *when* and all tokens aligned to *when* by SymGIZA++. Only tokens with the highest score and a p-value of 0 were kept as they were and did not undergo the next steps;

3. using spaCy's (Honnibal and Montani 2017) English model `en_core_web_sm`, the English source text was automatically annotated for syntactic dependency to identify the head of the token *when*. This allowed for the *verb* of the *when*-clause to be extracted and the parallel verb in the translation to be identified. This choice was informed by the observation that languages marking subordination on the verb itself (i.e. non-lexified WHEN-clauses) are much more likely to have an empty token <NOMATCH> aligned to English *when* rather than the verb itself, so that the latter must be included in the search for meaningful character n-grams associated with *when*;

4. associations were identified between *when* and n-grams of any size between 2 and 9 for all remaining tokens aligned to either *when* or its head verb;

5. the top-scoring 200 n-grams (by χ^2) were then sorted by the number of times *when* was found to cooccur with the n-gram. The top-scoring 20 n-grams among the latter were then extracted as potentially meaningful n-grams;

6. the 20 extracted n-grams were clustered to attempt capturing groups of n-grams that are likely to be allomorphs of the same morpheme. Clustering was done using DBSCAN after converting the list of n-grams to a matrix of TF-IDF features. DBSCAN was selected after comparison with several other clustering algorithms (i.e. K-Means, K-Means++, Agglomerative Clustering, and Gaussian Mixture Modelling).

7. Each cluster of n-grams was assigned the placeholder label `ngram_1`...`ngram_N`, where `N` is the number of potentially meaningful n-gram clusters found for any given language.

Geostatistical interpolation Ordinary Kriging was then used to interpolate the linguistic items (i.e. the parallel token, if any, to *when*, or the n-gram placeholder label) used in each data point by each language in the dataset, to look for semantically relevant cross-linguistic dimensions. The Kriging model was implemented using the PyKrige library (Müller et al. 2023), with a Gaussian variogram model, a single averaging bin for the variogram (`nlag`), and `coordinates_type` set to `geographic`. The optimal `range`, `sill`, and `nugget` values for the Kriging models were set through a trial-and-error calibration process. Different combinations of these parameters were tested, and the ones used to produce the maps presented in Section 3 were chosen based on the interpretability of the resulting contour maps, with particular attention to the map for Huichol, thanks to the additional automatic annotation performed on the language using external knowledge bases. The contour levels generated through Kriging were normalized between 0 and 1 to facilitate the interpretation of the relative intensity of a linguistic means in the semantic space so that the closer the contour level to 1, the more intense the concentration of the respective means in the area. In the maps in Section 3, contours are plotted at all levels between 0.8 and 1.

The advantage of employing a geostatistical approach, such as Ordinary Kriging, for mapping language patterns is its ability to account for spatial autocorrelation (cf. Getis 2008), which facilitates the nuanced weighting of variables based on their prevalence and intensity across geographical space. While one linguistic means might be more widespread in terms of raw occurrence count in a given region of the semantic map, Kriging allows us to discern the spatial intensity of competing means. This, in turn, can clarify whether other means, despite being less prevalent overall, are more concentrated in that area and therefore more directly representative of the meaning associated with the respective space in the semantic map.

In the Kriging maps in Section 3, the placeholders for the n-grams are used instead of the actual list of n-grams.[8]

3 Results

Huichol Figure 4 shows the Kriging map generated from the Huichol data automatically refined with the n-gram search method. This can be compared with the labeled map in Figure 3, which, as explained in the previous section, is instead based on the Huichol data directly annotated with switch-reference markers as presented in typological descriptions of the language.

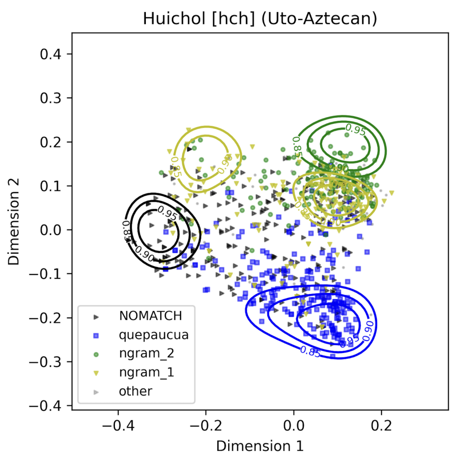


Figure 4: Kriging map of WHEN for Huichol.

Kriging detected relatively clearly separate areas (i.e. contexts or usage points) for lexified means (*quepaucua*), clustering at the bottom right of the map, and non-lexified means, corresponding to *ngram_1* and *ngram_2* in the map and clustering at the top of the map. *NOMATCH* indicates the absence of a parallel to English *when*, which suggests either a misalignment or the usage of a non-subordinate construction (e.g. an independent clause or a prepositional phrase, e.g. 'during dinner'). It is clear that the two automatically identified groups of n-grams, *ngram_1* and *ngram_2*, in the Huichol map correspond to DS and SS markers respectively. The *ngram_1* group includes *u*, *su*, *usu*, *cusu*, *icusu*, *ricusu*, *ericusu*, whereas *ngram_2* includes *ca*, *aca*, *eca*, *ieca*, *yaca*, *iyaca*, *xeiyaca*, *eiyaca*, *nieca*, which match the known switch-reference markers *-ku* and *-ka* (spelled as *-cu* and *ca* in our dataset) for DS and and SS respectively (Comrie 1983).

Based on the Huichol results, automatic word-alignment combined with the n-gram search method achieves a precision of 0.90, recall of 0.99, and F1-score of 0.94, calculated upon comparison with another manually annotated random sample of 300 English-Huichol WHEN-clauses with added switch-reference distinctions (i.e. English *when* was manually aligned to either *quepaucua* 'when', DS, or SS).

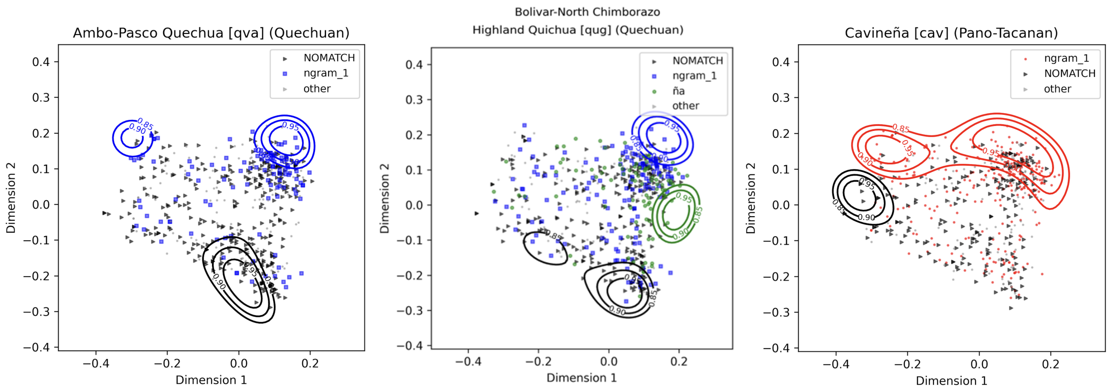
Switch-reference languages A clear validation of our method comes from the Quechuan languages in our dataset. According to van Gijn (2016, 168-169), all Quechuan languages have switch-reference marking, albeit with some differences in the markers used and their semantic scope. A closer inspection of the maps reveals that all Quechuan languages in our dataset show, in fact, a clear division of labor between the bottom and top of the map. Most commonly, the former is a *NOMATCH* area, whereas the top areas are instead most clearly under the scope of switch-reference markers. This is clearly the case, for example, in Ambo-Pasco Quechua (Figure 5a), from Peru, where the *ngram_1* group at the top of the map includes *r*, *ar*, *ur*, *cur*, *ycur*, *aycur*, *car*, all of which contain the distinctive *-r* SS marker of some Quechuan I subgroups (cf. van Gijn 2016, 168).

Another example is the map for Bolivar-North Chimborazo Highland Quichua (Figure 5b), from Ecuador. In this case, an *ngram_1* Kriging area was detected alongside a potentially lexified subordinator *ña*. The n-gram group includes *aca*, *paca*, *hpaca*, *shpaca*, *ushpaca*, *ashpaca*, where the Quechuan II SS marker, /ʃ/, spelled *sh*, can be discerned (van Gijn 2016, 171).[9]

A similar split, where the top area of the map is dominated by a n-gram group, is also found outside of Quechuan. This is the case, for instance, of Cavineña (Figure 5c), a Pano-Tacanan language of

[8] The reader can find which n-grams each group contains for any given language in the associated repository.

[9] The *-ca* ending is, in all likelihood, a personal ending that is particularly frequent in the source text.



(a) (b) (c)

Figure 5: Kriging maps of WHEN for three Latin American languages.

the Amazonian plains of northern Bolivia, where *ngram_1* includes *u*, *su*, *tsu*, *atsu*, *aatsu*, *catsu*, *baatsu*, *acatsu*, *bacatsu*, *itsu*, where the SS marker *-tsu* (cf. Guillaume 2008, 2011) can be seen.

The semantic maps for several other varieties from different language families show a division of labor similar to the Huichol one, between lexified means at the bottom of the map and n-gram groups (i.e. likely morphologically encoded WHEN-clauses) at the top of the map, as in Chuy (Mayan, Guatemala; Figure 6a), Comaltepec Chinantec (Otomanguean, Mexico; Figure 6b), or Terena-Kinikinao-Chane (Arawakan, Bolivia; Figure 6c).

Beyond switch-reference The integration of character n-grams to the semantic map of when was primarily driven by the aim of capturing morphological means of marking generic temporal subordination, which these examples from Latin American languages indicate as promising, especially in light of the known switch-reference markers captured in the maps. However, as mentioned in Section 1, there is great linguistic variation in the Latin American and Caribbean region and the new semantic maps helped capture more than just n-gram groups overlapping with the switch-reference markers in Huichol or Quechuan languages. Several languages, for instance, show an inverted pattern to the Huichol one, with a lexicalized means at the top of the map and an n-gram area at the bottom, as in Ticuna (Ticuna-Yuri, Western Amazon; Figure 6d) or Lomeriano-Ignaciano Chiquitano (Chiquitano, Bolivia; Figure 6e).

Yet others only use one n-gram for both the bottom and top areas, as in Tabasco Chontal (Mayan, Mexico; Figure 6f), or only lexified means, as in San Mateo del Mar Huave (Huavean/Isolate, Mexico; Figure 6g), Nivaclé (Matacoan, Argentina and Paraguay; Figure 6h), Kaqchikel (Mayan, Guatemala; Figure 6i), Guerrero Amuzgo (Otomanguean, Mexico; Figure 6j), Pichis Ashéninka (Arawakan, Peru; Figure 6k), and Chamacoco (Zamucoan, Paraguay; Figure 6l).

4 Conclusion & Future Work

Summary and findings This paper has presented probabilistic semantic maps of WHEN-clauses based on a parallel corpus of New Testament translations in Latin American and Caribbean languages. The rationale behind this study was the observation that WHEN-clauses in the Latin American region are often encoded *morphologically* (exclusively or predominantly so, i.e. in addition to lexified subordinators), which in previous token-based experiments (i.e. based only on full-token correspondences between languages) represented one of the main hurdles for the detection of systematic cross-linguistic variation in the expression of generic temporal subordination.

It built on previous approaches based on correspondences between a source word (English *when*) and character n-grams, using association measures to detect meaningful groups of n-grams that are likely to represent a particular morphological marker encoding temporal subordination in each target language. The approach has yielded results that are clearly helpful in identifying morphologically-encoded WHEN-clauses in languages where switch-reference markers (same-subject or different-subject marking) are employed to mark a predicate as subordinate to their matrix clause. The identification of groups of n-grams as switch-reference markers in some of the languages in the corpus was achieved by consulting descrip-

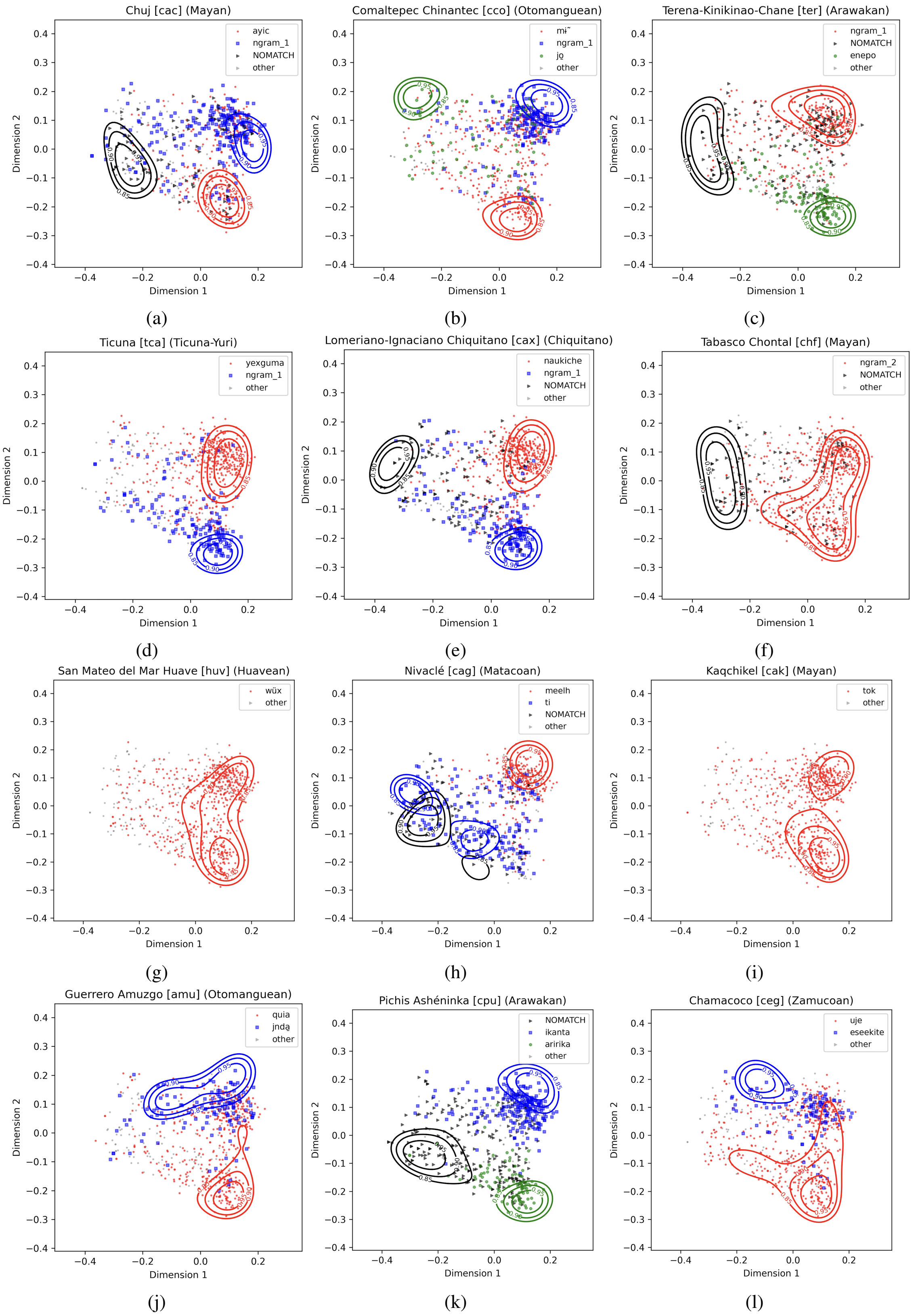


Figure 6: Kriging maps of WHEN showing some of the systematic variation in the dataset.

tive grammars and language-specific typological studies (e.g. on the Quechuan morphological system), but also because of the use of Huichol, a Mexican language with switch-reference morphology, as a point of reference to build a small benchmark and optimize hyperparameters during the generation of the semantic maps.

Future research Future studies may want to experiment with different n-gram sizes and different association measures and Kriging parameters, as well as use languages other than Huichol as benchmarks for the calibration of the Kriging models. Languages showing an opposite pattern to that of Huichol (i.e. a lexified means where Huichol has a morphological means, and vice versa) would particularly benefit from a close-reading evaluation to ascertain whether the method did manage to capture morphologically-expressed *when*-clauses as accurately as their opposite pattern.

Finally, the semantic dimensions in the maps have not been fully analyzed, and future studies will take a systematic approach to identifying clusters of observations that are frequently co-expressed, whether morphologically or lexically, across the languages of the corpus, and will establish whether such clusters represent cross-linguistically relevant gram types.

Limitations

The main limitation of this experiment is that evaluation, including hyperparameter optimization for the Kriging models, was based on one particular language, Huichol, because of the well-studied subordination system and the presence of a lexified subordinator in addition to the widely employed morphological means (switch-reference). Moreover, not only is switch-reference only one of the several attested morphological means to convey generic temporal subordination cross-linguistically, but there are also major differences between switch-reference systems (both in terms of the set of markers available to a language, but also their range of functions). The hyperparameters tuning based on Huichol has likely introduced some bias towards languages that have a similar system (i.e. one lexified counterpart to English *when* alongside switch-reference morphology), potentially obscuring other relevant typological dimensions (e.g. systematic clause-bridging marking).

The n-gram approach identifies *groups* of character n-grams, but does not yet provide a straightforward way of selecting one particular set of characters as the representative morpheme from a series of potential allomorphs. A tentative solution could be extracting the shortest allomorph, or the allomorph representing the common denominator among all n-grams in a set. However, this has not been tested and we have simply numbered each group of n-grams while keeping track of what forms each group contains for subsequent easier retrieval and inspection, if needed.

Acknowledgements

This work was supported by the Ecosystem Leadership Award under the EPSRC Grant EP/X03870X/1 & The Alan Turing Institute, particularly the Turing Research Fellowship scheme under that grant.

References

Ehsaneddin Asgari and Hinrich Schütze. 2017. Past, present, future: A computational investigation of the typology of tense in 1000 languages. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 113–124, Copenhagen, Denmark. Association for Computational Linguistics.

Stefanie Ramos Bierge. 2017. *Clause Types and Transitivity in Wixárika (Huichol): a Uto-Aztecan Language*. Ph.D. thesis, University of Colorado.

Bernard Comrie. 1982. Grammatical relations in Huichol. In *Studies in Transitivity*, pages 95–115. Brill, Leiden, The Netherlands.

Bernard Comrie. 1983. Switch-reference in Huichol: A typological study. In John Haiman and Pamela Munro, editors, *Switch Reference and Universal Grammar*, pages 17–38. John Benjamins, Amsterdam/Philadelphia.

Sonia Cristofaro. 1998. Deranking and Balancing in Different Subordination Relations: a Typological Study. *Sprachtypologie und Universalienforschung*, 51:3–42.

Sonia Cristofaro. 2019. *Subordination*. Oxford University Press, Oxford.

William Croft. 2002. *Typology and Universals*, 2nd edition. Cambridge University Press, Cambridge.

William Croft and Keith T. Poole. 2008. Inferring universals from grammatical variation: Multidimensional scaling for typological analysis. *Theoretical Linguistics*, 34(1):1–37.

Östen Dahl and Bernhard Wälchli. 2016. Perfects and iamitives: two gram types in one grammatical space. *Letras de Hoje*, pages 325–348.

Barrie Evans. 2017. A teaching grammar of Pular. Ms.

Arthur Getis. 2008. A history of the concept of spatial autocorrelation: A geographer's perspective. *Geographical Analysis*, 40(3):297–309.

Antoine Guillaume. 2008. *A Grammar of Cavineña*. Mouton de Gruyter, Berlin.

Antoine Guillaume. 2011. Subordinate clauses, switch-reference, and tail-head linkage in Cavineña narratives. In Pieter Muysken Rik van Gijn, Katharina Haude, editor, *Subordination in Native South American languages*, pages 109–140. John Benjamins, Amsterdam/Philadelphia.

John Haiman and Pamela Munro. 1983. Introduction. In John Haiman and Pamela Munro, editors, *Switch Reference and Universal Grammar*, pages ix–xv. John Benjamins, Amsterdam.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2023. *Glottolog 4.8*. Max Planck Institute for Evolutionary Anthropology, Leipzig.

Dag Haug and Nilo Pedrazzini. 2023. The semantic map of *when* and its typological parallels. *Frontiers in Communication*, 8.

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing.

Marcin Junczys-Dowmunt and Arkadiusz Szał. 2012. SyMGiza++: Symmetrized Word Alignment Models for Machine Translation. In *Security and Intelligent Information Systems (SIIS)*, volume 7053 of *Lecture Notes in Computer Science*, pages 379–390, Warsaw, Poland. Springer.

Stephan Keine. 2013. Deconstructing switch-reference. *Natural Language and Linguistic Theory*, 31:767–826.

Natalia Levshina. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology*, 23(3):533–572.

Natalia Levshina. 2022. Corpus-based typology: applications, challenges and some solutions. *Linguistic Typology*, 26(1):129–160.

Thomas Mayer and Michael Cysouw. 2014. Creating a massively parallel Bible corpus. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3158–3163, Reykjavik, Iceland. European Language Resources Association (ELRA).

Andrew McKenzie. 2012. *The role of contextual restriction in reference-tracking*. Ph.D. thesis, Amherst: University of Massachusetts.

Andrew McKenzie. 2015a. A survey of switch-reference in North America. *International Journal of American Linguistics*, 81:409–448.

Andrew McKenzie. 2015b. Austinian situations and switch-reference: The role of context in reference-tracking. Ms.

Sebastian Müller, Roman Yurchak, Benjamin Murphy, nannau, Malte Ziebarth, Sudipta Basak, Marcelo Albuquerque, Mark Vrijlandt, Matthew Peveler, Daniel Mejía Raigosa, Harry Matchette-Downes, Jordan Porter, Rhilip, Scott Staniewicz, Will Chang, and kvanlombeek. 2023. Geostat-framework/pykrige: v1.7.1.

Sebastian Nordhoff and Harald Hammarström. 2011. Glottolog/Langdoc: Defining dialects, languages, and language families as collections of resources. In *Proceedings of ISWC 2011*.

Simon E. Overall. 2014. Clause-chaining, switch-reference and nominalisations in Aguaruna (Jivaroan). In Rik van Gijn, Jeremy Hammond, Dejan Matić, Saskia van Putten, and Ana Vilacy Galucio, editors, *Information Structure and Reference Tracking in Complex Sentences*, pages 309–340. John Benjamins, Amsterdam/Philadelphia.

Simon E. Overall. 2016. Switch-reference and case-marking in Aguaruna (Jivaroan) and beyond. In Rik van Gijn and Jeremy Hammond, editors, *Switch Reference 2.0*, page 453–472. John Benjamins, Amsterdam/Philadelphia.

Nilo Pedrazzini. 2023. *A quantitative and typological study of Early Slavic participle clauses and their competition*. Ph.D. thesis, University of Oxford.

John R. Roberts. 2017. A Typology of Switch Reference. In Alexandra Y. Aikhenvald and R. M. W.Editors Dixon, editors, *The Cambridge Handbook of Linguistic Typology*, Cambridge Handbooks in Language and Linguistics, page 538–573. Cambridge University Press.

Leon Stassen. 1985. *Comparison and Universal Grammar*. Basil Blackwell, Oxford.

Lesley Stirling. 1993. *Switch-Reference and Discourse Representation*. Cambridge University Press.

Rik van Gijn. 2012. Switch-attention (aka switch-reference) in South-American temporal clauses. *Linguistic Discovery*, 10(1):112–27.

Rik van Gijn. 2016. Switch reference in Western South America. In Rik van Gijn and Jeremy Hammond, editors, *Switch Reference 2.0*, page 153–206. John Benjamins, Amsterdam/Philadelphia.

Rik van Gijn, Katharina Haude, and Pieter Muysken, editors. 2011. *Subordination in native South American languages*. John Benjamins, Amsterdam/Philadelphia.

Bernhard Wälchli and Michael Cysouw. 2012. Lexical typology through similarity semantics: Toward a semantic map of motion verbs. *Linguistics*, 50:671–710.

Comparing LLM prompting with Cross-lingual transfer performance on Indigenous and Low-resource Brazilian Languages

David Ifeoluwa Adelani
University College London, UK
d.adelani@ucl.ac.uk

A. Seza Doğruöz
Universiteit Gent, LT3, IDLab, Gent, Belgium
as.dogruoz@ugent.be

André Coneglian
Federal University of Minas Gerais, Brazil
coneglia@ufmg.br

Atul Kr. Ojha
Insight SFI Research Centre for Data Analytics, University of Galway, Ireland
atulkumar.ojha@insight-centre.org

Abstract

Large Language Models are transforming NLP for a variety of tasks. However, how LLMs perform NLP tasks for low-resource languages (LRLs) is less explored. In line with the goals of the AmericasNLP workshop, we focus on 12 LRLs from Brazil, 2 LRLs from Africa and 2 high-resource languages (HRLs) (e.g., English and Brazilian Portuguese). Our results indicate that the LLMs perform worse for the part of speech (POS) labeling of LRLs in comparison to HRLs. We explain the reasons behind this failure and provide an error analysis through examples observed in our data set.

1 Introduction

Despite numerous advancements in the NLP research due to Large Language Models (LLMs), available resources mainly cover 20 out of the estimated 7,000 languages (Magueresse et al., 2020). As a result, majority of world languages could still be considered as "low-resource".

Being a low-resource language (LRL) encompasses different types of inadequacies with respect to the availability of data for creating language technologies (Gupta, 2022). Focusing on multilingual linguistic scene in South America, we test the performance of LLMs for annotating part-of-speech (POS) tagging for 12 LRLs from Brazil, make a comparison with 2 LRLs from Africa and 2 high resource languages (HRLs) (e.g., English and Brazilian Portuguese) through human evaluation.

The evaluation is challenging for two reasons. First, there is a lack of benchmark datasets for the LRLs in Brazil in general. The ones we were able to find in universal dependencies (UD) data base, [1] do not have the training data to fine-tune multilingual language models. Hence, we can only leverage prompting LLMs or cross-lingual transfer through multilingual language models. Secondly, there is a lack of large monolingual data to benefit from effective multilingual and cross-lingual transfer techniques (Pfeiffer et al., 2020; Ansell et al., 2022; Alabi et al., 2022). We could only find the Bible corpora with less than 35K sentences for 7 out of the 12 languages.

We perform the evaluation on 12 Brazilian LRLs by prompting GPT-4 LLM and cross-lingual transfer individually from English and Brazilian Portuguese leveraging XLM-R. We preferred GPT-4 because the other open multilingual models (e.g., mT0 (Muennighoff et al., 2022), AYA (Ustun et al., 2024)) do not support the LRLs in this study. The results of both methods indicate low performance (less than 34.0% while high-resource languages achieved over 90.0%). However, GPT-4 leads to better results and Brazilian Portuguese performs better than English in zero-shot evaluation. Furthermore, to boost the performance of cross-lingual transfer, we perform language adaptation using XLM-R on each language, before fine-tuning Brazilian Portuguese, and evaluating on that language. This boosts the performance by $+3$ to $+12.0$ points on six out of seven languages. Our findings suggest that cross-lingual transfer to these languages is very challenging and having few training examples may further boost the performance. Therefore, there is a need for building NLP resources across different tasks for these LRLs.

1.1 Multilingualism in Brazil

Brazil is the 5th largest country of the world (qua land area) with a population of 203 million[2] and

[1] https://universaldependencies.org/

[2] Instituto Brasileiro de Geografia e Estatística. 2023. https://www.ibge.gov.br/en/cities-and-states.html. Accessed: 2023-12-15

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 34–41
June 21, 2024

it is highly multilingual. Although (Brazilian) Portuguese is the official language, there are approx. 160 native/indigenous as well as sign and immigrant languages.[3]

Following Rodrigues (1986), the two macro-language families among Brazilian native languages are Tupi (8 language families, 52 languages), and Macro-Jê (7 language families, 39 languages). There are also several large language families (e.g., Karib (21 languages), Arawak (20 languages), Arawá (7 languages), Tukano, Maku, and Yanomami), six smaller language families to the south of the Amazon river (e.g., Guaikurú (1 language), Nambikwára (3 languages), Txapakura (3 languages), Pano (13 languages), Múra (2 languages), and Matukína (4 languages)) and approx. 10 languages which are not part of any these families.

These languages share grammatical properties due to family inheritance or areal contact (Aikhenvald, 2002). In terms of morphology, most of these languages are polysynthetic, head-marking, and agglutinating with little fusion (Dixon and Aikhenvald, 1999; Hengeveld et al., 2007). In term of syntax, there is quite some variation in terms of word order among these languages(Campbell, 2012).

2 Literature Overview

In terms of labelled datasets for Brazilian LRLs, we only found datasets from the UD tasks: Gerardi et al. (2022) developed for TUDET UD treebanks covering 8 Tupian languages, other languages covered in UD are Apurina (Hämäläinen et al., 2021), Bororo, Madi-Jarawara, and Xavante (contributed by the TUDET team). For the monolingual data, we found seven Bible corpora on the eBible corpus (Akerman et al., 2023) that are freely available. All languages lack a large monolingual corpus which makes it very challenging for cross-lingual transfer and multilingual pre-training of LLMs.

In terms of evaluation, some studies have already shown the potential of prompting LLMs in multilingual settings (Ahuja et al., 2023a; Lai et al., 2023), including some LRLs (Ojo et al., 2023; Ahuja et al., 2023b). However, evaluations covering Brazilian LRLs are lacking. To the best of our knowledge, our study is the first to fill this gap.

3 Experimental setup

We focus our evaluation of POS tagging (a subtask of universal dependencies (UD)) on Brazilian LRLs due to the simplicity of the task, its popularity, and the availability of the test evaluation datasets in UD [4].

3.1 Evaluation Datasets

We evaluated 12 Brazilian LRLs and 2 African languages for a comparison across other regions with low-resource languages. Finally, we added 2 HRLs (i.e., English and Brazilian Portuguese). Our definition of HRL is based on the size of unlabelled data on the web. The larger their size are, the more likely they are included in pre-training of the LLMs [5] and multilingual pre-trained LMs (Conneau et al., 2020). While UD (Zeman et al., 2023) covers many languages, most LRLs only have a test set because of their limited sizes (less than 10k tokens). The Brazilian LRLs we evaluated on have also less than 13k tokens (except Nheengatu with 12,621 tokens).

Table 1 shows the languages in our evaluation, their language family, availability of monolingual corpus or Bible corpus in that language, UD dataset, and sizes. We collected the Bible corpus from the eBible website and used it for language adaptation. We have two test sets in our evaluation: (1) **Test set A**: the original test set in the UD benchmark (2) **Test set B** the subsample of Test set A where we removed sentences that GPT-4 fails to provide predictions for (mostly due to not properly identifying the language). We added this information for a fair comparison of the methods (i.e. using the same number of sentences in evaluation).

3.2 Models

For the experiments, we consider three approaches that are popular in the zero-shot setting since we lack the training data for the Brazilian languages (see Appendix A for details).

Prompting GPT-4 We prompt GPT-4 using a similar prompt provided by Lai et al. (2023) where the model is provided a task description before the input (seeAppendix B for details).

Cross-lingual transfer We trained a POS tagger individually for English and Portuguese, and per-

[3]PIB. 2023. https://pib.socioambiental.org/pt/Linguas. Accessed: 2023-12-15

[4]https://universaldependencies.org/

[5]https://help.openai.com/en/articles/8357869-chatgpt-language-support-alpha-web

Language	Language family	Monolingual data size	UD dataset name	Train	Dev	Test set A	Test set B
high-resource languages							
English (en)	Indo-European/West Germanic	not collected	en_ewt	12,544	2,001	2,007	-
Portuguese (pt)	Indo-European/Romance	not collected	pt_gsd	9,616	1,204	1,200	-
Brazilian languages							
Apurina (apu)	Arawakan	Bible (8,729)	apu_ufpa	-	-	152	134
Akuntsu (aqz)	Tupian	N/A	aqz_tudet	-	-	343	267
Karo (arr)	Tupian	N/A	arr_tudet	-	-	674	172
Bororo (bor)	Macro-Jê	Bible (8,254)	bor_bdt	-	-	371	161
Guajajara (gub)	Tupian	Bible (33,757)	gub_tudet	-	-	1,182	914
Madi-Jarawara (jaa)	Arawan	Bible (8,606)	jaa_jarawara	-	-	20	18
Makurap (mpu)	Tupian	N/A	mpu_tudet	-	-	37	8
Munduruku(myu)	Tupian	Bible (8,430)	myu_tudet	-	-	158	82
Tupinamba (tpn)	Tupian	N/A	tpn_tudet	-	-	581	458
Kaapor (urb)	Tupian	Bible (8,535)	urb_tudet	-	-	83	20
Xavante (xav)	Macro-Jê	Bible (8,213)	xav_xdt	-	-	148	128
Nheengatu (yrl)	Tupian	N/A	yrl_complin	-	-	1239	-
African languages							
Wolof (wo)	Niger-Congo/Senegambian	not collected	wo_wtb	1188	449	470	470
Yoruba (yo)	Niger-Congo/Volta-Niger	not collected	yo_ytb	-	-	318	318

Table 1: **UD-POS datasets in our evaluation:** We provide the training, validation and test splits we used for experiments. Test set A are the original test set in UD, the Test set B is a subset of A where we removed sentences that GPT-4 is not able to run inference for due to non-identification of the language.

form the zero-shot transfer on other languages. We used the XLM-R-large (or simply, XLM-R) (Conneau et al., 2020) for training the models.

Language Adaptive Fine-tuning (LAFT) We leverage LAFT for an effective cross-lingual transfer by first adapting XLM-R-large model to a new language with limited amount of monolingual data (Alabi et al., 2020; Pfeiffer et al., 2020; Chau and Smith, 2021; Alabi et al., 2022). We make use of the Bible data as the fine-tuning corpus since it is the largest one for these languages and we only found 7 (out of 12 Brazilian languages) languages which have a Bible corpus. Similar to Ebrahimi and Kann (2021), we examine the effectiveness of this small pre-training corpus with 8K-34K sentences. According to Pfeiffer et al. (2020), this approach can significantly boost cross-lingual transfer. However, it is not parameter-efficient like the MAD-X they proposed. On the other hand, Ebrahimi and Kann (2021) argued that simple adaptation to a new language is more effective than MAD-X especially when using the Bible corpus for adaptation and we follow this recommendation in our evaluation.

4 Results

Table 2 shows the result of our evaluation on POS tagging with the following key findings:

Zero-shot evaluation results While POS tagging has a performance of 98% (e.g. for English and Portuguese) when training data are available (especially for HRLs), the performance decreases while performing zero-shot transfer to other languages because POS tagging is language-specific. The transfer performance is low for both Brazilian and African languages (probably) because they are not typologically related whereas English and Portuguese are slightly related (i.e., being in the same Indo-European family) and covered by XLM-R, thus achieving an impressive transfer performance ($> +83\%$).

GPT-4 vs. basic cross-lingual transfer GPT-4 performed slightly better than the zero-shot transfer from other languages in our experiments indicating better abilities of LLMs for this task. For English and Portuguese, the performance reaches to 90% (although it is not on par with fully-supervised setting). For African languages, the performance was lower than the HRLs, but it was still decent (64.8-75.4) probably because the LLMs were exposed to some African languages during pre-training. The struggle of GPT-4 for Brazilian LRLs can be explained with the fact that these languages were probably not included during the pre-training. The generation is often not useful for some examples, where GPT-4 declines to give answers like "*As an AI, I'm unable to provide the POS tags for words in languages I'm not programmed to understand.*". Thus, we had to remove such examples from our evaluation. However, this was not the case for African LRLs and the HRLs.

Language adaptation for cross-lingual transfer performance We performed LAFT training on the Bible corpus individually for the *apu*, *bor*, *gub*, *jaa*, *myu*, *urb*, and *xav*. Our results indicate an im-

Language	XLM-R Test set A Full-sup.	XLM-R (zero-shot cross-lingual transfer) Test set A en→ xx	pt→ xx	LAFT + pt→ xx	Test set B en→ xx	pt→ xx	LAFT + pt→ xx	GPT-4 Test set B 0-shot
high-resource languages								
en_ewt	**98.0**	**98.0**	83.6			91.9		
pt_gsd	**97.8**	90.0	**97.8**			92.4		
Brazilian languages								
apu_ufpa	-	37.5	40.6	**44.9**	36.8	40.2	**44.7**	42.6
aqz_tudet	-	31.9	37.8		31.3	36.8		**49.5**
arr_tudet	-	3.9	14.9		6.3	19.8		**27.7**
bor_bdt	-	19.0	23.5	27.3	18.4	23.0	26.4	**41.3**
gub_tudet	-	26.5	30.2	36.0	27.8	32.1	**37.1**	36.2
jaa_jarawara	-	28.2	28.4	**34.5**	27.2	27.9	33.6	**33.0**
mpu_tudet	-	4.9	**9.0**		0.0	0.8		0.0
myu_tudet	-	21.2	27.1	**30.3**	10.8	14.8	16.5	18.2
tpn_tudet	-	39.1	41.9		38.9	41.8		**47.2**
urb_tudet	-	7.8	11.8	21.2	9.2	9.5	21.6	**32.3**
xav_xdt	-	26.5	29.0	28.2	27.3	29.9	29.3	**36.5**
yrl_complin	-	28.9	31.5		29.0	31.7		**41.2**
African languages								
wo_wtb	**87.6**	29.3	35.6					64.8
yo_ytb	-	22.5	31.5					**75.4**
Average (Brazilian languages)	-	23.0	27.1		21.9	25.7		33.8

Table 2: **POS accuracy results for Brazilian languages:** We compare the accuracy of GPT-4 to zero-shot cross-lingual transfer from English language and Portuguese leveraging XLM-R-large multilingual pre-trained language model. Test set A is the original test set found on UD while Test set B are the ones GPT-4 could automatically detect their language to run inference.

provement in accuracy on 6 out of the 7 languages, except for *xav*. The performance improvement is quite large for *urb* (+7.2 on test A, and 12.1 on Test B), and moderate improvement of +3 to +6 for other languages. This experimental result shows that with sufficient monolingual texts, we can increase the performance of the cross-lingual transfer results. However, for the LRLs, such data is scarce. A more effective approach is perhaps to annotate few examples (e.g. 10 or 100 sentences) for training POS taggers to boost the performance (cf. (Lauscher et al., 2020; Hedderich et al., 2020) for a larger boost in performance for token classification tasks in this few-short setting). Regardless, there is a need for better methods to leverage small monolingual data sets.

5 Error analysis

In this section, we provide examples from 2 Brazilian languages (Karo and Guajajara) where the LLMs made errors with the POS tagging. The first line refers to the original sentence, the second line refers to the gold-standard UD POS tag; the third line refers to the GPT-4 POS tag.)

In example (1), the auxiliary verb (in Karo) has the same orthographic form as the English interjection *okay*. In example (2), the Guajajara verb has (partially) the same orthographic form as the English interjection (*oh*). Due to these similarities, GPT-4 seems to tag the POS for these words according to English instead of the POS tagged in UD for Karo and Guajajara.

1. awero toba **okay**
   NOUN VERB **AUX**
   NOUN NOUN **INTJ**
2. **Oho** kaapii rehe .
   VERB NOUN ADP PUNCT
   INT VERB ADV PUNCT

6 Discussion & Conclusions

In our study, we explored how LLMs perform the NLP task of POS tagging for 12 LRLs in Brazil and compared this performance with 2 LRLs in Africa and 2 HRLs (English, Brazilian Portuguese). POS is a well established NLP task and it provides insights about the linguistic structures of the different languages especially when only limited data is available, such linguistic annotations have been shown to improve language understanding and generation for endangered languages (Zhang et al., 2024). Our results indicate that the LLMs (GPT-4) perform worse for LRLs on this task in general but older approaches like language adaptive fine-tuning that leverage multilingual encoder models provides some improvements. However, with the lack of available data, any improvements across methods are limited. Although we focused on 12 Brazilian LRLs, there are many other LRLs which we were not able to cover. Future work can expand this evaluation to more tasks and to other LRLs not only from Brazil but from other regions around the world as well.

7 Limitations

Due to limited space, we only focused on POS tagging for this paper but there is a need to explore how LLMs perform other NLP tasks for LRLs. We only evaluated ChatGPT in the zero-shot learning setting but we do not have comparisons with other recent multilingual LLMs, e.g., BLOOM (Scao et al., 2022), and Gemini, in various other learning scenarios. While some of these models are currently less accessible for large-scale evaluations, our plan is to include more models and learning settings along the way to strengthen our evaluations and comparisons in the future. Finally, the current work only evaluates ChatGPT in terms of performance over NLP tasks in different languages. To better characterize ChatGPT and LLMs, other evaluation metrics should also be investigated to report more complete perspectives for multilingual learning, including but not limited to adversarial robustness, biases, toxic/harmful content, hallucination, accessibility, development costs, and interoperability.

8 Ethics Issues

Since we used publicly available data sets, we do not foresee any major issues in terms of ethical concerns.

Acknowledgements

Atul Kr. Ojha would like to acknowledge the support of the Science Foundation Ireland (SFI) as part of Grant Number SFI/12/RC/2289_P2 Insight_2, Insight SFI Centre for Data Analytics and CA21167 COST Action UniDive (by COST (European Cooperation in Science and Technology). David Adelani acknowledges the support of DeepMind Academic Fellowship programme.

References

David Ifeoluwa Adelani, Jade Abbott, Graham Neubig, Daniel D'souza, Julia Kreutzer, Constantine Lignos, Chester Palen-Michel, Happy Buzaaba, Shruti Rijhwani, Sebastian Ruder, Stephen Mayhew, Israel Abebe Azime, Shamsuddeen H. Muhammad, Chris Chinenye Emezue, Joyce Nakatumba-Nabende, Perez Ogayo, Aremu Anuoluwapo, Catherine Gitau, Derguene Mbaye, Jesujoba Alabi, Seid Muhie Yimam, Tajuddeen Rabiu Gwadabe, Ignatius Ezeani, Rubungo Andre Niyongabo, Jonathan Mukiibi, Verrah Otiende, Iroro Orife, Davis David, Samba Ngom, Tosin Adewumi, Paul Rayson, Mofetoluwa Adeyemi, Gerald Muriuki, Emmanuel Anebi, Chiamaka Chukwuneke, Nkiruka Odu, Eric Peter Wairagala, Samuel Oyerinde, Clemencia Siro, Tobius Saul Bateesa, Temilola Oloyede, Yvonne Wambui, Victor Akinode, Deborah Nabagereka, Maurice Katusiime, Ayodele Awokoya, Mouhamadane MBOUP, Dibora Gebreyohannes, Henok Tilaye, Kelechi Nwaike, Degaga Wolde, Abdoulaye Faye, Blessing Sibanda, Orevaoghene Ahia, Bonaventure F. P. Dossou, Kelechi Ogueji, Thierno Ibrahima DIOP, Abdoulaye Diallo, Adewale Akinfaderin, Tendai Marengereke, and Salomey Osei. 2021. MasakhaNER: Named entity recognition for African languages. *Transactions of the Association for Computational Linguistics*, 9:1116–1131.

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023a. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

Sanchit Ahuja, Divyanshu Aggarwal, Varun Gumma, Ishaan Watts, Ashutosh Sathe, Millicent Ochieng, Rishav Hada, Prachi Jain, Maxamed Axmed, Kalika Bali, et al. 2023b. Megaverse: Benchmarking large language models across languages, modalities, models and tasks. *arXiv preprint arXiv:2311.07463*.

Alexandra Y. Aikhenvald. 2002. *Language contact in Amazonia*. Oxford University Press.

Vesa Akerman, David Baines, Damien Daspit, Ulf Hermjakob, Tae Young Jang, Colin Leong, Michael Martin, Joel Mathew, Jonathan Robie, and Marcus Schwarting. 2023. The ebible corpus: Data and model benchmarks for bible translation for low-resource languages. *ArXiv*, abs/2304.09919.

Jesujoba Alabi, Kwabena Amponsah-Kaakyire, David Adelani, and Cristina España-Bonet. 2020. Massive vs. curated embeddings for low-resourced languages: the case of Yorùbá and Twi. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2754–2762, Marseille, France. European Language Resources Association.

Jesujoba O. Alabi, David Ifeoluwa Adelani, Marius Mosbach, and Dietrich Klakow. 2022. Adapting pre-trained language models to African languages via multilingual adaptive fine-tuning. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4336–4349, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Alan Ansell, Edoardo Ponti, Anna Korhonen, and Ivan Vulić. 2022. Composable sparse fine-tuning for cross-lingual transfer. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1778–1796, Dublin, Ireland. Association for Computational Linguistics.

Lyle Campbell. 2012. Typological characteristics of south american indigenous languages. *The indigenous languages of South America: A comprehensive guide*, pages 259–330.

Ethan C. Chau and Noah A. Smith. 2021. Specializing multilingual language models: An empirical study. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 51–61, Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

R. M. W. Dixon and Alexandra Y. Aikhenvald. 1999. *The Amazonian Languages*. Cambridge University Press.

Abteen Ebrahimi and Katharina Kann. 2021. How to adapt your pretrained multilingual model to 1600 languages. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4555–4567, Online. Association for Computational Linguistics.

Fabrício Ferraz Gerardi, Stanislav Reichert, Carolina Aragon, Lorena Martín-Rodríguez, Gustavo Godoy, and Tatiana Merzhevich. 2022. Tudet: Tupían dependency treebank.

Akshat Gupta. 2022. On building spoken language understanding systems for low resourced languages. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–11, Seattle, Washington. Association for Computational Linguistics.

Mika Hämäläinen, University of Helsinki, Niko Partanen, and Khalid Alnajjar, editors. 2021. *Multilingual Facilitation*. University of Helsinki.

Michael A. Hedderich, David Adelani, Dawei Zhu, Jesujoba Alabi, Udia Markus, and Dietrich Klakow. 2020. Transfer learning and distant supervision for multilingual transformer models: A study on African languages. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2580–2591, Online. Association for Computational Linguistics.

Kees Hengeveld et al. 2007. Parts-of-speech systems and morphological types. *ACLC Working Papers Volume 2, issue*, page 31.

Viet Lai, Nghia Ngo, Amir Pouran Ben Veyseh, Hieu Man, Franck Dernoncourt, Trung Bui, and Thien Nguyen. 2023. ChatGPT beyond English: Towards a comprehensive evaluation of large language models in multilingual learning. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 13171–13189, Singapore. Association for Computational Linguistics.

Anne Lauscher, Vinit Ravishankar, Ivan Vulić, and Goran Glavaš. 2020. From zero to hero: On the limitations of zero-shot language transfer with multilingual Transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4483–4499, Online. Association for Computational Linguistics.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges.

Niklas Muennighoff, Thomas Wang, Lintang Sutawika, Adam Roberts, Stella Biderman, Teven Le Scao, M Saiful Bari, Sheng Shen, Zheng Xin Yong, Hailey Schoelkopf, Xiangru Tang, Dragomir R. Radev, Alham Fikri Aji, Khalid Almubarak, Samuel Albanie, Zaid Alyafeai, Albert Webson, Edward Raff, and Colin Raffel. 2022. Crosslingual generalization through multitask finetuning. *ArXiv*, abs/2211.01786.

Benjamin Muller, Antonios Anastasopoulos, Benoît Sagot, and Djamé Seddah. 2021. When being unseen from mBERT is just the beginning: Handling new languages with multilingual language models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 448–462, Online. Association for Computational Linguistics.

Jessica Ojo, Kelechi Ogueji, Pontus Stenetorp, and David I Adelani. 2023. How good are large language models on african languages? *arXiv preprint arXiv:2311.07978*.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Aryon Dall'Igna Rodrigues. 1986. *Línguas Brasileiras: para o conhecimento das línguas indígenas*. São Paulo: Edições Loyola.

A. Ustun, Viraat Aryabumi, Zheng-Xin Yong, Wei-Yin Ko, Daniel D'souza, Gbemileke Onilude, Neel Bhandari, Shivalika Singh, Hui-Lee Ooi, Amr Kayid, Freddie Vargus, Phil Blunsom, Shayne Longpre, Niklas

Muennighoff, Marzieh Fadaee, Julia Kreutzer, and Sara Hooker. 2024. Aya model: An instruction finetuned open-access multilingual language model. *ArXiv*, abs/2402.07827.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Salih Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Erik Andersen, Lene Antonsen, Tatsuya Aoyama, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, and et al. 2023. Universal dependencies 2.13. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Kexun Zhang, Yee Man Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. Hire a linguist!: Learning endangered languages with in-context linguistic descriptions. *ArXiv*, abs/2402.18025.

A Models

For the experiments, we consider three approaches that are popular in the zero-shot setting since we lack training data for the Brazilian languages.

Prompting GPT-4 GPT-4 [6] is a large language model developed by pre-training on a large amount of texts and code from the web, followed by instruction prompt tuning based on human feedback. We prompt GPT-4 using a similar prompt provided by Lai et al. (2023) where the model is provided a task description before the input. We provide the details in Appendix B.

Cross-lingual transfer We trained a POS tagger individually for English and Portuguese, and perform zero-shot transfer on other languages. We make use of the XLM-R-large (or simply, XLM-R) (Conneau et al., 2020) for training the models. XLM-R has been pre-trained on 100 languages of the world with over 2TB pre-training corpus size but this corpus does not include any indigenous Brazilian languages.

Language Adaptive Fine-tuning (LAFT) We leverage LAFT for an effective cross-lingual transfer by first adapting XLM-R-large model to a new language with limited amount of monolingual data (Alabi et al., 2020; Pfeiffer et al., 2020; Chau and Smith, 2021). This method was proven to be very effective for low-resource languages (Adelani et al., 2021; Muller et al., 2021). We make use of the Bible data as the fine-tuning corpus since it is the largest we found for these languages. We only found 7 (out of 12 Brazilian languages) languages with the Bible corpus. Similar to Ebrahimi and Kann (2021), we examine the effectiveness of this small pre-training corpus with 8K-34K sentences. Pfeiffer et al. (2020) showed that this approach can significantly boost cross-lingual transfer. However, it is not parameter-efficient like the MAD-X they proposed. On the other hand, Ebrahimi and Kann (2021) argued that simple adaptation to a new language is more effective than MAD-X especially when using the Bible corpus for adaptation and we follow this recommendation in our evaluation.

Hyper-parameter of experiments For the cross-lingual and LAFT experiments, we used HuggingFace transformers (Wolf et al., 2020) and A100 Nvidia GPU for fine-tuning the models. For the LAFT, we train for 3 epochs on one GPU while for cross-lingual, we fine-tune English and Portuguese individually using a batch size of 64, with gradient accumulation of 2, and a training epoch of 10.

B Prompt Template

Table 3 provides the prompt template we used for GPT-4 evaluation.

[6] https://chat.openai.com/

	Prompt
Task Description	Please provide the POS tags for each word in the input sentence. The input will be a list of words in the sentence. The output format should be a list of tuples, where each tuple consists of a word from the input text and its corresponding POS tag label from the tag label set: [”ADJ”, ”ADP”, ”ADV”, ”AUX”,”CCONJ”, ”DET”, ”INTJ”, ”NOUN”, ”NUM”,”PART”, ”PRON”, ”PROPN”, ”PUNCT”,”SCONJ”, ”SYM”, ”VERB”, ”X”].
Note	Your response should include only a list of tuples, in the order that the words appear in the input sentence, with each tuple containing the corresponding POS tag label for a word.
Input	[”What”, ”if”, ”Google”, ”Morphed”, ”Into”, ”GoogleOS”, ”?”]
Output	[(”What”, ”PRON”), (”if”, ”SCONJ”), (”Google”, ”PROPN”), (”Morphed”, ”VERB”), (”Into”, ”ADP”), (”GoogleOS”, ”PROPN”), (”?”, ”PUNCT”)]

Table 3: **Prompt template used for POS tagging** based on Lai et al. (2023). An example prediction by GPT-4

Analyzing Finetuned Vision Models for Mixtec Codex Interpretation

Alexander R. Webber, Gabriel Ayoubi, Justin Witter, Zachary Sayers,
Amy Wu, Elizabeth Thorner, and **Christan Grant**
{alexwebber, gabriel.ayoubi, jwitter, z.sayers, amy.wu, ethorner, christan}@ufl.edu
UF Data Studio
University of Florida

Abstract

Throughout history, pictorial record-keeping has been used to document events, stories, and concepts. A popular example of this is the Tzolk'in Maya Calendar. The pre-Columbian Mixtec society also recorded many works through graphical media called codices that depict both stories and real events. Mixtec codices are unique because the depicted scenes are highly structured within and across documents. As a first effort toward translation, we created two binary classification tasks over Mixtec codices, namely, gender and pose. The composition of figures within a codex is essential for understanding the codex's narrative. We labeled a dataset with around 1300 figures drawn from three codices of varying qualities. We finetuned the Visual Geometry Group 16 (VGG-16) and Vision Transformer 16 (ViT-16) models, measured their performance, and compared learned features with expert opinions found in literature. The results show that when finetuned, both VGG and ViT perform well, with the transformer-based architecture (ViT) outperforming the CNN-based architecture (VGG) at higher learning rates. We are releasing this work to allow collaboration with the Mixtec community and domain scientists.

1 Introduction

Vast amounts of historical and cultural documents are encoded in pictographic systems (Sampson, 2015). Representations such as Egyptian hieroglyphics use pictorial representations corresponding to words and sub-word components to express concepts. Other pictorial systems that display scenes that evoke a known narrative have also been used throughout the world. Rules govern the depiction of years, dates, names, class, ceremonies, and gender (Jansen, 1988). **The implicit grammatical rules can contribute to a deterministic interpretation of these ancient narratives.** Mixtec codices are highly structured and have fairly rigid conventions for the representation of people (Boone, 2000), such as loincloths on men and skirts on women. Consequently, the depiction of persons in these codices follows consistent patterns. Unfortunately, due to the ravages of time and conflict, only a few of these codices are presently available. Computational analyses of the codices and their underlying structures may help researchers better understand the remaining works. In this paper, we explore how models such as VGG-16 (Simonyan and Zisserman, 2015) and ViT-16 (Dosovitskiy et al., 2021) perform when used to classify these low-resource patterns and understand the features they find important in this task.

2 Mixtec Codices

The researchers labeled data from three popular sources: The Codices Vindobonensis Mexicanus I (Lehmann and Smital, 1929; Unbekannt, 1449), Selden (Caso, 1964; Bakewell and Hamann, 2023), and Zouche-Nuttall (Nuttall, 1902; Forstmann, 2023). Codex Vindobonensis Mexicanus I describes both the mythological and historical founding of the first Mixtec kingdoms, Codex Selden follows the founding of the kingdom of *Jaltepec* and its ruler, *Lady 6 Monkey*, and Codex Zouche-Nuttall primarily illustrates the life and conquests of *Lord 8 Deer Jaguar Claw*, but also details the histories of his ancestors. Other Mixtec codices

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 42–49
June 21, 2024

Codex	Total	Gender		Pose		Quality		
		Man	Woman	Standing	Not Standing	a	b	c
Nuttall	264	256	8	101	163	263	1	0
Selden	307	74	233	32	275	254	46	7
Vindobonensis Mexicanus I	714	573	141	253	461	569	123	22
Totals	*1285*	*903*	*382*	*386*	*899*	*1086*	*170*	*29*

Table 1: The counts of figures from each of the source Mixtec codices and in total, the number of `man` and `woman` labels per codex and in total, numbers of `standing` and `not standing` labels per codex and in total, and the **a**, **b**, and **c** labeled data items per codex and in total.

are extant, but their condition is degraded and not amenable to our current machine-learning pipeline. Each codex is made of deerskin folios, and each folio comprises two pages. The Codex Vindobonensis Mexicanus I contains 65 pages, Selden 20 pages, and the Zouche-Nuttall facsimile edition 40 pages. We chose to use the Zouche-Nuttall facsimile edition over the complete 84-page edition because of its restored quality and high-quality scans available.

2.1 Data Processing

We used the Segment Anything Model (SAM) (Kirillov et al., 2023) from Facebook AI Research to extract individual figures from the three source codices[1]. Figures are representations of people or gods in Mixtec mythology and are composed of different outfits, tools, and positions. Their names are represented by icons placed near their position on a page. Each figure was annotated according to the page it was found, its quality as either `a`, `b`, or `c`, and its order within the page. An `a` quality rating indicated the entire figure was intact, regardless of minor blemishes or cracking, and could be classified by a human annotator as man or woman, standing or not. A `b` rating means that while the previous characteristics of the figure could be determined, significant portions of the figures were missing or damaged. The `c` rated figures were missing most of the definable characteristics humans could use to classify the sample.

2.2 Labeling Procedures

After figure segmentation and grading, we added classification labels to each figure. Literature describes representations of gender and poses in Mixtec codices to guide our classifications (Boone, 2000; Smith, 1973; Jansen, 1988; Williams, 2013; Lopez, 2021). We propose two binary classification tasks: Gender (`man`/`woman`) and Pose (`standing`/`not standing`). These two categories represent meaningful distinctions in Mixtec codices and allow for the exploration of deeper, more complex investigations into the structure of these documents. We refer to research on Mixtec codices to guide our human evaluation of figures. The criteria used by our human evaluators to determine gender class membership were loincloths and anklets for `men`, and dresses and braided hair for `women`. For the *standing* and *not standing* task, if the figure is clearly on two feet and in an upright position, it is labeled *standing*, and any other position is labeled `not standing`. Two team members tagged the images for both categories independently and then verified the results with each other using the process of inter-rater reliability (Hallgren, 2012).

2.3 Dataset Statistics

Codex Vindobonensis Mexicanus I represents the largest proportion of the 1285 figures with 714, Codex Selden has 307, and Codex Zouche Nuttall is the smallest with 264. Codex Vindobonensis Mexicanus I contains 573 men and 141 women, Selden 74 men and 233 women, and Zouche-Nuttall 256 men and 8 women. This imbalance in each dataset can be attributed to the fact that each codex is centered on a different figure. The Pose category follows a similar proportion split, however, a `not standing` position outweighs `standing`, for each codex. The reason for this is unclear, although given the number of ceremonies that each codex describes, which entails a seated or kneeling position, this balance intuitively makes sense. The quality of the figures is largely dominated by the `a` classification with 1086 figures, followed distantly by `b` at 170 figures, and `c` com-

[1]Each codex we used were high-quality and designated as free for non-commercial use or provided by national libraries

prising only 29 figures. Of these totals, the Zouche-Nuttall accounts for 263 a, only one b designation, and zero c figures. The Selden contains 254 a classifications, 46 marked with b, and 7 c. Finally the Vindobonensis Mexicanus I has 568 a figures, 123 b, and 22 c. Given the small number of c samples across all three codices, we use all three categories in the model training and testing pipelines. These numbers can be viewed in Table 1.

3 Experiment

We describe the preprocessing, finetuning, and execution steps of this pipeline. We explore the hyperparameter space for each model first to find the optimal configuration to use during execution.

3.1 Preprocessing

For our model pipeline preprocessing, the figures are moved to tensors and then normalized to 224x224 pixels. We bias the loss function by weighting each class in the loss function by its inverse. Finally, due to the overall limited number of figures, and to prevent overfitting, we augmented the entire dataset by using random flips and blocking to increase the number of samples for training. The dataset is then split into training, testing, and validation sets, 60%, 20%, and 20% respectively. We set aside eight reference images to monitor which features of gender and pose are prevalent in activation and attention maps throughout training.

3.2 Models

Both CNNs and transformers are used in image classification (Lu et al., 2021). We fine-tuned popular vision models VGG-16 and ViT-16 to perform classification tasks and improve computational efficiency. We imported the models and their pretrained weights from the PyTorch library. We then unfroze the last four layers and heads of each model for training, as they are responsible for learning complex features specific to our classification tasks (Olah et al., 2017). Finally, the fully connected layer of each model was replaced by one matching our binary classification task.

3.2.1 Hyperparameters

Next, we explored the number of epochs, batch size, and learning rate of each of our models. We experimented with different batch sizes, ranging from 32 to 128, and opted for an average value of 64 as no size significantly outperformed the others. Once we finalized the hyperparameter space, we selected the loss function and optimizer according to the best practices associated with our pretrained models, VGG and ViT.

4 Model Evaluation

ViT performs consistently higher than VGG for these different learning rates, however, both returned strong results for each metric. The testing results for both ViT and VGG were high with a small standard deviation, around 98% and 1% standard deviation for both (see Table 2). Additional model evaluation results are listed in Appendix C.

Model	Task	Test Accuracy ± (stddev)
VGG-16	Gender	$0.978 \pm (0.009)$
VGG-16	Pose	$0.978 \pm (0.01)$
ViT-16	Gender	$0.977 \pm (0.009)$
ViT-16	Pose	$0.974 \pm (0.009)$

Table 2: Testing accuracy and their standard deviations for VGG-16 and ViT-16.

5 Discussion

The purpose of the experiments is to explore two research questions, namely: *Can CNN and transformer-based models be finetuned to classify figures from a Mixtec Codices dataset?* and *Does the model identify the same features experts do?* To answer the first question, we analyze and compare the performances of both the pretrained ViT and VGG models. Both models achieve great results across training, validation, and testing phases when using an appropriate learning rate. Smaller learning rates require more epochs to converge, as the steps are smaller, but are less likely to miss a minimum loss. On the other hand, larger learning rates require fewer epochs, but may not converge. As we can see in Figure 3, ViT converges for almost all learning rates, and so could be used in environments where compute resources are lacking.

Features and Literature. We assigned reference images for each class (man and woman, and standing/not standing) to understand which features each model learned, as well as to compare these learned features to those highlighted by experts. During training, we generated visualizations of activation and attention per pixel to view how the models learned important features over time. In the left image in Figure 1, the ViT model assigned higher attention to areas corresponding to

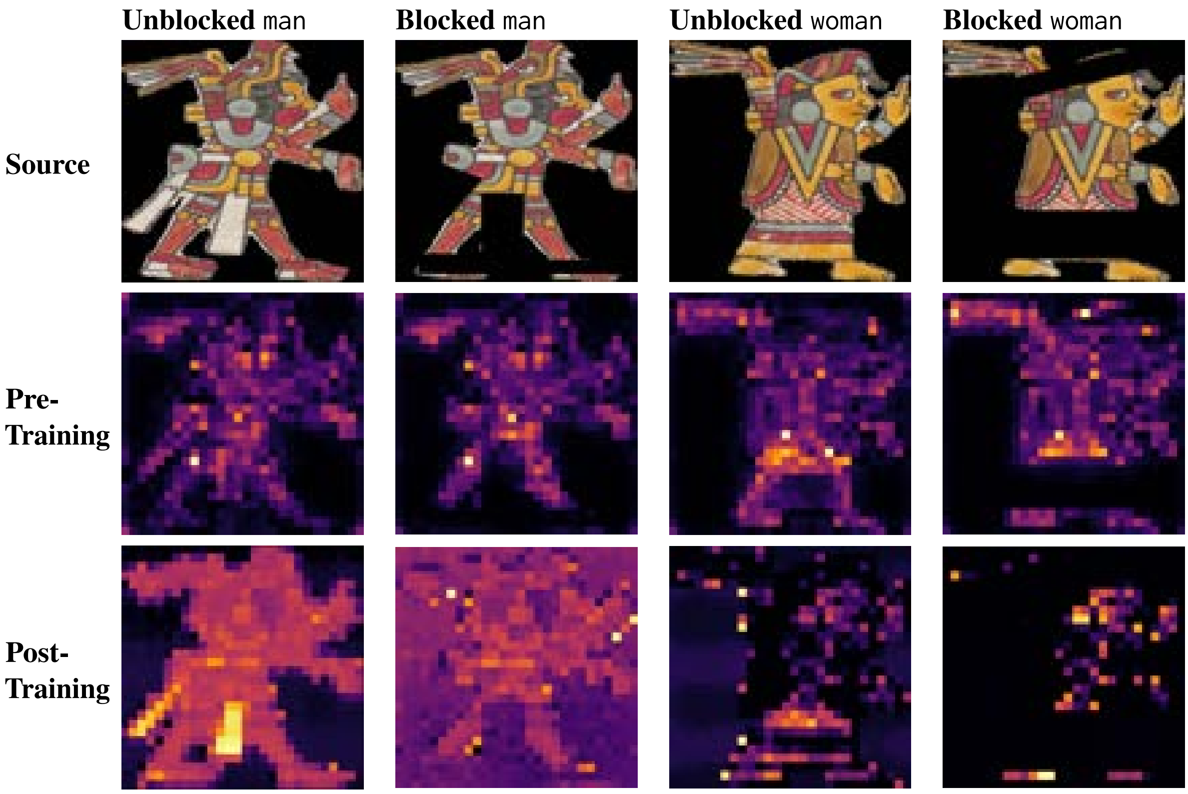


Figure 1: ViT-16 Mean Attention Maps for man and woman. The top row shows original reference images for both blocked and unblocked conditions. The next row shows attention maps extracted before the first epoch of training for man, and woman (right), for both blocked and unblocked conditions. The bottom row contains attention maps after the last epoch of training. The model shows increased attention in the loincloth area for an unblocked man, and the skirt area for an unblocked woman, which follows expert opinion. In the blocked conditions, different areas than the noted features are highlighted (woman), or do not converge to any particular area at all (man).

loincloths on man. On the right, ViT shows increased attention to the poncho area on a woman. We confirm that these are both features noted by domain experts (Boone, 2000). To verify that the model is indeed identifying the same features noted in literature, we masked attributes on the reference images. These features were earlier noted as discriminatory for human evaluators labeling gender: loincloths and anklets for man, and braided hair and dresses for woman. We extended our reference image set by adding three variations to each image: either blocked hair, blocked skirt, or both for woman. This process was replicated for the two features indicative of man. We then tested the fine-tuned models on the unblocked and blocked reference images and generated class activation and attention maps. ViT correctly predicted 100% of the unblocked reference images, 79% of the singly blocked images, and 63% of the double blocked images. Figure 1 shows the activation maps of the doubly blocked images. The model fails to find defined areas of attention. This again verifies that the model is learning features defined in literature.

6 Summary

In this paper, we presented a low-resource dataset of figures from three Mixtec codices: Zouche-Nuttall, Selden, and Vindobonensis Mexicanus I. We extracted the figures using Segment Anything Model and labeled them according to gender and pose, two critical features used to understand Mixtec codices. Using this novel dataset, we finetuned the last few layers of CNN and transformer-based foundational models, VGG-16 and ViT-16 respectively, to classify figures as either man or woman and standing or not standing. We found that both models have high accuracy with this task, but that ViT-16 may be more reliable for varying learning rates. We confirmed that the models are learning the features said to be relevant by experts using class activation maps and targeted blocking of said features. Given that these models can reliably classify figures from a low-resource dataset, this research opens the door for further processing and analysis of Mixtec Codices. The codices themselves are highly structured and carry a narrative woven through each scene. Finetuned state-of-the-art models could be combined to classify segmented figures within a scene, as well as classify the relationship between figures. These relationships would then be used to extract the narrative from a codex, as defined by domain experts.

7 Limitations

The Mixtec civilization produced many of the available codices, however, conquest and the passage of time have left us with only a few remaining high-quality samples (Boone, 2000). Fortunately, many of the surviving codices still contain examples of scenes and can be used to build a digitized corpus for machine processing. We chose popular models to demonstrate our method. We believe other architectures would have similar results. The quality results in both models show a specialized architecture is not required for accuracy. We have not yet explored more environmentally efficient models. Both models we adopt use pretrained classifiers, each trained on data not specific to our domain. The models inherit all biases previously encoded in the model. We have not investigated how these biases may affect downstream tasks. The finetuned models generated few errors in our investigation, however, we are unaware of how these biases may result in unintended effects.

We selected classification tasks that are well understood within the Mixtec research community, namely: `man` and `woman`, and `standing` and `not standing`. Many experts disagree on the interpretation of scenes across codices. For instance, some early 20th-century scholars have stated cannibalism and human sacrifice are depicted within the codices (Pohl, 1994), while others contend that these scenes should be understood as metaphorical interpretations (Lopez, 2021; Lopez and Collver, 2022). This work is an initial investigation into Mixtec and low-resource, semasiographic languages. We are prohibited from deeper explorations until we align our research direction with present communal, cultural, and anthropological needs. Support from Mixtec domain experts and native Mixtec speakers is essential for continued development.

References

Liza Bakewell and Byron Hamann. 2023. Codex selden.

Elizabeth Hill Boone. 2000. *Stories in red and black: Pictorial histories of the Aztecs and Mixtecs*. University of Texas Press.

Alfonso Caso. 1964. Códice selden. *Sociedad Mexicana de Antropología, Mexico City*.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale.

Sam Forstmann. 2023. Codex zouche-nuttall.

Kevin A Hallgren. 2012. Computing inter-rater reliability for observational data: an overview and tutorial. *Tutorials in quantitative methods for psychology*, 8(1):23.

Maarten Jansen. 1988. The art of writing in ancient mexico: an ethno-iconological perspective. *Visible Religion*, 6:86–113.

Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. 2023. Segment anything.

Walter Lehmann and Ottokar Smital. 1929. Codex vindobonensis mexic. 1. *Faksimileausgabe der mexikanischen Bilderhandschrift der Nationalbibliothek in Wien. Verlag von Anton Schroll & Co, Vienna*.

Felicia Rhapsody Lopez. 2021. *Women, Childbirth, and the Sticky Tamales: Nahua Rhetoric and Worldview in the Glyphic Codex Borgia*, chapter 4. University of New Mexico Press.

Felicia Rhapsody Lopez and Jordan Collver. 2022. How to read an aztec comic: Indigenous knowledge, mothers' bodies, and tamales in the pot. *Bobcat Comics*.

Kangrui Lu, Yuanrun Xu, and Yige Yang. 2021. Comparison of the potential between transformer and cnn in image classification. In *ICMLCA 2021; 2nd International Conference on Machine Learning and Computer Application*, pages 1–6. VDE.

Zelia Maria Magdalena Nuttall. 1902. Facsimile.

Chris Olah, Alexander Mordvintsev, and Ludwig Schubert. 2017. Feature visualization. *Distill*, 2(11):e7.

John Pohl. 1994. *The politics of symbolism in the Mixtec codices*. Vanderbilt University.

Geoffrey Sampson. 2015. Writing systems: methods for recording language. In *The Routledge handbook of linguistics*, pages 47–61. Routledge.

Karen Simonyan and Andrew Zisserman. 2015. Very deep convolutional networks for large-scale image recognition.

Mary Elizabeth Smith. 1973. *Picture writing from ancient southern Mexico; Mixtec place signs and maps.*, [1st ed.]. edition. The Civilization of the American Indian series, 124. University of Oklahoma Press, Norman.

Unbekannt. 1449. Bilderhandschrift: Sog. codex mexicanus bzw. codex yuta tnoho.

Robert Lloyd Williams. 2013. *The Complete Codex Zouche-Nuttall*. University of Texas Press.

A Example Codex Pages

Figure 2 show example pages from the three codices. The Codex Vindobonensis Mexicanus I and Facsimile Edition of the Codex Zouche-Nuttall (Nuttall, 1902; Forstmann, 2023) we reference were digitized by The Austrian National Library (Lehmann and Smital, 1929; Unbekannt, 1449), and the Codex Selden was digitized within the Mesolore (Caso, 1964; Bakewell and Hamann, 2023).

B Model Execution

Model training and inference were performed on an Nvidia A100 on the HiPerGator cluster using PyTorch 2.1 and CUDA 11. For both VGG and ViT, each run took up to 25 minutes to complete. Before the first and after the last epoch of training, an activation map for VGG and an attention map for ViT is output for each reference image. We then ran the testing phase of the model pipeline using the optimal hyperparameters found during training and validation. Testing is run 30 times for each model and classification task and the performance scores are averaged to measure the reliability of the model.

C Model Evaluation

For each training and validation run, we collected metrics such as accuracy, F1, recall, loss, and precision. The accuracy results from training for varying levels of learning rates are presented in Figure 3 for both VGG and ViT and both classification conditions. Hyperparameter investigations revealed that the accuracy for training and validation converged around 100 epochs and the ideal learning rate was 0.00025.

D Reference Images

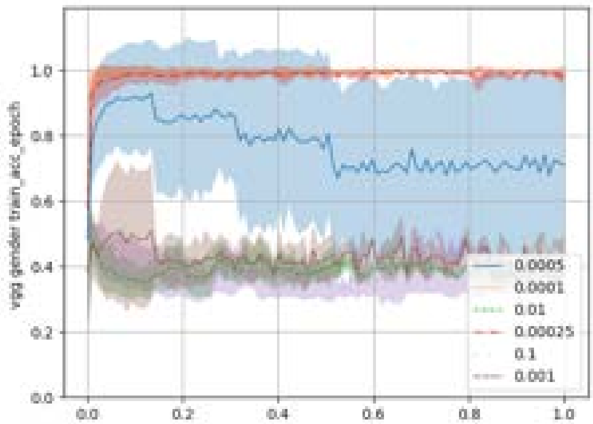
To observe the model's feature identification throughout the development process, we set aside a group of reference images with equal numbers

(a) Page 45 of the Codex Zouche-Nuttall.

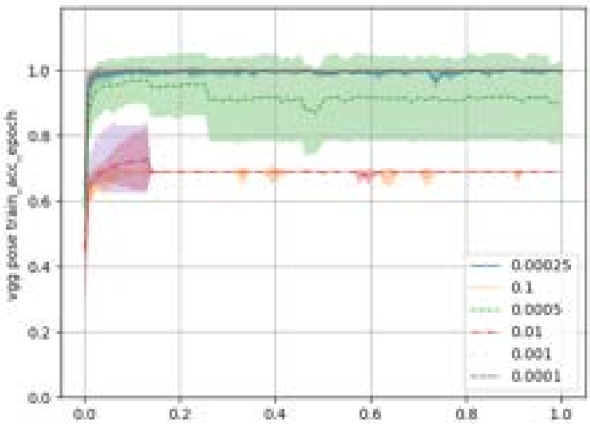
(b) Page 4 of the Codex Selden

(c) Page 13 of the Codex Codex Vindobonensis Mexicanus

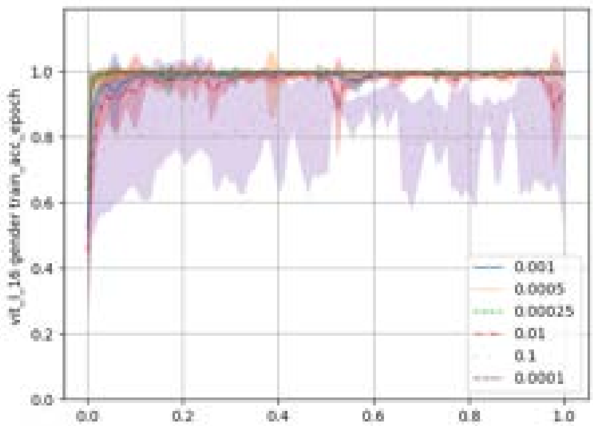
Figure 2: Sample pages from each of the three source codices: Codex Zouche-Nuttall, Codex Selden, and Codex Vindobonensis Mexicanus I.



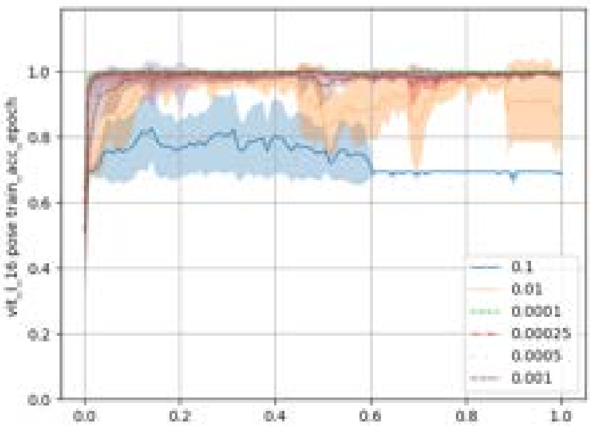
(a) VGG-16 model Gender accuracy of the training set.



(b) VGG-16 model Pose accuracy of the training set.



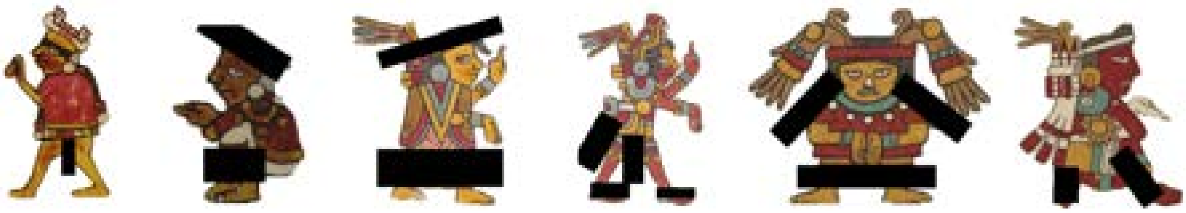
(c) ViT-16 model Gender accuracy of training set.



(d) ViT-16 model Pose accuracy of training set

Figure 3: Training accuracy vs. percentage to completion for a given run. Graphs execution across learning rates. Smaller learning rates converged faster across all runs, while some larger learning rates failed to converge.

(a) Unblocked Reference images.



(b) Blocked Reference images.

Figure 4: Six reference images sourced from our three references codices. From left to right the image show a `man standing`, `woman not standing`, `woman standing`, `man standing`, `woman standing`, `man not standing`. The bottom row shows the reference images with the full blocking available for each image.

labeled `man`/`woman` and `standing`/`not standing` as shown in Figure 4a. Each codex is represented equally in the set of reference images. We then created at most three variations for each image. The first two variations were generated by blocking one of the defining features, and the last involved blocking both. If a figure did not have one of the features, (i.e. a `man` without anklets, or a `woman` without braided hair) then only one variation was created. Before and after model training and validation, we used model inference on the reference images and output class activation and attention maps for VGG-16 and ViT-16 respectively. Figure 4b shows examples of both unmasked and fully masked images.

E Code & Data

Our source code and data for these experiments can be found in a GitHub repository `https://github.com/ufdatastudio/mixteclabeling`.

A New Benchmark for Kalaallisut-Danish Neural Machine Translation

Ross Deans Kristensen-McLachlan
Center for Humanities Computing
Department for Linguistics,
Cognitive Science, and Semiotics
Aarhus University
rdkm@cc.au.dk

Johanne Sofie Krog Nedergård
Department of Nordic Studies
and Linguistics
University of Copenhagen
jskn@hum.ku.dk

Abstract

Kalaallisut, also known as (West) Greenlandic, poses a number of unique challenges to contemporary natural language processing (NLP). In particular, the language has historically lacked benchmarking datasets and robust evaluation of specific NLP tasks, such as neural machine translation (NMT). In this paper, we present a new benchmark dataset for Greenlandic to Danish NMT comprising over 1.2m words of Greenlandic and 2.1m words of parallel Danish translations. We provide initial metrics for models trained on this dataset and conclude by suggesting how these findings can be taken forward to other NLP tasks for the Greenlandic language.

1 Introduction

Greenlandic (Kalaallisut) is an Inuit-Yupik-Unangan language spoken by around 60,000 people in Greenland and Denmark. While Greenlandic is classified as 'Vulnerable' according to the Endangered Languages Project.[1], it is nevertheless relatively healthy. Local governance, media, and schooling up to university level are conducted either purely in Greenlandic, the sole official language since 2009, or in both Greenlandic and Danish, the colonial language (Compton, 2024).

Greenlandic has a number of comparatively rare linguistic features, not least of which is the prominent use of polysynthesis (Fortescue, 2007). Morphological complexity not only manifests in word-final inflections (mood, person, number, and so on) but also the theoretically indefinite productivity of adding morphemes to stems and concatenating other morphemes. For example, *palasi* means 'priest', *palasi-nngor-poq* means 'he becomes a priest', *palasi-nngor-tip-paa* means 'he causes him to become a priest', and *palasi-nngor-tit-si-neq* means 'the act of causing someone to become a priest'.

Additionally, there is widespread morphophonological assimilation or fusion at the morpheme boundaries (Fortescue, 1980). While these changes are to a large extent predictable, they can make it difficult - both for machine learning models and humans who are not native speakers of Greenlandic - to analyze precisely which morphemes comprise any given word.

1.1 From linguistics to NLP

Greenlandic hence poses a number of specific challenges to contemporary machine learning-based approaches to NLP. For example, most contemporary NLP systems use sub-word tokenization strategies such as Byte Pair Encoding (BPE, (Zouhar et al., 2023). Given its morphological complexities, sub-word tokenization seems unsuited to working with Greenlandic data, and an informal consensus among experts in the language has been that it is hence not amenable to contemporary NLP techniques. Does this mean that the language is excluded from the fruits of recent advances in contemporary neural language technology?

We contend that this is not the case. It is true that there is still lacking scientific investigation into foundational aspects of how easily modern NLP methods can be applied to a morphologically complex, low-resource, indigenous language like Greenlandic. However, a rapidly growing body of research already exists for languages as morphologically diverse as Nahuatl, Raramuri, Shipibo-Konibo, and Wixarika (Mager et al., 2022) and indigenous languages closely related to Greenlandic (Liu et al., 2020a; Schwartz et al., 2020). Recent work has provided systematic analysis of challenges and methods involved the creation of NMT systems for these kinds of languages (Mager et al., 2023).

This paper aims to move the Greenlandic NLP more in this direction by introducing a benchmark dataset for Greenlandic to Danish and provide the

[1] https://www.endangeredlanguages.com/

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 50–55
June 21, 2024

first set of metrics on model performance. We specifically choose to translate to Danish, since this is a meaningful task given the complex social history between these two languages and cultures (Olsen, 2011; Kleeman-Andersen, 2021). In what follows, we outline the various steps taken to construct this dataset and the results of initial simple experiments.

1.2 Current state of Greenlandic NMT

Until recently, the only available tool for machine translation for Greenlandic to Danish was Nutserut, a rule-based approach to machine translation developed and maintained by Oqaasileriffik, the Language Secretariat of Greenland.[2] Research into Greenlandic NMT is thin on the ground. Earlier work exists on Greenlandic to English NMT but this work is hampered by the synthetic nature of the training data (Jones, 2022). There exists a growing body of work on related languages such as Inuktitut which investigates whether adding Greenlandic data to training pipelines might increase performance on English–Inuktitut NMT (to mixed results) (Roest et al., 2020). Beyond this, though, a survey of the existing literature suggests that there has to date been detailed empirical studies on the prospects and limitations of Greenlandic NMT.

Since we first started work on our project, a number of interested stakeholders have moved into this space, including the largest media house in Greenland[3] and Oqaasileriffik itself.[4] This is a positive development, since greater investment and engineering is likely to lead to growth and broader adoption of contemporary machine learning for Greenlandic. However, these tools are closed-source and do not provide transparent quantitative metrics for evaluating model performance. We are currently unaware of any work which has provided quantifiable metrics for Greenlandic to Danish NMT, meaning that the results presented here are the first such results on a benchmark dataset.

1.3 The problem of data

Thanks to the work done by custodians of Greenlandic, the language punches above its weight in terms of linguistic resources. For example, Oqaasileriffik have to date developed searchable text corpora; lexical resources such as dictionaries and terminology banks; and practical tools such as spellcheckers and text-to-speech models.[5] However, the language is still greatly under-resourced relative to other languages globally. This lack is most apparent in the context of well-designed Greenlandic-Danish parallel corpora. Currently there are no gold standard corpora in this area which can be used as a reliable benchmark for NMT.

This data scarcity has meant that rule-based approaches have dominated, since these approaches resolutely *do not* require large quantities of data. Nevertheless, rule-based approaches to language are now regularly replaced by or integrated alongside machine learning developed for high performance in low-data environments (Torregrosa et al., 2019; Huang et al., 2020). To ensure that Greenlandic is not left behind, it is necessary to explore all possibilities and to make the most of the available resources, even where they might not be ideally suited for the task.

For our experiments, we collected data with permission by scraping the public facing website of *Kalaallit Nunaata Radioa* (KNR), Greenland's national public broadcasting organization.[6] As a public broadcaster, KNR's data were freely accessible, and they have an official language policy necessitating that all texts published on their websites are published in both Greenlandic *and* Danish.

2 Methods

2.1 Data preperation

In May 2023, we scraped full articles for both Greenlandic and Danish versions of all articles stretching back to the first available digital texts. This created an initial corpus of roughly 72k Greenlandic language articles and around 63k Danish language articles. This is due to the fact that, as one goes further back in KNR's archives, it appears that earlier articles were not regularly translated into Danish, with the official dual translation policy only coming into effect in 2010.[7] Different translations of the same article are linked by a unique identifier, meaning that we were able to remove Greenlandic texts for which there was no translation. Scraped HTML files were consistently structured across the translations, meaning that it was simple task to automatically extract the main body text from each document creating a raw text

[2] https://nutserut.gl/pre
[3] https://www.sermitsiaq.ag/
[4] https://nutserut.gl/
[5] https://oqaasileriffik.gl/en/
[6] https://knr.gl/kl
[7] https://knr.gl/da/sprogpolitik-i-knr

corpus of parallel documents.

To create a parallel sentence corpus, we made use of a crude and efficient alignment algorithm. Documents were first split into sentences by tokenizing on common end-of-sentence punctuation such as periods, exclamation marks, and question marks. This resulted in each document being transformed into a list of sentences. We then compared the the overall number of sentences in each article between the Greenlandic and Danish version of the document. In the case of a mismatching number of sentences per document pair, we discarded this pair of articles from the corpus. If the number of sentences in an article matched, we assumed that there was a one-to-one mapping between sentences in the different translations of the text.

The final corpus hence comprises the sentence pairs from all of those articles which have the same number of sentences per article. While this approach is of course naïve, it was necessary given the lack of available resources to otherwise create a useable parallel corpus. We expand on this problem below in Limitations.

The final result of this process is a parallel corpus of around 73k sentence pairs, comprising around 1.2m words of Greenlandic and 2.1m words of Danish. This is comparable to previous studies working in similar linguistic contexts (Schwartz et al., 2020). Of this data, a randomly drawn sample of 1k sentences were held back as test data for evaluating model performance.

2.2 Model creation

All models were trained using *OpenNMT* with a *PyTorch* backend (Klein et al., 2017)[8]. BPE tokenizers were trained using *pyonmtok,* a wrapper for OpenNMT's tokenizer.[9]

Each experimental condition used the same Bi-LSTM encoder-decoder architecture adopting the default hyperparameters outlined in OpenNMT's documentation.[10] The only exceptions are the using of the Adam for optimization and an initial learning rate of 0.001. Each model ran for 100k training steps with model checkpoints saved after every 10k steps.

Alongside the custom RNN models outlined, we also tested the performance of state-of-the-art LLMs on this task. Using the OpenAI API, we performed zero-shot testing of *GPT-3.5-turbo* and *GPT-4* with the following prompt:

> *Translate this text from Greenlandic to Danish, without any additional comments or explanations: {text}*

3 Experiment

3.1 Hardware considerations

Local models were trained on a machine running Ubuntu (18.04.6 LTS) with 40 Intel(R) Xeon(R) Silver 4210 CPU cores and four Quadro RTX 8000 GPUs.

3.2 Evaluation metrics

Evaluating machine translation quantitatively is a notoriously fraught endeavour, with a number of different metrics proposed to quantify exactly how well any given model is performing (Popović, 2015; Chatzikoumi, 2020; Rei et al., 2020). Since no one metric robustly measures translation quality in a way which is entirely in line with expectations of human readers, we evaluate model performance using of a range of standard metrics.

Surface similarity is measured via n-gram overlap using BLEU (Papineni et al., 2002) and via character overlaps with chrF (Popović, 2015). Both sets of evaluations were performed using open-source and publicly available implementations of these algorithms.[11,12]

While both the BLEU and chrF metrics evaluate slightly different aspects of the generated translations, they are both ultimately based on the amount of string overlap between machine generated text and human references. However, it is of course true that any given sentence can be translated a seemingly indefinite number of ways while retaining the same meaning. In order to capture aspects of the *semantic similarity* machine generated translations and their references, we make use of BERTScore (Zhang et al., 2020).[13] This has been applied and shown to perform well in Danish contexts, such as evaluating abstractive text summarization (Kolding et al., 2023).

3.3 Results

The results for all models are shown in Table 1 below. We see that the Bi-LSTM model with the

[8] https://github.com/OpenNMT/OpenNMT-py
[9] https://github.com/OpenNMT/Tokenizer/
[10] https://opennmt.net/OpenNMT-py/options/train.html
[11] https://github.com/mjpost/sacrebleu
[12] https://github.com/google-research/google-research/tree/master/rouge
[13] https://github.com/Tiiiger/bert_score

Model	BERTscore (F1)	BLEU	chrF2
Bi-LSTM + 5k BPE	**0.74**	**16**	32.3
Bi-LSTM + 10k BPE	0.73	13	**32.5**
Bi-LSTM + 30k BPE	0.72	12	29.9
Bi-LSTM + 50k BPE	0.71	11	27.9
GPT-3.5-turbo	0.64	3	25.4
GPT-4	0.68	4	28.3

Table 1: Results across all model types

fewest number of joins performs best on this data, with the second highest scoring model being the model with the second lowest number of joins. In general, we see that increasing the number of BPE merges descreases performance in a proportional and linear way.

Perhaps contrary to expections, the LLM solutions perform notably worse than all of our much simpler, custom RNN models. However, this is likely due to the zero-shot nature of the task; additional experimentation is necessary to test the limits of LLMs for this particular tasks.

4 Discussion

4.1 What does this show?

The most striking takeaway from these preliminary results is that models generally perform reasonably well when evaluated using BERTScore, while scores tend to be much poorer for the n-gram and character-based metrics. Put simply, it would seem as though the translations produced by these models tend to be semantically close to the human generated sentences but are otherwise lexically or stylistically divergent from the human references.

Nevertheless, the numbers presented here are not widely different from research into similar indigenous linguistic contexts, such as reported BLEU scores for Yup'ik to English (≈13, (Liu et al., 2020a)) and Inuktitut to English (≈28, (Schwartz et al., 2020); see also (Nicolai et al., 2021)). Despite the widespread perception of the linguistic uniqueness of Greenlandic, it would seem that the language is nevertheless amenable to NMT.

Crucially, though, we also demonstrate that a smaller, simpler Bi-LSTM model currently outperforms more sophisticated LLM solutions. With a few-shot prompting regime and additional fine-tuning this could likely be improved, but it does provide a note of caution against immediately adopting "state-of-the-art" models without detailed testing and robust scientific evaluation.

4.2 Where next?

Our initial NMT experiments with Greenlandic to Danish are limited by our use comparatively simple architectures. Immediate next steps will be to experiment with more sophisticated neural network architectures such as transformer-based models, as well as the applicability of pre-trained multilingual embeddings such as *mT5* (Xue et al., 2021) or *mBART* (Liu et al., 2020b).

This opens up a wide range of possibilities, including practical technologies such as speech-to-text models and improved research methods for linguistic analysis and language modelling. This has the potential to contribute substantially to the scientific study of Greenlandic from the perspectives of cognitive science and language psychology, such as considering the relationship between sub-word tokenization and human morphological segmentation.

Finally, we aim for this to be a stepping stone towards collaboration with researchers currently engaged with similar work on other Inuit-Yupik-Unangan languages. Given the similarities between these languages, we believe that pooling resources could lead to substantial progress in language technology for languages in this region of the world.

5 Conclusions

This paper is a preliminary step towards training neural language technology for Greenlandic and, crucially, empirically testing both the possibilities and limitations of this approach. We present a benchmark dataset for Greenlandic-Danish NMT as well as providing initial metrics from simple models trained on this dataset. These initial experiments are not intended to provide complete, industry-strength machine translation for Greenlandic to Danish. Improvements in the area of

Greenlandic NMT and NLP more generally requires greater emphasis on the curation and stewardship of high quality training data. We believe that this focus would contribute greatly to Greenland's already rich linguistic cultural heritage.

Limitations

While we view these results positively, these trained models are far from production-ready or of practical use. The process for creating the parallel corpus is crude and involves a number of pragmatic decisions by the authors, neither of whom are native Greenlandic speakers. The collection algorithm outlined in Section 2.1 was designed as a "good enough" solution for initial experiments. However, greater quality control with more human intervention is required for future work to ensure that the corpus is in fact aligned.

KNR texts have some well-known limitations (Duus, 2012a,b; Hussain, 2018; Kleeman-Andersen, 2020). Several of the texts are originally written in Danish (largely by monolingual Danish-speaking or bilingual journalists) and subsequently translated to Greenlandic. Texts hence tend to be quite "literal" or non-idiomatic translations and thus appear somewhat unnatural to a Greenlandic speaker. The quality of the Greenlandic texts at KNR generally is a point of heated public debate in Greenland with many complaining about grammatical errors, repetitive expressions, and too much influence from Danish.

References

Eirini Chatzikoumi. 2020. How to evaluate machine translation: A review of automated and human metrics. *Natural Language Engineering*, 26(2):137–161.

Richard Compton. 2024. *38 Inuit-Yupik-Unangan: An overview of the language family*, pages 843–874. De Gruyter Mouton, Berlin, Boston.

Søren Duran Duus. 2012a. Journalister: Gør noget ved sproget i vores medier. *Sermitsiaq*.

Søren Duran Duus. 2012b. Knr efter sprogkritik: Vi kan blive bedre. *Sermitsiaq*.

Michael D. Fortescue. 1980. Affix ordering in west greenlandic derivational processes. *International Journal of American Linguistics*, 46(4):259–278.

Michael D. Fortescue. 2007. The typological position and theoretical status of polysynthesis. 5:1–27.

Jin-Xia Huang, Kyung-Soon Lee, and Young-Kil Kim. 2020. Hybrid translation with classification: Revisiting rule-based and neural machine translation. *Electronics*, 9(2).

Naimah Hussain. 2018. *Journalistik i små samfund: Et studie af journalistisk praksis på grønlandske nyhedsmedier*. Ph.D. thesis, Roskilde University.

Alex Jones. 2022. Finetuning a kalaallisut-english machine translation system using web-crawled data. *CoRR*, abs/2206.02230.

Camilla Kleeman-Andersen. 2020. Plastikblomster og tungeløse grønlændere - følelser i sprogdebatten 2009-2019.

Camilla Kleeman-Andersen. 2021. Den evigt nærværende koloniale fortid. In *Sprogs status i Rigsfællesskabet 2031*, pages 38–40. Københavns Universitets Humanistiske Fakultet.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. Association for Computational Linguistics.

Sara Kolding, Katrine Nymann, Ida Hansen, Kenneth Enevoldsen, and Ross Kristensen-McLachlan. 2023. DanSumT5: Automatic abstractive summarization for Danish. In *Proceedings of the 24th Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 248–264, Tórshavn, Faroe Islands. University of Tartu Library.

Christopher Liu, Laura Dominé, Kevin Chavez, and Richard Socher. 2020a. Central Yup'ik and machine translation of low-resource polysynthetic languages. *CoRR*, abs/2009.04087.

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

Manuel Mager, Rajat Bhatnagar, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2023. Neural machine translation for the indigenous languages of the Americas: An introduction. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 109–133, Toronto, Canada. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Elisabeth Mager, Katharina Kann, and Thang Vu. 2022. BPE vs. morphological segmentation: A case study on machine translation of four polysynthetic languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 961–971, Dublin, Ireland. Association for Computational Linguistics.

Garrett Nicolai, Edith Coates, Ming Zhang, and Miikka Silfverberg. 2021. Expanding the JHU Bible corpus for machine translation of the indigenous languages of North America. In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 1–5, Online. Association for Computational Linguistics.

Carl Chr. Olsen. 2011. Sproglovgivning under grønlands selvstyre. *Sprog i Norden*, 42(1):25–30.

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.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. 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.

Christian Roest, Lukas Edman, Gosse Minnema, Kevin Kelly, Jennifer Spenader, and Antonio Toral. 2020. Machine translation for English–Inuktitut with segmentation, data acquisition and pre-training. In *Proceedings of the Fifth Conference on Machine Translation*, pages 274–281, Online. Association for Computational Linguistics.

Lane Schwartz, Francis M. Tyers, Lori S. Levin, Christo Kirov, Patrick Littell, Chi-kiu Lo, Emily Prud'hommeaux, Hyunji Hayley Park, Kenneth Steimel, Rebecca Knowles, Jeffrey Micher, Lonny Strunk, Han Liu, Coleman Haley, Katherine J. Zhang, Robbie Jimerson, Vasilisa Andriyanets, Aldrian Obaja Muis, Naoki Otani, Jong Hyuk Park, and Zhisong Zhang. 2020. Neural polysynthetic language modelling. *CoRR*, abs/2005.05477.

Daniel Torregrosa, Nivranshu Pasricha, Maraim Masoud, Bharathi Raja Chakravarthi, Juan Alonso, Noe Casas, and Mihael Arcan. 2019. Leveraging rule-based machine translation knowledge for under-resourced neural machine translation models. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 125–133, Dublin, Ireland. European Association for Machine Translation.

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, Online. Association for Computational Linguistics.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with BERT. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Vilém Zouhar, Clara Meister, Juan Gastaldi, Li Du, Tim Vieira, Mrinmaya Sachan, and Ryan Cotterell. 2023. A formal perspective on byte-pair encoding. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 598–614, Toronto, Canada. Association for Computational Linguistics.

Morphological Tagging in Bribri Using Universal Dependency Features

Jessica Acton Karson
Dartmouth College
jess.a.karson.23@dartmouth.edu

Rolando Coto-Solano
Dartmouth College
rolando.a.coto.solano@dartmouth.edu

Abstract

This paper outlines the Universal Features tagging of a dependency treebank for Bribri, an Indigenous language of Costa Rica. Universal Features are a morphosyntactic tagging component of Universal Dependencies, which is a framework that aims to provide an annotation system inclusive of all languages and their diverse structures (Nivre et al., 2016; de Marneffe et al., 2021). We used a rule-based system to do a first-pass tagging of a treebank of 1572 words. After manual corrections, the treebank contained 3051 morphological features. We then used this morphologically-tagged treebank to train a UDPipe 2 parsing and tagging model. This model has a UFEATS precision of 80.5 ± 3.6, which is a statistically significant improvement upon the previously available FOMA-based morphological tagger for Bribri. An error analysis suggests that missing TAM and case markers are the most common problem for the model. We hope to use this model to expand upon existing treebanks and facilitate the construction of linguistically-annotated corpora for the language.

Resumen

Etiquetado morfológico del Bribri usando rasgos de Universal Dependencies. Este artículo presenta un experimento para el etiquetado automático de la morfología de las palabras en una colección de árboles sintácticos de dependencia en bribri, un idioma indígena de Costa Rica. El esquema *Universal Features* es un componente de etiquetado morfológico de *Universal Dependencies*, un estándar para el análisis sintáctico de oraciones. Este esquema busca poder etiquetar cualquier lengua del mundo y sus diversas estructuras (Nivre et al., 2016; de Marneffe et al., 2021). Empezamos el proyecto usando un sistema basado en reglas para etiquetar automáticamente una colección de árboles sintácticos con 1572 palabras. Después de una corrección manual, la colección tenía un total de 3051 etiquetas morfológicas. Esta nueva colección de árboles se usó para entrenar un modelo de UDPipe 2 que pudiera hacer etiquetado y análisis sintáctico automáticamente. Este modelo tiene una precisión de UFEATS de 80.5 ± 3.6, lo cual es una mejora estadísticamente significativa con respecto a los etiquetadores basados en FOMA disponibles para el bribri. Un análisis de errores sugiere que el principal problema para el modelo fue el no poder producir algunas etiquetas de TAM y de caso en la salida. Esperamos usar este modelo para expandir las colecciones de árboles ya existentes, y así facilitar la construcción de corpus anotados lingüísticamente para esta lengua.

1 Introduction

It is essential that the fields of linguistics and Natural Language Processing dedicate time and resources towards smaller, Indigenous, and minority languages. Building annotated and tagged corpora for smaller languages supports the expansion of NLP capabilities in processing them, and could potentially expand the languages' domain of usage and help create tools that aid in language revitalization and normalization. In this paper we worked on one small building block of future NLP tools: the morphological tagging of corpora in the Bribri language, an Indigenous language from Costa Rica. In section 1 we review the process of morphological tagging and describe the Bribri language's vitality and context. In section 2, we describe an algorithm for rule-based tagging, and how we used this for our first attempt at automatic tagging. After correcting any resulting errors, we trained a deep-learning based model to perform future tagging. Section 3 describes the tags applied to the treebank, compares the model's performance to a previously available tagger, and describes the errors that the model is making in its tagging output. Finally, section 4 describes some limitations of the tagging scheme when describing Bribri data, as well as directions of future work.

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 56–66
June 21, 2024 ©2024 Association for Computational Linguistics

1.1 Morphological Analysis and Tagging

Morphological analysis is the systematic breakdown of words into smaller pieces that reflect units of meaning (i.e. morphemes). For example, the input `cats` would return the output `cat-s`. In the context of natural language processing, morphological analysis can be paired with the task of morphological tagging. In morphological tagging, a word like `cats` would produce an output like `cat+[N;PL]`, `NN2` or `Number=Plur`. These three examples, which use different standards, indicate in different ways that the word is plural. This tagging can support the building of annotated corpora, which in turn allows for more advanced linguistic research, but also for more advanced NLP tasks such as lemmatization and disambiguation tasks.

Morphological analysis is undertaken using different standards and can use language-specific or language family-specific differentiations. The UCREL CLAWS7 tagset (UCREL, 2011), for example, is made for English and uses a one-tag-per-word system which labels both the part of speech and some related morphological characteristics (e.g. `cats` → `NN2` 'common noun plural'). The UniMorph standard (McCarthy et al., 2020) attempts to describe all languages using the same tags, and it uses a one-to-many system where one word can have several tags depending on its part of speech and its morphemes (e.g. `cats` → `cat+[N;PL]`). The Universal Dependencies' (Nivre et al., 2020) Universal Features schema (UFEATS) also attempts to offer coverage for the morphology of every language. It uses its own set of tags, leaving out the part of speech but including one or more morphological tags per word (e.g. `cats` → `Number=Plur`). This standard is used to annotate numerous treebanks in Universal Dependencies, including an existing one for Bribri (see section 2.1 below). Because of this, and because it would provide an additional way to query the existing treebank for specific morphemes, the UFEATS schema will be used in this work.

1.2 Automatic Morphological Analysis and Indigenous Languages

Morphological analysis for under-resourced Indigenous languages presents unique challenges for several reasons. The limited availability of data complicates progress when determining meaningful connections between words or units within words. Additionally, the input of the language data can have inconsistencies due to lack of standardization in orthography[1] and unaccounted-for variation in data collection.

Despite these challenges, there has been work for Universal Features tagging in languages of the Americas. There are Universal Features tagged datasets for Tupí languages (Rodríguez et al., 2022), K'iche (Tyers and Howell, 2021; Tyers and Henderson, 2021) and Yupik (Park et al., 2021). There are also languages tagged using UniMorph, such as Kanien'kéha (Kazantseva et al., 2024), Plains Cree, Gitksan, Asháninka, Aymara, Seneca, Dakota, Otomí, Mixtec, Chatino, Zapotec and Tohono O'odham (Batsuren et al., 2022). There is also work on using finite state transducers to do morphological tagging and segmentation. Some languages where such taggers exist are Haida (Lachler et al., 2018), Michif (Davis et al., 2021), Cree (Snoek et al., 2014), Lushootseed (Rueter et al., 2023), Wixarika (Mager et al., 2018a), Nahuatl (Pugh and Tyers, 2021) and Guaraní (Kuznetsova and Tyers, 2021). Languages where custom methods have been used for morphological tagging and segmentation include Inuktitut (Khandagale et al., 2022; Le and Sadat, 2021), Seneca (Liu et al., 2021), Quechua (Llitjós et al., 2005), Shipibo-Konibo (Mercado-Gonzales et al., 2018) and Mapugundun (Molineaux, 2023). In addition to these papers, Mager et al. (2018b) document additional efforts to work on morphological analysis and tagging of Indigenous languages in the Americas.

1.3 Bribri Language

Bribri is a language of the Chibchan family with approximately 7000 speakers in Costa Rica (INEC, 2011). Bribri is closely related to the Cabécar language also spoken in Costa Rica, and it is more distantly related to other Chibchan languages like Malecu and Ngäbe (Quesada, 2007). Bribri is a vulnerable language (Sánchez Avendaño, 2013), which means that there are children in the community who only speak Spanish. As Bribri is a low-resource language, documentation and natural

[1] We do not necessarily advocate standardization here. This is a decision that needs to be taken by the community. Moreover, very valuable materials are being published in Bribri using orthographic conventions unique to each author (MEP, 2013; García Segura, 2016; Jara Murillo and García Segura, 2022). This is a relatively common situation in writing within Indigenous communities, and pursuing a single standard might be detrimental to language revitalization (De Korne and Weinberg, 2021). We believe that this input issue is the engineers' problem to solve, not necessarily the communities.

language processing applications for the language are limited and difficult to make. Moreover, the particularities of Bribri morphosyntax make transfer learning from large-resource languages difficult. For example, the language is morphologically ergative, it has numerical classifiers and a complex deictic system, and it has a verbal system where "now" is not the locus of division between tenses, but rather "the night before". Examples of these phenomenona are presented in section 4.

2 Methodology

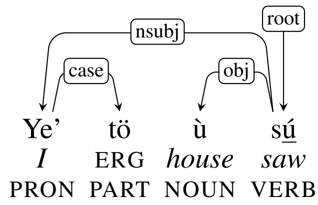
Our overall goal is to improve morphological tagging for Bribri. In this section we will explain how we used a rule-based algorithm to tag the existing treebank using Universal Dependencies Features. After manual correction, we tested these features by using them to train parsing models. When those models were trained we compared their performance to that of a pre-existing morphology analysis system for Bribri.

2.1 Bribri Data and Pre-existing Algorithms

There is relatively little unlabeled data for Bribri. The main data source is the oral corpus by Flores-Solórzano (2017a,b), which contains both text and audio for Bribri conversations. There are some printed materials which could provide written data, such as textbooks (Constenla et al., 2004; Jara Murillo and García Segura, 2013), a grammar book (Jara, 2018), two dictionaries (Margery, 2005; Krohn, 2021) and several educational books (Sánchez Avendaño et al., 2021a,b). Using this data there has been progress in NLP, in subfields such as speech recognition (Coto-Solano, 2021), forced alignment (Coto-Solano and Solórzano, 2016; Solórzano and Coto-Solano, 2017; Coto-Solano et al., 2022), machine translation (Feldman and Coto-Solano, 2020; Kann et al., 2022; Jones et al., 2023) and the study of semantics through embeddings (Coto-Solano, 2022). The work also includes the development of tools to extend the usage of the language, such as keyboards (Solórzano, 2010) and digital dictionaries (Krohn, 2020).

There are a few labelled datasets for the language (e.g. Ebrahimi et al. 2021), and one of them is a dependency treebank (Coto-Solano et al., 2021) tagged with Universal Dependencies v2 (Nivre et al., 2020) and stored in the CoNLL-U format. This treebank contains 315 sentences (1572 tokens) from some of the unlabeled sources above, and it includes information on part-of-speech and dependency arcs and labels. Figure 1 shows an example parse from this treebank.

(1) *Ye' tö ù sú̱* 'I saw the house'



Morphological analysis is one of the areas where there has been previous NLP research for Bribri. The state-of-the-art tagger is the Flores-Solórzano (2017b) FOMA-based tagger, which was built to tag the oral corpus (Flores-Solórzano, 2017a). It uses a finite state transducer (FST) which takes one word at a time, processes its characters one at a time, and follows a path that will ultimately lead the FST to an end node with a list of possible morphological features. Table 1 shows the morphological features for example sentence 1. The first word, *ye'* 'I' is correctly predicted as a first person singular pronoun. The second word *tö*, has three possible predictions: it could be a verb, a conjunction, or the ergative postposition. Here the third option is the correct one, but the FST does not output probabilities, so knowing this would require a human determination or an additional module. The third word, *ù* 'house' is correctly predicted as a noun (*sustantivo* in Spanish), but the tag does not specify the absolutive case that the noun is in. Finally, the fourth word *sú̱* 'saw' only has +? as its morphological tag. This means that the FST could not find the word amongst its states, and therefore cannot provide any morphological information.

Word	Features
ye'	`+1PSg`
tö	`te+V+Imp1Intran`
	`+Conj[subordinada]`
	`+Posp[Erg]`
ù	`u+Sust`
sú̱	`+?`

Table 1: FOMA-based morphological features for the Bribri sentence *Ye' tö ù sú* 'I saw the house'.

2.2 Assignment of Universal Features

The challenge we are trying to solve is to improve existing morphological tagging so that it can tag any word in the Bribri language, not just those in the existing FST. In order to do this, we wrote a series of rules (regular expressions) to tag the sentences using Universal Dependency Features. These regular expressions were created based off orthological patterns determined by the researchers with support from previously established character patterns noted in resources such as Flores-Solórzano's work with verbal conjugation (Flores-Solórzano, 2017b). In this paper we will focus our tagging on some of the major parts of speech: verbs, nouns, pronouns, adjectives and copulas. Here we detail the rules for the parts of speech we selected for this work.

2.2.1 First Pass of the Verbs

The first step in processing this data is to compile a list of the verbs present in the text, and specify which were transitive and which were intransitive. Given a transitive or intransitive verb in the surrounding context, we can then determine the case of nouns and pronouns, such as the ergative and absolutive cases which are vital to the structure of Bribri.

After this verb list is compiled, we used a series of regular expressions to assign Universal Dependency morphological features to each of the verbs. Once a `VERB` label is found in the part of speech column of the CoNLL-U, a regular expression would be used to find the conjugation of the verb and find its morphological features. For example, the regular expression `r".*r$"` is used to find the imperfective middle voice verbs. These tags would then be inserted into the 5th position of a new CoNLL-U file. For example, the verb *tkër* 'to be sitting' matches the previous regular expression, and so it would receive the tags `Aspect=Imp|Mood=Ind|Tense=Pres|VerbForm=Fin|Voice=Mid`.

2.2.2 Nouns

After the first pass of the verbs, nouns were analyzed for their plurality using the regular expression `r".*pa$"`, which triggers the tagging of that `NOUN` with `Number=Plur`. Then the cases of the nouns were determined by the presence of transitive or intransitive verbs either directly after or two words after the noun. `NOUN` subjects near verbs in the transitive verb list would receive the `Case=ERG` tag. `NOUN` subjects near verbs in the intransitive verb list would receive the `Case=ABS` tag. Finally, if the noun was an object, it would receive the `Case=ABS` tag as well.

2.2.3 Adjectives

For the most part, Bribri adjectives do not show number agreement with their nouns. However, there are a few adjectives which have irregular plural forms. For example, the word *tsîr* 'small' has the plural form *tsítsi*. We manually assessed the adjectives in the treebank and tagged the irregular plurals as `Number=Plur`.

2.2.4 Pronouns

Pronouns were analyzed in the same way as the nouns and checked for `Case` and `Number`. However, unlike nouns, the pronouns were also tagged for `Person` and for `Type`, such as personal and reciprocal pronouns. The first person plural pronouns were also tagged for `Clusivity` (i.e. *se'* 'inclusive we' and *sa'* 'exclusive we').

Possessive pronouns were also tagged. They are phonologically the same as the personal pronouns (compare *ye'* 'I' with *ye' ù* 'my house') so their `Poss=Yes` status was determined their by position directly preceding `NOUN` tokens.

2.2.5 Second Verb Pass and Copulas

A second round of verb analysis was then completed so that the newly tagged `Person` features of nouns and pronouns could be used to determined the `Person` value for verbs that appeared in the immediate context of those nominals. If the `VERB` has a `NOUN` subject, then the tag `Person=3` is assigned to the `VERB` automatically. If a `PRON` is the subject, then the `Person` of the `VERB` is directly copied from that of the `PRON`.

The copula *dör* is a special part of speech. Copulas do not behave morphologically like verbs: They don't have TAM suffixes like most action verbs, and they don't have plural forms like most positional verbs. Copulas, however, are obligatory in equative sentences, and pronoun subjects can present weak forms next to both verbs and copulas. Because of this similarity to verbs, copulas were tagged for `Person` in the same way as verbs, associating their `Person` to the that of the surrounding nouns or pronouns.

2.3 Training and Statistical Comparisons

The procedure described above was used to tag the treebank automatically. After the first and second passes, a manual revision was carried out by the researchers to correct the errors of the rule-based predictions. Approximately 24% of the 330 verbs were not recognized by the regular expressions, and so they were corrected manually by the authors, using the Flores-Solórzano (2017b) verbal description, the Jara (2018) grammar, the Constenla et al. (2004) textbook and the Krohn (2021) dictionary as our main references. The surrounding context of the verb was also referenced to support this manual correction process. All of the nouns and copulas were tagged correctly as predicted by the rules, but some of the possessive pronouns needed manual correction and this was undertaken in the same fashion as the aforementioned manual correction of some verbs. The irregular plural adjectives *tsítsi* 'small.PL' and *tsîrala'ralar* 'tiny.PL' were tagged for number manually because fitting regular expressions were not developed for these forms.

We used this new, morphologically-tagged CoNLL-U file to train twenty parsing models using UDPipe 2 (Straka, 2018). We trained these using a cloud-based system with a V100 GPU. Each model took approximately 1.5 hours to train and test, for a total of 30 hours of processing. The hyperparameters can be found in Appendix A. Once these models were trained, we calculated the precision of the feature tagging for each of them and used this information to compare the system's performance with that of the FOMA-based tagging.

3 Results

At the end of the tagging process, a word would have its Universal Dependencies' morphological features in the corresponding CoNLL-U column. Table 2 shows an example of a sentence and its features.

Word	POS	Features
ye'	PRON	`Case=ERG|Number=Sing|Person=1|PronType=Prs`
tö	PART	`_`
ù	NOUN	`Case=ABS|Number=Sing`
sṳ́	VERB	`Aspect=Perf|Mood=Ind|Tense=Past|VerbForm=Fin|Voice=Act`

Table 2: Universal Dependency Features morphological features (UFEATS) for *Ye' tö ù sṳ́* 'I saw the house'

3.1 Tags after Correction

After the manual corrections, there was a total of 3051 morphological features in the annotated treebank. Table 3 shows the total of features for each part of speech in the annotated dataset. The majority of the tags were dedicated to the verbs (n=1504, 49%), in particular the tense-mood-aspect (TAM) markers. There are also numerous tags for the distinction between active and middle voice, which is crucial in the description of Bribri grammar.

Part-of-Speech	Morphological Feature	n
Verb	Aspect=Imp	138
	Aspect=Prosp	45
	Aspect=Perf	65
	Mood=Des	1
	Mood=Imp	3
	Mood=Ind	245
	Person=1	47
	Person=2	16
	Person=3	32
	Polarity=Neg	3
	Tense=Pres	152
	Tense=Past	97
	VerbForm=Inf	63
	VerbForm=Fin	267
	Voice=Mid	62
	Voice=Act	268
Noun	Case=ABS	69
	Case=ERG	5
	Number=Plur	10
	Number=Sing	246
Adjective	Number=Plur	3
Pronoun	Case=ABS	121
	Case=ERG	19
	Clusivity=Ex	8
	Clusivity=In	11
	Number=Sing	274
	Number=Plur	51
	Person=1	136
	Person=2	38
	Person=3	131
	Poss=Yes	39
	PronType=Dem	16
	PronType=Int	12
	PronType=Prs	307
	PronType=Rcp	4
	Reflex=Yes	7
Copula	Person=1	14
	Person=2	6
	Person=3	20

Table 3: Part-of-Speech and Tagged Universal Features

Error	n	
TAM missing	10	28%
`Case` missing	9	18%
Hallucinated features	7	14%
`Person` missing	6	12%
`Number` missing	5	10%
Others	12	25%

Table 4: Types of errors in the output for morphological features in an example UDPipe 2 model (total n=49)

Pronouns were the next category in importance (n=1174, 38%). Most of the tags were for person and number, followed by case tags for those pronouns that were either the syntactic ergative or absolutive in the sentence. Importantly, the 1st person plural pronouns were also marked for clusivity (i.e. exclusive or inclusive), and the non-personal pronouns were marked for their function (e.g. demonstrative, interrogatives, reciprocals and reflexives). Nouns had the third most features (n=330, 11%). Like in the case of the pronouns, they were marked for number, and for case if they occupied a core argument (ergative or absolutive) position in the sentence.

Copula features (n=40, 1%) only have tags for the person that the copula refers to. Finally, the three irregular plural adjectives in the corpus were tagged with the corresponding plural feature.

3.2 Parsing Model Tests

Once the dataset was tagged, we used it to train a series of UDPipe 2 models in order to test whether this relatively small dataset could be used to expand our morphological tagging capabilities. We used the 315 sentences in the annotated treebank to create twenty random train/dev/test partitions (80%, 10%, 10%) and train the models. The average precision for the Universal Features (UFEATS) was 80.5 ± 3.6.

After this we randomly selected one of the models and analyzed the errors it produced. The test set contained 304 features, and 49 of these were predicted inaccurately (16%). Table 4 shows a summary of the errors produced by the model in the output hypotheses for the test set.

The most frequent errors are missing features that the model couldn't predict. Out of all of the errors, 28% were those where the TAM features was missing. 18% of the errors were the result of a missing absolutive or ergative case marking in the output, and a further 22% were because either the person (first, second or third) or the number of a word were missing.

There are also a few errors in the system where the word is assigned the correct part of speech but a wrong feature. For example, a verb could be assigned the past tense when it was actually in the present, or the verb could be tagged as having a 1st person subject, even though the gold standard had it as a 3rd person.

Finally, it is worth noting that the model does have a tendency to hallucinate morphological features. There were some parts of speech, like numerical classifiers and determiners, whose morphological tagging was not included in this work (more on this in the discussion below). However, the system would produce features for them. In the case of a numerical classifier like *bö̀k* 'two [round things]' the model treated this as a `VERB` and gave it features for finiteness and active voice. In the case of the determiner *i'* 'this one', the system misclassified the part of speech as a pronoun and then gave it features for singular number and tagged it as a demonstrative, probably because of its phonetic (but not syntactic) similarity with the demonstrative pronoun *i'* 'this one'.

3.3 Comparison with FOMA Parsing

While a direct comparison between the FOMA tags and the UFEATS is not possible due to the difference in their tagging conventions, we can estimate their difference in providing a tentative tag for unseen Bribri data.

In order to calculate an error for the FOMA, we devised the following test. We took the test sets from each of the twenty random samplings of the treebank. We took the words in those test sets and tagged them individually using Flores-Solórzano's (2019) FOMA tagger. This can only be done word by word because the system is based on an FST, and cannot get information from preceding or subsequent words. Then, we classified FOMA's responses into three possibilities. First, if FOMA produces no output (+?), then we consider this an error. Second, if FOMA produces more than one output (e.g. saying that the word *dör* is both the ergative marker and a copula), we consider this an error. This is because the system has no probabilistic information in its output, and it would be impossible to determine which of the two tags is correct without an additional module that considered context. The third condition is if the FOMA provides only one answer (e.g. labeling the word *ye'* 'I' as +1PSg). We assume this is a correct answer because of the FST nature of the FOMA system: it identifies a word directly and then it has a pre-programmed set of morphological outputs for it. Importantly, we calculate this for all the tokens in the UDPipe2 predictions, including those that are tagged with an empty response _, which is a correct gold-standard answer for words that don't have tags yet (e.g. postpositions and numerical classifiers).

When we calculate the results according to these three conditions, we get that, for the twenty runs, the average precision of the FOMA system is 59.5 $\pm$ 4.2. This is lower than the 80.5 $\pm$ 3.6 result for the UDPipe 2 model; this comparison is shown in figure 1. In fact, a paired t-student test revealed that the deep learning system performs significantly better with the same test sets (t(19)=16.5, p<0.00001)[2].

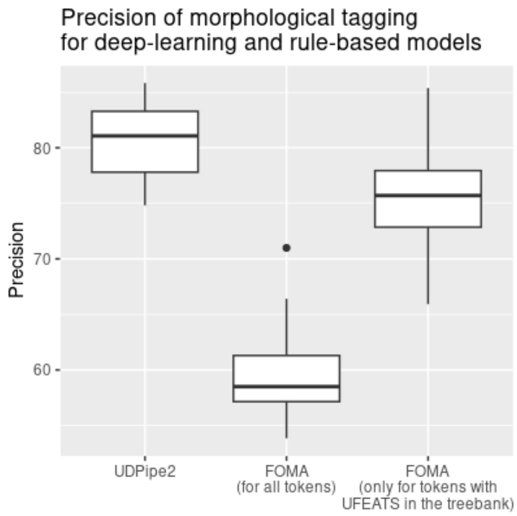


Figure 1: Precision for morphological tagging for a deep learning model (UDPipe 2), a rule-based FST model (FOMA) and a rule-based model that only looks at words with existing UFEAT tags in the gold standard.

We conducted a second, more rigorous test. In this test, we only considered words that actually have features in the gold standard treebank. As mentioned above, there are many words, for example postpositions, particles and numerical classifiers, which only have the marker _ in their feature column. In this second test we will only include tokens if the original treebank had actual UFEATS in it. After this modification, the precision of the FOMA system increases to 75.7 $\pm$ 4.7. Figure 1 shows the distribution of the twenty samples, under the condition *FOMA (only for tokens with UFEATS in the treebank)*. While this precision is higher than the FOMA for all the tokens, it is still significantly lower than the precision of the UDPipe 2 model (t(19)=3.4, p<0.005).

These results confirm that the deep learning model trained from our tagged treebank shows improvement in the state of the art for morphological tagging in the Bribri language.

4 Discussion

Overall, the rule-based tagging of the verbs was difficult due to their morphological complexity, and numerous manual corrections were needed. We had specific regular expressions for over 80 verbs, and so the rules described in section 2.2 would not be easily transferable to larger segments of written Bribri. However, our objective in using these rules was to create a new system which could accept forms it hasn't seen before as its input. The morphological feature tags we have introduced to the treebank produce acceptable results during inference. Our future work is to take this new treebank and use it to make morphological and syntactic parsings of unseen sentences of Bribri in order to expand existing corpora.

The most immediate item of future work is to expand the tags for the remaining parts of speech. For example, Bribri's deictic system includes pronouns that refer to distance from the speaker (near, far) and vertical position from the reference point (above, even and below). For example, the word *aí* means 'that one (above, near)', and the word *dià* means 'that one (below, far)'. It also includes deictics which need the feature Deixis=NVis (not visible), like the word *ñe̱'* 'that one (that can be heard but not seen)'. These are tags that already exist in the Universal Features, and should be easy to expand upon.

There are also places where the parts of speech treated here could be expanded. For example, Bribri has several diminutive morphemes for nouns and adjectives (e.g. *a̱mì̱* 'mother' versus *a̱mí̱la* 'mommy'). These would take the feature Degree=Dim, but this was not included in the present work. These morphemes are important

[2]A Shapiro-test confirms that all distributions meet the normality assumption: $p_{UDPipe2}$=0.25, $p_{FomaAll}$=0.06, $p_{FomaUfeats}$=0.99.

for the studying of Bribri discourse, and so their tagging is necessary in the future.

More complex to tag are numerical classifiers. These classifiers contain the number, but also semantic information about the geometry of the object. Some examples are: (i) *buà bö̀töm* 'two[long] iguanas', (ii) *apë' bö̂l* 'two people', (iii) *àshali bö̀k* 'two[round] oranges', and (iv) *kua'kua bö̀t* 'two[flat] butterflies'. There are at least 8 of these classes, and their tagging cannot be described with the features in Universal Features. That additional information would have to be included separately.

Finally, there is additional information about the verbs that also needs to be saved separately. For example, Bribri verbs distinguish between "recent" and "remote" past perfect tenses. For example, the sentence *ye' shka'* means 'I walked (sometime yesterday, before I went to bed last night, or further back in the past)'. On the other hand, the sentence *ye' shké* is also perfect, but it covers both the immediate present perfect (e.g. 'I will walk'), and a perfect aspect action that has occurred in the recent past, after the last time one went to bed (e.g. 'I walked (sometime today, in the recent past)'. This recent tense has also been called the *hodiernal* tense in literature (Dahl, 1983). This distinction cannot be described in the Universal Features, and would have to be stored separately as well.

One piece of future work is to make a system that performs automatic morphological segmentation. Such a system would get the input *Shkàne̱* 'There was walking', and would be able to produce the output `shk-àn-e̱`, with the root `shk` 'walk', the middle voice suffix `-àn` and the remote past tense perfect aspect suffix `-e̱`. We hope that the feature tagging described in this paper will be helpful in making such a segmentation system, which would further contribute to the creation of annotated corpora.

5 Conclusions

In this paper we have presented a new morphological tagger for the Bribri language. We automatically tagged an existing treebank with Universal Dependencies' Universal Features. We hand-corrected any errors during the tagging process, and then used this new treebank to train a parsing model. This model has significantly better performance than the previous FST-based analyzer. We will continue to expand upon this work, using these tools to aid in the annotation of corpora for the language.

Limitations

The system is limited to written Bribri, which might hinder its usability for other applications, as most speakers of Bribri do not write the language. and much of the data we ultimately want to tag is oral narratives. Moreover, the writing system represented in the dataset is only one of the orthographies currently in use for the language, and so an input system that can easily accept all orthographies would need to be deployed alongside this tagger in the future.

Ethics Statement

The work was done using openly available materials published by Costa Rican institutions (e.g. University of Costa Rica). The models will be used to work on corpora construction, in collaboration with Bribri community members who work on the linguistics of the language.

References

Khuyagbaatar Batsuren, Omer Goldman, Salam Khalifa, Nizar Habash, Witold Kieraś, Gábor Bella, Brian Leonard, Garrett Nicolai, Kyle Gorman, Yustinus Ghanggo Ate, et al. 2022. UniMorph 4.0: universal morphology. *arXiv preprint arXiv:2205.03608*.

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

Rolando Coto-Solano. 2021. Explicit tone transcription improves ASR performance in extremely low-resource languages: A case study in Bribri. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 173–184.

Rolando Coto-Solano. 2022. Evaluating word embeddings in extremely under-resourced languages: A case study in Bribri. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4455–4467.

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

Rolando Coto-Solano, Sally Akevai Nicholas, Brittany Hoback, and Gregorio Tiburcio Cano. 2022. Managing data workflows for untrained forced alignment: examples from Costa Rica, Mexico, the Cook Islands, and Vanuatu. *The Open Handbook of Linguistic Data Management*, 35.

Rolando Coto-Solano and Sofía Flores Solórzano. 2016. Alineación forzada sin entrenamiento para la anotación automática de corpus orales de las lenguas indígenas de Costa Rica. *Kánina*, 40(4):175–199.

Östen Dahl. 1983. Temporal distance: Remoteness distinctions in tense-aspect systems. *Linguistics*, 21(1).

Fineen Davis, Eddie Antonio Santos, and Heather Souter. 2021. On the computational modelling of Michif verbal morphology. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2631–2636.

Haley De Korne and Miranda Weinberg. 2021. "I Learned That My Name Is Spelled Wrong": Lessons from Mexico and Nepal on Teaching Literacy for Indigenous Language Reclamation. *Comparative Education Review*, 65(2):288–309.

Marie-Catherine de Marneffe, Christopher D Manning, Joachim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, pages 255–308.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir, et al. 2021. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. *arXiv preprint arXiv:2104.08726*.

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.

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

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

Sofía Flores-Solórzano. 2019. La modelización de la morfología verbal bribri - Modeling the Verbal Morphology of Bribri. *Revista de Procesamiento del Lenguaje Natural*, 62:85–92.

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

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

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

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

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

Alex Jones, Rolando Coto-Solano, and Guillermo González Campos. 2023. TalaMT: Multilingual Machine Translation for Cabécar-Bribri-Spanish. In *Proceedings of the 3rd Workshop on Multi-lingual Representation Learning (MRL)*, pages 106–117.

Katharina Kann, Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, John E Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo A Giménez-Lugo, et al. 2022. AmericasNLI: Machine translation and natural language inference systems for Indigenous languages of the Americas. *Frontiers in Artificial Intelligence*, 5:995667.

Anna Kazantseva, Karin Michelson, Jean-Pierre Koenig, et al. 2024. Fitting a Square Peg into a Round Hole: Creating a UniMorph dataset of Kanien'kéha Verbs. In *Proceedings of the Seventh Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 39–51.

Sujay Khandagale, Yoann Léveillé, Samuel Miller, Derek Pham, Ramy Eskander, Cass Lowry, Richard Compton, Judith L Klavans, Maria Polinsky, and Smaranda Muresan. 2022. Towards unsupervised morphological analysis of polysynthetic languages. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 334–340.

Haakon Krohn. 2020. Elaboración de una base de datos en XML para un diccionario bribri–español español–bribri en la web. *Porto das Letras*, 6(3):38–58.

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

Anastasia Kuznetsova and Francis Tyers. 2021. A finite-state morphological analyser for Paraguayan Guaraní. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 81–89.

Jordan Lachler, Lene Antonsen, Trond Trosterud, Sjur Moshagen, and Antti Arppe. 2018. Modeling Northern Haida Verb Morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Ngoc Tan Le and Fatiha Sadat. 2021. Towards a low-resource neural machine translation for indigenous languages in Canada. *Traitement Automatique des Langues*, 62(3):39–63.

Zoey Liu, Robert Jimerson, and Emily Prud'Hommeaux. 2021. Morphological segmentation for Seneca. In *First Workshop on Natural Language Processing for Indigenous Languages of the Americas*.

Ariadna Font Llitjós, Roberto Aranovich, and Lori Levin. 2005. Building Machine Translation Systems for Indigenous Languages. In *Second Conference on the Indigenous Languages of Latin America (CILLA II), Texas, USA*.

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

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza. 2018b. Challenges of language technologies for the indigenous languages of the Americas. *arXiv preprint arXiv:1806.04291*.

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

Arya D McCarthy, Christo Kirov, Matteo Grella, Amrit Nidhi, Patrick Xia, Kyle Gorman, Ekaterina Vylomova, Sabrina J Mielke, Garrett Nicolai, Miikka Silfverberg, et al. 2020. UniMorph 3.0: Universal Morphology. In *Proceedings of The 12th language resources and evaluation conference*, pages 3922–3931. European Language Resources Association.

MEP. 2013. *Los Bribris y Cabécares de Sulá - Tomo 1 - Minienciclopedia de los Territorios Indígenas de Costa Rica*. Ministerio de Educación Pública de Costa Rica.

Rodolfo Mercado-Gonzales, José Pereira-Noriega, Marco Antonio Sobrevilla Cabezudo, and Arturo Oncevay. 2018. ChAnot: An intelligent annotation tool for indigenous and highly agglutinative languages in Peru. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Benjamin Molineaux. 2023. The Corpus of Historical Mapudungun: Morpho-phonological parsing and the history of a Native American language. *Corpora*, 18(2):175–191.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 2016. Universal Dependencies v1: A Multilingual Treebank Collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Jan Hajič, Christopher D Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An evergrowing multilingual treebank collection. *arXiv preprint arXiv:2004.10643*.

Hyunji Park, Lane Schwartz, and Francis Tyers. 2021. Expanding universal dependencies for polysynthetic languages: A case of St. Lawrence island Yupik. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*.

Robert Pugh and Francis Tyers. 2021. 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.

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

Lorena Martín Rodríguez, Tatiana Merzhevich, Wellington Silva, Tiago Tresoldi, Carolina Aragon, and Fabrício F Gerardi. 2022. Tupían language ressources: Data, tools, analyses. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 48–58.

Jack Rueter, Mika Hämäläinen, and Khalid Alnajjar. 2023. Modelling the Reduplicating Lushootseed Morphology with an FST and LSTM. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 40–46.

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

Conor Snoek, Dorothy Thunder, Kaidi Lõo, Antti Arppe, Jordan Lachler, Sjur Moshagen, and Trond Trosterud. 2014. Modeling the noun morphology of Plains Cree. In *Proceedings of the 2014 Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 34–42.

Sofía Flores Solórzano. 2010. Teclado Chibcha: Un software lingüístico para los sistemas de escritura de las lenguas bribri y cabécar. *Revista de Filología y Lingüística de la Universidad de Costa Rica*, pages 155–161.

Sofía Flores Solórzano and Rolando Coto-Solano. 2017. Comparison of Two Forced Alignments Systems for Aligning Bribri Speech. *CLEI Electronic Journal*, 20(1):2–1.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies*, pages 197–207.

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

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

Francis Tyers and Robert Henderson. 2021. A corpus of K'iche'annotated for morphosyntactic structure. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 10–20.

Francis Tyers and Nick Howell. 2021. A survey of part-of-speech tagging approaches applied to K'iche'. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 44–52.

UCREL. 2011. UCREL CLAWS7 Tagset.

A UDPipe2 Hyperparameters

```
batch_size: 32
beta_2: 0.99
char_dropout: 0
cle_dim: 256
clip_gradient: 2.0
dropout: 0.5
epochs: [(3, 0.001), (3, 0.0001)]
exp: None
label_smoothing: 0.03
max_sentence_len: 120
min_epoch_batches: 300
parse: 1
parser_deprel_dim: 128
parser_layers: 1
predict: False
predict_input: None
predict_output: None
rnn_cell: LSTM
rnn_cell_dim: 512
rnn_layers: 2
rnn_layers_parser: 1
rnn_layers_tagger: 0
seed: 42
single_root: 1
tag_layers: 1
tags:['UPOS','XPOS','FEATS','LEMMAS']
threads: 4
variant_dim: 128
we_dim: 512
wembedding_model: bert-base-multilingual
    -uncased-last4
word_dropout: 0.2
```

LLM-Assisted Rule Based Machine Translation for Low/No-Resource Languages

Jared Coleman
Bhaskar Krishnamachari
Khalil Iskarous
University of Southern California
{jaredcol,bkrishna,kiskarou}@usc.edu

Ruben Rosales
rubenrosales@ieee.org

Abstract

We propose a new paradigm for machine translation that is particularly useful for no-resource languages (those without any publicly available bilingual or monolingual corpora): LLM-RBMT (LLM-Assisted Rule Based Machine Translation). Using the LLM-RBMT paradigm, we design the first language education/revitalization-oriented machine translator for Owens Valley Paiute (OVP), a critically endangered Indigenous American language for which there is virtually no publicly available data. We present a detailed evaluation of the translator's components: a rule-based sentence builder, an OVP to English translator, and an English to OVP translator. We also discuss the potential of the paradigm, its limitations, and the many avenues for future research that it opens up.

1 Introduction

Large Language Models like OpenAI's GPT series (OpenAI, 2023) have shown remarkable capability at an impressively wide range of natural language tasks (Bubeck et al., 2023) including machine translation (Hendy et al., 2023). These models work because they are trained on vast amounts of natural language data, primarily from the internet (OpenAI, 2023). For languages that are low-resource (languages for which there is little publicly available data) or no-resource (languages for which there is *no* publicly available data), models like these do not perform well on their own (Chowdhery et al., 2022; Robinson et al., 2023). There have been many efforts in improving machine translation for low-resource languages (see (Ranathunga et al., 2023) for a comprehensive survey), but no-resource languages have received much less attention in the literature.

Over past decades, researchers and community leaders have led many remarkable efforts in endangered language revitalization (Coronel-Molina and McCarty, 2016; Taylor and Kochem, 2022) and reclamation (Baird, 2016). In this paper, we begin to explore how the impressive general-purpose language skills of LLMs might be helpful for these kinds of efforts by introducing a new paradigm for low/no-resource machine translation: LLM-RBMT (LLM-Assisted Rule-Based Machine Translation). The intuition behind our approach is that, given the correct tools, humans are good at translating simple sentences even for languages they do not know. For example, a common task for students in a language class is to translate sentences given a set of words and rules. Given a conjugation table and a set of verbs a motivated student could probably translate a list of simple subject-verb (e.g., "I eat", "you sing", etc.) sentences with extremely high accuracy. Of course, they are limited to only translating specific kinds of sentences with a limited vocabulary. Still, the idea is interesting: if we provide enough context to an LLM like ChatGPT, which has been shown to exhibit human-level performance on many natural language tasks, we may not need it to actually *know* (exhibit general fluency in) the target language we want to translate to/from. While rule-based translation can likely never achieve the quality of modern ML translators for high-resource languages, this technique has a few potential advantages for low/no-resource languages. First and most importantly, it requires no parallel corpus. Also, it is amenable to *partial translation*. Humans (and, as we'll show in this paper, LLMs) are capable of giving partial translations when they don't have all of the necessary vocabulary. For example, a Spanish student that doesn't know the word for "dog" might still be able to translate the rest of the sentence "the dog ate the apple yesterday" by saying "el [dog] comió la manzana ayer".

In this paper, we leverage LLMs (specifically, gpt-3.5-turbo and gpt-4 from OpenAI's GPT series) to break down natural language sentences into

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 67–87
June 21, 2024

structured simple sentences compatible with hand-crafted rule-based translators. We also use them to turn sentence structured information (in JSON format) into natural language sentences. Using this approach, the LLM never interacts directly with the target language. Rather, we rely on the LLM to tell us how to use the simple, rule-based translators to provide a translation as close as possible to the user's original input. We use this technique to build and evaluate two machine translation tools for Owens Valley Paiute (OVP, also called Eastern Mono or Monache in linguistics literature, `ISO 639-3 mnr` (SIL International, 2024)), a critically endangered Indigenous American language in the Uto-Aztecan language family (Moseley, 2010). The first is a selection-based OVP to English simple sentence translator and the second is an English to OVP translator that uses available vocabulary to construct translations of arbitrary user input sentences. The translators are oriented toward language teaching and revitalization. They are not designed to be general-purpose translators, but rather as tools to help no-resource language learners express ideas using simple sentence constructions.

1.1 Contributions

The main contributions of this work are[1]:

1. An extensible OVP sentence builder for constructing valid subject-verb and subject-verb-object sentences in Owens Valley Paiute.

2. An LLM-assisted OVP to English translator that translates OVP sentence builder sentences to English with high accuracy.

3. An LLM-assisted English to OVP translator that translates English sentences to Owens Valley Paiute using the sentence-builder and OVP to English translation tools.

4. A novel methodology for the design and evaluation of no-resource language translators.

The rest of this paper is organized as follows. We discuss related work in Section 2. We present the sentence building tool and OVP to English translation system in Section 3. Then, in Section 4, we present the English to OVP translation system and report results on translation quality for different types of input sentences using embeddings models to measure semantic similarity. We conclude the paper with a summary of contributions and discussion of future research directions in Section 5.

[1]The code behind all contributions is open-source: `https://github.com/kubishi/kubishi_sentences`

2 Related Work

The landscape of low-resource machine translation is vast and constantly growing. A comprehensive survey on this subject is provided by (Ranathunga et al., 2023), which outlines the current techniques, guidance for selecting which techniques to use for a given language, and future directions for research. Of particular interest within this survey is the examination of unsupervised neural machine translation. While there exists promising research on constructing translators from minimal corpora, these methods invariably require *some* natural language data and thus have limited applicability to no-resource languages (such as OVP). The survey also discusses data augmentation strategies, including word or phrase replacement-based augmentation and Back-Translation-based Data Augmentation, both of which could potentially be integrated with some of the solutions presented in this paper (in particular the sentence builder and OVP to English translator to be presented in Section 3). Such an exploration is an interesting topic for future work. Other approaches discussed in the survey such as supervised learning, transfer learning, and semi-supervised learning are inapplicable to our scenario due to the absence of bilingual or monolingual corpora.

Contrary to the prevailing assumption in the literature that rule-based machine translation (RBMT) is a relic of the past, there remains active research and development in RBMT systems tailored for the most under resourced of languages (Khanna et al., 2021; Pirinen, 2019; Torregrosa et al., 2019). Recent work has also explored the utilization of Large Language Models (LLMs) for enhancing translation capabilities in low-resource languages through fine-tuning techniques (Lankford et al., 2023). Although this approach has shown promise in improving LLM translation quality for low-resource languages like (e.g., Irish), its reliance on bilingual corpora make it infeasible for no-resource languages like OVP. Recently, semantic similarity has been used to evaluate the quality of Machine Translation systems (Cao et al., 2022; Song et al., 2021). We observe that this technique is particularly useful for evaluating the quality of the English to OVP translator presented in this paper, due to the lack of parallel corpora. See (Muennighoff et al., 2023) for a comprehensive benchmark of different embeddings models used for computing semantic similarities.

3 OVP to English Translation

In this section, we present an LLM-assisted selection-based OVP to English translator. The first piece of this translation system is a selection-based OVP sentence builder. The sentence builder allows users to select each of the following parts of speech from a list of choices until they form a valid sentence:

- Subject: The subject of the sentence.
- Subject Suffix: In OVP, noun subjects are always suffixed with either **-ii** (if the subject is proximal to the speaker) or **-uu** (if the subject is distant to the speaker).
- Verb: the verb of the sentence.
- Verb Suffix: In OVP, verb suffixes express the tense/aspect of the verb.
- Object: The object of the sentence (disallowed if the selected verb is intransitive, optional otherwise).
- Object Suffix: In OVP, noun objects are always suffixed with either **-(n)eika** (if the subject is proximal to the speaker) or **-(n)oka** (if the subject is distant to the speaker).
- Verb Object Pronoun Prefix: In OVP, object pronouns are prefixed to the verbs they are the object of. Object pronouns are required for all transitive verbs. Even when a noun object is specified, the pronoun is still required and should match the object suffix (**-(n)eika** matches **a-** or **ma-** and **-(n)oka** matches **u-**).

Not all of these parts of speech are required to create a valid sentence in OVP, though. In fact, some of them are incompatible. For example, if an intransitive verb is chosen, then it cannot have an object or object pronoun. In other words, the valid choices a user can make is a function of the choices they have already made. In our python implementation for this translator, we process each user selection and change the list of valid options for each part-of-speech based on their current selections to ensure they always create a valid OVP sentence. This is the *rule* part of LLM-Assisted *Rule*-Based Machine Translation that requires expert knowledge in the target language to implement. The entire vocabulary available for the users to select among fits on a single page (and can be found in Appendix C).

After the user creates a valid OVP sentence, we translate it by first encoding the following sentence information into an English-only (using vocabulary definitions) *structured* simple sentence:

- Subject: noun or pronoun subject in English
- Subject Proximity: **proximal** if user selected subject suffix **-ii** or **distant** if the user selected subject suffix **-uu**
- Object: noun or pronoun object in English
- Object Proximity: **proximal** if user selected object suffix **-(n)eika** or **distant** if the user selected object suffix **-(n)oka**
- Verb: verb in English
- Verb Tense/Aspect: one of **past**, **present**, **future**, **past-continuous**, **present-continuous**, **present-perfect**

Then, we use few-shot prompting to encourage an LLM to transform the structured English data into a natural language sentence. Consider the example in Figure 1 where few-shot training examples (colored black) tell the LLM how to respond to the actual structured data for the randomly generated sentence for “Wo’ada-ii pagwi-noka u-zawa-dü.” (colored blue). Observe, the LLM is prompted to translate using only the English, structured version of the selected sentence.

To evaluate the accuracy of the translator, we generated 100 random valid OVP sentences by iteratively selecting a random choice among available choices for each of the parts of speech until the sentence is valid. Of the 100 random sentences generated, 98 were translated into English accurately using gpt-3.5-turbo model from OpenAI’s GPT-series. Translations and their accuracy labels can be found in Appendix A. While impressively accurate, this translator has many disadvantages. It only works for simple subject-verb and subject-verb-object sentences that use the nouns and verbs available in the system. Also, since pronouns and suffixes in OVP encode temporal/spatial information, translations don’t always capture full meaning of the sentence. The English translations are correct, but may be missing useful information. For example "kamü-uu wo’abi-neika a-düka-ti" translates to "the jackrabbit is eating the worm", which is technically correct, but -uu also indicates that the jackrabbit is not present and the -neika indicates the worm *is* present. Then, since -ti can is used for both the present *and* past continuous tenses (*is*

Wo'ada-ii pagwi-noka u-zawa-dü.

system

You are an assistant for translating structured sentences into simple natural English sentences.

user

[{'part_of_speech': 'subject', 'positional': 'proximal', 'word': 'wood'}, {'part_of_speech': 'object', 'positional': 'proximal', 'word': 'dog'}, {'part_of_speech': 'verb', 'tense': 'present continuous (-ing)', 'word': 'see'}]

assistant

This wood is seeing this dog.

user

[{'part_of_speech': 'subject', 'positional': 'distal', 'word': 'pinenuts'}, {'part_of_speech': 'object', 'positional': 'distal', 'word': 'horse'}, {'part_of_speech': 'verb', 'tense': 'future (will)', 'word': 'see'}]

assistant

Those pinenuts will see that horse.

user

[{'part_of_speech': 'subject', 'word': 'mosquito', 'positional': 'proximal'}, {'part_of_speech': 'object', 'word': 'fish', 'positional': 'distal'}, {'part_of_speech': 'verb', 'word': 'cook', 'tense': 'present'}]

assistant

This mosquito is cooking that fish.

Figure 1: Few-shot examples for translating "Wo'ada-ii pagwi-noka u-zawa-dü." using gpt-3.5-turbo.

[x]-ing or *was [x]-ing*), a better translation would be "the jackrabbit was eating this worm". More advanced rules and better prompt-engineering may help mitigate this issue.

Despite some expected disadvantages, this translator has many advantages. First, it is the first machine translator for OVP. It is also easy to extend the tool with new nouns and verbs. Also, while implementing the rules required expert knowledge of what makes an OVP sentence valid, no expert knowledge of how the rules map to English was required (or needed to be programmed), thanks to the LLM. Finally, we believe this kind of translation system might be a useful educational tool that helps students learn how to build simple sentences. It also may be useful as a data augmentation technique for training neural machine translation models for low-resource languages.

4 English to OVP Translation

In this section, we present the first English to Owens Valley Paiute machine translator. Unlike the OVP to English translator presented in the previous section, the user can input any sentence in natural language. The translator works by first using an LLM to break the input sentence into a set of simple structured subject-verb and subject-verb-object sentences, discarding any adjectives, adverbs, prepositions, objects (except for direct objects), etc. We encourage (through few-shot prompt engineering) the LLM to preserve as much semantic meaning as possible between the original sentence and the set of simple sentences. Consider the example in Figure 2 where few-shot examples (colored black) tell the LLM how to respond to a given input sentence "We are playing and laughing." (colored blue)[2]. Then, we use these structured sentences and available vocabulary to build valid OVP sentences with the sentence-building tool described in Section 3. Once the sentence is built, we use the translator described in Section 3 to translate the OVP sentences back into English. The idea is that, while some meaning may have been lost between the original input sentence and the final output English sentences, the user can be fairly confident (given the accuracy of the OVP to English translator) that the final translations are correct. The entire English to OVP translation process is depicted in Figure 3.

[2]We also leverage OpenAI's function calling capability to ensure that responses are consistently formatted. We refer interested readers to the open-source implementation's documentation at https://github.com/kubishi/kubishi_sentences for full details.

We are playing and laughing.

system

You are an assistant that splits user input sentences into a set of simple SVO or SV sentences. The set of simple sentences should be as semantically equivalent as possible to the user input sentence. No adjectives, adverbs, prepositions, or conjunctions should be added to the simple sentences. Indirect objects and objects of prepositions should not be included in the simple sentences.

user

I am sitting in a chair.

assistant

[{'subject': 'I', 'verb': 'sit', 'verb_tense': 'present_continuous', 'object': null}]

user

I saw two men walking their dogs yesterday at Starbucks while drinking a cup of coffee

assistant

[{'subject': 'I', 'verb': 'see', 'verb_tense': 'past', 'object': 'man'}, {'subject': 'man', 'verb': 'walk', 'verb_tense': 'past_continuous', 'object': 'dog'}, {'subject': 'man', 'verb': 'drink', 'verb_tense': 'past_continuous', 'object': 'coffee'}]

user

We are playing and laughing.

assistant

[{'subject': 'we', 'verb': 'play', 'verb_tense': 'present_continuous', 'object': null}, {'subject': 'we', 'verb': 'laugh', 'verb_tense': 'present_continuous', 'object': null}]

Figure 2: Few-shot training examples for the English to OVP using gpt-3.5-turbo.

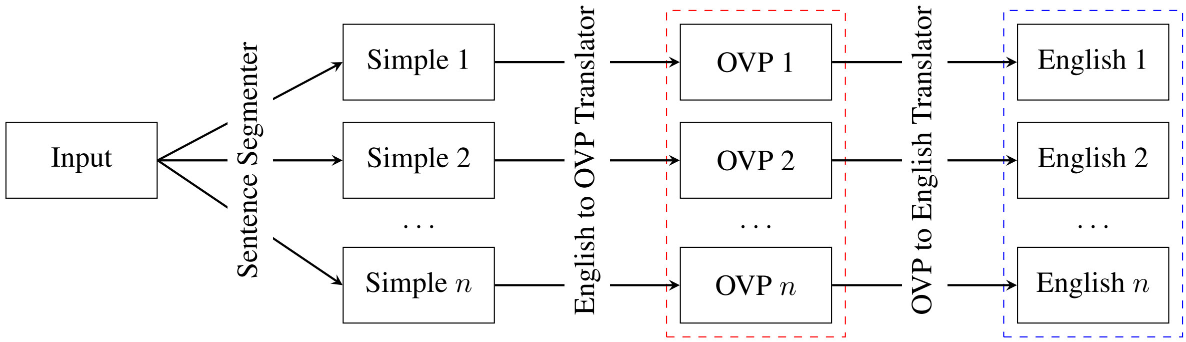


Figure 3: The entire English to OVP translation process. The box with a red, dashed border indicates the set of sentences in Owens Valley Paiute (the target language) and the box with a blue, dashed border indicates the set of English sentences they translate to. Ideally, the input sentence, simple sentences, and English output sentences will have equivalent or very similar semantic meaning.

We evaluate the system by translating a set of 125 sentences. There are five types of sentences in the dataset (25 of each):

- subject-verb (e.g., "I read" or "she sings")
- subject-verb-object (e.g., "Mom made dinner" or "John read a book")
- two-verb (e.g., "She sings and dances." or "I ate while watching TV.")
- two-clause (e.g., "My brother drove and I waited." or "Harry wrote and Ron read.")
- complex (e.g., "Rachel and Monica share an apartment." or "Romeo and Juliet loved deeply.")

We translated all 125 sentences using both the gpt-3.5-turbo and gpt-4 models from OpenAI's GPT-series, resulting in a total of 250 translations.

To measure the quality of each translation, we compute the semantic similarity between the input sentence and:

- the set of simple sentences generated by the LLM-powered segmenter (denoted *simple*).

- the set of simple sentences with unknown vocabulary removed (denoted *comparator*). Intuitively, this represents the best the translator can achieve given the vocabulary it has access to. For example, suppose the verb "wash" is not in our available vocabulary. Then the comparator sentence for the simple sentence "The woman is washing" would be "The woman is [VERB]-ing"
- the "round-trip" English translation (denoted *backwards*). This is the sentence produced using by translating the translated sentence (in OVP) to English using the method described in Section 3.

The usefulness of these measurements depends greatly on the function used to compute semantic similarity. We compute semantic similarity by generating embeddings for sentences (using some embeddings model) and computing the normalized cosine similarity between these embeddings. For our application, we want semantically similar sentences to have small normalized cosine similarity, independent of other linguistic features like syntax. For example, an ideal semantic similarity function would rank "an apple is eaten by a man" more similar to "a man eats apples" than the sentence "a woman drinks coffee", despite the latter sentence being essentially grammatically equivalent to the target sentence.

We evaluated seven different embeddings models for this purpose and measured the semantic similarity between twelve target sentences and a ranked list of 10 sentences for each ranging from most to least semantically similar (sentences can be found in Appendix B). For each target sentence, we compare the ground-truth ranking of the 10 sentences to the ranking determined by the semantic similarity scores yielded by a particular embeddings model. We measure the similarity between the two rankings using two metrics: average displacement (average distance between a sentence's position in the computed ranking and its position in the target ranking) and RBO (Rank-biased Overlap (Webber et al., 2010)). Table 1 tabulates the results of this evaluation. Results indicate that the all-MiniLM-L6-v2 embeddings model perform well with respect to both Average Displacement and RBO. For this reason, we run the rest of our experiments using this embeddings model for computing the semantic similarity between sentences.

We computed the semantic similarity between all pairs of sentences in the dataset to establish a baseline for comparison. The mean semantic similarity between a pair of unrelated sentences was $\mu \approx 0.574$ with a standard deviation of $\sigma \approx 0.061$. Furthermore, the distribution appears to be relatively Gaussian (a histogram can be found in Appendix E). Intuitively, this suggests that semantically unrelated sentences are very unlikely to have a semantic similarity score of above $\mu + 3\sigma$ (i.e., greater than 0.757).

A good translation, then, should score high on all three semantic similarity metrics. For example, the translation

I am swimming.

Simple	**Semantic Similarity**
I am swimming.	1
Comparator	
I am swimming.	1
Target	
Nüü pahabi-ti.	
Backwards	
I am swimming.	1

(model: gpt-3.5-turbo)

is perfect. There are other interesting cases too, though. For example, when the comparator score is low but simple and backwards scores are high, the translator appears to do well with the vocabulary available, but is only able to give a partial translation. For example, the translation

Birds will migrate and return.

Simple	**Semantic Similarity**
A bird will migrate.	
The bird will return.	0.955
Comparator	
A bird will [VERB].	
The bird will [VERB].	0.778
Target	
[migrate]-wei tsiipa-uu.	
[return]-wei tiipa-uu.	
Backwards	
That bird will migrate.	
That bird will return.	0.944

(model: gpt-4)

has a high simple score, since the "Birds will migrate and return" is quite similar (semantically) to "A bird will migrate. The bird will return.". The comparator score is much lower, however, since

Embeddings Model		Average Displacement mean	std	RBO mean	std
text-embedding-ada-002	(OpenAI, 2024a)	0.967	0.442	0.885	0.053
all-MiniLM-L6-v2	(Reimers and Gurevych, 2020)	0.933	0.323	0.884	0.050
text-embedding-3-small	(OpenAI, 2024b)	1.000	0.362	0.882	0.051
text-embedding-3-large	(OpenAI, 2024b)	0.917	0.463	0.882	0.054
paraphrase-MiniLM-L6-v2	(Reimers and Gurevych, 2020)	1.150	0.410	0.870	0.054
bert-base-uncased	(Reimers and Gurevych, 2019)	1.600	0.703	0.777	0.100
spacy/en_core_web_md	(Explosion, 2024)	1.833	0.466	0.760	0.090

Table 1: Quality of different embeddings models in measuring semantic similarity between sentences. A lower average displacement and higher RBO indicate a better embeddings model for this purpose.

the words for "migrate" and "return" are not available in the vocabulary. The backwards score is good because the sentence "[migrate]-wei tsiipa-uu. [return]-wei tiipa-uu.", when translated using the OVP to English translator described in Section 3, becomes "That bird will migrate. That bird will return", which is almost equivalent to the simple sentence. This example also highlights an advantage to our approach. No-resource language learners can use this system to understand how sentences should be structured even if it doesn't have all of the vocabulary necessary to generate the entire translation. Furthermore, users can be fairly confident that the OVP target sentence is accurately translated into the English backwards sentence, due to the accuracy of the OVP to English translator.

The simple score being low indicates that a significant portion of the input sentence's meaning is lost during segmentation (due, for example, to the input sentence containing adverbs, adjectives, prepositions, etc. that don't fit anywhere in a simple subject-verb or subject-verb-object sentence). For example, the translation

My brother and I went hiking.

	Semantic Similarity
Simple	
Brother went. I went.	0.794
Comparator	
[SUBJECT] went. I went.	0.608
Target	
mia-ku [brother]-uu. nüü mia-ku.	
Backwards	
The brother went. I went.	0.806

(model: gpt-3.5-turbo)

lost meaning in the first step segmenting the input sentence into simple sentences because it chose to use the verb "to go" instead of "hike" which is the main topic of the sentence. Perhaps a better way to have segmented this sentence would be: "Brother hiked. I hiked". It may be possible to encourage the LLM to prefer "topic" verbs through prompt engineering.

Another interesting case is when the simple and comparator scores are high and only the backwards score is lower. This is observed in cases where there is ambiguity in OVP where there is not in English. For example, in the translation

She is cooking.

	Semantic Similarity
Simple	
She is cooking.	1
Comparator	
She is cooking.	1
Target	
Uhu sawa-ti.	
Backwards	
He is cooking.	0.836

(model: gpt-4)

"she" turns to "he" in the backwards translation because OVP does not have gendered pronouns. Despite the lower backwards score, this translation is accurate.

In general, both gpt-3.5-turbo and gpt-4 models do well with respect to the simple and backwards semantic similarity scores. Table 2 summarizes the mean semantic similarity scores for each model and type of sentence. Figure 4 depicts results for subject-verb sentences. Plots for the rest of the results can be found in Appendix D. That the simple and backwards scores are consistently higher than the comparator scores suggests that the translator can be greatly improved simply by expanding its

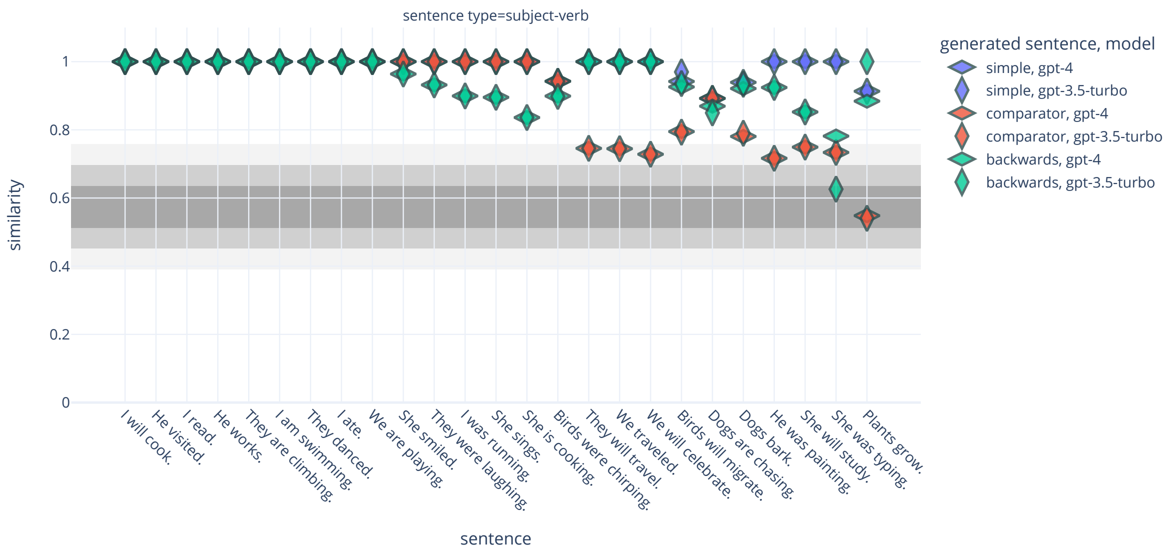


Figure 4: Results for subject-verb sentences. The dark, medium, and light gray bands represent the baseline similarity (between unrelated sentences in the dataset) +/- one, two, and three standard deviations, respectively.

Model	Type	Mean Sim.
gpt-3.5-turbo	subject-verb	0.941
	two-verb	0.906
	subject-verb-object	0.869
	two-clause	0.879
	complex	0.829
gpt-4	subject-verb	0.941
	subject-verb-object	0.866
	two-verb	0.905
	two-clause	0.877
	complex	0.830

Table 2: Translation qualities by model and sentence type: mean semantic similarities between input sentence and the simple, comparator, and backwards sentences produced during translation.

vocabulary. It is also interesting to note that the cheaper, weaker model performs quite well[3].

5 Conclusion

In this paper, we present the first translation tool for Owens Valley Paiute (a critically endangered Indigenous American language) and, in doing so, propose a new methodology for low/no-resource machine translation: LLM-RBMT (LLM-Assisted Rule-Based Machine Translation). Due to a lack of bilingual or monolingual corpora, we use semantic similarity to measure the quality of translations, and observe interesting cases where the translator performs well, where it performs well given the vocabulary (primarily verbs and nouns) available to it, and where it performs poorly due to current implementation limitations.

This research opens up many directions for future work. First, expanding the translation tool (e.g., by adding more vocabulary, building versions for other languages, etc.) will make it more useful for the communities it is meant to serve. We are currently working to add more complex sentence structures to test the limits of this methodology. Building this system required expert knowledge of the target language. It may be interesting to explore how large language models might be able to help with this task as well (e.g., in finding patterns and implementing RBMT logic) to make it easier for non-experts (and even experts in the language who are less comfortable writing code) to build these kinds of translators. Finally, other methodologies for leveraging LLMs for low/no-resource machine translation should be explored. For example, a RAG (retrieval augmented generation) (Lewis et al., 2020) approach that searches a knowledge-base for known sentence structures, vocabulary, grammar rules, etc. and uses them to perform zero-shot translation might work well. The remarkable general-purpose language skills that LLMs exhibit make them a promising tool in helping revitalize critically endangered languages.

[3]The cost to run generate all translations was \$0.09 using gpt-3.5-turbo and \$5.48 using gpt-4.

References

Jessie Little Doe Baird. 2016. Wopanaak language reclamation program: bringing the language home. Journal of Global Indigeneity, 2(2).

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. 2023. Sparks of Artificial General Intelligence: Early experiments with GPT-4.

Jialun Cao, Meiziniu Li, Yeting Li, Ming Wen, Shing-Chi Cheung, and Haiming Chen. 2022. Semmt: A semantic-based testing approach for machine translation systems. *ACM Trans. Softw. Eng. Methodol.*, 31(2).

Aakanksha Chowdhery et al. 2022. PaLM: Scaling Language Modeling with Pathways.

Serafín M. Coronel-Molina and Teresa L. McCarty. 2016. Indigenous language revitalization in the americas.

Explosion. 2024. Industrial-Strength Natural Language Processing. `https://spacy.io/`. Accessed: 3 Mar 2024.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation.

Tanmai Khanna, Jonathan N. Washington, Francis M. Tyers, Sevilay Bayatlı, Daniel G. Swanson, Tommi A. Pirinen, Irene Tang, and Hèctor Alòs i Font. 2021. Recent advances in Apertium, a free/open-source rule-based machine translation platform for low-resource languages. *Machine Translation*, 35(4):475–502.

Séamus Lankford, Haithem Afli, and Andy Way. 2023. adaptmllm: Fine-tuning multilingual language models on low-resource languages with integrated LLM playgrounds. *Inf.*, 14(12):638.

Patrick S. H. Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual.*

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

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2023. MTEB: massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 2006–2029. Association for Computational Linguistics.

OpenAI. 2023. GPT-4 Technical Report.

OpenAI. 2024a. New and Improved Embedding Model. `https://openai.com/blog/new-and-improved-embedding-model`. Accessed: 3 Mar 2024.

OpenAI. 2024b. New embedding models and API updates. `https://openai.com/blog/new-embedding-models-and-api-updates`. Accessed: 3 Mar 2024.

Tommi A Pirinen. 2019. Workflows for kickstarting RBMT in virtually no-resource situation. In *Proceedings of the 2nd Workshop on Technologies for MT of Low Resource Languages*, pages 11–16, Dublin, Ireland. European Association for Machine Translation.

Surangika Ranathunga, En-Shiun Annie Lee, Marjana Prifti Skenduli, Ravi Shekhar, Mehreen Alam, and Rishemjit Kaur. 2023. Neural machine translation for low-resource languages: A survey. *ACM Comput. Surv.*, 55(11):229:1–229:37.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Nathaniel R. Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for High- (but not Low-) Resource Languages.

SIL International. 2024. 639 Identifier Documentation: mnr. Accessed: 11 Mar 2024.

Yurun Song, Junchen Zhao, and Lucia Specia. 2021. Sentsim: Crosslingual semantic evaluation of machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021*, pages 3143–3156. Association for Computational Linguistics.

Joshua Taylor and Timothy Kochem. 2022. Access and empowerment in digital language learning, maintenance, and revival: a critical literature review. *Diaspora, Indigenous, and Minority Education*, 16(4):234–245.

Daniel Torregrosa, Nivranshu Pasricha, Maraim Masoud, Bharathi Raja Chakravarthi, Juan A. Alonso, Noe Casas, and Mihael Arcan. 2019. Leveraging rule-based machine translation knowledge for under-resourced neural machine translation models. In *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks, MT-Summit 2019, Dublin, Ireland, August 19-23, 2019*, pages 125–133. European Association for Machine Translation.

William Webber, Alistair Moffat, and Justin Zobel. 2010. A similarity measure for indefinite rankings. *ACM Trans. Inf. Syst.*, 28(4):20:1–20:38.

Appendices

A OVP to English Translator Results

sentence	translation	label
tei-zawa-ku ihi	This cooked us.	1
isha'-uu tüba-neika ai-hibi-pü	That coyote has drunk these pinenuts.	1
tüba-uu tsibui-dü	Those pinenuts climb.	1
ta-w̃ui-gaa-wei mahuw̃a	"They are going to write to us, you and I."	1
wo'ada-uu aaponu'-oka u-zawa-ti	The mosquito is cooking the apple.	1
tübbi-uu tüwoobü-neika a-yadohi-pü	The rock has talked to the earth.	1
payahuupü-uu katü-ti	That river is sitting.	1
kwisha'i-wei üü	You will sneeze.	1
isha'-ii tübbi-neika mai-w̃ui-gaa-wei	This coyote is going to write these rocks.	1
toni-uu wünü-ti	The wickiup is standing.	1
isha'pugu-neika ihi mai-dama'i-ku	This found these dogs.	1
wo'ada-neika ihi mai-dama'i-gaa-wei	This will find these mosquitoes.	1
tabuutsi'-uu tüba-noka u-buni-ku	The cottontail saw those pinenuts.	1
maishibü-neika uhu ai-nia-ti	He/she/it is reading these corn.	1
koopi'-ii wükada-noka ui-nia-ku	This coffee read those bird snakes.	1
tuunapi-uu waakü-pü	That food has worked.	1
katünu-ii koopi'-oka ui-nobini-wei	This chair will visit those coffees.	1
aingwü-neika mahu ma-nia-dü	He/she/it reads the squirrel.	1
maishibü-uu wükihaa-gaa-wei	That corn is going to smile.	1
isha'-uu isha'-oka ui-zawa-wei	That coyote will cook those coyotes.	1
aingwü-ii tsibui-dü	This squirrel climbs.	1
katünu-ii tübbi-neika ma-buni-wei	This chair will see this rock.	1
wükada-uu tsibui-ku	The bird snake climbed.	1
wo'ada-uu paya-neika ma-hibi-ti	The mosquito is drinking the water.	1
pagwi-neika mahu ma-w̃ui-pü	He/she/it has written/is writing this fish.	1
tabuutsi'-uu isha'pugu-neika mai-nobini-gaa-wei	That cottontail is going to visit those dogs.	1
paya-neika mahu mai-hibi-gaa-wei	He/she/it is going to drink this water.	1
nishua'i-pü nüü	I am laughing.	1
aaponu'-ii küna-neika a-düka-pü	This apple has eaten this wood.	1
katü-dü uhu	He/she/it sits.	1
mukita-uu isha'pugu-noka u-naki-ti	The lizard is chasing the dog.	1
isha'-oka üü ui-dama'i-ku	You found those coyotes.	1
pahabichi-uu wükihaa-dü	That bear smiles.	1
pahabichi-ii wo'abi-noka ui-naka-dü	This bear hears those worms.	1
habi-ku ihi	This lay down.	1
tümui-ku taa	You and I wrote.	1
üwi-ku ihiw̃a	These slept.	1
tübbi-uu ta-naka-ku	"That rock heard us, you and I."	1
wo'ada-uu tei-gwana-dü	That mosquito smells us.	1
tümui-dü mahuw̃a	They write.	1
üwi-dü nüügwa	We are sleeping.	1
tsibui-ti mahu	He/she/it is climbing.	1
mukita-uu tsibui-pü	That lizard has climbed.	1
payahuupü-uu toyabi-neika ma-zawa-gaa-wei	The river is going to cook the mountain.	1

nobi-uu kwisha'i-ku	That house sneezed.	1
kamü-uu wükihaa-dü	That jackrabbit smiles.	1
toni-uu katü-wei	That wickiup will sit.	1
aingwü-uu katünu-noka u-zawa-gaa-wei	The squirrel is going to cook that chair.	1
paya-uu pasohobü-neika ai-buni-wei	That water will see those trees.	1
toyabi-ii tsibui-ku	The mountain climbed.	1
tsibui-wei taa	You and I will climb.	1
pugu-uu wo'abi-neika ai-naki-ku	That horse chased those worms.	1
mukita-uu wai-noka u-nobini-gaa-wei	The lizard is going to visit the rice.	1
wükihaa-ti mahu	He/she/it is smiling.	1
tüsüga-ii tüwoobü-neika ma-naki-gaa-wei	This weasel is going to chase this earth.	1
yadoha-ku uhuw̃a	They talked.	1
pahabichi-ii pugu-noka ui-nia-ku	The bear read those horses.	1
paya-uu katünu-noka ui-yadohi-gaa-wei	Water is going to talk to those chairs.	1
pagwi-ii wo'abi-noka ui-düka-ti	This fish is eating those worms.	1
tabuutsi'-uu tübinohi-ku	That cottontail played.	1
tünia-ku nüü	I read.	1
poyoha-pü ihiw̃a	These are running.	0
mukita-uu yotsi-wei	That lizard will fly.	1
tabuutsi'-uu tübbi-neika ma-buni-pü	That cottontail has seen this rock.	1
isha'pugu-ii tüba-neika ai-nobini-ku	This dog visited these pinenuts.	1
isha'-uu katü-dü	That coyote sits.	1
pasohobü-ii tsiipa-noka ui-naka-ku	The tree heard those birds.	1
kamü-uu mukita-neika ma-zawa-wei	That jackrabbit will cook this lizard.	1
tuunapi-uu tümui-gaa-wei	That food is going to write.	1
wai-uu aingwü-neika ai-naka-pü	That rice has heard those squirrels.	1
tsiipa-uu pugu-noka ui-naka-pü	The bird has heard those horses.	1
pagwi-ii wükihaa-gaa-wei	These fish are going to smile.	1
tei-nobini-dü uhuw̃a	They visit us.	1
paya-neika ihi ma-dama'i-dü	This finds this water.	1
isha'pugu-neika nüügwa ma-düka-ku	We ate this dog.	1
tabuutsi'-uu hubiadu-dü	The cottontail sings.	1
kwadzi-ii yadoha-wei	This tail will talk.	1
isha'pugu-uu mukita-neika ai-naka-ti	That dog is hearing these lizards.	1
pasohobü-noka üü u-zawa-gaa-wei	You are going to cook that tree.	1
tsiipa-noka uhu u-buni-pü	He/she/it has seen or is seeing that bird.	1
tuunapi-neika mahu mai-gwati-ku	He/she/it hit the foods here.	0
isha'-eika nüügwa ai-w̃ui-gaa-wei	We are going to write coyotes.	1
pasohobü-uu toyabi-neika ma-hibi-pü	The tree has drunk the mountain.	1
aingwü-uu aaponu'-eika ai-naka-ti	The squirrel is hearing these apples.	1
tabuutsi'-uu wai-noka u-yadohi-pü	The cottontail has talked to the rice.	1
katünu-noka mahuw̃a ui-gwati-dü	They are hitting those chairs.	1
tüwoobü-neika uhuw̃a mai-buni-ti	They are seeing these earths.	1
koopi'-uu hukaw̃ia-ti	Coffee is walking.	1
küna-ii ni-naka-ti	This wood is hearing us.	1
toyabi-neika taagwa ma-buni-gaa-wei	We are going to see this mountain.	1
isha'pugu-ii üwi-dü	This dog sleeps.	1
pagwi-neika ihi mai-naka-gaa-wei	This will hear these fish.	1
kwadzi-ii toni-neika ai-gwati-gaa-wei	This tail is going to hit those wickiups.	1

paya-ii tuunapi-noka u-düka-ti	This water is eating that food.	1
mukita-uu tümui-gaa-wei	That lizard is going to write.	1
pahabichi-uu küna-neika ma-zawa-ku	The bear cooked the wood.	1
isha'pugu-uu tabuutsi'-eika a-zawa-dü	That dog is cooking this cottontail.	1
katünu-uu pahabichi-noka u-naki-ku	The chair chased the bear.	1
küna-uu waakü-gaa-wei	That wood is going to work.	1
pugu-neika mahu ai-naka-ku	He/she/it heard these horses.	1

Table 3: One hundred randomly generated sentences (using the OVP sentence building tool described in Section 3) and translated (using the OVP to English translator described in Section 3) labeled 1 if the translation is accurate and 0 otherwise.

B Semantic Similarity Sentences

Base Sentence	**Other Sentences** **(in order of most to least semantically similar to Base Sentence)**
She sings.	He sings. He/she/it sings. She performs a song. A song is being sung by her. She hums a tune. She listens to music. She dances. She eats. The cat sleeps. Mountains echo silently.
The dog fell.	The dog fell yesterday. A dog stumbled. The puppy tripped over. The cat is running. An animal is in motion. The bird flies. Leaves fall in autumn. He reads a book. Clouds cover the sky. Apples on the moon are hungry.
The man ate an apple.	The apple was eaten by the man. A man consumes a fruit. The boy nibbles on an apple. Someone is eating. He drinks water. The woman ate a pie. A cat chases a mouse. Trees grow in the forest. The car is red. Stars twinkle at night.
The sun rises in the east.	The east welcomes the sunrise. Sunrise occurs in the east. Day breaks in the east. The moon sets in the west. The stars shine at night.

	Clouds gather before rain. The wind changes direction. Leaves fall in autumn. Snow covers the mountains. A book rests on the table.
Birds fly south for the winter.	For winter, birds head south. Migratory birds travel south when it gets cold. Birds migrate to warmer climates during winter. Fish swim upstream. Bears hibernate in winter. Flowers bloom in spring. The earth orbits the sun. Trees lose their leaves in fall. The sky is blue. A cat sleeps on the couch.
I read a book yesterday.	Yesterday, I finished reading a book. A book was read by me yesterday. I watched a movie last night. I'll visit the library tomorrow. She writes a letter. He cooks dinner. They are painting a house. The sun sets in the evening. A dog barks at night. The car needs fuel.
The cake was delicious.	Delicious was the cake. The dessert tasted great. We enjoyed the tasty cake. The pie is sour. Coffee complements breakfast. Leaves rustle in the wind. A bird sings outside. Children play in the park. Traffic is heavy today. The phone is ringing.
Lightning precedes thunder.	Thunder follows lightning. First comes lightning, then comes thunder. The storm brings lightning and thunder. Rain refreshes the earth. The sun warms the ground. A river flows to the sea. Mountains reach towards the sky. A cat chases a mouse. Books fill the shelf. The clock ticks steadily.
She painted a beautiful picture.	A beautiful picture was painted by her. The painting she created is beautiful. She sketches a portrait. He writes a poem. They are filming a movie. Birds nest in spring.

	Flowers wilt in the heat. Kids play video games. Cars fill the parking lot. The sun sets late in summer.
The computer is broken.	A broken state afflicts the computer. The machine isn't working. We need to repair the computer. The phone's battery is dead. Lights flicker during a power outage. A book lies open on the desk. Water boils at 100 degrees Celsius. A cat purrs contentedly. The door creaks when opened. Birds migrate in autumn.
He solved the puzzle quickly.	The puzzle was quickly solved by him. Quickly, he found the solution to the puzzle. She completes the crossword. The mystery remains unsolved. A race against time. Flowers are sold at the market. The river cuts through the valley. A key unlocks the door. Leaves turn red in autumn. The train arrives at noon.
The stars twinkle at night.	At night, the stars shimmer. Twinkling stars fill the night sky. Night unveils a sky full of stars. The moon glows brightly. Clouds mask the moon. The sun sets, stars appear. A comet streaks through the sky. Fireflies glow in the dark. Crickets chirp in the evening. A candle flickers in the window.

Table 4: Base sentences and other sentences ordered by their semantic similarity to the base sentence (as determined by authors).

C Vocabulary

tüka	eat
puni	see
hibi	drink
naka	hear
kwana	smell
kwati	hit
yadohi	talk to
naki	chase
tsibui	climb
sawa	cook
tama'i	find
nia	read
mui	write
nobini	visit

(a) Transitive Verbs

katü	sit
üwi	sleep
kwisha'i	sneeze
poyoha	run
mia	go
hukawã	walk
wünü	stand
habi	lie down
yadoha	talk
kwatsa'i	fall
waakü	work
wükihaa	smile
hubiadu	sing
nishua'i	laugh
tsibui	climb
tübinohi	play
yotsi	fly
nüga	dance
pahabi	swim
tünia	read
tümui	write
tsiipe'i	chirp

(b) Intransitive Verbs

isha'	coyote
isha'pugu	dog
kidi'	cat
pugu	horse
wai	rice
tüba	pinenuts
maishibü	corn
paya	water
payahuupü	river
katünu	chair
toyabi	mountain
tuunapi	food
pasohobü	tree
nobi	house
toni	wickiup
apo	cup
küna	wood
tübbi	rock
tabuutsi'	cottontail
kamü	jackrabbit
aaponu'	apple
tüsüga	weasle
mukita	lizard
wo'ada	mosquito
wükada	bird snake
wo'abi	worm
aingwü	squirrel
tsiipa	bird
tüwoobü	earth
koopi'	coffee
pahabichi	bear
pagwi	fish
kwadzi	tail

(c) Nouns

ku	completive (past)
ti	present ongoing (-ing)
dü	present
wei	future (will)
gaa-wei	future (going to)
pü	have x-ed, am x-ed

(d) Object Suffixes

nüü	I
uhu	he/she/it
uhuw̃a	they
mahu	he/she/it
mahuw̃a	they
ihi	this
ihiw̃a	these
taa	you and I
nüügwa	we (exclusive)
taagwa	we (inclusive)
üü	you
üügwa	you (plural)

(e) Subject Pronouns

ii	(proximal)
uu	(distal)

(f) Subject Suffixes

i	me
u	him/her/it (distal)
ui	them (distal)
ma	him/her/it (proximal)
mai	them (proximal)
a	him/her/it (proximal)
ai	them (proximal)
ni	us (plural, exclusive)
tei	us (plural, inclusive)
ta	us (dual), you and I
ü	you (singular)
üi	you (plural), you all

(g) Object Pronouns

eika	(proximal)
oka	(distal)

(h) Object Suffixes

Table 5: Vocabulary available in sentence building system.

D English to OVP Translation Results

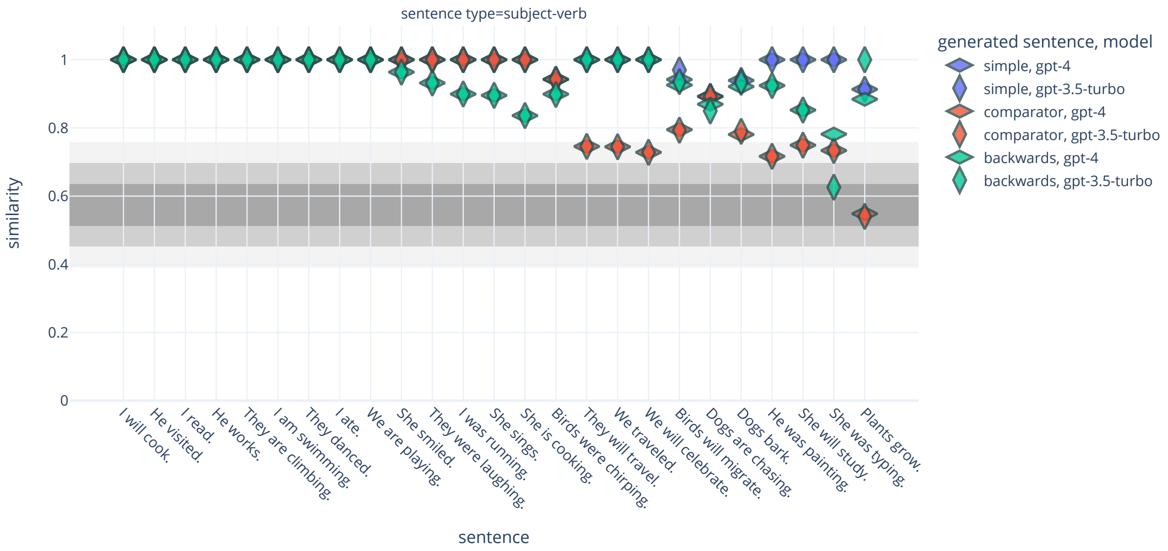


Figure 5: Results for subject-verb sentences. The dark, medium, and light gray bands represent the baseline similarity (between unrelated sentences in the dataset) +/- one, two, and three standard deviations, respectively.

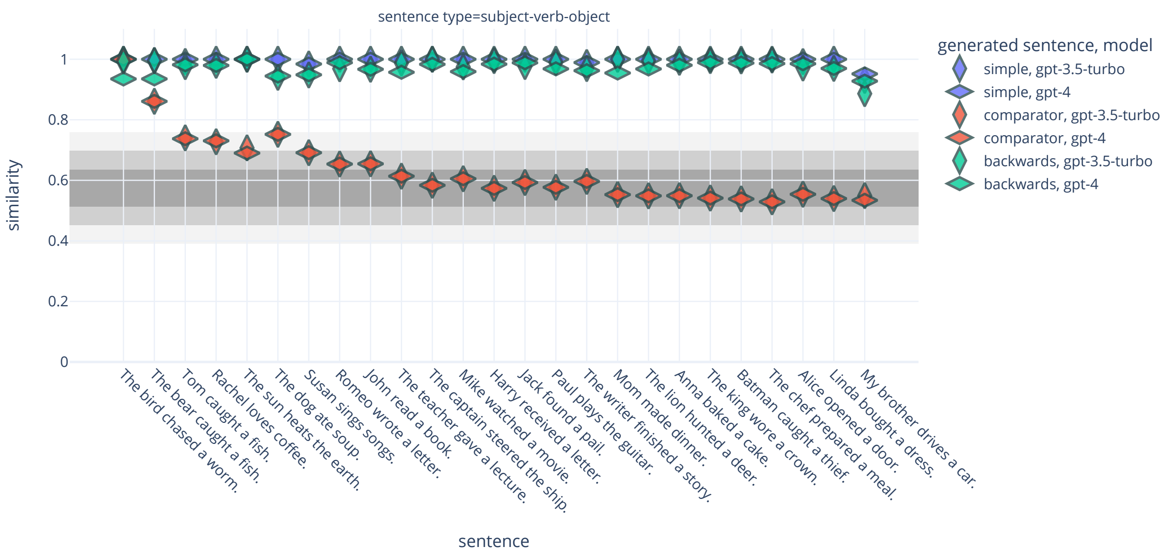


Figure 6: Results for subject-verb-object sentences. The dark, medium, and light gray bands represent the baseline similarity (between unrelated sentences in the dataset) +/- one, two, and three standard deviations, respectively.

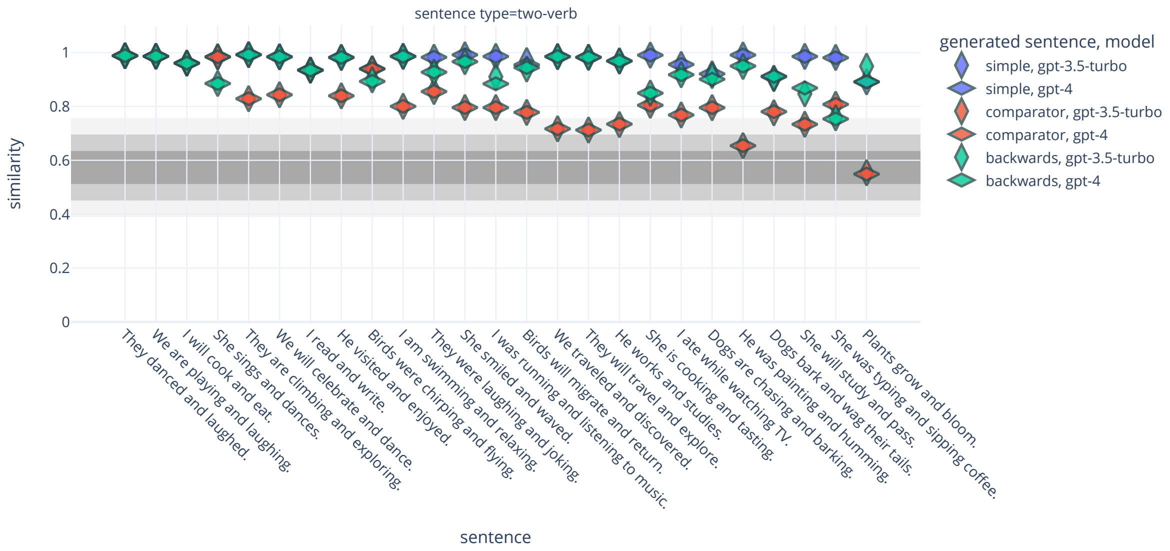


Figure 7: Results for two-verb sentences. The dark, medium, and light gray bands represent the baseline similarity (between unrelated sentences in the dataset) +/- one, two, and three standard deviations, respectively.

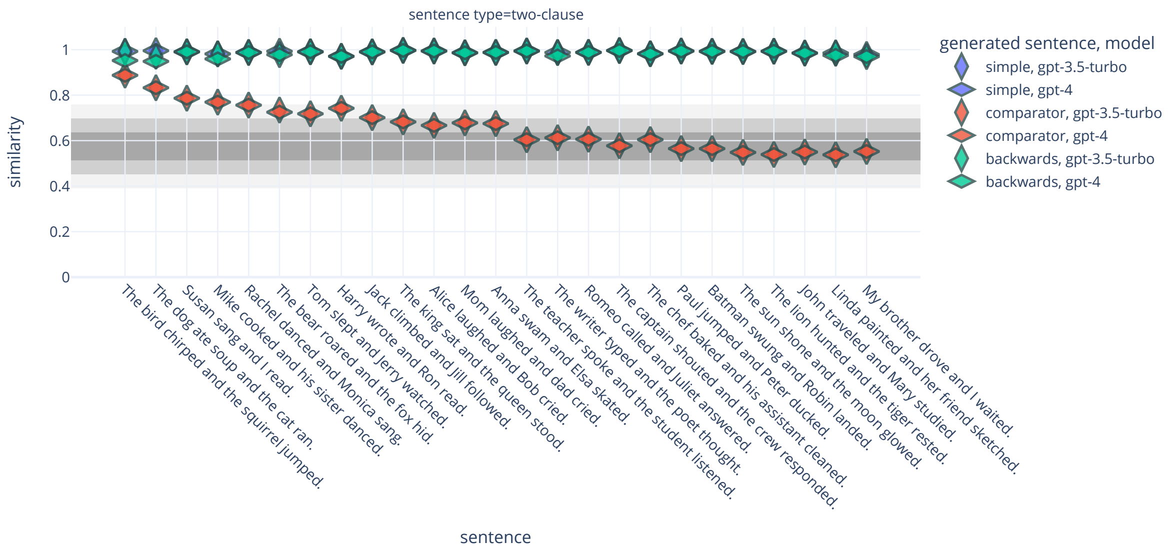


Figure 8: Results for two-clause sentences. The dark, medium, and light gray bands represent the baseline similarity (between unrelated sentences in the dataset) +/- one, two, and three standard deviations, respectively.

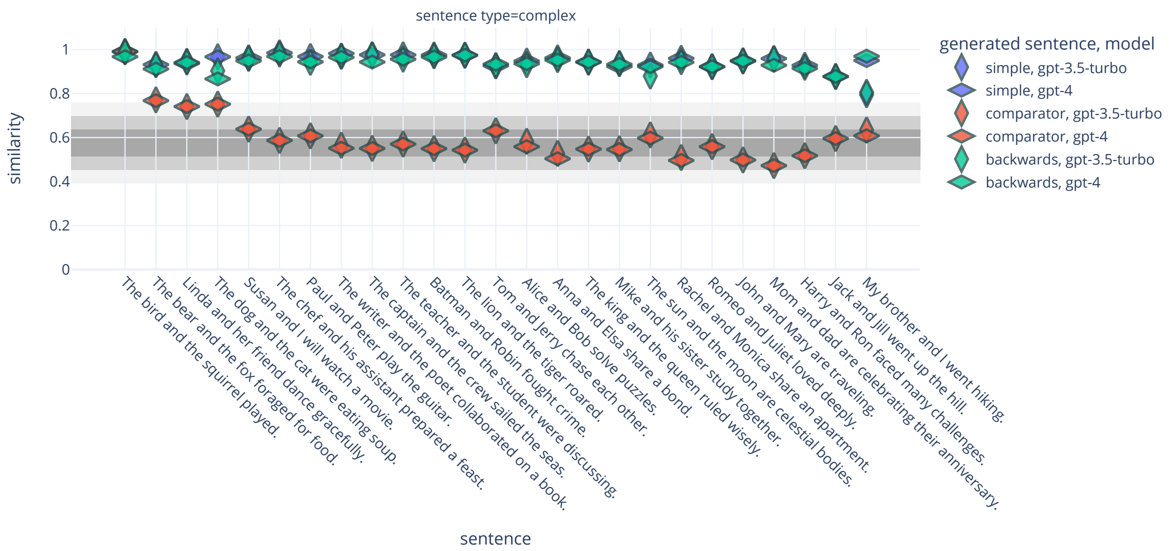


Figure 9: Results for complex sentences. The dark, medium, and light gray bands represent the baseline similarity (between unrelated sentences in the dataset) +/- one, two, and three standard deviations, respectively.

E Semantic Similarity Baseline Score

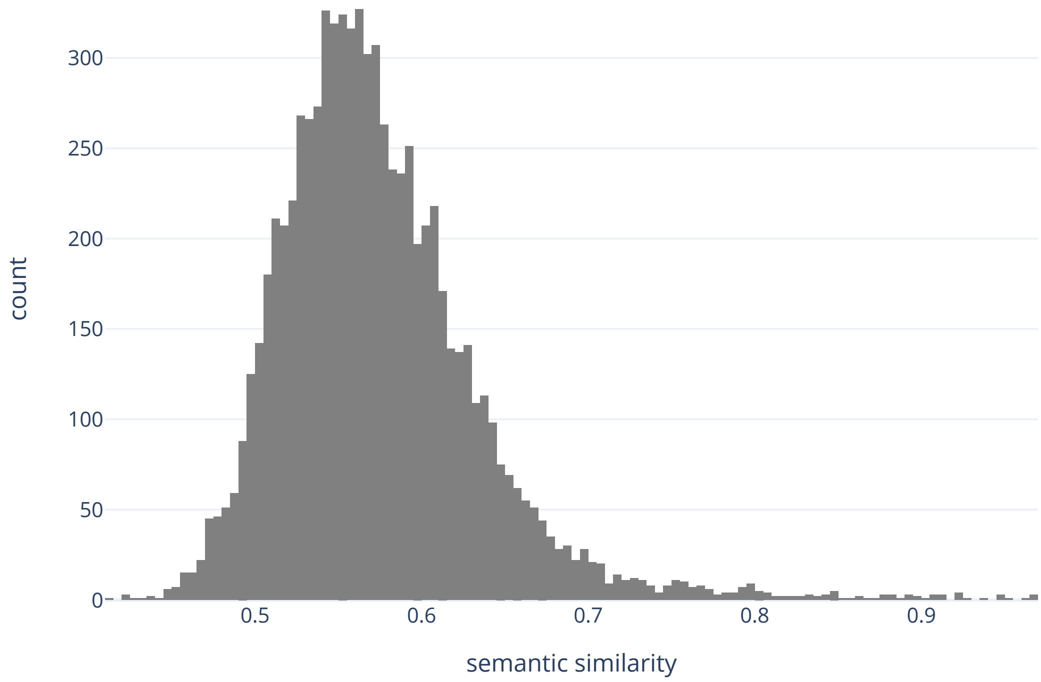


Figure 10: Distribution of semantic similarity scores between all pairs of sentences in the dataset.

A Concise Survey of OCR for Low-Resource Languages

Milind Agarwal
George Mason University
magarwa@gmu.edu

Antonios Anastasopoulos
George Mason University
antonis@gmu.edu

Abstract

Modern natural language processing (NLP) techniques increasingly require substantial amounts of data to train robust algorithms. Building such technologies for low-resource languages requires focusing on data creation efforts and data-efficient algorithms. For a large number of low-resource languages, especially Indigenous languages of the Americas, this data exists in image-based non-machine-readable documents. This includes scanned copies of comprehensive dictionaries, linguistic field notes, children's stories, and other textual material. To digitize these resources, Optical Character Recognition (OCR) has played a major role but it comes with certain challenges in low-resource settings. In this paper, we share the first survey of OCR techniques specific to low-resource data creation settings and outline several open challenges, with a special focus on Indigenous Languages of the Americas. Based on experiences and results from previous research, we conclude with recommendations on utilizing and improving OCR for the benefit of computational researchers, linguists, and language communities.

1 Introduction

Latin America is home to a linguistically diverse set of hundreds of indigenous languages. Many of these are low-resource in terms of text and audio resources, and generally lack basic natural language applications such as spell checkers, part of speech (POS) taggers, etc. However, these languages have a large number of digital resources (not machine-readable) in the form of recordings, plays, stories, and dictionaries. One major repository of such materials is the Archive of the Indigenous Languages of Latin America (AILLA).[1] Of the documents in AILLA's collection, particularly interesting to NLP researchers are linguistic materials such as grammars, dictionaries, ethnographies,

[1] A joint effort of the LLILAS Benson Latin American Studies and Collections and UT Austin.

Figure 1: We highlight 10 Indigenous Languages from Central and South America with large amounts of undigitized resources to anchor our survey and workflow recommendations for researchers and linguists.

and field notes, that can serve as training data for NLP applications and Optical Character Recognition (OCR). Releasing digitized versions of such a repository of hundreds of datasets can preserve invaluable linguistic materials and accelerate research in NLP. Modern OCR can extract text from such documents, but this requires accurate layout detection and post-processing to make the extracted text usable for downstream NLP tasks (Bustamante et al., 2020). OCR is a well-established field, with its advances mostly drawing from innovations in Computer Vision. More recently, OCR has been increasingly used for resource-creation for low-resource languages in NLP contexts (Ignat et al., 2022a). There are also several excellent surveys and tutorials (Nguyen et al., 2021; Neudecker et al., 2021; Memon et al., 2020) on building and using OCR for broad applications, however, there is a dearth of specialized surveys for low-resource language OCR. Therefore, the aim of this paper is to fill this gap and acquaint researchers and language

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 88–102
June 21, 2024

Language	693-3	Family	Main Country	Speakers	Pages	Undigitized Resource
S. Bolivian Quechua	QUH	Quechuan	Bolivia	1.6M	216	Kalt (2016)
Mísqito	MIQ	Misumalpan	Nicaragua, Honduras	150K	61	Bermúdez Mejía (2015)
Mam	MAM	Mayan	Guatemala	600K	144	England (1972-1985)
Chuj	CAC	Mayan	Guatemala	60K	564	Hopkins (1964)
Chimalapa Zoque	ZOH	Mixe-Zoquean	Mexico	<10K	3744	Johnson (2000-2005)
Chiquián Quechua	QXA	Quechuan	Peru	100K	29	Proulx (1968)
Sharanahua	MCD	Panoan	Peru	<10K	209	Déléage (2002)
Tzeltal	TZH	Mayan	Mexico	600K	38	Kaufman (1960-1993)
Baniwa	BWI	Maipurean	Brazil, Venezuela	12K	310	Wright et al. (2000)
Ixil	IXL	Mayan	Guatemala	120K	2	Adell et al. (2016)

Table 1: A brief description of the 10 languages that we focus on to highlight the amount of data in Indigenous Languages of the Americas that requires high-quality OCR. We include their ISO 693-3 codes, primary country, number of speakers, and references to the resource that requires digitization. Overall, this data includes 5317 pages to be transcribed which *if digitized* can be sufficient to train many downstream NLP tasks.

communities with techniques and adaptions necessary for high-quality digitization in low-resource settings. To summarize, this paper makes the following contributions:

1. Highlights undigitized resources in 10 American Indigenous languages (§2).
2. First concise survey of OCR for low-resource settings and languages (§3).
3. Discussion on major open problems in scaling digitization for low-resource languages (§4).
4. Recommendations for researchers, linguists, and language communities on the entire resource curation and digitization pipeline (§5).

2 Undigitized Data

Over the past decade, many researchers, linguists, and consortiums have worked closely with native speakers and language communities to create datasets including digitized text, audio, transcriptions, translations, stories, etc. Some of these resources may not be machine-readable but include extremely valuable resources from an NLP perspective, such as multilingual lexicons, pronunication guides, plain text from a wide-variety of domains such as stories, essays, plays, news, linguistics etc. A comprehensive guide with resources in all low-resource languages (over 6000+) would be valuable but is out of the scope of this paper, so we focus on highlighting relevant resources in 10 American Indigenous languages. The AILLA collection contains several textual corpora in non-machine-readable image format for the selected languages in Table 1, as well as in hundreds of other indigenous languages of the Americas. The selected languages together cover over 5000 pages of undigitized data in these 10 languages. Each page contains multilingual textual data that needs high-quality extraction.

A large majority of OCR work for low-resource settings includes preservation and digitization of historical data, early printed books (Reul et al., 2017), palm-leaf manuscripts (Prusty et al., 2019; Sharan et al., 2021; Alaasam et al., 2019) etc. Preexisting repositories (such as PubMed or arXiv) are also widely used for training OCR systems (Zhong et al., 2019; Blecher et al., 2023), however, note that this approach is not scalable to low-resource settings which often lack such ready-to-use datasets.

For the Americas, due to widespread adoption of extended Latin alphabets in writing, texts from the last couple of centuries are often typed, but several collections include partially or completely written handwritten documents and annotations. Historically used typing fonts may be challenging to decipher or out of use due to orthographic reforms (Naoum et al., 2019; Klaiman and Lehne, 2021; Jiang et al., 2019), and handwriting varies widely across individuals, making extraction challenging (Déjean and Meunier, 2019; Alaasam et al., 2019; Sharan et al., 2021). Over time, language communities may even adopt new orthographies, which might require researchers to build new keyboards and transcription systems to make the digitized corpora readable by community members (Rijhwani et al., 2023). Digitizing these resources can allow for more accessible linguistic research, training language models, translation systems, POS taggers, etc. The AILLA collection constitutes of a healthy mix of both typed and handwritten text. As evident from Table 1, the highlighted languages will require sustained OCR efforts to digitize their respective resources. With access to machine-readable text, downstream NLP tools can then begin to be built.

Note that in our concise survey paper, our aim is not to digitize these specific books - that would warrant separate carefully designed studies as each resource is bound to have unique challenges and is connected to language communities with possibly different language technology needs. This paper highlights, for researchers unfamiliar with these languages and domain, different resources available for experimenting with OCR modeling approaches and recommended workflows for achieving such digitization.

3 A Concise Survey of OCR

Now that we've seen the data resources available for our 10 selected Indigenous languages (§2), we will highlight useful and practical OCR adaptations and innovations necessary for digitization of such low-resource language data. We cover techniques in four major parts of the digitization pipeline: preparation of the data and model, active training, decoding or generation, and post-processing. To ground the following discussion, we will define an example dataset $\mathcal{C}$, with K pages, where p_i represents separate pages. L represents the paired labels for each $p_i \in C$ (with each l_i representing the the ground-truth words and characters for the page p_i).

$$\mathcal{C} = \{p_i\}_{i=1}^{K}; \mathcal{L} = \{l_i\}_{i=1}^{K}$$

For an OCR experimental setup, we would usually have four different datasets: C_{pretrain} (unlabeled pages), C_{train} (with labels L_{train}), C_{val} (validation/development set used for evaluation during training along with labels L_{val}), C_{test} (for reporting model performance along with labels L_{test}).

3.1 Preparation: Setting the Stage

Data Augmentation Due to lack of data in low-resource indigenous languages, data augmentation should be the first step for any digitization pipeline, to increase the utility of the small labeled gold dataset (Shorten and Khoshgoftaar, 2019). For an OCR system, this means that the images themselves must go through several transformations such as skewing, binarization, scaling, cropping, blurring, etc. to ensure that the final model can handle such variations in-the-wild and still be able to extract text from the image. Data augmentation is well-studied in literature (Liu et al., 2018; Khan et al., 2021) and incorporating it into OCR pipelines has shown to increase robustness and performance by making the best use of a small training set (Storchan and Beauschene, 2019; Namysl and Konya, 2019).

More precisely, a set of augmentation operations, $O = \{o_1, o_2, .., o_j\}$ where j denotes the number of operations can be applied to each image p_i. o can denote functions like binarization, greyscale, gaussian blur, cropping etc. C_{train} can be augmented using any combination of operations from set O, to generate a new set $C_{\text{train-aug}}$, which would serve as the newly expanded training corpus. For each new augmented page, $p_{i,j} = o_j(p_i)$ and it's label would be $l_i \in L_{\text{train}}$.

Pretraining with General Unlabeled Data For data that is not labeled, self-supervised pretraining techniques are often used to better initialize the network (Li et al., 2023; Bugliarello et al., 2021). In case of encoder-decoder models, pretraining has been applied to both components separately and has been shown to be successful (Lyu et al., 2022; D'hondt et al., 2017), when large amounts of unlabeled images or text are available. Similarly, in case of the in-house pretraining set C_{pretrain}, ground-truth *text* labels are not available. So, the images from this pretraining set can be used to pretrain the OCR model, and the first-pass text can incorporated into pretraining the post-correction model (with learned denoising rules) (Rijhwani et al., 2020).

Transfer Learning from Related Languages For certain minority and low-resource languages, it is shown that a base system that is trained to identify a similar language or similar character-set generally leads to performance increases downstream(Lin et al., 2019; Zhuang et al., 2021; Rijhwani et al., 2019). As an illustration, in our selection of 10 American Indigenous languages, choosing a corpus in a high-resource language of Central and South America i.e. Spanish or Portuguese, might be appropriate for transfer learning. In the OCR domain, transfer learning has been applied to better enhance the quality of low-resource OCR output at the decoding step (Todorov and Colavizza, 2020; Jaramillo et al., 2018). However, Tjuatja et al. (2021) investigates transfer learning for OCR post-correction for indigenous and endangered languages and points at mixed results. They say that for downstream performance improvements, transfer learning is not straightforward and may require getting data from a larger set of domains. Gunna et al. (2021) investigated transfer at the text detection level for Indian languages, and observed positive outcomes when transferring from other Indian languages that look *visually* similar, even if they are from different language families.

3.2 Training: Learning Quickly and Better

For training OCR systems, supervised techniques are usually preferred in low-resource settings. Unsupervised methods have shown some promise recently (Gupta et al., 2021; Dong and Smith, 2018; Garrette and Alpert-Abrams, 2016), but they often require larger datasets for training. Since our focus is on low-resource indigenous languages, we restrict our discussion to supervised techniques. In this setup, there are usually two options - using off-the-shelf systems like Google Vision, Tesseract etc. or training from scratch. For Indigenous languages of the Americas, using off-the-shelf OCR systems can give an excellent starting point (Rijhwani et al., 2023), and since they are the focus languages of our paper, we will discussing training strategies on top of the first-pass OCR output obtained from such systems. Post-OCR processing aims to rectify mistakes made by OCR systems in text extraction, and can be extremely valuable for low-resource languages. Post-processing is valuable because it makes little to no assumptions about the first-pass OCR system itself (helpful when the system is commercial or closed-source) and instead focuses on improving the quality of the output (Kolak and Resnik, 2005).

First-Pass OCR For the first-pass, a high-quality OCR system, such as Google Vision or Tesseract, that is known to work well on endangered-language documents (Fujii et al., 2017; Rijhwani et al., 2020) is commonly used. Performing OCR on page p_i gives us a first-pass output, f_i in the form of n_i bounding boxes x and the texts within them a. Each x contains the set of coordinates for the bounding box, and the corresponding string a represents the text within the box.

$$f_i = [(x_1, a_1), (x_2, a_2), .., (x_{n_i}, a_{n_i})]$$

Text Corrections An ideal post-OCR text correction algorithm would model the error distribution of the OCR algorithm's output text and systematically correct it (Berg-Kirkpatrick et al., 2013; Schulz and Kuhn, 2017). This can be an extremely valuable tool when digitizing indigenous language documents because the OCR pipeline's decoder language model is often of low-quality due to the low-resource nature of indigenous and endangered languages. Across the digitization efforts that we've highlighted and amongst others, it is quite common to perform text-based semi-automatic or human post-correction (Maxwell and Bills, 2017; Cordova and Nouvel, 2021; Rijhwani et al., 2021). For every first-pass page f_i, we output a corrected page:

$$q_i = [(x_1, b_1), (x_2, b_2), .., (x_{n_i}, b_{n_i})]$$

where x indicates the boxes from the first-pass, and b indicates corresponding corrected text. In human post-correction, an annotator (preferably a speaker of the language), would edit the first-pass OCR output to match with the ground-truth text as evident from the image. In semi-automatic setups, several consistent OCR errors may be identified from a small number of corrections and automatically applied to the remaining first-pass prediction to reduce the burden on the annotator.

Coverage Mechanism Since OCR is seen a generation task, it can be important for the model's attention distribution to pay attention to different parts of the input string. To ensure that this happens, a coverage mechanism is often introduced (Tu et al., 2016; Mi et al., 2016). This mechanism has empirically been shown to greatly improve OCR accuracy and seq2seq performance (See et al., 2017; Rijhwani et al., 2021; Klaiman and Lehne, 2021). A coverage vector at at time step t will be

$$c_t = \sum_{t'=0}^{t'=t-1} \alpha_{t'}^{a}$$

where α_t^a represents the attention distribution for the input a at time step t. This coverage vector c_t can be weighted and included in the attention computation for the next α_{t+1}, and be added to the base cross entropy loss as follows:

$$\sum_t \sum_{i=0}^{\mathrm{len}(a)} \min(\alpha_{t,i}^{a}, c_{t,i})$$

Diagonal Attention Since post-correction from first-pass OCR output is mostly a copying step and reordering rarely occurs (Schnober et al., 2016), the model can mostly focus on generating the elements close to the diagonal. Therefore, under this paradigm, off-diagonal entries outside a certain radius are penalized more heavily by including them in the training loss (Cohn et al., 2016). This simplifies the decoding step and encourages the model to maximize attention on items within the diagonal attention range. The modified loss function at time step t for a diagonal range d and attention distribution α would be:

$$\sum_{t'=1}^{t-d} \alpha_{t,t'}^{a} + \sum_{t'=t+d}^{\text{len}(a)} \alpha_{t,t'}^{a}$$

Diagonal attention shown to empirically improve OCR performance for low-resource languages (Rijhwani et al., 2021, 2020) and can be easily incorporated in OCR post-correction modeling.

Active Learning Data labeling is an expensive task for low-resource languages and especially so for a non-trivial annotation task such as OCR correction or image labeling. To select only those pages to annotate that would help the OCR model the most, a systematic paradigm called Active Learning can been utilized (Settles, 2012). For the low-resource OCR domain and for layout analysis, active learning has shown to be empirically quite valuable (Reul et al., 2018; Shen et al., 2022; Monteleoni and Kaariainen, 2007; Abdulkader and Casey, 2009; Gupta et al., 2016). It can help select which part of the C_{pretrain} to annotate and add into C_{train} using *query by committee* which trains several learner models on the current C_{train} and each model casts its vote/prediction on a set of V unlabeled examples from C_{pretrain}. In the equations below, $\text{uq}(\cdot)$ counts the number of unique characters in a list of predictions, M represent the independently trained models (m in total), s_v represents the vth sentence in C_{pretrain}, and $V = \text{len}(C_{\text{pretrain}})$.

$$\text{ag}_{s_v} = \text{uq}([M_1(s_v), M_2(s_v), ..., M_m(s_v)])$$
$$v* = \text{argmax}_{v=0}^{V}(\text{ag}_{s_v})$$

Sample $v* \in C_{\text{pretrain}}$ is the sample that models disagree on most so it is actively added into the training set C_{train} (principle of maximal disagreement) since it would benefit from human annotation and improve the OCR model the most (Settles, 2012).

3.3 Decoding: To Generate or Not?

In this subsection, we'll discuss some recently proposed and empirically useful strategies to improve OCR decoding under low-resource settings.

Copy Mechanism Since at the decoding step, it is highly likely that most of the corrected text would be identical to the input, it is shown to be useful (Gu et al., 2016) to have two different probability distributions for decoding - *copy* and *generation*. At decoding, the model can choose, whether to sample from the attention distribution (P_{copy}) or generate the output through generation (See et al., 2017; Sutskever et al., 2014).

$$P_{\text{copy}}(y_t) = \sum_{t'=0}^{t} \alpha_{t,t'}$$

This can reduce the OCR character and word error rates by 2-5 times under low-resource settings (Rijhwani et al., 2020; Gu et al., 2016). Krishna et al. (2018) also use a copying mechanism for Sanskrit OCR and gain about 10% points over the base model with the copy mechanism, demonstrating that incorporating copying into an OCR pipeline for low-resource indigenous languages can be extremely beneficial. The copy probability can be weighted for each time step based on a $p_{\text{copy}} \in (0, 1)$ which can be generated as a weighted sum of the context vector, decoder state, and the previous time step's decoder probability. Therefore, we get the following copy-generation probability for a particular time step t and output string y:

$$p(y_t) = (1 - p_{\text{copy}}) * P(y_t) + p_{\text{copy}} * P_{\text{copy}}(y_t)$$

Lexical Decoding In order to counter the noise that self-training from the previous training step is bound to introduce i.e. reinforcing the errors from the first-pass, *lexical* adaptations have been successfully introduced in the OCR decoding step to improve the quality of the prediction (Schulz and Kuhn, 2017; Rijhwani et al., 2021). This proposed approach has shown to empirically benefit the decoding because it assumes that the correct forms of a word appear more frequently (assuming OCR errors to be inconsistent) and biases the output towards such observed forms.

3.4 Evaluation: How to Measure Progress?

Prediction Scoring and Evaluation Metrics When building an OCR system from scratch, mean-average-precision (mAP) and intersection-over-union (IoU) are the most commonly used metric to evaluate the quality of the bounding boxes. For the predicted bounding boxes $P = \{x_1, x_2, ..., x_e\}$, researchers commonly use IoU over all pairs of boxes to generate a ranked list of the best possible bounding box prediction and reference pairs (Girshick, 2015; Prasad et al., 2019; Prieto and Vidal, 2021). Then, a range of IoU thresholds can be used generate a confusion matrix from which we can get a pair of precision and recall values for that threshold. Plotting these two values for all thresholds, we can get a precision-recall curve, the area under which is called AP i.e. average precision. We can get an AP for each reference box x_e, and averaging them all will generate a mAP for that page. This can indicate the quality of alignment of the predictions P with the true reference labels.

However, for many Indigenous Languages of the Americas, off-the-shelf systems and commercial systems will produce a reasonable first-pass prediction since they use extensions of the Latin alphabet (Rijhwani et al., 2020). In this case, evaluation needs to match two text strings: the prediction and the gold reference. For this, character error rate (CER) and word error rate (WER) are the most popular evaluation metrics. Depending on the language, both CER and WER may not be indicative - for polysynthetic languages where a large amount of vocabulary would be unseen at test-time, character-level error rate has been shown to be more indicative of OCR performance (Rijhwani et al., 2023).

$$\text{CER} = \frac{s_c + d_c + i_c}{n_c}; \text{WER} = \frac{s_w + d_w + i_w}{n_w}$$

where s, d, and i represent substitutions, deletions, and insertions at the $c$haracter or $w$ord level over the reference text which has n characters/words.

Loss Functions If using an off-the-shelf system for first-pass output, researchers only need to train post-correction models. In this case, a cross entropy loss is essential in addition to several other adaptive losses discussed in §3.2 such as diagonal loss and coverage loss (Cohn et al., 2016; Tu et al., 2016). To optimize a combination of these losses, common optimizers like Stochastic Gradient Descent (SGD) or Adam are often used (Rijhwani et al., 2020). In situations where the OCR system needs to be trained from scratch, per-pixel sigmoid or softmax losses are employed due to the pixel-level nature of the predictions from common models like Mask R-CNN and Fast R-CNN (Girshick, 2015; He et al., 2017). Multiple losses are generally computed if different branches of a network analyze and predict different aspects of the recognition task, and total loss in such cases can be computed by using a convex combination of these individual losses (Prusty et al., 2019).

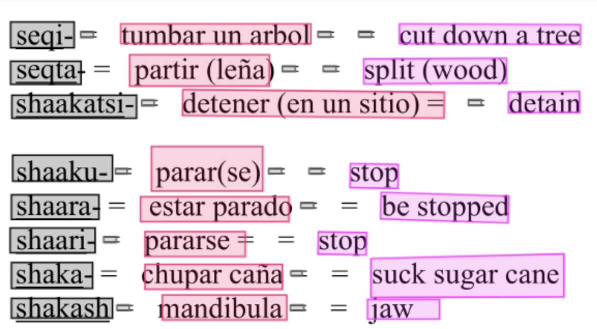


Figure 2: A post-corrected OCR document in Chiquián Quechua (multilingual with Spanish and English) from the AILLA collection (§2). Here, the annotator readjusted he detected bounding boxes, corrected the textual errors in the new boxes, and colored boxes belonging to the 3 languages differently.

4 Open Problems

Layout Preservation One of the most pressing issues remaining largely unsolved in OCR literature is that of structure preservation. OCR tools, especially those off-the-shelf may not good preserve the layout of the page in the output OCR text accurately and might require manual post-OCR alignment (Tafti et al., 2016; Rijhwani et al., 2020). The detected bounding boxes may not follow a logical layout as would be expected by human inspection. This means that researchers need to perform some level of alignment after getting the OCR outputs (Xie and Anastasopoulos, 2023), before applying OCR models (Ignat et al., 2022a), or cropping each image into separate line-level images (may be financially impractical if using commercial systems). From a resource-creation perspective for indigenous languages, preserving structure in the final output is extremely important, so we recommend that researchers think about how to design their experiments early on to address this issue.

To the best of our knowledge, while previous work has focused on layout detection as a first-step (Bustamante et al., 2020), it has not been explored as a post-processing step, primarily due to a lack of ground-truth structural data. Previously, two major studies (Blecher et al., 2023; Zhong et al., 2019) have used existing large-scale corpora like arXiv to extract large-scale ground truth (source-code); but, this approach is not scalable to resource-creation efforts involving low-resource languages. To build such a structure post-correction model, annotators would be required to not only correct the text in the OCR but also structurally correct the first-pass OCR outputs in some kind of graphical user interface (as shown in Figure 2). This would involve scaling, translating, merging, or splitting bounding boxes, while keeping the text within faithful to the each box's new coordinates. Such a task could be framed as follows: for every text-corrected page q_i, we output a corrected page

$$r_i = [(y_1, c_1), (y_2, c_2), .., (y_{m_i}, c_{m_i})]$$

where m_i denotes the number of new bounding boxes after post-correction (may be different from n_i). We consider human-corrected r_i as the ground-truth text and layout. Note that while this step mainly transforms the structure, it would also involve transferring the initially corrected text (b_i, b_{i+1}, etc) from the first-pass boxes that now make up the corrected box y_i, and therefore, the texts are labeled as c_i. However, since such structural post-correction ground-truth data may be expensive to obtain, researchers may also consider getting this ground-truth from a dedicated layout detection model automatically, and post-correcting output from the best first-pass OCR system to adapt to this automatically-extracted desired layout.

Atypical Characters, Fonts, and Words Modeling historical orthographic variations with modern-day LMs, trained on current spelling conventions, can prove challenging during the decoding step (Poncelas et al., 2020). Work on better text extraction from historical documents from the printing press era resulted in the development of the popularly used unsupervised Ocular model (Berg-Kirkpatrick et al., 2013). Synthetic data has been successfully used before to offset the effect of atypical characters and typefaces (Borenstein et al., 2023; Drobac et al., 2017), and unsupervised techniques have been used to automatically learn the font style of a document in the context of historical document recognition and OCR (Berg-Kirkpatrick and Klein, 2014). However, research is still limited in the low-resource domain and researchers would need to ensure that their fonts and character sets are supported by their chosen OCR model (if training one from scratch) or are reconstructed using recent work in visual representation learning (Srivatsan et al., 2021; Vogler et al., 2022). For off-the-shelf systems (interfaced through APIs), it is not possible to directly include support for unique characters and researchers will need to add a post-correction step that includes a mix of post-processing scripts for easily resolvable errors and dedicated trained supervised models to correct first-pass output.

Linguistic Diversity For researchers working with several low-resource or indigenous languages at the same time, it can be desirable to train one model that is capable of handling different writing systems, diacritics, differing image qualities, and unique document formatting (Joshi et al., 2020). While one approach may be to develop 'language-agnostic' methods, previous work has shown that in practice such models are far from language-agnostic (Joshi et al., 2020; Bender, 2011) and tend to have high-performance only for a handful of languages. High OCR accuracy is usually desirable for all the low-resource languages under consideration, and in such a scenario, it may be best to train separate OCR or post-correction models.

5 Workflow Recommendations

In this section, we share recommendations based on the most successful strategies followed in the surveyed papers. This can serve as a starting point for computational researchers, linguists, and students new to the low-resource domain. We acknowledge that these recommendations, while grounded in our survey, are still subjective and researchers may need to modify some elements to suit their specific cases.

Language and Document Selection To anchor our survey, we selected 10 languages that have permissive licenses, use the Latin alphabet, whose special diacritics were available on the English keyboard, and which had typed documents in common fonts. Similarly, when selecting documents for their languages of interest, researchers should consider licensing, need for special keyboards, quality of the document, and layout diversity.

Evaluation Techniques For evaluation of the OCR quality, we recommend simply looking at the final output i.e. text. Provided there is a reference or gold text, these predictions can be compared to them and a CER/WER can be obtained. There is no gold standard target CER/WER so researchers will have to inspect the quality of the output and decide, with feedback from language community members, the CER/WER they would like to target in any potential modeling. Note that a line-aligned version of the extracted text will be required and this may either be obtained by only OCRing at the line-level (after cropping) or by aligning the predicted text with the reference text using metrics such as Levenshtein distance.

Preliminary Experiments For preliminary experiments, we recommend that 1-2 lead researchers manually annotate and audit a modest sample of the dataset themselves. This can help ensure that the researchers and language community members are familiar with the annotation workflow and can better guide any future annotators. Conducting some annotation before running OCR experiments is crucial because there needs to be some standard set for evaluation of all the models we will now experiment with. Once a few pages have been annotated, researchers can begin the OCR process using a common off-the-shelf OCR method like Google Vision or Ocular.

Data Annotation Now that the researchers have an idea of the quality of off-the-shelf OCR systems, we recommend that they recruit annotators to annotate a larger sample of the data if the OCR quality was found to be *low*. Annotators don't need to be speakers of the indigenous languages selected; however, they should have basic pattern recognition, data annotation, and typing skills. Previous work has shown that annotators without knowledge of the indigenous language are fairly adept at performing OCR corrections, if they can read the language's script and distinguish between any new diacritics (Rijhwani et al., 2023). Annotators should be trained to use the annotation platform using standardized guidelines, and a manual audit should be conducted by lead researchers to ensure compliance.

Post-Correction If the performance from preliminary experiments is satisfactory, we recommend post-correcting to further improve the results. A post-correction model should ideally help reduce character-level errors down to less than 5% (Maxwell and Bills, 2017; Cordova and Nouvel, 2021; Rijhwani et al., 2021). As discussed in §3, we recommend that researchers use a combination of copy mechanism, coverage, diagonal/positional attention, and active learning to improve performance. Rijhwani et al. (2021) implements most of the necessary post-correction features and their code can be used directly to train post-correction models for low-resource settings.

Training from Scratch On the other hand, if the preliminary experiments reveal that the error rates are quite high, researchers can consider simply training a custom OCR system from scratch. This will require a sizeable amount of annotated pages for training in addition to computational expertise in settling on the best hyperparameters and setting up the training pipeline. Human-annotations collected in the data annotation phase can be used to train OCR models from scratch and first-pass outputs can be used to train further with post-correction models. We recommend using open-source tools like Tesseract (Smith, 2007) or Ocular (Berg-Kirkpatrick et al., 2013) to train custom models due to their efficiency, optimizations, and active user community. More advanced researchers may also consider writing their own architecture and training pipeline from scratch. However, note that training systems from scratch is not straightforward, and researchers are bound to run into challenges. For instance, Tesseract has a high setup time and learning curve, doesn't have any graphical user interface, and requires high-quality images which may not be available for certain low- languages or image collections.

Deployment and Improvements For in-house use, the final trained model can be used directly to digitize the entire corpus and any other collections in that language. We recommend that computational researchers and language community members stay in touch throughout the training, annotation, and deployment process, and flag any issues with OCR quality and modeling. In some cases, if sufficient OCR quality is not being achieved despite trying the aforementioned techniques, some concessions and further data selection and annotation may be required. For instance, the quality of the data itself might need improvement (redoing scanning of the original source text), another phase of annotation may need to be conducted for substantially more data, or some unique algorithmic techniques may need to be developed to achieve quality OCR for the particular documents. We recommend using LabelStudio (Tkachenko et al., 2020-2022), which is an open-source labeling and annotation platform. The user interface is high-quality, user-friendly, and quite simple to setup and share with collaborators and annotators. There is also an active LabelStudio Slack where issues get resolved relatively quickly.

6 Related Work

Optical Character Recognition OCR has been studied as a research problem for decades, and today, commercial and open-source OCR systems can extract text quite accurately from most images and can even be used in real-time due to test-time efficiency (Smith, 2007; Blecher et al., 2023; Berg-Kirkpatrick et al., 2013). OCR can involve extracting characters, words, paragraphs, and even preserving the layout of text on a page or in an image. OCR is widely used in the digital humanities (Reul et al., 2017; Rijhwani et al., 2021, 2020) and in businesses since it is a necessary step for digitization of rare manuscripts, books, linguistic field notes, invoices, business documents etc. It is also an invaluable technique in creating new data for low-resource languages for downstream NLP tasks and applications (Ignat et al., 2022b). In the last two decades, several excellent surveys from the computer vision community have been published covering OCR developments (Nguyen et al., 2021; Neudecker et al., 2021; Memon et al., 2020). In the low-resource domain, Hedderich et al. (2021)'s survey covers broad NLP advances but it does not cover optical character recognition. To the best of our knowledge, other than ours, no previous work has surveyed OCR for low-resource languages.

Resource Creation Text or image-based datasets and corpora are most commonly created by scraping or crawling the web; however, we would like to highlight a few additional OCR-created datasets, especially those that work with American indigenous languages other than those reported in Table 1. Cordova and Nouvel (2021) addresse the lack of resources for Central Quechua, since resources exist mostly in the dominant Southern variety, using OCR technologies and share a successfully digitized corpus. Hunt et al. (2023) digitizes an Akuzipik (indigenous language spoken in Alaska and parts of Russia) dictionary parallel with Russian text, which was shown to be very valuable for downstream NLP tasks. Other relevant but non-OCR dataset creation efforts include Guarani-Spanish news articles' (Góngora et al., 2021), Nahuatl speech translation (Shi et al., 2021), and Mazatec and Mixtec translations (Tonja et al., 2023), which can serve as valuable pretraining corpora for OCR.

7 Conclusion

In this paper, we have presented a concise survey of optical character recognition (OCR) techniques shown to be most applicable to low-resource languages in the OCR literature. The survey is focused and similar work has not been published before due to the small community of OCR researchers working with low-resource and Indigenous language communities. We also highlight undigitized datasets in 10 Central and South American Indigenous languages, mostly from the AILLA collection, that can be extremely valuable to digitize for downstream NLP applications. Based on our own experiences and on findings from our literature review, we conclude with recommendations on utilizing and improving OCR for the benefit of computational researchers, linguists, and language communities. We hope that our paper can be used as a starting point for researchers or language community members wishing to digitize their resources but unaware of what OCR adaptations have become absolutely necessary to move towards a high-quality OCR output as well as what the open challenges in the field are.

Limitations

We acknowledge that even with the page limit provided by a long paper, fitting all details even for a focused topic like ours is not possible. Where possible, we have included the most relevant details, including mathematical equations, figures, and tables, in order to keep the survey concise and relevant to the AmericasNLP community. In addition, experimental results for the 10 Indigenous languages we selected to anchor our survey are out of the scope of this paper and would easily warrant a separate study.

Ethics Statement

The raw data resources shared for the 10 selected Indigenous languages are entirely hosted by AILLA. The data is freely available to the general public, with some files shareable through request. The data can be used without asking for permission, and without paying any fees, as long as the resource and collection is cited appropriately. We acknowledge the linguists, native and heritage speakers, and the AILLA team for creating such a valuable repository of raw data in indigenous languages of Latin America. An ethical implication of this work is that it will allow for more sustainable and equitable work in language resource creation and natural language processing. However, we don't foresee any negative ethical concerns with our work, which hopes to encourage open-source development of OCR models to allow researchers to move away from relying on commercial systems to process low-resource and Indigenous language data.

Acknowledgments

This work was generously supported by the National Endowment for the Humanities under award PR-276810-21. The authors are also grateful to the anonymous reviewers for their valuable suggestions, feedback, and comments.

References

Ahmad Abdulkader and Mathew R. Casey. 2009. Low cost correction of ocr errors using learning in a multi-engine environment. In *2009 10th International Conference on Document Analysis and Recognition*, pages 576–580.

Eric Adell, Antonio Moisés Toma Cruz, and Edelmira Catarina Sánchez Toma. 2016. Kam nib'anax tu ma'l xhemaana (a brief description of a typical week). In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID: ailla:119533. Accessed March 21, 2024. IXIL-CTZ-DES-EST-2016-06-23-0507.

Reem Alaasam, Berat Kurar, and Jihad El-Sana. 2019. Layout analysis on challenging historical arabic manuscripts using siamese network. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 738–742. IEEE.

Emily M. Bender. 2011. On achieving and evaluating language-independence in nlp. In *Linguistic Issues in Language Technology*.

Taylor Berg-Kirkpatrick, Greg Durrett, and Dan Klein. 2013. Unsupervised transcription of historical documents. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 207–217, Sofia, Bulgaria. Association for Computational Linguistics.

Taylor Berg-Kirkpatrick, Greg Durrett, and Dan Klein. 2013. Unsupervised transcription of historical documents. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 207–217. The Association for Computer Linguistics.

Taylor Berg-Kirkpatrick and Dan Klein. 2014. Improved typesetting models for historical OCR. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 118–123, Baltimore, Maryland. Association for Computational Linguistics.

Tulio Bermúdez Mejía. 2015. Miskitu dance, food, and traditions: traditional miskitu food, dance, songs, festivities. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID ailla:119700. Accessed February 15, 2024. Other Contributors include Waldan Peter, Wanda Luz (Speaker), Bermúdez Mejía, Tulio (Transcriber), Waldan Peter, Wanda Luz (Translator).

Lukas Blecher, Guillem Cucurull, Thomas Scialom, and Robert Stojnic. 2023. Nougat: Neural optical understanding for academic documents.

Nadav Borenstein, Phillip Rust, Desmond Elliott, and Isabelle Augenstein. 2023. PHD: Pixel-based language modeling of historical documents. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 87–107, Singapore. Association for Computational Linguistics.

Emanuele Bugliarello, Ryan Cotterell, Naoaki Okazaki, and Desmond Elliott. 2021. Multimodal pretraining unmasked: A meta-analysis and a unified framework of vision-and-language BERTs. *Transactions of the Association for Computational Linguistics*, 9:978–994.

Gina Bustamante, Arturo Oncevay, and Roberto Zariquiey. 2020. No data to crawl? monolingual corpus creation from PDF files of truly low-resource languages in Peru. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2914–2923, Marseille, France. European Language Resources Association.

Trevor Cohn, Cong Duy Vu Hoang, Ekaterina Vymolova, Kaisheng Yao, Chris Dyer, and Gholamreza Haffari. 2016. Incorporating structural alignment biases into an attentional neural translation model. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 876–885, San Diego, California. Association for Computational Linguistics.

Johanna Cordova and Damien Nouvel. 2021. Toward creation of Ancash lexical resources from OCR. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 163–167, Online. Association for Computational Linguistics.

Hervé Déjean and Jean-Luc Meunier. 2019. Table rows segmentation. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 461–466. IEEE.

Eva D'hondt, Cyril Grouin, and Brigitte Grau. 2017. Generating a training corpus for OCR post-correction using encoder-decoder model. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1006–1014, Taipei, Taiwan. Asian Federation of Natural Language Processing.

Rui Dong and David Smith. 2018. Multi-input attention for unsupervised OCR correction. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2363–2372, Melbourne, Australia. Association for Computational Linguistics.

Senka Drobac, Pekka Kauppinen, and Krister Lindén. 2017. OCR and post-correction of historical Finnish texts. In *Proceedings of the 21st Nordic Conference on Computational Linguistics*, pages 70–76, Gothenburg, Sweden. Association for Computational Linguistics.

Pierre Déléage. 2002. Sharanahua language collection of pierre déléage. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public.Accessed February 15, 2024.

Nora England. 1972-1985. Mam language stories and grammars. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID ailla:119520, ailla:119520,ailla:119520,ailla:119520,ailla:119520. Accessed February 15, 2024.

Yasuhisa Fujii, Karel Driesen, Jonathan Baccash, Ash Hurst, and Ashok C. Popat. 2017. Sequence-to-label script identification for multilingual OCR. In *14th IAPR International Conference on Document Analysis and Recognition, ICDAR 2017, Kyoto, Japan, November 9-15, 2017*, pages 161–168. IEEE.

Dan Garrette and Hannah Alpert-Abrams. 2016. An unsupervised model of orthographic variation for historical document transcription. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 467–472, San Diego, California. Association for Computational Linguistics.

Ross B. Girshick. 2015. Fast R-CNN. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 1440–1448. IEEE Computer Society.

Santiago Góngora, Nicolás Giossa, and Luis Chiruzzo. 2021. Experiments on a Guarani corpus of news and social media. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 153–158, Online. Association for Computational Linguistics.

Jiatao Gu, Zhengdong Lu, Hang Li, and Victor O.K. Li. 2016. Incorporating copying mechanism in sequence-to-sequence learning. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1631–1640, Berlin, Germany. Association for Computational Linguistics.

Sanjana Gunna, Rohit Saluja, and C. V. Jawahar. 2021. Transfer learning for scene text recognition in indian languages. In *Document Analysis and Recognition, ICDAR 2021 Workshops, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part I*, volume 12916 of *Lecture Notes in Computer Science*, pages 182–197. Springer.

Anshul Gupta, Ricardo Gutierrez-Osuna, Matthew Christy, Richard Furuta, and Laura Mandell. 2016. Font identification in historical documents using active learning. *CoRR*, abs/1601.07252.

Harsh Gupta, Luciano Del Corro, Samuel Broscheit, Johannes Hoffart, and Eliot Brenner. 2021. Unsupervised multi-view post-OCR error correction with language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8647–8652, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kaiming He, Georgia Gkioxari, Piotr Dollár, and Ross B. Girshick. 2017. Mask R-CNN. In *IEEE International Conference on Computer Vision, ICCV 2017, Venice, Italy, October 22-29, 2017*, pages 2980–2988. IEEE Computer Society.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568, Online. Association for Computational Linguistics.

Nicholas Hopkins. 1964. A dictionary of the chuj (mayan) language language community. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID ailla:119647. Accessed February 15, 2024.

Benjamin Hunt, Lane Schwartz, Sylvia Schreiner, and Emily Chen. 2023. Community consultation and the development of an online akuzipik-English dictionary. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 134–143, Toronto, Canada. Association for Computational Linguistics.

Oana Ignat, Jean Maillard, Vishrav Chaudhary, and Francisco Guzmán. 2022a. OCR improves machine translation for low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1164–1174, Dublin, Ireland. Association for Computational Linguistics.

Oana Ignat, Jean Maillard, Vishrav Chaudhary, and Francisco Guzmán. 2022b. OCR improves machine translation for low-resource languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1164–1174, Dublin, Ireland. Association for Computational Linguistics.

José Carlos Aradillas Jaramillo, Juan José Murillo-Fuentes, and Pablo M. Olmos. 2018. Boosting handwriting text recognition in small databases with transfer learning. In *16th International Conference on Frontiers in Handwriting Recognition, ICFHR 2018, Niagara Falls, NY, USA, August 5-8, 2018*, pages 429–434. IEEE Computer Society.

Zhaohui Jiang, Zheng Huang, Yunrui Lian, Jie Guo, and Weidong Qiu. 2019. Integrating coordinates with context for information extraction in document images. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 363–368. IEEE.

Heidi Anna Johnson. 2000-2005. A grammar of san miguel chimalapa zoque. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID: ailla:119500 Accessed February 15, 2024.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 6282–6293. Association for Computational Linguistics.

Susan Kalt. 2016. Entrevista con tomas castro v y santusa quispe de flores. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID ailla:119707, ailla:119707 . Accessed February 15, 2024. Other Contributors include Waldan Peter, Wanda Luz (Speaker), Bermúdez Mejía, Tulio (Transcriber), Waldan Peter, Wanda Luz (Translator).

Terrence Kaufman. 1960-1993. Colección de idiomas mayenses de terrence kaufman. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID ailla:119707, ailla:119707 . Accessed February 15, 2024.

Umar Khan, Sohaib Zahid, Muhammad Asad Ali, Adnan Ul-Hasan, and Faisal Shafait. 2021. Tabaug: Data driven augmentation for enhanced table structure recognition. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part II*, volume 12822 of *Lecture Notes in Computer Science*, pages 585–601. Springer.

Shachar Klaiman and Marius Lehne. 2021. Docreader: Bounding-box free training of a document information extraction model. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part I*, volume 12821 of *Lecture Notes in Computer Science*, pages 451–465. Springer.

Okan Kolak and Philip Resnik. 2005. OCR post-processing for low density languages. In *HLT/EMNLP 2005, Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing, Proceedings of the Conference, 6-8 October 2005, Vancouver, British Columbia, Canada*, pages 867–874. The Association for Computational Linguistics.

Amrith Krishna, Bodhisattwa P. Majumder, Rajesh Bhat, and Pawan Goyal. 2018. Upcycle your OCR: Reusing OCRs for post-OCR text correction in Romanised Sanskrit. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 345–355, Brussels, Belgium. Association for Computational Linguistics.

Minghao Li, Tengchao Lv, Jingye Chen, Lei Cui, Yijuan Lu, Dinei A. F. Florêncio, Cha Zhang, Zhoujun Li, and Furu Wei. 2023. Trocr: Transformer-based optical character recognition with pre-trained models. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023, Thirty-Fifth Conference on Innovative Applications of Artificial Intelligence, IAAI 2023, Thirteenth Symposium on Educational Advances in Artificial Intelligence, EAAI 2023, Washington, DC, USA, February 7-14, 2023*, pages 13094–13102. AAAI Press.

Yu-Hsiang Lin, Chian-Yu Chen, Jean Lee, Zirui Li, Yuyan Zhang, Mengzhou Xia, Shruti Rijhwani, Junxian He, Zhisong Zhang, Xuezhe Ma, Antonios Anastasopoulos, Patrick Littell, and Graham Neubig. 2019. Choosing transfer languages for cross-lingual learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3125–3135, Florence, Italy. Association for Computational Linguistics.

Manfei Liu, Zecheng Xie, Yaoxiong Huang, Lianwen Jin, and Weiyin Zhou. 2018. Distilling gru with data augmentation for unconstrained handwritten text recognition. In *2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR)*, pages 56–61.

Pengyuan Lyu, Chengquan Zhang, Shanshan Liu, Meina Qiao, Yangliu Xu, Liang Wu, Kun Yao, Junyu Han, Errui Ding, and Jingdong Wang. 2022. Maskocr: Text recognition with masked encoder-decoder pretraining. *CoRR*, abs/2206.00311.

Michael Maxwell and Aric Bills. 2017. Endangered data for endangered languages: Digitizing print dictionaries. In *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 85–91, Honolulu. Association for Computational Linguistics.

Jamshed Memon, Maira Sami, Rizwan Ahmed Khan, and Mueen Uddin. 2020. Handwritten optical character recognition (ocr): A comprehensive systematic literature review (slr). *IEEE Access*, 8:142642–142668.

Haitao Mi, Baskaran Sankaran, Zhiguo Wang, and Abe Ittycheriah. 2016. Coverage embedding models for neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 955–960, Austin, Texas. Association for Computational Linguistics.

Claire Monteleoni and Matti Kaariainen. 2007. Practical online active learning for classification. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8.

Marcin Namysl and Iuliu Konya. 2019. Efficient, lexicon-free ocr using deep learning. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 295–301.

Andrew Naoum, Joel Nothman, and James R. Curran. 2019. Article segmentation in digitised newspapers with a 2d markov model. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 1007–1014. IEEE.

Clemens Neudecker, Konstantin Baierer, Mike Gerber, Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. 2021. A survey of ocr evaluation tools and metrics. In *Proceedings of the 6th International Workshop on Historical Document Imaging and Processing*, HIP '21, page 13–18, New York, NY, USA. Association for Computing Machinery.

Thi Tuyet Hai Nguyen, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. 2021. Survey of post-ocr processing approaches. *ACM Comput. Surv.*, 54(6).

Alberto Poncelas, Mohammad Aboomar, Jan Buts, James Hadley, and Andy Way. 2020. A tool for facilitating OCR postediting in historical documents. In *Proceedings of LT4HALA 2020 - 1st Workshop on Language Technologies for Historical and Ancient Languages*, pages 47–51, Marseille, France. European Language Resources Association (ELRA).

Animesh Prasad, Hervé Déjean, and Jean-Luc Meunier. 2019. Versatile layout understanding via conjugate graph. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 287–294. IEEE.

José Ramón Prieto and Enrique Vidal. 2021. Improved graph methods for table layout understanding. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part II*, volume 12822 of *Lecture Notes in Computer Science*, pages 507–522. Springer.

Paul Proulx. 1968. Chiquian quechua vocabulary. In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public.Accessed February 15, 2024.

Abhishek Prusty, Sowmya Aitha, Abhishek Trivedi, and Ravi Kiran Sarvadevabhatla. 2019. Indiscapes: Instance segmentation networks for layout parsing of historical indic manuscripts. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 999–1006. IEEE.

Christian Reul, Uwe Springmann, and Frank Puppe. 2017. LAREX - A semi-automatic open-source tool for layout analysis and region extraction on early printed books. *CoRR*, abs/1701.07396.

Christian Reul, Uwe Springmann, Christoph Wick, and Frank Puppe. 2018. Improving OCR accuracy on early printed books by combining pretraining, voting, and active learning. *J. Lang. Technol. Comput. Linguistics*, 33(1):3–24.

Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. 2020. OCR Post Correction for Endangered Language Texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5931–5942, Online. Association for Computational Linguistics.

Shruti Rijhwani, Daisy Rosenblum, Antonios Anastasopoulos, and Graham Neubig. 2021. Lexically aware semi-supervised learning for OCR post-correction. *Transactions of the Association for Computational Linguistics*, 9:1285–1302.

Shruti Rijhwani, Daisy Rosenblum, Michayla King, Antonios Anastasopoulos, and Graham Neubig. 2023. User-centric evaluation of OCR systems for kwak'wala. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 19–29, Remote. Association for Computational Linguistics.

Shruti Rijhwani, Jiateng Xie, Graham Neubig, and Jaime G. Carbonell. 2019. Zero-shot neural transfer for cross-lingual entity linking. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 6924–6931. AAAI Press.

Carsten Schnober, Steffen Eger, Erik-Lân Do Dinh, and Iryna Gurevych. 2016. Still not there? comparing traditional sequence-to-sequence models to encoder-decoder neural networks on monotone string translation tasks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 1703–1714, Osaka, Japan. The COLING 2016 Organizing Committee.

Sarah Schulz and Jonas Kuhn. 2017. Multi-modular domain-tailored OCR post-correction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2716–2726, Copenhagen, Denmark. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Burr Settles. 2012. *Active Learning*. Synthesis Lectures on Artificial Intelligence and Machine Learning. Morgan & Claypool Publishers.

Prema Satish Sharan, Sowmya Aitha, Amandeep Kumar, Abhishek Trivedi, Aaron Augustine, and Ravi Kiran Sarvadevabhatla. 2021. Palmira: A deep deformable network for instance segmentation of dense and uneven layouts in handwritten manuscripts. In *16th International Conference on Document Analysis and Recognition, ICDAR 2021, Lausanne, Switzerland, September 5-10, 2021, Proceedings, Part II*, volume 12822 of *Lecture Notes in Computer Science*, pages 477–491. Springer.

Zejiang Shen, Weining Li, Jian Zhao, Yaoliang Yu, and Melissa Dell. 2022. OLALA: Object-level active learning for efficient document layout annotation. In *Proceedings of the Fifth Workshop on Natural Language Processing and Computational Social Science (NLP+CSS)*, pages 170–182, Abu Dhabi, UAE. Association for Computational Linguistics.

Jiatong Shi, Jonathan D. Amith, Xuankai Chang, Siddharth Dalmia, Brian Yan, and Shinji Watanabe. 2021. Highland Puebla Nahuatl speech translation corpus for endangered language documentation. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 53–63, Online. Association for Computational Linguistics.

Connor Shorten and Taghi M. Khoshgoftaar. 2019. A survey on image data augmentation for deep learning. *J. Big Data*, 6:60.

R. Smith. 2007. An overview of the tesseract OCR engine. In *9th International Conference on Document Analysis and Recognition (ICDAR 2007), 23-26 September, Curitiba, Paraná, Brazil*, pages 629–633. IEEE Computer Society.

Nikita Srivatsan, Si Wu, Jonathan Barron, and Taylor Berg-Kirkpatrick. 2021. Scalable font reconstruction with dual latent manifolds. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 3060–3072, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Victor Storchan and Jocelyn Beauschene. 2019. Data augmentation via adversarial networks for optical character recognition/conference submissions. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 184–189.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems 27: Annual Conference on Neural Information Processing Systems 2014, December 8-13 2014, Montreal, Quebec, Canada*, pages 3104–3112.

Ahmad Pahlavan Tafti, Ahmadreza Baghaie, Mehdi Assefi, Hamid R. Arabnia, Zeyun Yu, and Peggy L. Peissig. 2016. OCR as a service: An experimental evaluation of google docs ocr, tesseract, ABBYY finereader, and transym. In *Advances in Visual Computing - 12th International Symposium, ISVC 2016, Las Vegas, NV, USA, December 12-14, 2016, Proceedings, Part I*, volume 10072 of *Lecture Notes in Computer Science*, pages 735–746. Springer.

Lindia Tjuatja, Shruti Rijhwani, and Graham Neubig. 2021. Explorations in transfer learning for ocr post-correction. In *Fifth Widening Natural Language Processing Workshop (WiNLP)*, volume 6.

Maxim Tkachenko, Mikhail Malyuk, Andrey Holmanyuk, and Nikolai Liubimov. 2020-2022. Label Studio: Data labeling software. Open source software available from https://github.com/heartexlabs/label-studio.

Konstantin Todorov and Giovanni Colavizza. 2020. Transfer learning for historical corpora: An assessment on post-ocr correction and named entity recognition. In *Proceedings of the Workshop on Computational Humanities Research (CHR 2020), Amsterdam, The Netherlands, November 18-20, 2020*, volume 2723 of *CEUR Workshop Proceedings*, pages 310–339. CEUR-WS.org.

Atnafu Lambebo Tonja, Christian Maldonado-sifuentes, David Alejandro Mendoza Castillo, Olga Kolesnikova, Noé Castro-Sánchez, Grigori Sidorov, and Alexander Gelbukh. 2023. Parallel corpus for indigenous language translation: Spanish-mazatec and Spanish-Mixtec. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 94–102, Toronto, Canada. Association for Computational Linguistics.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 76–85, Berlin, Germany. Association for Computational Linguistics.

Nikolai Vogler, Jonathan Allen, Matthew Miller, and Taylor Berg-Kirkpatrick. 2022. Lacuna reconstruction: Self-supervised pre-training for low-resource historical document transcription. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 206–216, Seattle, United States. Association for Computational Linguistics.

Robin M. Wright, Manuel da Silva, and José Felipe Aguiar. 2000. Baniwa history: Uapui cachoeira, aiary river (1970s - 2000). In *The Archive of the Indigenous Languages of Latin America, ailla.utexas.org*. Access: public. PID: ailla:119657 Accessed March 21, 2024.

Ruoyu Xie and Antonios Anastasopoulos. 2023. Noisy parallel data alignment. In *Findings of the Association for Computational Linguistics: EACL 2023*, pages 1501–1513, Dubrovnik, Croatia. Association for Computational Linguistics.

Xu Zhong, Jianbin Tang, and Antonio Jimeno-Yepes. 2019. Publaynet: Largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition, ICDAR 2019, Sydney, Australia, September 20-25, 2019*, pages 1015–1022. IEEE.

Fuzhen Zhuang, Zhiyuan Qi, Keyu Duan, Dongbo Xi, Yongchun Zhu, Hengshu Zhu, Hui Xiong, and Qing He. 2021. A comprehensive survey on transfer learning. *Proc. IEEE*, 109(1):43–76.

Unlocking Knowledge with OCR-Driven Document Digitization for Peruvian Indigenous Languages

Shadya Sánchez Roberto Zariquiey Arturo Oncevay
Chana Research Group, Pontificia Universidad Católica del Perú, Perú
{shadya.sanchez,rzariquiey,arturo.oncevay}@pucp.edu.pe

Abstract

The current focus on resource-rich languages poses a challenge to linguistic diversity, affecting minority languages with limited digital presence and relatively old published and unpublished resources. In addressing this issue, this study targets the digitalization of old scanned textbooks written in four Peruvian indigenous languages (Asháninka, Shipibo-Konibo, Yanesha, and Yine) using Optical Character Recognition (OCR) technology. This is complemented with text correction methods to minimize extraction errors. Contributions include the creation of an annotated dataset with 454 scanned page images, for a rigorous evaluation, and the development of a module to correct OCR-generated transcription alignments.

1 Introduction

Natural Language Processing (NLP) has prompted the development of diverse language technologies, including machine translation, spell checkers, and information extraction tools. Given this impact, there is an urgent need to democratize these technologies, making them available for speakers of the more than 7,000 languages spoken worldwide.

Currently, such technologies are restricted to languages with ample linguistic resources that are easily exploitable (Ataa Allah et al., 2023). This presents a challenge for minority languages due to their limited digital presence and the prevalence of their resources in less accessible formats, hindering their incorporation into the development of these technologies (Bustamante et al., 2020). Consequently, speakers of minority languages are forced to adopt languages with greater technological access, leading to a loss of cultural, historical, and linguistic knowledge.

To address this situation, multiple efforts are underway to diversify these technologies to minority languages and their speakers, who face the challenge of overcoming data availability limitations.

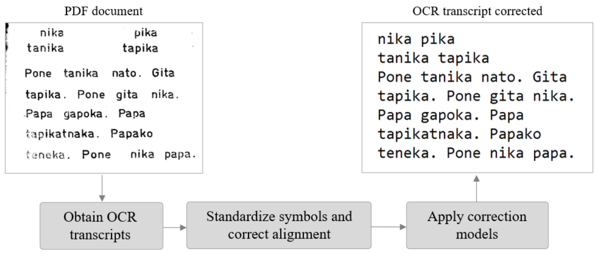


Figure 1: OCR process

In some cases, synthetic data has been generated (Oncevay et al., 2022), translations to languages with more resources have been utilized (Ko et al., 2021; Rijhwani et al., 2020), and technologies have been adapted for dataset extraction and processing (Bustamante et al., 2020).

However, identifying digital sources of knowledge for endangered languages is a very challenging task, as they are not usually available on the web (Bustamante et al., 2020), which is the case for several indigenous languages of Peru. In this context, Optical Character Recognition (OCR) models have been useful to extract information and new resources for endangered language texts (Rijhwani et al., 2020, 2021).

For this reason, we extend the application of OCR for digitizing old documents, with typewritten texts, in four Peruvian languages (Asháninka, Shipibo-Konibo, Yanesha, and Yine), using Optical Character Recognition (OCR), and followed by correction methods to minimize extraction errors (see Figure 1).

2 Language context

According to official statistics, 48 languages are spoken in Peru. 44 out of these 48 languages are Amazonian languages (de Educación del Peru, 2013). The four languages this paper is focused on (Shipibo-Konibo, Asháninka, Yanesha, and Yine) are Amazonian languages.

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 103–111
June 21, 2024

Asháninka, belonging to the Nijagantsi branch of the Arawak language family, is primarily spoken in the central Peruvian Amazonia, along the Low Perené, Tambo, Ene, Urubamba, and Apurímac rivers (Pedrós, 2018). Although the Asháninka population is estimated to be around 70,000 speakers (Pedrós, 2018), it remains unclear if this count includes speakers of Ashéninka, a closely related language.

Yine and Yanesha are also languages of the Arawak family. Yine is spoken by approximately 3,000 people living near the Ucayali and Madre de Dios rivers. Yanesha, in turn, is spoken by 1,142 people in the Peruvian department of Pasco. Yanesha people generally express concern for their language, since very few children speak it and speakers are mostly over 30 years old. Both Yine and Yanesha are classified as "definitely endangered" according to the UNESCO Atlas of the World's Languages in Danger (Moseley, 2010).

With an estimated 40,000 speakers, Shipibo-Konibo is by large the most vital language in the Pano language family. It is predominantly spoken in the Peruvian regions of Ucayali and Loreto, along the Ucayali river and its tributaries (Valenzuela, 2003). It is important to mention that there is a relatively large Shipibo-Konibo community in Lima.

3 Related work

The correction of OCR transcripts has seen the application of various methodologies, ranging from manual and resource-intensive approaches to more recent and prevalent machine learning models, particularly those based on neural networks (see Nguyen et al. (2021) for further details). The effectiveness of applications such as language models, translation models, and spell checkers in rectifying OCR errors is well-established. For instance, Afli et al. (2016) employed a statistical machine translation (SMT) model, while Schulz and Kuhn (2017) combined such models with spell checkers.

Furthermore, sequence-to-sequence neural networks have emerged as successful models in correcting OCR transcripts, especially in scenarios with limited data availability. Rijhwani et al. (2020) developed a model that effectively learned from limited data for languages like Ainu, Griko, and Yakkha by leveraging existing translations. This approach was further enhanced in Rijhwani et al. (2021) through the incorporation of lexical decoding and self-training strategies, achieving significant improvements (up to 29%). For Sanskrit texts, Maheshwari et al. (2022) obtained favorable results by considering both phonetic encoding and the language's official alphabet.

4 Methodology for dataset creation

4.1 Data selection

We sourced documents from the SIL International[1] repository, targeting materials written in four languages: Asháninka, Shipibo-Konibo, Yanesha and Yine. These languages were chosen for their availability of resources within the repository compared to other Peruvian languages [2].

The documents, primarily in PDF format, present a wide range of contents, attributes, and layouts, including typewritten and handwritten text, tables, and images. Additionally, the content may be organized in multiple columns and vary in font sizes, sometimes presented in multiple languages.

To ensure dataset consistency, we focused on a subset of monolingual, typewritten documents with uniform font sizes. From each document, we selected a sample of pages (10%) for annotation and evaluation, based on criteria such as readability, resolution, tilt, and content alignment.

4.2 Data annotation

We manually annotated the documents following the workflow depicted in Figure 2. This process involved two key roles: an annotator and a reviewer. The annotator, possessing prior annotation experience, utilized a free online OCR tool[3] to generate a preliminary transcription. This initial step facilitated the annotation process, minimizing the time, effort, and potential errors associated with manual transcription.

Subsequently, the preliminary transcription was rectified by the annotator and reviewed by another annotator, who double-checked the annotations and addressed any discrepancies or errors encountered.

4.3 Data preprocessing

Although most document pages generally exhibit good quality, certain defects, such as ink stains overlapping with characters and low scanning resolution, can significantly affect fine text details,

[1] SIL International: https://www.sil.org/

[2] See Figure 4 in Appendix B for details about the data availability.

[3] Free online OCR tool: https://www.onlineocr.net/es/

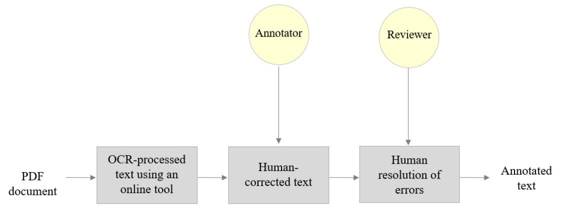


Figure 2: Annotation Process.

degrading page quality, and potentially impacting OCR performance. To address these issues, we conducted preprocessing steps including noise and attribute removal, as well as image enhancement, aimed at improving OCR accuracy.

Noise removal We manually cleaned documents using the Nitro PDF[4] tool to eliminate elements adding noise, such as images and page numbers.

Attribute removal Text delimited by boxes can impact text recognition (see Figure 5 in the Appendix). Consequently, we categorized pages into two groups: those containing only text (Group 1) and those with text delimited by boxes (Group 2).

For Group 2 pages, we applied image correction to automatically detect and remove the boxes when needed (for one of the tools we experimented with, it did not provide any benefit).

Image enhancement After converting the pages to Portable Network Graphics (PNG) format, we applied corrections focused on removing irregularities and improving the contour of the characters to achieve more effective segmentation.

Language	NoP	NoS	NoT	NoUT	NoTP
Asháninka	134	2239	8103	2309	61
Shipibo-Konibo	89	1495	7251	1685	82
Yanesha	91	1468	6315	1574	70
Yine	140	2246	9754	2449	70

Table 1: Corpora description: NoP = Number of pages, NoS = Number of sentences, NoT = Number of tokens, NoUT = Number of unique tokens, NoTP = Number of tokens per page.

4.4 Dataset description

The resulting dataset comprises 454 scanned pages from 89 books written in the four indigenous languages: Asháninka, Shipibo-Konibo, Yanesha, and Yine. This dataset[5] comprises 31,423 tokens, distributed almost equally across the languages (see Table 1). Significantly, compared to the dataset previously generated by Bustamante et al. (2020), our work expands the vocabulary by incorporating an average of 3,110 unique tokens per language.

Beyond standard alphanumeric characters, the dataset includes digits (0-9), diacritics, punctuation marks, and various compound characters like **m̃**, **c̈**, and **t̃**. Approximately 36% of the characters appear fewer than 10 times. Moreover, nearly 45% of the employed characters deviate from the contemporary official alphabets of these languages.[6] This issue arises from the fact that the analyzed documents were written before the establishment of the official alphabets for these languages.

5 OCR Process

We employed two OCR systems, Google Vision[7] (version 3.4.4) and Tesseract[8] (version 5.3.3), to generate initial text transcriptions. Although neither system directly supports the languages studied, they recognize the common Latin script shared by these languages. Previous research has demonstrated their effectiveness in low-resource language settings, including Sanskrit (Maheshwari et al., 2022), Ainu, Griko, Yakka (Rijhwani et al., 2020), Tamil, and Sinhala (Vasantharajan et al., 2022).

After the initial OCR transcriptions, a two-step preprocessing stage was implemented to enhance output quality. We evaluated the results using two standard metrics: Character Error Rate (CER) and Word Error Rate (WER). These metrics quantify OCR accuracy based on the Levenshtein distance, which measures the minimum number of edit operations (substitutions, deletions, insertions) required to transform the original text into the OCR-generated text (Neudecker et al., 2021).

5.1 Preprocessing of initial OCR transcripts

Both OCR systems faced challenges in differentiating between various forms of similar punctuation marks, such as hyphens (-, —, _) and single quotation marks (', ’). To address this ambiguity, we standardized analogous punctuation marks and converted all text to lowercase while preserving the original content.

[4] Nitro PDF: https://www.gonitro.com/

[5] The dataset is available at https://github.com/iapucp/OCR-Peru

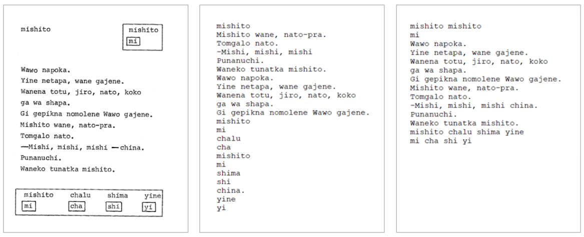
[6] The complete list of characters are included in Table 3 in the Appendix.

[7] Google Vision OCR: https://cloud.google.com/vision/docs/ocr

[8] Tesseract OCR: https://tesseract-ocr.github.io/

5.2 Alignment correction

Additionally, we observed that Google Vision OCR recognizes texts but fails to maintain the correct order, particularly affecting the text recognition of Group 2 pages, as depicted in Figure 3. To address this challenge, we developed a module to automatically align the initial transcriptions based on their vertical and horizontal positions, resulting in a reduction of approximately 9% in CER and 12% in WER. This enhancement was unnecessary for Tesseract OCR transcripts, as it effectively detects text order using text block segmentation.



(a) Scanned document

(b) Google Vision OCR transcription

(c) Google Vision OCR transcription aligned using the developed automatic alignment module

Figure 3: Alignment of Google Vision OCR transcriptions

5.3 Types of errors

We identified three main types of errors in the OCR predictions:

Misprediction of characters with diacritics Characters such as b̃, c̈, m̃, p̃, t̃, and s̈ primarily found in ancient texts, were frequently misrecognized. This likely stems from their relative scarcity in modern Latin-script training data used in the OCR model training. This limited exposure led to inefficient recognition, contributing to approximately 60% of OCR errors, particularly in Shipibo-Konibo and Yanesha languages.

Insertion of non-existent characters Both OCR engines introduced orthographically similar characters not present in the original texts. Tesseract was more prone to this error (2.7 times more frequently), adding an average of 65 additional characters compared to Google Vision's 24. Tesseract showed repetitive patterns in adding characters, combining similar ones like **cç** and **ií**, often at sentence boundaries. Additionally, it misrecognizes small stains as characters. In contrast, Google Vision demonstrated better stain filtering but tended to replace similar characters like **š** with **ŝ**. This error type represented approximately 12% of the errors made by Google Vision OCR and 20% by Tesseract.

Incorrect word boundary detection Predominantly observed in Google Vision OCR, this involved adding extra spaces between words. It accounted for 47% of text identification errors in Asháninka and Yine languages but only 8% in Shipibo-Konibo and Yanesha languages.

6 Post-OCR process

6.1 Correction models

We applied five post-OCR methods to correct the errors made by the OCR systems:

SingleSource (Rijhwani et al., 2020) A sequence-to-sequence model tailored to effectively learn from limited data. We employed the single-source model.

Denorm (Oncevay et al., 2022) A spell checker trained to correct misspelling errors in Asháninka, Shipibo-Konibo, Yanesha, and Yine languages, normalizing sentences according to each language's grammar and norms.

Ensemble (Oncevay et al., 2022) An ensemble spell checker addressing five types of errors: character replacements, insertions, or deletions; errors from using a QWERTY keyboard; errors due to syllable similarity or ambiguity between phonemes and graphemes; and characters not included in the standardized alphabets of the languages.

SingleSource+Denorm A cascaded approach applying the SingleSource model followed by the Denorm model.

SingleSource+Ensemble A cascaded approach applying the SingleSource model followed by the Ensemble model.

6.2 Model training

Since only the model proposed by Rijhwani et al. (2020) required training, we trained it for each language using the basic hyperparameters configuration suggested. We employed five different random initializations on a system with 45 GB of RAM and 8 CPUs, and it required a total of 98 hours to complete. Subsequently, we evaluated all models and tools using the annotated test set.

OCR	Model	CER				WER			
		Asháninka	Shipibo-Konibo	Yanesha	Yine	Asháninka	Shipibo-Konibo	Yanesha	Yine
Tesseract	Baseline OCR	1.65	4.30	8.11	1.61	13.42	19.74	41.89	9.53
	SingleSource	1.34	1.55	3.75	**1.12**	9.85	8.21	20.66	**6.83**
	Denorm	7.32	15.83	11.32	3.68	35.84	50.33	52.28	15.25
	Ensemble	4.92	11.85	9.91	4.49	29.35	44.5	48.63	18.77
	SingleSource + Denorm	6.95	14.52	8.31	2.87	32.74	42.91	43.17	12.46
	SingleSource + Ensemble	4.2	9.48	6.85	3.74	25.81	34.17	38.8	16.28
Google Vision	Baseline OCR	**0.76**	2.61	5.53	1.49	9.00	12.98	39.16	10.26
	SingleSource	0.92	**0.88**	**2.32**	1.45	**8.61**	**4.85**	**17.41**	7.74
	Denorm	6.86	14.61	10.01	3.45	33.19	44.77	54.1	15.69
	Ensemble	4.04	10.55	8.09	4.47	25.52	39.07	50.46	20.09
	SingleSource + Denorm	6.56	13.63	7.09	3.02	31.42	40.13	41.89	12.76
	SingleSource + Ensemble	3.83	8.75	5.66	4.01	24.48	31.52	37.89	17.16

Table 2: Results of applying the correction methods to the transcripts of the Tesseract and Google Vision OCRs

7 Results

Table 2 presents the results of applying correction methods to OCR transcripts. The SingleSource model proved most effective in rectifying OCR errors due to several factors. Firstly, pre-training the model with the languages' characters facilitated the removal of non-existent characters from the transcripts. Secondly, it reduced errors from incorrect word boundary identification by 33% in Asháninka and Yine languages. Lastly, it significantly enhanced the recognition of characters with diacritics by 35% for Shipibo-Konibo and 65% for Yanesha, achieving a 99% accuracy in identifying these characters.

Regarding errors introduced by this model, they primarily involved character deletion but were significantly fewer compared to successfully corrected words. The ratio of successfully corrected words to unsuccessfully corrected words was 5:1 for Shipibo-Konibo, 4:1 for Yanesha, and 2:1 for Yine. However, in the case of Asháninka, this ratio shifted to 2:1 only for the Tesseract OCR-generated transcripts, but reversed to 3:5 for Google Vision OCR-generated transcripts. This reversal led to a higher number of degraded words than enhanced ones, evidenced by a 0.16 increase in CER attributed to minimal errors in OCR transcription that the correction model cannot rectify. An important consideration of this model arises when the text contains low-frequency characters. Uneven distribution during dataset partitioning may result in some characters being absent from training but present in evaluation sets, impacting performance.

On the other hand, the correction models based on spell-checkers approached OCR transcription errors by standardizing the texts. Given that most of these texts were old and did not adhere to the official alphabet and language rules, this method was ineffective, resulting in more errors introduced by the model than words successfully corrected. These errors primarily consisted of omitted diacritics and character replacements aimed at conforming the text to standardization norms. More analysis about the standardization is discussed in Appendix A.

8 Conclusions and future work

This work digitized textbooks in four Peruvian languages using OCR systems. We contributed an annotated dataset to assess the performance of Google Vision and Tesseract OCRs. Google Vision demonstrated higher accuracy in character recognition, while Tesseract excelled in maintaining text order across multiple columns. To address Google Vision's limitation in maintaining text order, we developed an alignment module. Additionally, we evaluated five error correction methods and found that the SingleSource model, designed for learning from limited data, was the most effective, particularly in correcting characters with diacritics.

Future efforts aim to optimize the hyperparameters of the SingleSource model and implement the multi-source model by Rijhwani et al. (2020) to leverage Spanish translations available for 89% of the books.

References

Haithem Afli, Loïc Barrault, and Holger Schwenk. 2016. Ocr error correction using statistical machine translation. *Int. J. Comput. Linguistics Appl.*, 7:175–191.

Michael Arrigo, Stephanie Strassel, Nolan King, Thao Tran, and Lisa Mason. 2022. CAMIO: A corpus for OCR in multiple languages. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 1209–1216, Marseille, France. European Language Resources Association.

Fadoua Ataa Allah, Siham Boulaknadel, and Seth Darren. 2023. New trends in less-resourced language processing: Case of amazigh language. 12.

Gina Bustamante, Arturo Oncevay, and Roberto Zariquiey. 2020. No data to crawl? monolingual corpus creation from PDF files of truly low-resource languages in Peru. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2914–2923, Marseille, France. European Language Resources Association.

Ministerio de Educación del Peru. 2013. *Documento nacional de lenguas originarias del Perú*. Lima.

Wei-Jen Ko, Ahmed El-Kishky, Adithya Renduchintala, Vishrav Chaudhary, Naman Goyal, Francisco Guzmán, Pascale Fung, Philipp Koehn, and Mona Diab. 2021. Adapting high-resource nmt models to translate low-resource related languages without parallel data.

Ayush Maheshwari, Nikhil Singh, Amrith Krishna, and Ganesh Ramakrishnan. 2022. A benchmark and dataset for post-ocr text correction in sanskrit.

Christopher Moseley, editor. 2010. *Atlas of the world's languages in danger*, 3rd edition. UNESCO, Paris, France.

Clemens Neudecker, Konstantin Baierer, Mike Gerber, Christian Clausner, Apostolos Antonacopoulos, and Stefan Pletschacher. 2021. A survey of ocr evaluation tools and metrics. In *Proceedings of the 6th International Workshop on Historical Document Imaging and Processing*, HIP '21, page 13–18, New York, NY, USA. Association for Computing Machinery.

Thi Tuyet Hai Nguyen, Adam Jatowt, Mickael Coustaty, and Antoine Doucet. 2021. Survey of post-ocr processing approaches. *ACM Comput. Surv.*, 54(6).

Arturo Oncevay, Gerardo Cardoso, Carlo Alva, César Lara Ávila, Jovita Vásquez Balarezo, Saúl Escobar Rodríguez, Delio Siticonatzi Camaiteri, Esaú Zumaeta Rojas, Didier López Francis, Juan López Bautista, Nimia Acho Rios, Remigio Zapata Cesareo, Héctor Erasmo Gómez Montoya, and Roberto Zariquiey. 2022. SchAman: Spell-checking resources and benchmark for endangered languages from amazonia. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 411–417, Online only. Association for Computational Linguistics.

Toni Pedrós. 2018. Ashéninka y asháninka: ¿de cuántas lenguas hablamos? *Cadernos de Etnolingüística*, 6(1):1–30.

Shruti Rijhwani, Antonios Anastasopoulos, and Graham Neubig. 2020. OCR Post Correction for Endangered Language Texts. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5931–5942, Online. Association for Computational Linguistics.

Shruti Rijhwani, Daisy Rosenblum, Antonios Anastasopoulos, and Graham Neubig. 2021. Lexically aware semi-supervised learning for ocr post-correction.

Sarah Schulz and Jonas Kuhn. 2017. Multi-modular domain-tailored OCR post-correction. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2716–2726, Copenhagen, Denmark. Association for Computational Linguistics.

Pilar M. Valenzuela. 2003. *Transitivity in Shipibo-Konibo Grammar*. Ph.D. thesis, Eugene: University of Oregon, Eugene, OR.

Charangan Vasantharajan, Laksika Tharmalingam, and Uthayasanker Thayasivam. 2022. Adapting the tesseract open-source ocr engine for tamil and sinhala legacy fonts and creating a parallel corpus for tamil-sinhala-english. In *2022 International Conference on Asian Language Processing (IALP)*. IEEE.

A Spell checker assessment with standardized texts

Due to the spell checker's limitations in correcting OCR transcripts using the standardization approach, we assessed a small set of 50 sentences. This evaluation compared the spell checker's corrections with the original texts in their standardized versions. We manually standardized 25 sentences in both Yanesha and Shipibo-Konibo languages with the support of native speakers to ensure accuracy. Despite this fair comparison to the standardized texts, we noted no improvement in the CER and WER values. Moreover, opportunities for enhancement remain in both the Denorm and Ensemble models proposed in Oncevay et al. (2022).

B Dataset additional information

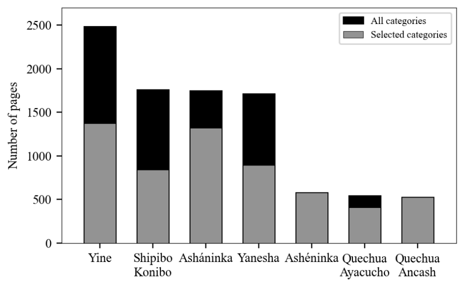
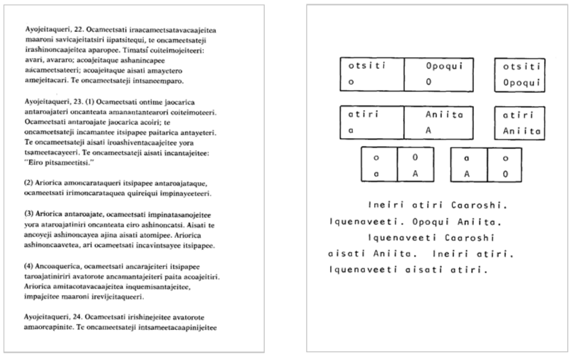


Figure 4: Number of pages available in the SIL repository for documents written in Peruvian languages. The selected categories exclude handwritten and unconstrained size texts.



(a) Group 1: Only text

(b) Group 2: Text delimited by boxes

Figure 5: Examples of pages from Group 1 and Group 2. Group 1 consists of text-heavy documents, whereas Group 2 presents either the entire text or portions of the text within tables.

(a) Multi-column and multilingual

(b) Text-heavy documents

(c) Handwritten text

(d) Unconstrained size text

(e) Text with images

(f) Text within tables and guidelines

Figure 6: Examples of document pages considering the classification made by Arrigo et al. (2022).

Language	Official characters	Unofficial characters	Punctuation marks	Digits
Asháninka	a, b, ch, e, i, j, k, m, n, ñ, o, p, r, s, sh, t, ts, y	á, c, d, é, f, g, í, l, ó, q, u, ú, v, x, z	!, ", (,), ,, -, ., /, :, ;, ?, —,¡, ¿	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
Shipibo-Konibo	a, b, ch, e, i, j, k, m, n, o, p, r, s, sh, t, ts, y	á, c, d, é, e , f, g, h, í, l, ñ, ó, q, š, u, ú, v, z	!, ", (,), ,, -, ., :, ;, =, ?, ¡, ¿, —	0, 1, 2, 3, 4, 5, 6, 7, 8, 9
Yanesha	a, b, ch, e, ë, g, j, k, ll, m, n, ñ, o, p, r, rr, s, sh, t, ts, y	á, ä, b̃, c, c̈, d, f, h, i, í, l, m̃, ó, p̃, q, t̃, u, ú, v, z	", ', (,), ,, -, ., /, :, ?, ¿, —, '	0, 1, 2, 3, 4, 7, 8, 9
Yine	a, ch, e, g, i, j, k, l, m, n, o, p, r, s, sh, t, ts, u, w, x, y	á, b, c, d, é, f, í, ú, ü, v	", (,), *, ,, -, ., /, :, ;, ?, ¿, —, _, !, ¡	0, 1, 2, 3, 4, 5, 6, 7, 8, 9

Table 3: Characters present in the documents. Official characters: Belonging to the official alphabet of the language. Unofficial characters: Not belonging to the official alphabet of the language.

C Resource collection

Title	Language
Shitsa pajitachari anquilostoma aisati ameba	Asháninka
Campa 2	Asháninka
Ocantacota nonampi	Asháninka
Nantayetiri nonampiqui aisati noquemayetiri	Asháninka
Shiquiri	Asháninka
Naturaleza y vida social 1, 2	Asháninka
Naturaleza y vida social 1, 2	Asháninka
Quenquetsarentsi	Asháninka
Ina	Asháninka
Timayetatsiri quipatsiqui	Asháninka
Ompiquiri 6	Asháninka
Tsame aneanatacoteri Caaroshi (Vamos a leer sobre Carlos: Libro 3 para la lectura y escritura)	Asháninka
Tuberculosis (Libro de ciencias naturales 6)	Asháninka
Timatsi cameetsatatsiri acoajeitaqueri: Te oncameetsateji intsaneemparo	Asháninka
Jaoca ancantajeari antecatsijeitantajeari maaroni	Asháninka
Jaoca icanteetirori aamaacoventearo ajipee	Asháninka
Icantacota peeraniniri	Asháninka
Avatsa (El cuerpo humano: Libro de ciencias naturales 2)	Asháninka
Icantacota Shintsia	Asháninka
Jaoca acanteriri ameneri cameetsa vaca	Asháninka
Ameneri cameetsa aparoni jananequi	Asháninka
Tsame aneanatacoteri ompiquiri (Vamos a leer acerca de Ompíquiri)	Asháninka
Campa 3	Asháninka
Gigkanni Pirana	Yine
Gitaklu pirana ga wa prachi	Yine
Giyoliklu pirana	Yine
Gwacha Ginkakle	Yine
Jitomta 3	Yine
Lima pirana	Yine
Mgenoklumta	Yine
Muchikawa kewenni pirana ga wa pimri ginkaklukaka 10	Yine
Naturaleza y vida social 1: Manual para los cursos de naturaleza y vida social, y práctica de Castellano, para primer año	Yine
Naturaleza y vida social 2	Yine
Naturaleza y vida social 3	Yine
Nopra kina 2	Yine
Papa nikchi gijga	Yine
Papa-mta 1	Yine
Papisho 5	Yine
Pejri-mta 4	Yine
Walo-mta 7, 8	Yine
Yine ginkaklekaka 12 (Cartilla 12)	Yine
Yine sana kamruta	Yine
Yineru tokanu 3a	Yine
Yineru tokanu 3b	Yine
Yineru tokanu IX	Yine

Title	Language
Cuentos de la zorra y el zorro	Shipibo-Konibo
Ëa- tapaman caní 5	Shipibo-Konibo
Japari peoquin yoyo ati quirica	Shipibo-Konibo
Jascaaquin baqueshocobo coiranhanan, jahuequiamati yoii ica	Shipibo-Konibo
Jatibiainoa joni coshibaon, jascaas̈hon jacon jahuequi aresti jonibaon jahuequescamabi itiaquin shinana	Shipibo-Konibo
Moa peoquin yoyo ati: Quirica 4	Shipibo-Konibo
Moa peoquin yoyo ati: Quirica 5	Shipibo-Konibo
Moa peoquin yoyo ati: Quirica 6	Shipibo-Konibo
Nanbonyabi nacan, noa isin meniai yoia	Shipibo-Konibo
Naturaleza y vida social 2	Shipibo-Konibo
Naturaleza y vida social 3	Shipibo-Konibo
Non paron ja jahuequibo 1 (Nuestros recursos naturales: Guía didáctica 1 de ciencias naturales)	Shipibo-Konibo
Quimisha Incabo ini yoia (Leyendas de los shipibo-conibo sobre los tres Incas)	Shipibo-Konibo
Quirica 10 (Afianzamiento de lectura 10: Animales del mundo)	Shipibo-Konibo
Quirica 4	Shipibo-Konibo
Quirica 5	Shipibo-Konibo
Quirica 6	Shipibo-Konibo
Quirica 7	Shipibo-Konibo
Quirica 8	Shipibo-Konibo
Quirica 9 (Libro 9: Afianzamiento para la lectura)	Shipibo-Konibo
Ach	Yanesha
Ahuat̃ serraparñats at̃o eñalleta atsne'ñam̃a arrorr	Yanesha
Amuesha 7 - SHAñE'	Yanesha
Apa ñam̃a ach (Papá y mamá: Libro 3 para la lectura y escritura)	Yanesha
Atet̃cha'yecue'cheshat̃oll	Yanesha
At̃o'yepotamperra Meshtaso ñam̃a po'poñ serrparñats	Yanesha
Berroc̈hno ñeñt̃ Africo'marnesha'	Yanesha
Cartilla 9 (Besllom̃)	Yanesha
Chom - Amuesha 8	Yanesha
Homenaje a la Declaración Universal de Derechos Humanos en su 40 aniversario 1948-1988	Yanesha
Ma’yarr poyoc̈her ñam̃a po’poñec̈hno serrparñats	Yanesha
Manual de ganadería	Yanesha
Naturaleza y vida social 3	Yanesha
Nochcar (Mi perro: Libro 6 para la lectura y escritura)	Yanesha
Ot̃ec̈hno	Yanesha
Pa'namen alloch yechopene'champesyen	Yanesha
Pa'namen atsnañtsoc̈hno	Yanesha
Pepe	Yanesha
Pepe payara	Yanesha
Pepe ñam̃a ema'(Pepe y la niñita: Libro 4 para la lectura y escritura)	Yanesha
Posho'll (La ardilla: Libro 7 para la lectura y escritura)	Yanesha
Tempo pueserrpareñ	Yanesha
YANESHA'	Yanesha
Yehuom̃cheña	Yanesha

Table 4: Resources utilized from the SIL repository.

Awajun-OP: Multi-domain Dataset for Spanish–Awajun Machine Translation

Oscar Moreno Veliz[†] **Yanua Liseth Atamain Uwarai**[‡] **Arturo Oncevay**[†]
[†]Chana Research Group, Pontificia Universidad Católica del Perú, Perú
[‡]Instituto de Defensa Legal del Ambiente y Desarrollo Sostenible – IDLADS, Perú
omoreno@pucp.edu.pe,yanua.atamain@gmail.com,arturo.oncevay@pucp.edu.pe

Abstract

We introduce a Spanish-Awajun parallel dataset of 22k high-quality sentence pairs with the help of the journalistic organization Ojo Público[1]. This dataset consists of parallel data obtained from various web sources such as poems, stories, laws, protocols, guidelines, handbooks, the Bible, and news published by Ojo Público. The study also includes an analysis of the dataset's performance for Spanish-Awajun translation using a Transformer architecture with transfer learning from a parent model, utilizing Spanish-English and Spanish-Finnish as high-resource language-pairs. As far as we know, this is the first Spanish-Awajun machine translation study, and we hope that this work will serve as a starting point for future research on this neglected Peruvian language. The dataset is released in the following URL: https://github.com/iapucp/Awajun-OP

1 Introduction

In the diverse linguistic landscape of the Americas, Peru boasts 47 native languages, including 4 Andean and 43 Amazonic languages (Zariquiey et al., 2019). Castilian Spanish, the primary official language spoken by the majority, starkly contrasts with these native languages. This vast language diversity is both a cultural treasure and a significant communication barrier. Although human translators have played a crucial role in bridging these gaps, their availability remains limited. Peru recognizes the pivotal role of translation in politics, striving for equal language rights through initiatives like the National Registry of Interpreters and Translators of Indigenous Languages (Vásquez, 2015).

Efforts to preserve Peru's languages have been insufficient, with most endangered and lacking documentation (Zariquiey et al., 2019). Mainly passed down orally, these languages pose significant computational challenges for machine translation due to scarce monolingual or parallel texts.

However, recent research has shown promise in Neural Machine Translation (NMT) for select Peruvian native languages like Quechua Ayacucho (quy), Quechua Cuzco (quz), Aymara (aym), Shipibo-Konibo (shp), and Asháninka (cni). Despite progress, the Awajun language (agr), with around 55,000 speakers, remains overlooked, lacking dedicated NMT research. Furthermore, Awajun's data in OPUS is limited to fewer than 7,000 sentences. Given this context, exploring alternative approaches is crucial to developing more effective machine translation systems for Awajun and creating new parallel corpora to support these efforts.

In this work, we aim to provide a comprehensive introduction to the Awajun language (see Appendix B), introduce a new parallel corpus for Spanish-Awajun (see §3), and experiment with transfer learning strategies for developing, as far as we know, the first NMT systems for the language pair Spanish-Awajun (see §5 for NMT experiments and Appendix A for related work).

2 Quispe Chequea

Ojo Público has developed a digital tool using artificial intelligence to produce journalistic verification content in multiple formats and up to three native languages of Peru: Quechua, Aymara, and Awajún. This platform automates text generation, translation, and conversion into audio messages, which can be broadcasted by radios in nine regions, including Loreto, Junín, Amazonas, Piura, San Martín, Ayacucho, Apurímac, Puno, and Tacna. Developed by a team of journalists, technologists, translators, and interpreters, the project aims to combat misinformation affecting citizens and communities in the Andes and the Amazon of Perú.

This study focuses on the translation component

[1]"Ojo Público" is a Peruvian media outlet. It is operated by a non-profit journalistic organization based in Lima

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 112–120
June 21, 2024

of the project, specifically from data compilation to training an NMT model from Spanish to Awajun.

3 Corpus Development

We first compared the OPUS dataset (currently available data) and the new corpora extracted that we call Awajun-OP. An official translator validated all sources for the corpora to ensure the same dialect is used and to verify the translation quality.

3.1 OPUS dataset

From OPUS, Christodouloupoulos and Steedman (2015) is a dataset with translations of the Bible, including the New Testament of the 1973 edition.

3.2 Awajun-OP : New parallel corpora

Sources of Awajun translations

1. Ebible: A curated corpus of parallel data derived from versions of the Bible provided by Ebible.org that includes the old and new Testaments(Ebible.org, 1997). Notably, a significant portion of the audited and web-scraped data from MADLAD-400(Kudugunta et al., 2024) originates from this source, as it underwent manual verification due to its comparable number of sentences. This process yields new parallel corpora while eliminating potential monolingual data.

2. Poems&Stories: The website *Cultura Awajun*[2] features poems, vocabulary, and common expressions in Awajun along with their Spanish translations (Yanua, 2015a, 2016, 2015b,c). Additionally, it hosts several ancestral stories in Awajun accompanied by their Spanish versions compiled by the National Fund for the Development of Peruvian Education (FONDEP, 2019).

3. Laws&Protocols: We have identified five official documents comprising laws and protocols translated into Awajun. The protocols include the documentation protocol for individuals belonging to indigenous peoples of the Peruvian Amazon (RENIEC, 2015) and the protocol for the care of people with disabilities (RENIEC, 2014). As for the laws, they encompass the Law on artisans and the development of artisanal activity (MINCETUR, 2020), the Right to prior consultation (MINCUL, 2013), and the Agreement 169 (Palomino, 2015a,b).

4. Guidelines: Various government institutions have translated and disseminated documents to facilitate community guidance, including those promoting awareness of universal health rights (SUSALUD, 2018), civil registration procedures (RENIEC, 2018), and the registration of acts and rights of native communities (SUNARP, 2023).

5. Handbook: To aid in the language acquisition of Awajun, the Amazon Center for Anthropology and Practical Application has published a handbook as an educational resource (Regan, 1991).

6. News by Ojo Público: Ojo Público, a Peruvian media outlet, has previously translated its news into Awajun (with a professional translator). Additionally, they have generated translations for short sentences about common domain knowledge.

Methodology for corpus creation The only sources extracted and aligned automatically using the document's dot, newline character, line break, or position were (MINCETUR, 2020; RENIEC, 2014, 2015; MINCUL, 2013). Paragraphs with more than one sentence that had an equal number of sentences as their translation was split into small sentences. The Ebible source can be automatically aligned using their repository[3]. All the additional sentences were extracted and aligned manually arranging the sentence breaks to separate each translation pair.

Data pre-processing It has been established that only the following symbols can be preserved: ".", ",", "!", "?", "," "%." . In addition, pairs of sentences containing empty sentences or only white spaces are excluded. Lastly, duplicates and sentences exceeding 50 words on the Spanish side are removed. It is worth noting that we refrain from considering the elimination of sentence pairs based on a word ratio criterion, given the unique characteristics of Awajun.

Corpora description We perform a large number or rare events (LNRE)[4] modeling to analyze the Ebible, Poems&Stories, Laws&Protocols, Guidelines, Handbook, News by Ojo Público and Opus-agr. The values are shown in Table 1.

In this study, we have opted not to incorporate OPUS-agr into our dataset Awajun-OP, considering Ebible already encompasses the New Testament. Most sentences stem primarily from Ebible, comprising nearly 74% of the compiled dataset. Poems&Stories and Handbook datasets exhibit the least volume of sentences. News by Ojo Público, conversely, experiences a notable reduction in the

[2] https://culturaawajun.blogspot.com/

[3] https://github.com/BibleNLP/ebible

[4] We used the LNRE calculator created by Kyle Gorman: https://gist.github.com/kylebgorman/

	S raw	S clean	$r_{\text{agr}\rightarrow\text{es}}$	N		V		V1		V/N		V1/N	
				es	agr	es	agr	es	agr	es	agr	es	agr
OPUS-agr	6,739	6,717	0.92	141,290	125,525	13,966	26,538	7,411	16,558	0.10	0.21	0.05	0.13
Ebible	16,945	16,591	0.85	400,118	330,404	21,556	51,581	10,583	31,974	0.05	0.16	0.03	0.10
Poems&Stories	178	173	0.75	1,388	880	633	630	475	499	0.46	0.72	0.34	0.57
Laws&Protocols	1,032	938	0.89	17,621	13,890	3,394	4,413	1,963	3,001	0.19	0.32	0.11	0.22
Guidelines	778	735	0.82	11,373	8,545	2,224	2,478	1,251	1,640	0.20	0.29	0.11	0.19
Handbook	366	364	0.62	1,720	1,009	713	752	513	616	0.41	0.75	0.30	0.61
News by Ojo Público	4,221	3,646	1.14	28,643	24,223	5,407	6,625	2,948	4,154	0.19	0.27	0.10	0.17
Total	23,520	22,447	0.89	460,863	378,951	27,758	60,127	13,563	37,366	0.06	0.16	0.03	0.10

Table 1: Corpora description: S = #sentences in corpus; $r_{\text{agr}\rightarrow\text{es}}$ = average of the ratio agr-es per sentence; N = number of tokens; V = vocabulary size; V1 = number of tokens occurring once (hapax); V/N = vocabulary growth rate; V1/N = hapax growth rate

number of sentences, attributed mainly to their extended length.

Although it was expected, the vocabulary size and tokens occurring only once are higher for Awajun, as this demonstrates its agglutinative property. We have observed that the Handbook and Poems&Stories datasets have a larger vocabulary and a higher number of tokens occurring only once (V1), even though they have fewer tokens per sentence (N). Moreover, the sentences in these datasets exhibit more agglutinative characteristics, as their $r_{\text{agr}\rightarrow\text{es}}$ are the lowest. On the other hand, the News by Ojo Público dataset has a $r_{\text{agr}\rightarrow\text{es}}$ greater than one, and it is the only dataset with more sentences in Spanish than in Awajun. However, this only happens because News by Ojo Público has approximately 57% sentences with less or equal to 4 words.

The following example illustrates this scenario:

***agr**: Distrito alcaldeji nuwa*

***es (en)**: Alcaldesa distrital (District Mayoress)*

In Awajun, the word "nuwa" is added to indicate the gender of the subject.

4 Datasets of High Resource Languages

Spanish-English dataset For pre-training, we used the EuroParl dataset for Spanish–English (1.9M sentences) (Koehn, 2005) and for validation and testing the WMT2007 dataset (Callison-Burch et al., 2007).

Spanish-Finnish dataset For pre-training, we used EuroParl (1.9M sentences) (Koehn, 2005), EUbookshop (1.8M) (Skadiņš et al., 2014), and TED2020 (44k) (Reimers and Gurevych, 2020) datasets for Spanish–Finnish, this excluding 3k sentences for validation, and for testing, the Tatoeba (9.9k) dataset (Ho and Simon, 2016).

5 Neural Machine Translation for Awajun

5.1 Data partition for evaluation

Understanding the distribution of a suitable dataset for development/testing is crucial to ensuring the adequacy of selected sentences. Given the Bible's predominant role as the primary source of sentences, it's essential to carefully determine the quantity and source of sentences to evaluate the model impartially.

We followed a similar methodology as described in Oncevay (2021) and collected a comprehensive sample from various domains including News by Ojo Público, Poems & Stories, and Handbooks. The sample consisted of 1012 sentences in total, out of which 200 sentences were from News by Ojo Público, ranging from more than 9 to less than 20 words in Spanish, 400 sentences were between 5 to 9 words, and 464 sentences were sampled from Poems & Stories and Handbooks. This sample set was divided into 25%-25%-50%, with the first two segments allocated for validation and testing. An additional 250 sentences were added to each of the two segments from a stratified sample of the available datasets (Ebible, Guidelines, etc). The remaining 50% and News by Ojo Público's dataset (excluding sentences with <4 words) was added to the training set and upsampled to form 20% of the training data, aiming to minimize the domain gap within the training data. The final distribution is shown in Table 6.

The primary metric used in this study is chrF (Popović, 2017), which evaluates character n-grams and is particularly useful for agglutinative languages like Awajun. Additionally, BLEU scores (Papineni et al., 2002) were reported, utilizing implementations of sacreBLEU (Post, 2018).

Parent	Dataset	Model	Validation BLEU	Validation Chrf	Test BLEU	Test Chrf
Es-En	Baseline (OPUS-agr)	Transformer	4.05	30.04	4.05	30.27
	Awajun-OP	Transformer	7.36	38.14	6.75	37.87
Es-Fi	Baseline (OPUS-agr)	Transformer	3.87	31.48	4.21	32.21
	Awajun-OP	Transformer	**7.97**	**38.72**	**7.03**	**38.79**
-	Awajun-OP	GPT - Babbage	1.77	29	1.52	29.41
-	Monolingual Curated Ebible	MADLAD-400 3B	0.69	9.60	0.67	9.56

Table 2: Results in BLEU and Chrf for all trained models in validation and test sets.

5.2 Subword segmentation

Subword segmentation is an important process when translating agglutinative languages such as Awajun. We used the Byte-Pair-Encoding (BPE; Sennrich et al., 2016) implementation in SentencePiece (Kudo and Richardson, 2018) with a vocabulary size of 16,000. To enhance our vocabulary, we trained a segmentation model incorporating all three languages: Spanish, English/Finnish, and Awajun. We upsampled the Awajun data to achieve an even distribution among the languages.

5.3 Procedure

For all experiments, we used a Transformer-based model (Vaswani et al., 2017) with default parameters from the Fairseq toolkit (Ott et al., 2019).

To improve the encoding capability on the Spanish side, we started by pre-training a Spanish-English model on the Europarl dataset. After that, we fine-tuned the pre-trained model on the Spanish-Awajun dataset. We repeated the same experiment, but this time we used Spanish-Finnish as the HRL.

6 Results and discussion

Table 2 presents the outcomes of transfer learning models using Awajun-OP and OPUS-agr as baselines. The most remarkable scores in BLEU and chrF were attained by Awajun-OP when utilizing the Spanish-Finnish model as its parent. These findings suggest that the agglutinative nature of Finnish may have contributed to Awajun's successful translation. Moreover, the close resemblance between validation and test results underscores the model's generalization capabilities.

It is also noted that employing Awajun-OP yielded a notable enhancement compared to the baseline, achieving an improvement in BLEU score of +2.98 and a +8.52 in chrF. Furthermore, utilizing Spanish-Finnish as the parent model resulted in a 0.28 increase in BLEU and 0.92 in chrF for the test set compared to the Spanish-English model. Table 3 shows a translation output from the best model.

Input (ES)	*Publicó el video original en su sitio web con el titular*
Input (EN)	*He posted the original video on his website with the headline*
Reference (Awajun)	*nagkamchaku video jiikbauwa nuna niina webjin titularan aputus jiikiu*
Output	*Video nagkamchaku jiikbauwa duka sitio webnum agagtmitkau*

Table 3: Translation example

Dataset	Good	Bad	Acc
Ebible	68	158	30%
Poems&Stories	5	24	17%
Laws&Protocols	5	11	31%
Guidelines	2	4	33%
Handbook	24	69	26%
News by Ojo Público	64	82	44%
Total	168	348	33%

Table 4: This table presents the results of the translator's examination, indicating both correct and incorrect translations. Accuracy is calculated as: Good/(Good+Bad).

In addition to the transfer learning experiment, we trained a GPT-Babbage model. However, the results were unsatisfactory, and we decided to stop training this type of model. Morevoer, we tested MADLAD-400 (Kudugunta et al., 2024), which contains part of the Ebible data for Awajun, but it underperformed as well.

BLEU scores only may not appear promising, which is similar to the results for other low-resource languages from the Americas. To complement the evaluation, a professional Awajun translator assessed a sample of the outputs of the best model. Table 4 showcases the translation ratings, categorized by dataset. News by Ojo Público attained the highest accuracy level at 44%, potentially attributed to sentence length. Poems & Stories and

Handbook results were less favorable, likely due to the limited sentences in these datasets. Overall, approximately one-third of translations were deemed of good quality. The translator noted that some sentences labeled as "Bad" possessed well-written content but differed in meaning from the reference.

7 Conclusion

In this study, we extracted and created new parallel corpora for Spanish-Awajun, which comes from different sources, such as stories, laws, protocols, or guidelines from the web, plus in-house translated news texts. This helped us to develop the first NMT models for Spanish-Awajun. Our work revealed that implementing transfer learning with Spanish-Finnish as a parent language resulted in better outcomes for both the baseline and Awajun-OP. Furthermore, we sought the assistance of a professional translator to validate our findings and obtain a human perspective on the quality of our model. Despite the limited availability of data, our research produced promising results.

We have taken the initial steps towards developing reliable translations in Awajun. For future work, we aim to acquire additional monolingual data for back-translation and fine-tune large multilingual models such as NLLB (Costa-jussà et al., 2022), among others.

Limitations

This paper aims to give an introduction to researchers, students, of interested community indigenous community members to the topic of Machine Translation for Indigenous languages of the Americas. Therefore, this paper is not an in-depth survey of the literature on indigenous languages nor a more technical survey of low-resource machine translation. We would point the reader to more specific surveys on these aspects

Ethical statement

We could not find any specific Ethical issue for this paper or potential danger. Nevertheless, we want to point to the reader that working with indigenous languages (in this case, MT) implies a set of ethical questions that are important to handle. For a deeper understanding of the matter, we suggest specialized literature to the reader https://aclanthology.org/2023.americasnlp-1.13.pdf

Acknowledgements

This work would not have been possible without the support of Ojo Público, a well-known non-profit Peruvian media outlet, and their initiative to provide factual and verified news to the communities in the Andes and Amazon of Peru. The author is thankful to the directors and members of Ojo Público, and in particular, to David Hidalgo, Jorge Miranda Jaime, and Gianella Tapullima Abriojo, who were part of the discussions for this study. Furthermore, the second author is an official translator for Awajun and a proud member of the Awajun community.

References

Honorio Apaza, Brisayda Aruhuanca, Mariela M. Nina, Anibal Flores, Carlos Silva, and Euler Tito. 2023. Neural machine translation for native language aymara to english. In *Proceedings of the Future Technologies Conference (FTC) 2022, Volume 3*, pages 565–576, Cham. Springer International Publishing.

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Baljekar, Xavier Garcia, Wolfgang Macherey, et al. 2022. Building machine translation systems for the next thousand languages. *arXiv preprint arXiv:2205.03983*.

Chris Callison-Burch, Philipp Koehn, Cameron Shaw Fordyce, and Christof Monz, editors. 2007. *Proceedings of the Second Workshop on Statistical Machine Translation*. Association for Computational Linguistics, Prague, Czech Republic.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the bible in 100 languages. *Language resources and evaluation*, 49:375–395.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

Ebible.org. 1997. Ebible.org. https://ebible.org/. Accessed: 2024-03-11.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E Ortega, Rolando Coto-Solano, et al. 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.

FONDEP. 2019. *Relatos ancestrales del pueblo Awajún - Cuentos, mitos y leyendas*, first edition.

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.

Trang Ho and Allan Simon. 2016. Tatoeba: Collection of sentences and translations.

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

Amir Hossein Kargaran, Ayyoob Imani, François Yvon, and Hinrich Schütze. 2023. Glotlid: Language identification for low-resource languages. *arXiv preprint arXiv:2310.16248*.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2024. Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36.

Alexandra Espichán Linares and Arturo Oncevay-Marcos. 2017. A low-resourced peruvian language identification model. In *CEUR Workshop Proceedings. CEUR-WS*.

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

MINCETUR. 2020. Versiones bilingües en lenguas originarias de la "ley del artesano y del desarrollo de la actividad artesanal - ley nº 29073". https://www.gob.pe/es/i/470746. Accessed: 2024-03-11.

MINCUL. 2013. Consulta previa. https://consultaprevia.cultura.gob.pe/materiales-informativos. Accessed: 2024-03-11.

Oscar Moreno. 2021. The REPU CS' Spanish–Quechua submission to the AmericasNLP 2021 shared task on open machine translation. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 241–247, Online. Association for Computational Linguistics.

Arturo Oncevay. 2021. Peru is multilingual, its machine translation should be too? In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 194–201, Online. Association for Computational Linguistics.

John E Ortega, Richard Castro Mamani, and Kyunghyun Cho. 2020. Neural machine translation with a polysynthetic low resource language. *Machine Translation*, 34(4):325–346.

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Simon Overall. 2010. Jerarquía y tratamiento de la primera persona plural en la gramática de aguaruna.

Marco A. Huaco Palomino. 2015a. Convenio 169 shiig antumain tibau. https://culturaawajun.blogspot.com/p/convenio-169-shiig-antumain-tibau-marco.html. Accessed: 2024-03-11.

Marco A. Huaco Palomino. 2015b. El convenio 169 de la oit dicho en otras palabras por marco a. huaco palomino. https://culturaawajun.blogspot.com/p/el-convenio-169-de-la-oit-dicho-en.html. Accessed: 2024-03-11.

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.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Jaime Regan. 1991. *Chichasájmi: Primer nivel*, volume 1. Centro Amazónico de Antropología y Aplicación Práctica.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation.

RENIEC. 2014. Protocolo para la documentacion de las personas perteneciente a los pueblos indigenas de la amazonia peruana. https://www.reniec.gob.pe/portal/html/registro-civil-bilingue/portalrcb2016/5-protocolos/05-protocolo-documentacion-awajun.pdf. Accessed: 2024-03-11.

RENIEC. 2015. Protocolo de atenciÓn a personas con discapacidad. https://www.reniec.gob.pe/portal/html/registro-civil-bilingue/portalrcb2016/5-protocolos/09-protocolo-atencion-discapacidad-awajun.pdf. Accessed: 2024-03-11.

RENIEC. 2018. Cartilla de atención - registro civil bilingüe. https://www.reniec.gob.pe/portal/html/registro-civil-bilingue/portalrcb2016/5-protocolos/Cartilla-atencion-awajun.pdf. Accessed: 2024-03-11.

Ketty Betsamar García Ruiz. 2020. Análisis semántico de términos binomiales de flora y fauna en la lengua awajún. Master's thesis, Pontificia Universidad Catolica del Peru (Peru).

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Aaron Serianni and Daniel Whitenack. Exploring transfer learning pathways for neural machine back translation of eskimo-aleut, chicham, and classical languages.

Raivis Skadiņš, Jörg Tiedemann, Roberts Rozis, and Daiga Deksne. 2014. Billions of parallel words for free: Building and using the eu bookshop corpus. In *Proceedings of LREC*, page 29. LREC.

SUNARP. 2023. Guía general de comunidades nativas. https://www.gob.pe/institucion/sunarp/informes-publicaciones/2454692-guia-general-de-comunidades-nativas. Accessed: 2024-03-11.

SUSALUD. 2018. Conocemos y promovemos nuestros derechos y deberes en salud y el aseguramiento universal. http://sistec.sis.gob.pe/fuente/files/pdf/Instructivo%20Rotafolio%20Formaci%C3%B3n%20de%20Formadores%20en%20Awajun.pdf. Accessed: 2024-03-11.

J Vásquez. 2015. La implementación de derechos lingüísticos para la mejora de servicios públicos y la recuperación y fortalecimiento de lenguas indígenas. XX Congreso Internacional del CLAD sobre la Reforma del Estado y de la ....

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Yanua. 2015a. Cultura awajun - jikamajame. https://culturaawajun.blogspot.com/2015/11/jikamajame.html. Accessed: 2024-03-11.

Yanua. 2015b. Cultura awajun - saludos y expresiones usuales. https://culturaawajun.blogspot.com/2015/09/saludos-y-expresiones-usuales.html. Accessed: 2024-03-11.

Yanua. 2015c. Cultura awajun - vocabulario awajun castellano. https://culturaawajun.blogspot.com/2015/11/vocabulario-awajun-castellano.html. Accessed: 2024-03-11.

Yanua. 2016. Cultura awajun - poema inia: Ju kashia dui. https://culturaawajun.blogspot.com/2016/06/poema-inia-ju-kashia-dui.html. Accessed: 2024-03-11.

Fei Yuan, Yinquan Lu, Wenhao Zhu, Lingpeng Kong, Lei Li, Yu Qiao, and Jingjing Xu. 2022. Lego-mt: Learning detachable models for massively multilingual machine translation. *arXiv preprint arXiv:2212.10551*.

Roberto Zariquiey, Harald Hammarström, Mónica Arakaki, Arturo Oncevay, John Miller, Aracelli García, and Adriano Ingunza. 2019. Obsolescencia lingüística, descripción gramatical y documentación de lenguas en el Perú: hacia un estado de la cuestión. *Lexis*, 43(2):271–337.

A Related Work

The Quechuan language family has been a primary focus in MT research. Notable studies include Ortega et al. (2020), which employed a sequence-to-sequence NMT model for Southern Quechua, utilizing transfer learning with Finnish due to its agglutinative characteristic. Similarly, Huarcaya Taquiri (2020) utilized the Jehovah Witnesses dataset, along with supplementary lexicon data, to train an NMT model for Quechua, achieving notable BLEU scores. However, it's important to note that the high results in both cases may be attributed to the development and test sets being drawn from the same religious domain and distribution as the training set. In addition to Quechuan languages, NMT models have been developed for Aymara (Apaza et al., 2023) and Shipibo-Konibo (Gómez Montoya et al., 2019), with Spanish as their paired language. Even in the private sector, Google Translator has expanded its language offerings to include Quechua and Aymara (Bapna et al., 2022).

Recent research within the AmericasNLP community has been dedicated to advancing Machine Translation (MT) for indigenous languages of the Americas. Workshops held in 2021 and 2023 focused on translating texts in 10 indigenous languages, including peruvian native languages such as Quechua Ayacucho (quy), Aymara (aym), Shipibo-Konibo (shp), and Ashaninka (cni) (Mager et al., 2021; Ebrahimi et al., 2023). These efforts aimed to explore various approaches, including utilizing high-resource bilingual systems like Spanish–English and Spanish–Finnish pretrained models, alongside Statistical Machine Translation (SMT) models. Additionally, researchers experimented with fine-tuning different multilingual architectures such as mT5, mBART, etc. Notably, the importance of clean data was emphasized, with studies showing improved results through the generation of additional clean data, particularly in the case of Quechua (Moreno, 2021).

Despite efforts in Neural Machine Translation (NMT) for Peruvian native languages, significant attention has not been directed towards Awajun (agr). In terms of MT models, (Serianni and Whitenack) showcased the utility of Transfer Learning, even when the related language does not perfectly align with the target domain, by employing Awajun alongside English with parallel data from the OPUS dataset. Kudugunta et al., 2024 compiled a massive audited monolingual dataset, which includes Awajun, and utilized it alongside publicly available datasets to train extensive multilingual models spanning 419 languages. Similarly, Yuan et al., 2022 delved into learning Detachable Models for Massively Multilingual Machine Translation for 433 languages using the OPUS dataset, with both studies integrating Awajun as one of the languages for translation. However, none of these investigations have specifically targeted the enhancement of MT performance in Awajun nor have they presented metrics for Awajun translation. Furthermore, research conducted by (Linares and Oncevay-Marcos, 2017) focused on language identification models using data from web and private repositories of 16 Peruvian native languages, while GlotLID, targeting low-resource languages, identified 1665 languages (Kargaran et al., 2023). In summary, limited work has been conducted on Spanish-Awajun MT, with data primarily sourced from the OPUS parallel dataset.

B Language specifics

Awajun (agr), also known as Aguaruna, belongs to the jíbaro family and it is the second most spoken language in the Amazon of Peru with approximately 55,000 native speakers. It is spoken in the peruvian regions of Amazonas, Cajamarca, San Martín, and Loreto (Ruiz, 2020). As with many of the native languages of Peru, it has different dialects depending on the geography of the speakers. Based on the National Registry of Interpreters and Translators of Indigenous Languages[5], at the moment of this study, there are 42 translators for all dialects. The dialect of the Marañon River is the most spoken and is the one chosen to recollect the data from this study.

Examples of the different variants are shown in table 5. For the word 'smile' in Awajun, it can be observed that the 'shiawai' subword is maintained for both dialects. Furthermore, the Marañon River dialect uses the endings with 'g' and the Nieva and Canepa River (NSR) dialect uses the 'j' endings. Although there are different dialects in Awajun, there are mainly minor differences in vocabulary and terminology.

Awajun exhibits a rich morphological structure

[5]It is a database that contains contact and registration information of citizens who have been trained by the Ministry of Culture of Peru, through Indigenous Language Interpreter and Translator Courses developed since 2012. Their website is: https://traductoresdelenguas.cultura.pe/.

English	Spanish	Awajun (NSR)	Awajun (Marañon)
smile	sonríe	yushiawai	dushiawai
brother	hermano	yatsuj	yatsug
sister	hermana	kaij	kaig

Table 5: Dialects in Awajun

characterized by agglutinative processes, primarily suffixation. In contrast to Spanish, Awajun exhibits a distinct word order, typically following a subject-object-verb (SOV) structure. This deviation poses a considerable contrast, as Spanish predominantly follows a subject-verb-object (SVO) order. This linguistic distinction not only presents a challenge in comprehension but also underscores the cultural and grammatical differences between the two languages. Furthermore, it employs a double marking system for grammatical categories, both in the head and the dependent elements. Awajun marks first or second-person objects with obligatory verbal suffixes, while nominal or pronominal objects are also marked with suffixes (Overall, 2010).

C Additional information

Additional tables and figures with information about the corpora creation and translation metrics.

Dataset	Train	Validation	Test
Ebible	16,591	226	226
Poems&Stories	263	29	28
Laws&Protocols	938	6	10
Guidelines	735	16	14
Handbook	720	93	93
News by Comp. C	6,686	146	145
Total	**25,933**	**277**	**278**

Table 6: Final distribution of datasets for train, validation, and test

Wav2pos: Exploring syntactic analysis from audio for Highland Puebla Nahuatl

Robert Pugh♠ and **Varun Sreedhar**◇ and **Francis Tyers**♠
pughrob@iu.edu, varunsreedhar14@gmail.com, ftyers@iu.edu
Indiana University, Bloomington
♠Department of Linguistics
◇Luddy School of Informatics, Computing, and Engineering

Abstract

We describe an approach to part-of-speech tagging from audio with very little human-annotated data, for Highland Puebla Nahuatl, a low-resource language of Mexico.[1] While automatic morphosyntactic analysis is typically trained on annotated textual data, large amounts of text is rarely available for low-resource, marginalized, and/or minority languages, and morphosyntactically-annotated data is even harder to come by. Much of the data from these languages may exist in the form of recordings, often only partially-transcribed or analyzed by field linguists working on language documentation projects. Given this relatively low-availability of text in the low-resource language scenario, we explore end-to-end automated morphosyntactic analysis directly from audio. The experiments described in this paper focus on one piece of morphosyntax, part-of-speech tagging, and builds on existing work in a high-resource setting. We use weak supervision to increase training volume, and explore a few techniques for generating word-level predictions from the acoustic features. Our experiments show promising results, despite less than 400 sentences of audio-aligned, manually-labeled text.

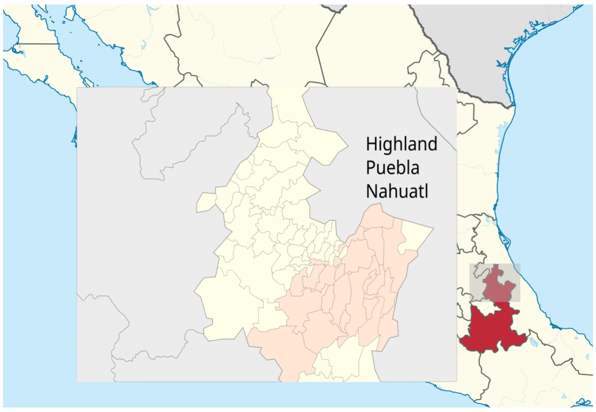


Figure 1: A map highlighting the 24 municipalities where HPN is spoken in the Sierra Norte de Puebla region of Mexico.

1 Introduction

Automatic morphosyntactic processing, such as morphological analysis or syntactic parsing, is an important task in Natural Language Processing (NLP) for the purposes of language documentation, feature-extraction for downstream NLP tasks (Sidorov, 2019; Wu et al., 2021; Sartakhti et al., 2021), and for quantitative corpus-based linguistic analysis (Tyers and Henderson, 2021; Kim et al., 2021).

The ample research exploring these tasks has, overwhelmingly, focused on textual data. However, text for low-resource, endangered, marginalized, and/or minority languages, which constitute a majority of the world's languages, is often sparsely available, if at all. Instead, much of the data from these languages may exist in the form of recordings, potentially only partially-transcribed or analyzed by field linguists working on language documentation projects. At the same time, recent progress in speech processing has resulted in powerful, pretrained speech representation models such as Wav2Vec2.0 (Baevski et al., 2020), which make it possible to achieve impressive ASR systems via fine-tuning on relatively little data (Yin et al., 2022). These same representations have also been shown to be useful in audio classification problems, such as speaker recognition (Vaessen and Van Leeuwen, 2022) and emotion detection (Pepino et al., 2021).

In the remainder of this paper, we explore end-to-end automated part-of-speech (POS) tagging directly from audio for an endangered Nahuatl variant, using a modest amount of transcribed data and only a few hundred sentences of annotated text. In light of a very limited data set.

[1]The code used for the work described here is available at https://github.com/VarunS9000/Wav2Pos

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 121–126
June 21, 2024 ©2024 Association for Computational Linguistics

2 Highland Puebla Nahuatl

Nahuatl is a polysynthetic, agglutinating Uto-Aztecan language spoken throughout Mexico and Mesoamerica, made up of 30 recognized varieties (INALI, 2009).

Highland Puebla Nahuatl, (or *Sierra Puebla Nahuatl*, also referred to by INALI as *Náhuatl del noreste central*, `ISO-639-3` *azz*, henceforth HPN) is a Nahuatl variant group spoken by about 70,000 people (Ethnologue's 2007 estimate) in the Northeastern Sierra region of the state of Puebla, Mexico (see Figure 1) in 24 municipalities (INALI, 2009).

HPN has been the subject of documentary and descriptive linguistic efforts (Key, 1960; Robinson, 1970; Key and Key, 1953; Key and Richie de Key, 1953; Cortez Ocotlán, 2017). Furthermore it is one of two Nahuatl variants with a free and open morphosyntactically-annotated corpus, in the form of a Universal Dependencies treebank.

As an indigenous language of Mexico, HPN is considered at risk of being lost (INALI, 2012).

3 Related work

While most effort in the area of morphosyntactic analysis has focused on textual input, some recent work explores the idea of performing natural language processing directly from audio. Pupier et al. (2022) perform end-to-end dependency parsing for French from audio, by extracting audio features, aggregating them into audio word embeddings using LSTMs, and performing dependency parsing using these embeddings. These experiments used a dataset size consisting of 169,500 training sentences. Omachi et al. (2022) describe a non-autoregressive (non-transformer) method for performing end-to-end ASR and downstream NLP tasks such as named-entity recognition, performing part-of-speech tagging on a large.[2] corpus of spoken Japanese, and NER on a corpus of English containing approximately 10k training sentences (Bastianelli et al., 2020).

Shi et al. (2021) create a speech translation corpus of HPN using the same dataset as in the present paper, leveraging the fact that the entire dataset has transcriptions and translations (dataset details presented in Section 4).

[2]The size, e.g. number of sentences of words, of the corpus is not reported in Omachi et al. (2022) The publication presenting the corpus, Maekawa et al. (2000), describes it as having 7 million morphemes.

4 Data

The speech files, the transcriptions, and the pertinent metadata were obtained via the dataset from Amith et al. (2019) (hereafter "OpenSLR corpus"), which consists of about 50 hours of audio transcribed in ELAN. We processed the ELAN files, splitting the audio into utterance-level chunks using the Pydub Python library.[3]

The labeled HPN part-of-speech data comes from recently-released Highland Puebla Nahuatl UD treebank, which consists of (1) a small subset of the OpenSLR corpus annotated for morphosyntax, (2) a subset of texts in the *azz* variant from the multi-variant parallel corpus Axolotl (Gutierrez-Vasques et al., 2016), and (3) technical publications by the Sociedad Mexicana de Física.[4] Only (1), which contains 399 sentences and 3,463 tokens, has corresponding audio, and is held out for system evaluation. (2) and (3), totaling 838 sentences and 6,671 tokens, are used in training a simple text-based part-of-speech tagger, with which we generate synthetic data as described in Section 5.1.

The remaining OpenSLR corpus data (i.e. all of the audio/transcriptions that has not been annotated) is partitioned into a larger dataset for fine-tuning Wav2Vec2 (about 40k sentences), and a smaller dataset for training the audio-based POS tagger (about 7k sentences). We chose to use the majority of the data for Wav2Vec2 fine-tuning in order to ensure a high-performing ASR model since without that a POS tagger would have no words to align its POS tags with.[5] We chose not to use overlapping data for training both Wav2Vec2 and the Wav2pos tagger to avoid overfitting and poor performance on unseen data.

5 Methodology

In this section, we describe our method for training a POS tagger (Wav2pos) directly from audio with less than 400 labeled examples. This process involves training an acoustic feature extractor, generating silver training data, and aggregating the acoustic features to word-level.

[3]`https://github.com/jiaaro/pydub`

[4]`https://site.inali.gob.mx/SMF/Libros2.0/nhtl/index.html`

[5]In hindsight, it likely would have been worthwhile to experiment with different ways to partition this data, e.g. less data for Wav2Vec2 training and more for POS tagger training.

Dataset	Contents	Sentences	Tokens
Wav2Vec2 train*	OSLR − HPN	32k	285k
Text POS Tagger train	(HPN − OSLR) ∪ WSPN	1.7k	17.6k
Wav2pos train*	OSLR − HPN	8k	71k
Test data	HPN ∩ OSLR	363	2.4k

Table 1: A description of the contents of the different datasets. OSLR = OpenSLR data; HPN = Highland Puebla Nahuatl UD treebank; WSPN = Western Sierra Puebla Nahuatl UD Treebank. *The Wav2Vec2 fine-tuning data and the Wav2pos training data both come from the set of OpenSLR transcriptions not contained in the HPN treebank, but they are non-overlapping.

5.1 Synthetic label generation

The total amount of labeled POS data with corresponding audio for HPN is very small (363 sentences, 2k tokens). We hold it out for this purpose.

In order to produce enough labeled data to train the models, we label otherwise-unannotated OpenSLR transcriptions using a simple tagger. Specifically, we train an averaged perceptron model on the remaining UD trees (those sentences without corresponding audio), about 600 sentences. Since this is quite small for a training set, we supplement it with a UD treebank for another Nahuatl variant, Western Sierra Puebla Nahuatl (WSPN, `ISO-639-3` *nhi*) (Pugh et al., 2022), which added about 1k training sentences. The decision to add data from another variant is motivated by other recent work on Nahuatl syntactic parsing.

The averaged perceptron model uses words, substrings, previous words, and previous predicted tags as features. Once trained, we use it to predict POS tags on the unannotated OpenSLR transcriptions, resulting in "silver" training and validation data.

5.2 Extracting acoustic features

We use the unannotated (but transcribed) OpenSLR audio (split into a training and development set) to fine-tune the pret-trained Wav2Vec2.0 model (Baevski et al., 2020) on an ASR task. Our resulting fine-tuned model achieves a WER of 39% and CER of 18% on the held-out transcriptions.

In addition to being useful for automatically generating transcriptions, this fine-tuned model also gives us access to audio embeddings, corresponding to the discretized audio input, which have been fine-tuned for HPN. This sequence has many more elements than there are words (or even characters) in the sentence. We take two approaches to converting the longer sequence of acoustic embeddings into a single, word-level prediction in order to generate the part-of-speech tags.

5.3 Aggregating audio word embeddings with a BiLSTM

In the first Wav2pos approach, we first identify the subsequence of the acoustic embeddings by separating them by predicted whitespace characters. We pass each sequence (corresponding to segments of a single word) through a Long Short-Term Memory network (LSTM). The final hidden state is, then, a vector corresponding to a word. The sequence of word vectors is POS-tagged with a separate, bidirectional LSTM.[6]

5.4 Character-based prediction approach

As an alternative approach, we first reformat our data so that the label sequence, instead of consisting of a single POS tag per word, has a POS tag corresponding to each character (where the character's POS tag is that of its word). For example, for a transcription like *kemah niyas*, which originally is tagged `[INTJ, VERB]`, the label sequence is converted to `[INTJ, INTJ, INTJ, INTJ, INTJ, SPACE, VERB, VERB, VERB, VERB, VERB]`, such that each character in the transcription has a corresponding POS tag (note the inclusion of the `SPACE` tag corresponding to the word boundary). We pass the entire sequence of acoustic embeddings (without splitting them up into predicted audio words) through a BiLSTM, and make a POS prediction at each time step. For this approach, we use CTC loss in order to optimally-align the predicted POS tags with the labels. During inference, for a given word we choose the most frequent of its character-based POS tags as its tag.

5.5 Experiments

Given our silver training and validation data, and gold, human-annotated evaluation dataset, we compare the performance of three systems (as described in the previous section):

[6] The two LSTMs are trained jointly.

		Micro			Macro		
System	**Accuracy**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**	**F1**
apt	69.7	71.7	69.8	70.7	64.5	64.8	61.5
wb	53.2	57.1	53.2	55.1	51.4	48.2	46.3
cb	70.1	71.8	70.1	70.9	74.6	64.0	63.2

Table 2: Results comparing three approaches to POS-tagging our corpus. `wb` and `cb` correspond to systems that make predictions directly from audio, whereas the `apt`represents a pipeline system, wherein a text-based POS-tagger is run on the transcriptions output from the ASR system. While the two Wav2pos systems vary widely in their performance, the performance of the *cb* system suggests that the acoustic representations in the Wav2Vec2 model do in fact contain sufficient syntactic information.

apt: Averaged perceptron run on the output of the ASR. This method allows us to ascertain whether there is any benefit to calculating the POS tags directly from the audio instead of chaining the tagging with the ASR system.

wb: Word-based aggregation, where each hidden vector corresponding to an acoustic word (defined by the model's whitespace predictions) are first aggregated into a single word vector via an LSTM, and the sequence of aggregated word vectors is passed through a BiLSTM to predict the sequence of POS tags. This system is described above in Section 5.3

cb: Character-based aggregation, where the POS tag of a word is predicted for each of its characters, as described in Section 5.4.

5.6 Evaluation methodology

Since Wav2pos is based on the acoustic representations of the Wav2Vec2 model fine-tuned for ASR, and the ASR model may mistranscribe some words, our evaluation only takes into consideration words that the ASR model correctly transcribed. Specifically, we create tuples from the Wav2pos prediction and the ASR output, and match the ASR output to words in the correct transcription. If the word is transcribed correctly (i.e. it is found in the gold transcription), we compare the POS tags.

6 Results

The results of our three experiments are reported in Table 2. We note the passable performance of character-based Wav2pos model (`cb`), which is slightly better than the pipeline approach of tagging the transcriptions with a text-based POS-tagger (`apt`). This result suggests that indeed there is recoverable syntactic information represented in the acoustic feature embeddings learned by the Wav2Vec2 model, even (or especially) when these embeddings correspond to only a small piece of the word, as in the character model.

There is a significant difference in performance between the two Wav2pos models, the the `wb` model much worse on all metrics. While without more detailed analysis we can only speculate, it appears as though the aggregation step, which involves passing a sequence of acoustic vectors through an LSTM to produce a single "audio word vector," may introduce too many additional parameters for the model to learn given the relatively small amount of training data.

While these results are certainly interesting, they raise more questions than they answer, primarily as the result of the constrained set of experiments we performed. For future work, we plan to replicate these experiments, but using the same, larger Wav2Vec2 training dataset to train the Wav2pos models. We would also like to explore the hyperparameter space of these models in more depth, and try using a stronger text-based tagger, such as a multilingual pretrained transformer-based model, to create the silver data.

Finally, given the promising results for POS tagging, we are interested in expanding these efforts to other aspects of syntactic analysis such as dependency parsing.

7 Concluding remarks

We have presented a preliminary investigation of automated morphosyntactic analysis from audio with no human-labeled training data. We leveraged a large set of transcribed audio to fine-tune a Wav2Vec2 acoustic feature-extraction model, and experimented with producing POS-tags directly from the acoustic embeddings. We created our training data by tagging unlabeled transcription data using a simple classifier model. The results showed that one of our audio-based POS-tagging models performed slightly better than using the text-based tagger to tag the transcriptions.

8 Acknowledgements

This research was supported in part by Lilly Endowment, Inc., through its support for the Indiana University Pervasive Technology Institute. We also thank the anonymous reviewers for their helpful feedback.

References

Jonathan D. Amith, Amelia Dominguez Alcántara, Hermelindo Salazar Osollo, Ceferino Salgado Castañeda, and Eleuterio Gorostiza Salazar. 2019. Audio corpus of Sierra Nororiental and Sierra Norte de Puebla Nahuat(l) with accompanying time-code transcriptions in ELAN.

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. *Advances in neural information processing systems*, 33:12449–12460.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. SLURP: A spoken language understanding resource package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7252–7262, Online. Association for Computational Linguistics.

Pedro Cortez Ocotlán. 2017. *Diccionario Nahuat–Español de la Sierra Nororiental del Estado de Puebla*. Tetsijtsilin, Tzinacapan, Cuetzalan.

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.

INALI. 2009. *Catalogo De Las Lenguas Indigenas Nacionales: Variantes Linguisticas De Mexico Con Sus Autodenominaciones Y Referencias Geoestadisticas*. Instituto Nacional de Lenguas Indigenas, México, D.F.

INALI. 2012. *México: Lenguas indígenas nacionales en riesgo de desaparición*. Instituto Nacional de Lenguas Indigenas, México.

Harold Key. 1960. Stem construction and affixation of Sierra Nahuat verbs. *International Journal of American Linguistics*, 28(2):130–145.

Harold Key and Mary Richie de Key. 1953. *Vocabulario Mejicano de la Sierra de Zacapoaxtla, Puebla*. Instituto Lingüístico de Verano, México, D.F.

Mary Key and Harold Key. 1953. The phonemes of sierra nahuat. *International Journal of American Linguistics*, 19(1):53–56.

Jongin Kim, Nayoung Choi, Seunghyun Lim, Jungwhan Kim, Soojin Chung, Hyunsoo Woo, Min Song, and Jinho D. Choi. 2021. Analysis of zero-shot crosslingual learning between English and Korean for named entity recognition. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 224–237, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Kikuo Maekawa, Hanae Koiso, Sadaoki Furui, and Hitoshi Isahara. 2000. Spontaneous speech corpus of Japanese. In *Proceedings of the Second International Conference on Language Resources and Evaluation (LREC'00)*, Athens, Greece. European Language Resources Association (ELRA).

Motoi Omachi, Yuya Fujita, Shinji Watanabe, and Tianzi Wang. 2022. Non-autoregressive end-to-end automatic speech recognition incorporating downstream natural language processing. In *ICASSP 2022 - 2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6772–6776.

Leonardo Pepino, Pablo Riera, and Luciana Ferrer. 2021. Emotion recognition from speech using wav2vec 2.0 embeddings. *arXiv preprint arXiv:2104.03502*.

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.

Adrien Pupier, Maximin Coavoux, Benjamin Lecouteux, and Jérôme Goulian. 2022. End-to-end dependency parsing of spoken french. In *Interspeech*.

Dow F. Robinson. 1970. *Aztec studies 2: Sierra Nahuat word structure*. Summer Institute of Linguistics.

Moein Salimi Sartakhti, Romina Etezadi, and Mehrnoush Shamsfard. 2021. Improving Persian relation extraction models by data augmentation. In *Proceedings of the Second International Workshop on NLP Solutions for Under Resourced Languages (NSURL 2021) co-located with ICNLSP 2021*, pages 32–37, Trento, Italy. Association for Computational Linguistics.

Jiatong Shi, Jonathan D. Amith, Xuankai Chang, Siddharth Dalmia, Brian Yan, and Shinji Watanabe. 2021. Highland Puebla Nahuatl speech translation corpus for endangered language documentation. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 53–63, Online. Association for Computational Linguistics.

Grigori Sidorov. 2019. *Syntactic n-grams in computational linguistics*. Springer.

Francis Tyers and Robert Henderson. 2021. A corpus of K'iche' annotated for morphosyntactic structure. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 10–20.

Nik Vaessen and David A Van Leeuwen. 2022. Fine-tuning wav2vec2 for speaker recognition. In *ICASSP 2022-2022 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 7967–7971. IEEE.

Sixing Wu, Ying Li, Minghui Wang, Dawei Zhang, Yang Zhou, and Zhonghai Wu. 2021. More is better: Enhancing open-domain dialogue generation via multi-source heterogeneous knowledge. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2286–2300, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Haiyan Yin, Dingcheng Li, and Ping Li. 2022. Learning to selectively learn for weakly supervised paraphrase generation with model-based reinforcement learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1385–1395, Seattle, United States. Association for Computational Linguistics.

From Field Linguistics to NLP: Creating a curated dataset in Amuzgo language

Antonio Reyes Pérez
Facultad de Lenguas y Letras, UAQ
antonio.reyesp@uaq.edu.mx

Hamlet Antonio García Zuñiga
Instituto Nacional de Antropología e Historia
hamlet_garcia@inah.gob.mx

Abstract

This article presents an ongoing research on one of the several native languages of the Americas: Amuzgo or $jny'on^3\ nda^3$. This language is spoken in Southern Mexico and belongs to the Otomanguean family. Although Amuzgo vitality is stable and there are some available resources, such as grammars, dictionaries, or literature, its digital inclusion is emerging (cf. Eberhard et al. (2024)). In this respect, here is described the creation of a curated dataset in Amuzgo. This resource is intended to contribute to the development of tools for scarce resources languages by providing fine-grained linguistic information in different layers: From data collection with native speakers to data annotation. The dataset was built according to the following method: i) data collection in Amuzgo by means of linguistic fieldwork; ii) acoustic data processing; iii) data transcription; iv) glossing and translating data into Spanish; v) semiautomatic alignment of translations; and vi) data systematization. This resource is released as an open access dataset to foster the academic community to explore the richness of this language.

1 Introduction

According to the facts reported in the survey *Analysis of the Language Technologies in Mexico* (ASTLM, 2018), Latin America is a linguistic region with a minimum development in the creation of digital resources, specifically, regarding native languages. One of the main causes, pointed out by the authors, is the scarcity of data. The researchers frequently face a lack of materials to study these languages. For instance, several native languages have not even been described, either because they have not been considered academically, or because their grammar is difficult, or data collection is complicated or, even, highly risky. This stresses a gradual loss of worldviews, as well as an augment of the digital divide, which will directly impact on the native communities by making inequalities and marginalization larger.

With respect to the Mexican context, specialized organisms such as the National Institute of Statistics and Geography or the National Institute of Indigenous Languages report an important linguistic wealth. According to the numbers registered in (INALI, 2008), apart from Spanish, more than 60 native languages coexist in the country. This diversity is classified in 11 linguistic families, 68 languages, and 364 dialectal varieties. In terms of their vitality, the languages with more speakers are Nahuatl (more than one million), Mayan (around 800,000), Mixtec, and Zapotec (over 400,000 speakers each). On the contrary, there exist an important number of languages with fewer than 1,000 speakers (INEGI, (2015). Sadly, despite this enormous linguistic diversity, there is still a lack of resources, tools, and even linguistic materials for the majority of these languages (surprisingly, some of them well studied and described).

Amuzgo or $jny'on^3\ nda^3$ is one of the languages cited in the previous reports. Currently, the language has a couple of grammars and several studies about its varieties (see (Buck, 2018; Smith and Tapia, 2002; Palancar and Feist, 2015; Hernández et al., 2017)). In accordance with these specialized works, Amuzgo is quite different from the major language in the country (Spanish), which makes it more complex to directly apply techniques or tools from other languages. In this respect, this work describes an ongoing research about the creation of a dataset in Amuzgo. The dataset contains curated linguistic data collected from colloquial speech in Amuzgo. These data are presented considering the following phases or levels: acoustic signal processing, transcription, glossed and human translation into Spanish, semiautomatic alignment of human translations, and annotation. This dataset is intended to contribute the development of tools for scarce resources languages.

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 127–131
June 21, 2024

The rest of the article is organized as follows: Section 2 describes the main linguistic characteristics of Amuzgo, focusing on their complexity. Section 3 presents the method to create the dataset. Section 4 points out some of the results and applications so far. Finally, Section 5 summarizes our findings and highlights the future work.

2 Language description

Amuzgo or $jny'on^3$ nda^3 is spoken in Southern Mexico. It has over 60,000 speakers, according to the data reported by (INEGI, (2015). Grammatically, the language belongs to the Otomanguean family, alongside languages such as chatino, zapotec, mazatec, or popolaca. It is characterized by a wide set of personal pronouns (Buck, 2000; Palancar and Feist, 2015), and lexical classes, which impacts on its verbal complexity (Smith and Tapia, 2002). In addition, this is a language with five tones to mark change of meaning in different levels (Hernández, 2019). For instance, Examples (1) and (2) show how the different tones impact, not only on the meaning, but in the linguistic level as well.

(1) a. $nkia^3$ "to hit"
b. $nkia^5$ "fearful"

(2) a. ba'^1 "his/her house"
b. ba'^4 "your house"

While in (1), the difference between the low level tone and the high level tone (superscripts 3 and 5, respectively) impacts on a change in the lexical level; i.e. the tonal contrast produces two different types of words, in (2), the change moves to the morphological level; i.e. the tonal change (rising tone, superscript 1, and mid level tone, superscript 4) generates a clear distinction with respect to who possesses the house.

2.1 Linguistic complexity

This couple of examples stress how complex the language is, and somehow, how difficult is to apply tools or techniques that have shown their usefulness in other languages. Now, to summarize the challenges that this language entails, both in terms of its linguistic description and its possible digital implementation, we provide below some of the mos salient linguistic particularities reported in the specialized literature.

The phonological system in Amuzgo has 15 consonants and 7 vowels. Some of these phonemes are product of the contact with Spanish. As noted previously, its tonal characteristics impact on different linguistic levels: From phonology to pragmatics. In this respect, it has been reported the existence of five tones: rising, falling, low level, mid level, and high level. However, some works report two more tones; i.e. five level tones and two contour tones (Hernández, 2019).

On the other hand, Amuzgo organizes the grammatical persons in two classes: Singular and plural. Each class contains three persons morphologically marked, being the particular interest the third person in plural (they), which differentiates between inclusion and exclusion regarding the hearer.

Finally, in terms of morphosyntactic and syntactic particularities, Amuzgo is a head-marking language, but for the third persons; the syntactic relations are given by juxtaposition; and the syntactic template both for transitive and intransitive verbs is VSO.

3 Method

In this section, we describe the method to build the dataset. Specifically, we highlight how the dataset is curated by means of adding fine-grained linguistic information to the gathered data. The method consist of the following phases: i) collecting acoustic data from fieldwork, ii) acoustic data processing with Praat, iii) data transcription, iv) data glossed and human translation, v) semiautomatic alignment of human translations, and vi) data systematization.

3.1 Data collection

The data was collected by means of traditional linguistic fieldwork in one of the Amuzgo communities. Some of the researchers traveled to the community to interview a set of speakers in their natural environment. The Amuzgo speakers were asked some questions about how frequent they speak the language, the contexts in which they use it, their linguistic skills, both in Amuzgo and Spanish, whether or not they consider themselves monolingual or bilingual speakers, as well as information about their age, gender, time of residence in the community, whether or not they were migrants, and so on. In addition, they had to tell a story in Amuzgo. All the stories were digitally recorded. The amount of hours of this oral material is around 8 hours. Finally, it is important to remark that all the participants were informed about the use of the material in accordance with the Mexican laws. If they agreed[3], then they should sign a document. A more detailed explanation about this process is provided in the project GitHub site (see Section 3.5).

3.2 Acoustic data processing

The dataset here presented is based on the 8 hours of recordings. This oral material is, thus, the raw data from which we build the dataset.

Prior to adding any linguistic annotation, the oral material was acoustically processed with the software Praat (Boersma, 2014). This was done in order to get unbiased phonological information to be further analyzed, especially, given the tonal characteristics of the language reported in Section 2.

It is worth stressing that the acoustic signal processing is work in progress. So far, only three recordings have been fully processed. Once this phase is complete, it will be a major outcome of the research.

3.3 Data transcription

The following phase consisted in making the transcription of the oral data. Firstly, the stories were divided in segments to facilitate the transcription; i.e. we did not transcribe each story at once; on the contrary, we attempted to recognize informative segments or sentences to achieve a coherent text. In this recognition process, the work performed with the tools for speech analysis in Praat were highly helpful, due to they were a visual guide to segment the recordings properly.

The resulting text was linguistically analyzed in order to corroborate the initial segmentation. In case of unnatural divisions, for instance, incomplete sentences, the transcription was modified to adapt as much as possible each segment with a clause. This decision impacts directly on the glosses and the human translation, yet it did not affect the language grammaticality.

The last process of the transcription consisted in registering the Amuzgo phonological features, as well as incorporating the tones. The following text is a sample of the resulting transcription.

Twe'^{3} nkwi^{3}xue^{12} m'an^{3} kwi^{3}ti'^{12}tyo^{3}ndye^{35} ts'a^{3} ti'^{12},
ts'ian^{5} jndë^{12}, tyua'^{12} ju'^{35} sku'^{35} ti'^{12} k'a^{35} ti'^{12} jndë^{12},
Mo'^{12} twe'^{3} nkwi^{3} xue^{5} t-ja^{3} ti'^{12},
të^{1}ki^{3}tsa^{35} ti'^{12} ts'ian^{5},
no'^{1} ya^{12} tje^{35} ti'^{12} tyua'^{12}je^{12},

3.4 Glosses and human translation

In these two processes, glosses and human translation, rely the major richness of the dataset. With respect to the former, it is well-known that a glossed process entails a syntactic segmentation (just like the one attempted in the previous phase), as well as a part of speech annotation. This linguistic information contributes to enhance the description and understanding of any language, due it provides an in-depth vision of its linguistic relations. In addition, the glosses provide formal elements that can be used to perform a better translation.

The glosses were done manually, i.e. a human expert in Amuzgo analyzed the transcriptions, made a re-alignment of the unnatural segments proposed in the previous phase, and performed the glossed process according to the Leipzig Glossing Rules (Comrie et al., 2008). Furthermore, the expert considered the following elements when generating the glosses: use of a consistent orthographic system (Hernández, 2019), distinction between phonological word and lexical entry (see Examples (1) and (2)), and clitic marking.

The following process was a human translation, which was guided by the linguistic information registered in the glosses. Figure 1 illustrates the result of the glossing and human translation.

3.5 Translation alignment

This phase was performed to enhance the scope of the dataset. Mainly, by creating a parallel corpus Amuzgo-Spanish, which was generated semi-automatically by applying the Gale-Church algorithm for translation alignment (Gale and Church, 1993). This algorithm is implemented in the CAT tool OmegaT and it considers two methods: Parsewise and heapwise. Each one makes the alignment taking into consideration different features in the texts. For instance, a possible syntactic parallelism regarding the parsewise method, or a global textual integration regarding the heapwise method. We tried with both methods and the result was unsatisfactory. This is obviously due to the huge linguistic differences between both languages.

Despite the poor performance, the results were used as a source to manually improve the alignments. This human improvement produced a better parallel corpus, which is freely available in this address: `https://github.com/areyesp-77/amuzgo-dataset.git`.

1. Twe’^{3} nkwi^{3}xue^{12} m’an^{3} kwi^{3}ti’^{12}tyo^{3}ndye^{35} ts’a^{3} ti’^{12}

T-we’^{3}	nkwi^{3}=xue^{12}	m’an^{3}	kwi^{3}	ti’^{12}=tyo^{3}ndye^{35}	ø-ts’a^{3}	ti’^{12}
CPL-haber.3SG	ART.INDEF.SG=día	HAB.estar.3SG	uno	compañero=zorro	PROG-hacer[3SG]	compañero

Hubo una vez un zorro

2. ts’ian^{5} jndë^{12}, tyua’^{12} ju’^{35} sku’^{35} ti’^{12} k’a^{35} ti’^{12} jndë^{12}.

ts’ian^{5}	jndë^{12}	tyua’^{12}	ø-ju’^{35}	sku’^{12}	ti’^{12}	ø-k’a^{35}	ti’^{12}	jndë^{12}
trabajo	monte	temprano	PROG-moler. 3SG	esposa[3SG]	compañero	HAB.ir[3SG]	compañero	monte

que trabajaba en el campo, temprano molía su esposa [y] él iba al monte.

3. Mo’^{12} twe’^{3} nkwi^{3} xue^{5} t-ja^{3} ti’^{12}

mo’^{12}	t-we’^{3}	nkwi^{3}	xue^{5}	t-ja^{3}	ti’^{12}
pero	CPL-haber.3SG	uno	día	CPL-ir[3SG]	compañero

Pero hubo un día

4. të^{1}ki^{3}tsa^{35} ti’^{12} ts’ian^{5}

të^{1}-ki^{3}-tsa^{35}	ti’^{12}	ts’ian^{5}
CPL-CAUS-hacer[3SG]	compañero	trabajo

que fue al trabajo

5. no’^{1} ya^{12} tje^{35} ti’^{12} tyua’^{12}je^{12}

no’^{1}	ya^{12}	t-je^{35}	ti’^{12}	tyua’^{12}=je^{12}
y	cuando	CPL-llegar[3SG]	compañero	temprano=INT

y cuando llegó más temprano a [su] casa

Figure 1: Sample of the glosses and human translation.

3.6 Data systematization

The last phase to create the dataset implies the integration of the linguistic information in a lexical resource to expand the possibilities of investigation, as well as to start applying some NLP techniques in tasks as diverse as machine translation, part of speech tagging, speech recognition, and so on. To this end, we have been working in systematizing and formalizing our data in a unique file to simplify, as much as possible, all the fine-grained linguistic information that we have.

In this respect, a beta version of this curated dataset with the following information has been released: lexical entry in Amuzgo, POS annotation, tone annotation, linguistic processes identified, translation into Spanish, and source (recording). This version is available at: www.geco.unam.mx.

4 Result and projection

The outcome of all the processes described so far is a dataset with fine-grained linguistic information in Amuzgo. Beyond the traditional linguistic analysis, this information could be used to foster and increase the efforts of the NLP community regarding scarce resources languages, such as $jny'on^3$ nda^3.

In addition, a small parallel corpus Amuzgo-Spanish has been generated to freely explore some characteristics of the language by using common corpus analyses, such as concordances, collocations, or keywords.

As an optimistic projection, the creation of resources like this dataset would allow the experimentation in different areas, the generation of new knowledge, the preservation of endangered languages, and the minimization of the digital divide and its negative consequences for the native communities.

5 Conclusions

In this article we have described the creation of a linguistic dataset in Amuzgo. The dataset contains curated information provided by specialists in that language. Such information covers different linguistic levels, such phonology, morphology, and syntax, as well as a human translation into Spanish. In addition, the dataset was used to generate a small parallel corpus Amuzgo-Spanish by applying a known algorithm in NLP. Both resources are freely available for academic purposes.

Although the materials exceed the 8 hours of recordings, the dataset here described contains only the data of one hour. Therefore, as further work, it is planned to process the remaining data to enhance the linguistic description and, accordingly, to improve the dataset.

References

ASTLM. 2018. *Análisis del sector de las Tecnologías del lenguaje en México*. Gobierno de España.

Paul Boersma. 2014. The Use of Praat in Corpus Research. In Ulrike Gut Jacques Durand, and

Gjert Kristoffersen, editor, *The Oxford Handbook of Corpus Phonology*. Oxford University Press.

M. Buck. 2000. *Gramática del amuzgo de San Pedro Amuzgos*. Instituto Lingüístico de Verano.

M. Buck. 2018. *Gramática del amuzgo de Xochistlahuaca*. Instituto Lingüístico de Verano.

B. Comrie, M. Haspelmath, and B. Bickel. 2008. *The Leipzig Glossing Rules: Conventions for Interlinear Morpheme-by-morphene Glosses*. Max Planck Institute for Evolutional Anthropology.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig (eds.). 2024. *Ethnologue: Languages of the World*. SIL International.

W. Gale and K. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

N. Hernández. 2019. *El sistema tonal en el amuzgo de San Pedro Amuzgos: Interacción entre el tono de la base nominal y los clíticos*. Tesis de Maestría en Lingüística Indoamericana, Ciudad de México.

N. Hernández, A. Mora, and H. García. 2017. Estructura de la frase nominal posesiva en amuzgo (otomangue). *UniverSOS. Revista de Lenguas Indígenas y Universos Culturales*, 14:63–82.

INALI. 2008. Catálogo de las lenguas indígenas nacionales: variantes lingüísticas de México con sus autodenominaciones y referencias geoestadísticas. *En Diario Oficial de la Federación*.

INEGI. (2015). Encuesta intercensal 2015. url =.

E. Palancar and T. Feist. 2015. Agreeing with subjects in number: The rare Split of Amuzgo verbal inflection. *Linguistic Typology*, 93(3):337–383.

T. Smith and F. Tapia. 2002. Amuzgo como lengua activa. pages 81–129. Del cora al maya yucateco. Estudios lingüísticos sobre algunas lenguas indígenas mexicanas.

Enenlhet as a case-study to investigate ASR model generalizability for language documentation

Éric Le Ferrand
Boston College
leferran@bc.edu

Raina Heaton
University of Oklahoma
rainaheaton@ou.edu

Emily Prud'hommeaux
Boston College
prudhome@bc.edu

Abstract

Although both linguists and language community members recognize the potential utility of automatic speech recognition (ASR) for documentation, one of the obstacles to using these technologies is the scarcity of data necessary to train effective systems. Recent advances in ASR, particularly the ability to fine-tune large multilingual acoustic models to small amounts of data from a new language, have demonstrated the potential of ASR for transcription. However, many proof-of-concept demonstrations of ASR in low-resource settings rely on a single data collection project, which may yield models that are biased toward that particular data scenario, whether in content, recording quality, transcription conventions, or speaker population. In this paper, we investigate the performance of two state-of-the art ASR architectures for fine-tuning acoustic models to small speech datasets with the goal of transcribing recordings of Enenlhet, an endangered Indigenous language spoken in South America. Our results suggest that while ASR offers utility for generating first-pass transcriptions of speech collected in the course of linguistic fieldwork, individual vocabulary diversity and data quality have an outsized impact on ASR accuracy.

1 Introduction

The fields of descriptive and documentary linguistics concentrate on collecting and analyzing language samples, particularly from understudied, Indigenous, and endangered languages. Typically, documentary linguists – who can be researchers or language community members – make audio recording of unscripted or prompted speech, followed by transcription, glossing, translation, and analysis. Transcription, however, often becomes bottleneck when dealing with large speech corpora, rendering only a fraction of the available speech data available for analysis or for language instruction (Himmelmann, 1998).

Automatic speech recognition (ASR) has emerged as a potential solution by providing first-pass transcripts that can be manually corrected (Mitra et al., 2016; Bird, 2021; Jimerson and Prud'hommeaux, 2018). The preferred approach for building an ASR system with scarce resources is to fine-tune a large multilingual model to whatever small amount of transcribed audio data is available for the target language. Many demonstrations of the efficacy of this approach, however, rely on corpora with relatively few speakers or with recordings made under the same condition (e.g., all read speech or broadcast news) (Jimerson et al., 2023). One problem is that models trained on a single uniform speech corpus may overfit to that corpus acoustically or lexically. It is not clear how such models will generalize to new data – whether that data is archival recordings or recent data from a different speaker population or data collected with different prompts.

In this paper, we address this question using a corpus of speech recordings for the language Enenlhet (ISO 639-3 code `tmf`; not to be confused with the related language Enlhet, ISO 639-3 code `enl`). Enenlhet is spoken by fewer than 2,000 people living in what is now Paraguay. While thousands of the world's languages, like Enenlhet, are endangered and have minimal written documentation, many of these languages lack three important characteristics that make Enenlhet an ideal language for exploring the utility of ASR for documentation of diverse speech data. First, the amount of available transcribed speech data – 5 hours – is relatively substantial for an endangered and primarily oral language. Second, the large quantity of untranscribed audio – over 100 hours – is highly unusual for any endangered language, offering the potential for unsupervised training and for experimentation with integrating ASR into the documentation pipeline. Third, the Enenlhet speakers who provided their voices have stated enthusiasm for generating new

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 132–137
June 21, 2024

documentation for their language in collaboration with outsiders.

Using two ASR architectures that support fine-tuning acoustic models to the task of ASR for small speech datasets, we explore to what extent an existing corpus can be used to train models that generalize well to new data. The dataset we use, while part of a single data collection effort, was collected over multiple years under varying conditions from a large number of speakers. We simulate introducing a new recording by holding out each speaker in turn, training on the remaining speakers, and testing on the held-out speaker. We find that the high degree of lexical diversity across speakers, as well as differences in audio and transcription quality, contribute to variability in word error rate (WER), findings we quantify with a regression analysis.

2 Related work

ASR has long been proposed as a solution to the "transcription bottleneck" challenge of language documentation, but there has been relatively little effort dedicated to practically using ASR for this purpose. The focus of much of the early work was phone-level transcription (DiCanio et al., 2013; Johnson et al., 2018; Zahrer et al., 2020). Other applications have involved keyword spotting (Le Ferrand et al., 2022) or the development of front-end tools for building ASR systems (Foley et al., 2018). Only recently has ASR been actually used in an active documentation pipeline (Prud'hommeaux et al., 2021; Gupta and Boulianne, 2020; Shi et al., 2021; Rodríguez and Cox, 2023). There is some prior work investigating data partitioning strategies (Liu et al., 2023), which we indirectly address in our work when we use a held-out-speaker approach to simulate testing a trained model on new data. Le Ferrand et al. (2023) applied a trained model for an under-resourced language to new data, yielding surprisingly weak results and indicating that models fine-tuned on small amounts of data may not generalize well to new data. We also note previous work on the impact of specific dataset characteristics, including OOV rate and audio quality which we explore here, on word error rate (Jimerson et al., 2023). This last paper includes data from the AmericasNLP 2022 ASR shared task (Ebrahimi et al., 2022). While the AmericasNLP datasets were extremely small (typically less than one hour), they contained fieldwork recordings with characteristics similar to those included in our study.

3 Data

The language of the corpus used in this study is known by a number of names (Cabanatit, Enenxet, Toba-Enenlhet), but following the preferences of the community, we will refer to it here as Enenlhet (ISO 639-3 code tmf). Enenlhet is a polysynthetic language spoken by fewer than 2,000 people living in the Paraguayan Chaco region. Migration and displacement have led to dramatic language loss; the current Ethnologue status of Enenlhet is 6b (Threatened). Enenlhet has remarkably little available documentation. Aside from a few short word lists compiled in the 1920s and 1960, there are no dictionaries, and there is only one available grammar (Unruh et al., 2003). A phone set of approximately 4 vowels and 15 consonants can be inferred from the dataset described below.

The data used here is part of a recent multi-year data collection effort, which has so far yielded over 120 hours of recordings with more than 40 individuals. The data was collected with university IRB approval and is archived with the Archive of Indigenous Languages of Latin America at the University of Texas. Ethical practices require a consultation with the language community before using language material for research purposes (Pirinen et al., 2024). Thus, while the data we use is publicly available[1], the co-author who collected the data gained express permission from the community for research purposes. She notes that the Enenlhet speakers who participated in the data collection were eager for their speech recordings to be used to support documentation and revitalization efforts.

Approximately 10 hours of the recordings from 16 speakers have been transcribed with utterance-level timestamps. The total quantity of speech data available in these recordings – after stripping out silence, segments in another language, and speech produced by the interviewer – is approximately 5 hours. Table 1 shows information for each speaker.

4 Methods

4.1 Experiments

We trained ASR models using two frameworks that support fine-tuning from a multilingual model: Whisper (Radford et al., 2022) and wav2vec (Baevski et al., 2020; Conneau et al., 2021). In the case of Whisper, we used the Whisper medium

[1] https://ailla.utexas.org/islandora/object/ailla%3A266554

	SSA	CA	ER	IF	PA	OM	HM	FF	TF	MRR	MM	AR	LM	BT	MR	LF
duration	80:16	65:00	5:40	20:32	18:51	26:18	10:45	39:00	5:20	12:21	4:53	7:00	4:21	2:42	3:00	2:48
tokens	6663	5828	434	1495	1571	2139	756	3192	362	868	509	814	349	209	253	201
types	2181	1769	228	513	847	965	437	1448	229	399	261	214	164	134	144	117

Table 1: Duration (MM:SS), token count, and type count for each of the 16 speakers in our dataset.

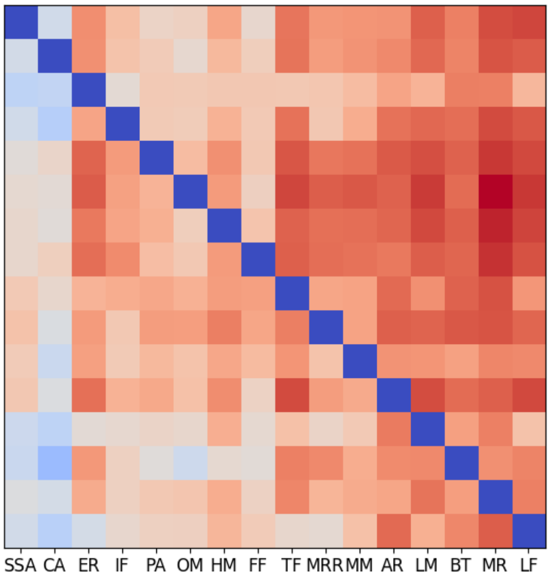


Figure 1: OOV heatmap: Blue indicates low OOV rate while red indicates high OOV rate, with the hue indicating more extreme values in each respective direction.

model, adhering to the hyperparameters specified in the main tutorial[2]. For wav2vec, we employed xlsr-53, following the hyperparameters of the main tutorial[3]. Regarding wav2vec decoding, we decoded with a language model trained on the transcripts of the relevant training data. Notably, we opted for default values for decoding parameters α and β given their minimal impact with small LMs.

Recall that the recordings used were collected over several years, in different recording conditions from different individuals. Our goal is not to create a robust ASR system, but rather to assess whether a model trained on existing data will generalize to a new speech recording or corpus. Initially, we train a baseline model by randomly partitioning the entire dataset into training and testing sets. Subsequently, we use a "leave speaker out" cross validation approach to simulate the testing of new data on an existing trained ASR model. For each speaker within our dataset, we train an ASR model using all recordings except for the those of the target speaker, whose data is reserved for testing purposes.

[2]https://huggingface.co/blog/fine-tune-whisper

[3]https://huggingface.co/blog/fine-tune-xlsr-wav2vec2

4.2 Analysis

The ASR experiments are evaluated using the traditional Word Error Rate (WER) metric. We then aim to identify factors that could impact system performance. First, we focus on the lexicon, examining two factors: the Out-of-Vocabulary (OOV) rate, which represents the proportion of tokens in the test set that did not appear in the training set (see Figure 1). The blue cells, corresponding to the longest recordings SSA and CA, have a substantially larger vocabulary that overlaps with the rest of the collection. We perform the same analysis with the types. We then calculate the duration of both the training and testing sets. Following this, we assess the audio quality in the test sets based on two measures: loudness and sharpness. Loudness is a measure designed to mimic sound perception in humans, while sharpness relates to the subjective perception of high-frequency content in a sound. Both sharpness and loudness are determined using the Zwicker method (Zwicker, 1960) with the Mosquito toolkit[4]. Finally, we evaluate the transcription quality by conducting a CTC-based alignment of the transcription and utilizing the resulting CTC posterior probabilities as a measure of transcription reliability. Our intuition is that low CTC probabilities indicate that the alignment algorithm had difficulty determining the alignment, perhaps because of noisy recordings or inconsistent transcriptions. To perform the alignment we used an ASR model trained in English (wav2vec2_ASR_base_960h).

5 Results

Results are presented in Figure 2. Each bar corresponds to a test conducted on a different speaker's data. The baselines are indicated by the dotted lines. First, we see across all scenarios, wav2vec performs systematically worse than Whisper. Second, we observe in all experiments, the baseline does not exhibit consistent inferior or superior performance in either architecture. We note that potential biases during experiments conducted on random splits do not significantly impact overall performance in one

[4]https://github.com/Eomys/MoSQITo/tree/master

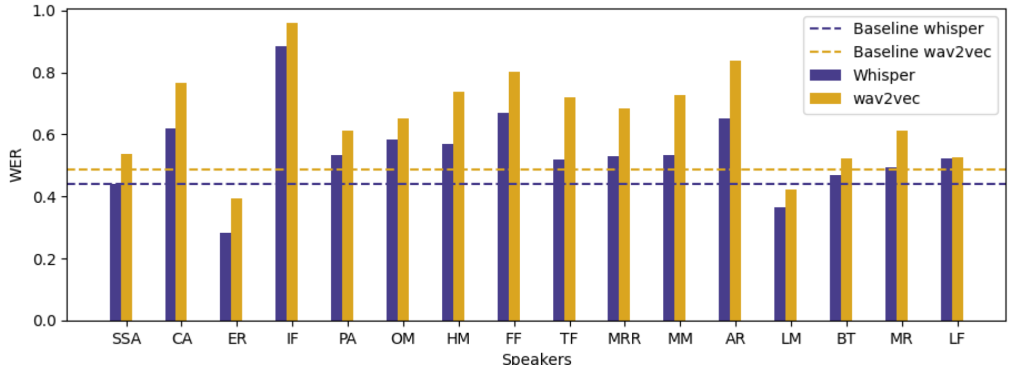


Figure 2: WER across all speakers. Baselines are derived from a random split across all speakers.

	OOV tokens	OOV types	train duration	test duration	loudness	sharpness	alignment score
Coeff.	0.79	0.36	-0.001	0.001	0.001	0.134	-1.588
95% CI	(0.07, 1.1)	(0.07,1.4)	(0.07,184.)	(0.07,12.3)	(0.07, 0.67)	(0.07,2.04)	(0.07,18.9)

Table 2: regression results with WER results derived from Whisper. CI stands for Confidence Interval.

	OOV tokens	OOV types	train duration	test duration	loudness	sharpness	alignment score
Coeff.	1.171	0.452	-0.001	0.001	0.000	0.301	-2.547
95% CI	(0.08, 1.2)	(0.08,1.6)	(0.08,210.)	(0.08,14.)	(0.08, 0.76)	(0.08,2.25)	(0.07,20.5)

Table 3: regression results with WER results derived from wav2vec. CI stands for Confidence Interval.

direction or the other. Secondly, performance is dependent on factors inherent to the test speaker's data. The regression analysis enabled us to ascertain the impact of these factors on WER.

The results of the regression analysis can be found in Table 2 for Whisper and Table 3 for wav2vec. A significantly positive coefficient value indicates that the factor leads to a higher (worse) WER while a significantly negative coefficient indicates that the factor leads to a lower (better) WER. Across both architectures, three factors do not have a significant influence on WER: duration of the training data and test data, and loudness. One of the most relevant factors is the OOV tokens and to a lesser extent, the OOV types. These factors happen to be much more salient for wav2vec than for Whisper which can perhaps explained by the use of a language model during decoding in wav2vec.

Two metrics were used evaluate the quality of the audio in the test sets. Loudness was found to have minimal impact, but sharpness appears to negatively impact WER. Additional experiments on larger datasets are necessary to validate the efficacy of this measure in assessing audio quality.

The posterior probabilities obtained from the CTC alignment exhibited a strong negative impact on WER, suggesting that a higher confidence score in the alignment corresponds to a lower WER. However, the confidence interval is relatively high, raising doubts about the reliability of this measure to evaluate the quality of the transcription. Examining specific examples, we verified the data quality for IF, where a very high WER was observed. It was discovered that the transcription for this speaker did not align with the audio; instead, it appeared to be a translation of the audio into a related language or dialect. However, the CTC alignment did not substantially differ from other speakers where the transcription matched well and the WER is much lower. This measure appears instead to be relevant for evaluating audio quality when there is not a significant mismatch with the transcription.

6 Conclusions

This paper explores how contemporary speech recognition architectures perform in a language documentation setting, focusing on the Enenlhet language as a case study. In order to simulate testing of new data using a model trained on a previous data collection corpus, we conducted training and testing of ASR models using a leave-one-out evaluation approach, where the models were trained on all Enenlhet speakers except one and tested on the one left out. Additionally, we performed a regression analysis to determine the factors that may influence WER.

The experimental results initially revealed that the leave-one-out evaluation approach neither outperforms or underperforms a random split approach for our specific case. Subsequently, we found that Out-of-Vocabulary (OOV) rates are the most significant factor in explaining the WER for a given test set. Lastly, both the sharpness measure and the CTC posterior probabilities show promise in assessing the quality of the speech signal, which could potentially correlate with the word error rates. Further analysis is necessary to confirm this correlation. These results suggest that in low-resource settings, ASR models may not always generalize well to new data, which could hamper the utility of ASR for language documentation.

References

Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A framework for self-supervised learning of speech representations. In *Advances in Neural Information Processing Systems*, volume 33, pages 12449–12460.

Steven Bird. 2021. Sparse transcription. *Computational Linguistics*, 46(4):713–744.

Alexis Conneau, Alexei Baevski, Ronan Collobert, Abdelrahman Mohamed, and Michael Auli. 2021. Unsupervised cross-lingual representation learning for speech recognition. In *Proceedings of Interspeech*, pages 2426–2430.

Christian DiCanio, Hosung Nam, Douglas H Whalen, H Timothy Bunnell, Jonathan D Amith, and Rey Castillo García. 2013. Using automatic alignment to analyze endangered language data: Testing the viability of untrained alignment. *The Journal of the Acoustical Society of America*, 134(3):2235–2246.

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

Ben Foley, Joshua T Arnold, Rolando Coto-Solano, Gautier Durantin, T Mark Ellison, Daan van Esch, Scott Heath, Frantisek Kratochvil, Zara Maxwell-Smith, David Nash, et al. 2018. Building speech recognition systems for language documentation: The coedl endangered language pipeline and inference system (elpis). In *SLTU*, pages 205–209.

Vishwa Gupta and Gilles Boulianne. 2020. Speech transcription challenges for resource constrained indigenous language Cree. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 362–367.

Nikolaus P. Himmelmann. 1998. *Documentary and descriptive linguistics*, volume 36. de Gruyter.

Robbie Jimerson and Emily Prud'hommeaux. 2018. Asr for documenting acutely under-resourced indigenous languages. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Robert Jimerson, Zoey Liu, and Emily Prud'hommeaux. 2023. An (unhelpful) guide to selecting the best asr architecture for your under-resourced language. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1008–1016.

Lisa M Johnson, Marianna Di Paolo, and Adrian Bell. 2018. Forced alignment for understudied language varieties: Testing prosodylab-aligner with tongan data.

Éric Le Ferrand, Fabiola Henri, Benjamin Lecouteux, and Emmanuel Schang. 2023. Application of speech processes for the documentation of kréyòl gwadloupéyen. In *Proceedings of the Second Workshop on NLP Applications to Field Linguistics*, pages 17–22.

Éric Le Ferrand, Steven Bird, and Laurent Besacier. 2022. Learning from failure: Data capture in an Australian aboriginal community. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4988–4998.

Zoey Liu, Justin Spence, and Emily Prud'hommeaux. 2023. Investigating data partitioning strategies for crosslinguistic low-resource asr evaluation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 123–131.

Vikramjit Mitra, Andreas Kathol, Jonathan D Amith, and Rey Castillo García. 2016. Automatic speech transcription for low-resource languages-the case of yoloxóchitl mixtec (mexico). In *INTERSPEECH*, pages 3076–3080.

Flammie Pirinen, Linda Wiechetek, Trond Trosterud, Sjur Moshagen, and Børre Gaup. 2024. Computel partnerships in practice: Giellalt. In *Proceedings of 7th the Workshop on Computational Methods for Endangered Languages*.

Emily Prud'hommeaux, Robbie Jimerson, Richard Hatcher, and Karin Michelson. 2021. Automatic speech recognition for supporting endangered language documentation. *Language Documentation and Conservation*, 15:491–513.

Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. 2022. Robust speech recognition via large-scale weak supervision. *arXiv preprint arXiv:2212.04356*.

Lorena Martín Rodríguez and Christopher Cox. 2023. Speech-to-text recognition for multilingual spoken data in language documentation. In *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 117–123.

Jiatong Shi, Jonathan D. Amith, Rey Castillo García, Esteban Guadalupe Sierra, Kevin Duh, and Shinji Watanabe. 2021. Leveraging end-to-end ASR for endangered language documentation: An empirical study on Yoloxóchitl Mixtec. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1134–1145.

Ernesto Unruh, Hannes Kalisch, and Manolo Romero. 2003. *Enenlhet apaivoma: nentengiai'a nengiangveiakmoho neliatekamaha enenlhet apaivoma, guía para el aprendizaje del idioma materno toba*. Nengvaanemkeskama Nempayvaam Enlhet.

Alexander Zahrer, Andrej Žgank, and Barbara Schuppler. 2020. Towards building an automatic transcription system for language documentation: Experiences from muyu. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2893–2900.

Eberhard Zwicker. 1960. Ein verfahren zur beredinung der lautstärke. *Acta Acustica united with Acustica*, 10(4):304–308.

Advancing NMT for Indigenous Languages: A Case Study on Yucatec Mayan and Chol

Julio C. Rangel and **Norio Kobayashi**
RIKEN Information R&D and Strategy Headquarters,
2-1 Hirosawa, 351-0198 Wakoshi, Japan
{juliocesar.rangelreyes, norio.kobayashi}@riken.jp

Abstract

This study leverages Spanish-trained large language models (LLMs) to develop neural machine translation (NMT) systems for Mayan languages. For that, we first compile and process a low-resource dataset of 28,135 translation pairs of Chol and Yucatec Mayan extracted from documents of the CPLM Corpus (Martínez et al.). Then we implement a prompt-based approach to train one-to-many and many-to-many models. By comparing several training strategies for two LLMs, we found that, on average, training multilingual models is better, as shown by the ChrF++ reaching 50 on the test set in the best case. This study reinforces the viability of using LLMs to improve accessibility and preservation for languages with limited digital resources. We share our code, datasets, and models to promote collaboration and progress in this field [1].

1 Introduction

In recent times, there has been a push towards creating NLP tools for the native languages of the Americas (Mager et al., 2023). Within this context, Mayan languages have not received attention in machine translation (NMT) studies, despite their deep linguistic roots and large speaker populations. Our study aims to bridge this gap by specifically developing and refining NMT systems for Mayan languages. By leveraging advancements in large language models (LLMs) pre-trained in Spanish, we aim to overcome the scarcity of a comprehensive parallel corpus for Mayan languages. As a result, building NMT systems for languages could greatly benefit these language communities by enabling them to access services and information related to law, healthcare, and finance in their mother tongues.

[1]https://github.com/RIKEN-DKO/iikim_translator

2 Languages

The Mayan languages form a family spoken by the Maya peoples, is primarily spoken across various regions in Central America. This family stands as among the most thoroughly researched and documented in the Americas (Campbell, 2000). It is believed that the contemporary Mayan languages originated from the Proto-Mayan language, which was likely spoken over 5,000 years ago. This ancient language eventually branched out into at least six distinct lineages: Huastecan, Quichean, Yucatecan, Qanjobalan, Mamean, and Ch'olan–Tzeltalan [2]

2.1 Yucatec Mayan

Yucatec Mayan, commonly referred to as Maya is a language spoken in the Yucatán Peninsula and the northern regions of Belize. Being one of the Mayan languages Yucatec Mayan plays a vital role, in connecting us to the diverse cultural and historical legacy of the Mayan civilization. Unlike indigenous languages Yucatec Mayan boasts a substantial number of speakers with an estimated count of approximately 800,000 individuals [3].

2.2 Chol

The Chol people, a group, in Mexico mainly live in the mountains of Chiapas. Being part of the Maya community they speak Ch'ol or Chol which belongs to the Mayan language group. Ch'ol has three dialects (Sabanilla, Tilá and Tumbalá), these dialects,often considered a single language showcasing the language's vitality and regional diversity. Had approximately 140,806 speakers, in 2000 including individuals who speak only this language[4].

This paper outlines the design and implementation of a comparative study on two sophisticated

[2]https://en.wikipedia.org/wiki/Mayan_languages
[3]https://en.wikipedia.org/wiki/Yucatec_Maya_language
[4]https://en.wikipedia.org/wiki/Chol_people

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 138–142
June 21, 2024

neural machine translation (NMT) models, T5S (T5 Spanish) and M2M100, specifically tailored for the translation of Yucatec Mayan and Chol languages. We focus mainly on comparing the models' accuracy in translating to the Mayan languages, aiming to determine the most effective approach for developing NMT systems that can serve as a starting point for future Mayan-based NMT systems in low and high resource instances.

3 Methodology

3.1 Dataset

In support of our research, we gathered a dataset comprising 28,135 translation pairs from Spanish to Chol languages using the CPLM (Parallel Corpus, for Mexican Languages) web tool[5]. The data extraction process involved downloading ZIP files, each potentially containing multiple files with parallel sentences in Spanish and one or more target languages. We utilized the *langdetect*[6] library to verify the presence of Spanish; ZIP files without Spanish were excluded. To identify relevant files for Yucatec Mayan and Chol, we looked for language codes 'yua' and 'MY' for Mayan and 'ctu' and 'CHL' for Chol. If codes were absent, we searched for language names such as 'maya' and 'chol'. Finally, we aligned the files to create Spanish-to-Chol and Spanish-to-Mayan parallel datasets. The number of parallel pairs per language is shown in Table 1.

3.2 Data Preparation

Inspired by previous NMT systems for Indigenous Languages (De Gibert et al., 2023), in the post-processing phase, we applied a length ratio filter to improve the quality of our translation pairs, removing any with a character length ratio exceeding 4. This filtering step was critical for maintaining a high-quality dataset by excluding pairs that could adversely affect translation accuracy. We then randomly divided the sentences into training, development, and testing sets. The results of this data preparation phase, including the final counts of translation pairs, are detailed in Table 1.

4 Models

Our selection criteria focused on recent models with extensive pretraining in Spanish, as evidence suggests this significantly aids in translating to native languages (Vázquez et al., 2021). Accordingly, we selected the T5S (T5 Spanish) and M2M100 (480M version) models for our translation tasks. While both models adopt the encoder-decoder architecture foundational to Transformer models, they are distinguished by their underlying philosophies and optimizations.

This research aims to compare two approaches to language models (LLMs): T5S, which is versatile for various NLP tasks, and M2M100, which is specialized for translation purposes. This comparison intends to evaluate how well a general model like T5S can handle low resource translation scenarios and determine the performance of M2M100 in translating between less commonly spoken languages. Through this method, we aim to identify which model design and training approach are most effective in creating Mayan NMT systems.

Language	Original # Pairs	Cleaned	Train	Dev/Test
maya-spanish	16149	13528	12176	1352
chol-spanish	11986	10660	9594	1066

Table 1: Summary of the dataset used for training and testing the NMT models.

4.1 T5S (T5 Spanish)

The T5 model, recognized for addressing a range of NLP tasks as text-to-text conversions—including translation, question answering, and classification—generates target text from input (Raffel et al., 2020). Its variant, IndT5, has been applied for translating Spanish into 10 Indigenous languages (Nagoudi et al., 2021). We utilize T5S (Araujo et al., 2023), an iteration adapted from T5.1.17, featuring an encoder-decoder structure with 12 layers, 12 attention heads, and 768 hidden dimensions. T5S was pretrained on Spanish data totaling approximately 674GB, comprising the OSCAR 21.09 corpus (160GB), mC4-es corpus (500GB), and SUC corpus (14GB).

4.2 M2M100

We use the M2M100 (480M) model, with 12 encoders, 16 decoder layers,a feed forward network (FFN) size of 4096, and embedding dimensions of 1024 that have been optimized for machine translation. It allows for translation among 100 languages, including Spanish, without requiring a language. Trained on a dataset of over 1.5 billion sequences (Fan et al., 2020) it aligns with our strategy of

[5] http://www.geco.unam.mx/
[6] https://github.com/Mimino666/langdetect

utilizing pre trained Spanish language models. Previous studies have showcased the effectiveness of M2M100 in translating to languages like Mixtec (Tonja et al., 2023) along with its performance in tasks such as the AmericasNLP 2024 Shared Task (Stap and Araabi, 2023).

5 Experiments and Results

This section outlines our training approach, experimental setup, and the results obtained from deploying various strategies on the T5S and M2M100 models.

5.1 Training Methodology

For our experiments, both models were trained until no improvement was observed for three consecutive epochs on the development set, with the best-performing checkpoint on this set being used for testing. To enable a single model to translate between multiple languages, we adopted a prompt format of "`{source_text} translate {source_lang} to {target_lang}: {source_text}`" before tokenization and training commenced. This approach facilitated the development of models capable of one-to-many (Spanish to Indigenous languages and vice versa) translations.

5.2 Experimental Setup and Results

Table 2 summarizes the experimental results, presenting both models' performances across different training configurations. "Mayan and Chol" refers to a one-to-many model trained on both languages. In contrast, "Mayan" and "Chol" indicate models trained exclusively on a single language. The "Zero shot" configuration evaluates model performance without fine-tuning. All models were trained for translation from Spanish to a native language, except those with prefixes 'bi', indicating bi-directional training.

In addition to the base dataset, we explored the impact of augmenting it with additional data from the Americas NLP2023 Shared Task[7] (AmeNLP), which introduces 11 more target languages. The inclusion of AmeNLP data initially led to a decrease in performance metrics for both models. However, implementing a uniform sampling strategy mitigated this degradation for combinations of Mayan languages with AmeNLP data but was less effective for Mayan and Chol alone. This observation suggests that the uniform sampling strategy is more advantageous when a model is trained across multiple datasets.

The M2M100 model outperformed T5S in translating Mayan and Chol languages, likely due to its specialization in translation tasks. For both models, the best average results were achieved when training with the Mayan and Chol datasets combined in a one-to-many approach. Interestingly, M2M100 showed a slight improvement on the Chol test set with the inclusion of "AmeNLP + uniform", suggesting that this strategy holds promise for enhancing multilingual model performance with additional data sources. The "biAmeNLP + uniform" strategy did not yield as positive results, possibly due to the requirement of specifying a target language tag for M2M100, which our Indigenous languages lack. Further investigation is needed to fully understand this aspect, despite indications that translation quality remains consistent irrespective of the chosen target language tag (Stap and Araabi, 2023).

For T5S, the "Maya and Chol" configuration was confirmed as the most effective strategy, with "biAmeNLP + uniform" emerging as the second-best approach. This suggests that for T5S, a bidirectional model is preferable, potentially because T5S does not necessitate explicit source or target language tags.

6 Conclusion and Future Work

This study has successfully demonstrated that the T5S and M2M100 models can be adapted for translation tasks between Spanish and Mayan languages, showcasing the potential of neural machine translation (NMT) in enhancing language preservation and accessibility. The M2M100 model, with its translation-focused architecture, excels in one-to-many translation scenarios. Conversely, the T5S model shows versatility in managing bidirectional translations, benefiting from its flexible design.

The incorporation of the "AmeNLP + uniform" strategy has emerged as a promising method to broaden the models' capabilities across multiple languages, though it introduces challenges that necessitate further exploration. Initial experiments have validated the potential of NMT for Yucatec Mayan and Chol, with both models performing effectively in low-resource settings. Despite the variation in translation quality, the results affirm the capacity of these models to acquire meaningful

[7] https://github.com/AmericasNLP/americasnlp2023

Table 2: Comparative performance metrics. **Bold** denotes overall best; underscore for best ST5 results.

Dataset	Set	Maya ChrF++	Maya BLEU	Chol ChrF++	Chol BLEU	Average ChrF++	Average BLEU
T5S							
Mayan and Chol	dev	32.69	10.17	34.6	11.51	33.645	10.84
	test	33.13	10.39	35.1	11.42	34.115	10.905
biAmeNLP + uniform	dev	29.53	8.23	32.92	10.35	31.225	9.29
	test	29.91	8.11	33.03	10.17	31.47	9.14
AmeNLP + uniform	dev	21.27	4.38	23.24	5.1	22.255	4.74
	test	21.1	4.01	23.42	5.02	22.26	4.515
Mayan	dev	27.63	7.09				
	test	27.52	6.99				
Chol	dev			28.1	7.83		
	test			28.7	8.03		
Zero shot	dev	7.68	0.13	7.55	0.1	7.615	0.115
	test	7.65	0.09	7.44	0.08	7.545	0.085
M2M100							
Mayan and Chol	dev	**50.56**	**27.5**	**48.22**	**23.92**	**49.39**	**25.71**
	test	51.48	28.85	48.88	25.11	**50.18**	**26.98**
AmeNLP + uniform	dev	49.11	25.48	47.85	24	48.48	24.74
	test	50.07	26.36	**48.89**	**25.17**	49.48	25.765
Mayan	dev	50.31	27.27				
	test	**51.55**	**29.13**				
Chol	dev			47.38	23.41		
	test			48.27	24.66		
biAmeNLP + uniform	dev	47.43	22.44	47.27	23.16	47.35	22.8
	test	47.99	22.98	48.21	24.22	48.1	23.6
Zero shot	dev	10.2	1.2	10.37	1.31	10.285	1.255
	test	9.89	0.62	10.26	1.25	10.075	0.935

translations from scant data.

This study marks the beginning of exploring NMT systems designed specifically for Mayan languages, highlighting both the possibilities and challenges of using NMT for languages, with resources. Moving forward, our future efforts will concentrate on expanding datasets and investigating active learning and few shot learning approaches. Furthermore, we plan to customize the M2M100 model for other native languages and delve into the nuances of tag selection to enhance translation accuracy. By progressing in these areas, we aim not only to improve the effectiveness of NMT systems for languages but also to contribute to the broader field of language technology and digital transformation (DX). Additionally, we plan to apply the techniques developed here to other low-resource scenarios, such as natural language to SPARQL translation.

References

Vladimir Araujo, Maria Mihaela Trusca, Rodrigo Tufiño, and Marie-Francine Moens. 2023. Sequence-to-sequence spanish pre-trained language models. (arXiv:2309.11259). ArXiv:2309.11259 [cs].

Lyle Campbell. 2000. *American Indian languages: the historical linguistics of Native America*, volume 4. Oxford University Press.

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)*, page 177–191, Toronto, Canada. Association for Computational Linguistics.

Angela Fan, Shruti Bhosale, Holger Schwenk, Zhiyi Ma, Ahmed El-Kishky, Siddharth Goyal, Mandeep Baines, Onur Celebi, Guillaume Wenzek, Vishrav Chaudhary, Naman Goyal, Tom Birch, Vitaliy Liptchinsky, Sergey Edunov, Edouard Grave, Michael Auli, and Armand Joulin. 2020. Beyond english-centric multilingual machine translation. (arXiv:2010.11125).

Manuel Mager, Rajat Bhatnagar, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2023. Neural machine translation for the indigenous languages of the americas: An introduction. (arXiv:2306.06804). ArXiv:2306.06804 [cs, stat].

Gerardo Sierra Martínez, Cynthia Montaño, Gemma Bel-Enguix, Diego Córdova, and Margarita Mota Montoya. Cplm, a parallel corpus for mexican languages: Development and interface.

El Moatez Billah Nagoudi, Wei-Rui Chen, Muhammad Abdul-Mageed, and Hasan Cavusogl. 2021. Indt5: A text-to-text transformer for 10 indigenous languages. (arXiv:2104.07483).

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21:1–67. ArXiv: 1910.10683 Citation Key: Raffel2020.

David Stap and Ali Araabi. 2023. Chatgpt is not a good indigenous translator. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, page 163–167, Toronto, Canada. Association for Computational Linguistics.

Atnafu Lambebo Tonja, Christian Maldonado-Sifuentes, David Alejandro Mendoza Castillo, Olga Kolesnikova, Noé Castro-Sánchez, Grigori Sidorov, and Alexander Gelbukh. 2023. Parallel corpus for indigenous language translation: Spanish-mazatec and spanish-mixtec. (arXiv:2305.17404). ArXiv:2305.17404 [cs].

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*, page 255–264, Online. Association for Computational Linguistics.

BSC Submission to the AmericasNLP 2024 Shared Task

Javier García Gilabert **Aleix Sant Savall** **Carlos Escolano**
Francesca De Luca Fornaciari **Audrey Mash** **Maite Melero**
Barcelona Super Computing Center (BSC),
{ javier.garcia1, aleix.santsavall, carlos.escolano,
francesca.delucafornaciari, audrey.mash, maite.melero}@bsc.es

Abstract

This paper describes the BSC's submission to the AmericasNLP 2024 Shared Task. We participated in the Spanish to Quechua and Spanish to Guarani tasks. In this paper we show that by using LoRA adapters we can achieve similar performance as a full parameter fine-tuning by only training 14.2% of the total number of parameters. Our systems achieved the highest ChrF++ scores and ranked first for both directions in the final results outperforming strong baseline systems in the provided development and test datasets.

1 Introduction

In this paper we present the submission of the Barcelona Super Computing Center to the Workshop on Natural Language Processing (NLP) for Indigenous Languages of the Americas (AmericasNLP) 2024 Shared Task on machine translation systems for Indigenous languages. We participated in two directions: Spanish to Guarani and Spanish to Quechua.

The Quechuan language family, comprising a diverse range of dialects, is spoken by over 6 million individuals across South America. It has a wide geographic spread, extending from Colombia through Ecuador, Peru, Bolivia, and Argentina. Quechua is an oral language, which results in a scarcity of textual resources. In this work we use an amalgamated data set for Quechua from Cuzco (quz) and Ayacucho (quy), which are the most widely spoken.

Guarani belongs to the Tupi–Guarani branch of the Tupian language family. It is one of the official languages of Paraguay and has more than 9 million speakers. The data used in this work contains a mix of dialects, from pure Guarani to more mixed Jopara.

In this work we fine-tuned various versions of NLLB-200 (Costa-jussà et al., 2022) as in previous year winning submission (Gow-Smith and Sánchez Villegas, 2023). Originally, NLLB-200 had been trained using parallel data in Quechua, Guarani, and Spanish. Our experiments aimed to answer the following questions: *(1)* How beneficial is fine-tuning with multilingual data in low resource scenarios? *(2)* Does translation quality improve with a greater model size for low resource languages? *(3)* Can we use LoRA for fine-tuning multilingual neural machine translation (MNMT) systems?

The remainder of this paper is organized as follows: Section 2 describes the data collection and preprocessing, Section 3 introduces our training methodology, Section 4 presents the results of our experiments and comments the results on the test sets, and finally, Section 5 summarizes the main findings.

2 Data collection and preparation

In this section we detail our data sources and the data cleaning pipeline used for cleaning training data. We utilize the parallel data provided by the organizers and additional sources for training our systems.

2.1 Data collection

AmericasNLP 2024 For Quechua, AmericasNLP training data included the Jw300 corpus (Agić and Vulić, 2019), MINEDU, and dict_misc. Additionally, English to Ayacucho Quechua (quy) and English to Cuzco Quechua (quz) translations of the Jw300 were also given. As for Guarani, the parallel data provided by the organizers was collected from web sources in a semi-automatic way and later supervised by a translator (Chiruzzo et al., 2020).

Helsinki The organizers made available parallel data from the Helsinki University (De Gibert et al., 2023). This data included resources from OPUS (Tiedemann, 2012). In addition, for Guarani they

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 143–149

also provided PYconst, News and Jojajovai. Furthermore, they generated synthetic data using two approaches: first, Spanish synthetic data was generated from Quechua and Guarani using Wikipedia monolingual corpus (Tiedemann, 2020), and second, by pivoting with English. They translated the English source sentences from the en-quy and en-quz Jw300 corpus into Spanish, resulting in synthetic es-quy and es-quz data, and translated the English source sentences from en-gn NLLB-seed corpus into Spanish, resulting in es-gn.

OPUS From the OPUS corpus collection, we get GNOME, Mozilla-I10n, Tatoeba, Ubuntu, and Wikimedia, all of which are available in Quechua and Guarani.

FLoRes-200 Developed by Costa-jussà et al., it is a widely-used multiparallel evaluation dataset which includes Guarani and Quechua. We extracted the corresponding Spanish to Guarani and Spanish to Quechua pairs from test and devtest sets, concatenated them, and utilized them as training data.

Hackathon-pln-es This dataset contains translations from Spanish to Quechua Ayacucho (quy). It is available at HuggingFace[1] and is a collection of webs and others datasets, namely: "Mundo Quechua", "Kuyakuykim (Te quiero): Apps con las que podrías aprender quechua", "Piropos y frases de amor en quechua", "Corazón en quechua", and parallel data from AmericasNLP 2021 Shared Task. We concatenated test set, validation set, and train sets.

Monolingual-Quechua The Monolingual-Quechua-IIC dataset (Zevallos et al., 2022) is a monolingual corpus of Quechua extracted from 50 different monolingual sources on the web which is available at HuggingFace[2]. Google Translate was employed to generate Spanish translations from this monolingual dataset, thereby obtaining parallel synthetic source generated data.

NLLB The NLLB corpus[3] is a dataset that was created by OPUS (Tiedemann, 2012) based on metadata for mined bitext released by the NLLB project (Costa-jussà et al., 2022). We extracted and incorporated pt-qu, pt-gn, en-qu and en-gn parallel data in our training dataset.

After collecting all the data, it was concatenated per language pair and then each set underwent the cleaning pipeline explained in the following subsection.

2.2 Data preprocessing

Given that training data was obtained from various sources and a significant part is synthetically-created, we employed a comprehensive data preprocessing pipeline to obtain good quality parallel data. In particular, we remove duplicates and near duplicates, we filter parallel sentences using heuristic-based and embedding-based methods and we remove off-target translations.

Cleaning step 1 We employed Moses (Koehn et al., 2007) standard preprocessing script to training data: `clean-corpus-n.perl` which removes sentences that have more than 150 tokens and removes sentence pairs that have a length ratio greater than 3. Then, Bifixer (Ramírez-Sánchez et al., 2020) was used to fix possible text issues.

Cleaning step 2 To remove duplicates and near-duplicates we used NLPDedup[4]. Then, we removed off-target translations using googletrans python library[5]. We found that there were a considerable number of near duplicates and off-target translations in the NLLB dataset collected from OPUS. In the English to Quechua direction, for example, 47% of the data were near duplicates, and 21.8% of the deduplicated data were off-target translations.

Cleaning step 3 We computed embedding-based similarity between a sentence pair using a sentence embedding model. We use SONAR (Duquenne et al., 2023) to embed both source and target sentences then compute a cosine similarity score between both. SONAR is a novel sentence embedding model that covers 200 languages and outperforms existing sentence embeddings such as LASER3 (Heffernan et al., 2022) and LabSE (Feng et al., 2022). It consists of an encoder-decoder architecture initialized from NLLB 1.3B dense model and trained by adding a pooling operation at the encoder's outputs to extract sentence representations.

[1] https://huggingface.co/datasets/hackathon-pln-es/spanish-to-quechua
[2] https://huggingface.co/datasets/Llamacha/monolingual-quechua-iic
[3] https://opus.nlpl.eu/NLLB/corpus/version/NLLB
[4] https://github.com/saattrupdan/NLPDedup
[5] https://github.com/ssut/py-googletrans

3 Methodology

3.1 Baseline Fine-tuning

We fine-tune separate NLLB models of 1.3B parameters for the Spanish to Quechua and Spanish to Guarani training data respectively. For each direction we experiment with different thresholds for choosing the training set according to the cosine similarity values computed using SONAR.

3.2 Multilingual Fine-tuning

Following similar lines as Gow-Smith and Sánchez Villegas, we fine tune NLLB-1.3B with multilingual data. We experiment with different variants of the training data to study whether a richer linguistic mixture of Quechua and Guarani is beneficial for the translation quality. Specifically, we include Portuguese and English directions to the Spanish ones. In table 1 we show the amount of parallel sentences collected for each translation pair.

Source	Quechua	Guarani
es `Original`	613,006	91,025
pt `Original`	2,071,571	684,883
en `Original`	2,874,090	2,959,122
es `Filtered`	132,884	42,504
pt `Filtered`	124,942	69,046
en `Filtered`	167,947	96,487
Total	427,773	208,037
Deduplicated	425,773	208,037

Table 1: Corpus statistics. `Filtered` refers to the number of pairs resulting from the data preprocessing pipeline.

Synthetic target generated data We study the impact of translation quality when adding synthetic generated data on the target side when fine tuning the model. Synthetic data was generated using Google Translate, translating Spanish monolingual data from TED2020 (Reimers and Gurevych, 2020) collected from OPUS website[6] which consists in 416,846 sentences in Spanish for about 4,000 TED talks covering a wide range of domains. We find that this data improves performance for the Spanish to Quechua direction (see Section 4).

Model size To investigate the effect of increasing the model capacity, we fine-tune the NLLB pre-trained model of 3.3B parameters with the best configuration found for the previous experiments with the model of 1.3B parameters.

3.3 LoRA Fine-tuning

Parameter-efficient fine-tuning (PEFT) techniques adapt pre-trained models by fine-tuning only a small subset of the model's parameters. The Low-Rank Adaptation (LoRA) technique (Hu et al., 2021) has been popularized for LLM training for its efficiency and often comparable results with full-parameter fine-tuning (Sun et al., 2023).

LoRA uses low-rank parametrized update matrices to reduce the number of trainable parameters. More specifically, given a pre-trained weight matrix $\mathbf{W}_0 \in \mathbb{R}^{d\times k}$, matrix $\mathbf{W}_0$ is updated with a low-rank decomposition as follows:

$$\mathbf{W}_0 + \frac{\alpha}{r}\Delta\mathbf{W} = \mathbf{W}_0 + \frac{\alpha}{r}\mathbf{BA}, \qquad (1)$$

where $\mathbf{B} \in \mathbb{R}^{d\times r}$, $\mathbf{A} \in \mathbb{R}^{r\times k}$, r is the rank of the trainable matrices and α is a scaling parameter that scales the learned weights. During the fine-tuning, the original matrix $\mathbf{W}_0$ remains unchanged and does not receive gradient updates. However, matrices $\mathbf{A}$ and $\mathbf{B}$ are updated during training.

We apply LoRA to the feed-forward and attention layers. The rank of the trainable matrices was set to 256 and α was set to 512.

3.4 Setup

For fine-tuning we use AdamW (Loshchilov and Hutter, 2019) optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.98$, $\epsilon = 10^{-6}$, $\lambda = 0.01$. We use the inverse square root scheduler with an initial learning rate of 2e-4 and 15,000 warmup steps. We set a batch size of 4 with 2 gradient accumulation steps, and train on 10 epochs. We use 1,000 sentences from the training dataset as the validation dataset in training. All models are trained using the Transformers[7] library on H100 GPUs. For LoRA we use the Peft[8] library. We saved best checkpoints every 4,000 steps according to the best ChrF++ metric in the validation set.

3.5 Inference

We limit the translation length to 512 tokens. For the case of Quechua we do not allow the model to generate an apostrophe as the Quechua Ayacucho variety was the one used in development and test

[6] https://opus.nlpl.eu/

[7] https://huggingface.co/

[8] https://huggingface.co/docs/peft/index

sets which does not handle pentavocalism. For inference, we use beam search and we experiment with different values of the beam size and the repetition penalty term (β). We find that adjusting the repetition penalty term yields major improvements in the case of Guarani.

3.6 Evaluation

We evaluate our fine-tuned models with the evaluation script provided by the organizers. Specifically, models were evaluated with two metrics: BLEU (Papineni et al., 2002) and ChrF++ (Popović, 2017) using SacreBLEU[9] implementation (Post, 2018).

We compare our model's performance against the baseline models released by the University of Helsinki (De Gibert et al., 2023) and the University of Sheffield (Gow-Smith and Sánchez Villegas, 2023). In addition, we provide the performance of Google Translate and the pre-trained NLLB 1.3B and 3.3B without any fine-tuning.

4 Results

4.1 Dev Set Results

Baseline Fine-tuning We first fine-tune the NLLB 1.3B models using bilingual data. Figure 1 shows the results of a full parameter fine-tuning of the model using different thresholds for selecting training data given the cosine similarity scores computed using SONAR. Based on the results we decide to use a threshold of 0.2 and 0.3 for Guarani and Quechua respectively for our next experiments. Note that we do not experiment with thresholds greater than 0.4 for Guarani as the number of sentence pairs between Spanish and Guarani given that threshold is relatively low (less than 19K pairs).

Multilingual Fine-tuning Table 3 show the BLEU and ChrF++ metrics evaluated on the development dataset for the Spanish to Quechua and Spanish to Guarani directions. As shown in the table, using synthetic target generated data combined with multilingual training improves translation quality for the Spanish to Quechua direction. In comparison to baseline models, we observe that our systems exhibit strong performance. In the development set we outperform Google Translate by +1.41 ChrF++ points and Sheffield's baseline by +6.05 points.

Regarding the Spanish to Guarani direction, we find that training with multilingual parallel data

[9]Word n-gram order was set to 2 for SacreBLEU implementation of ChrF++.

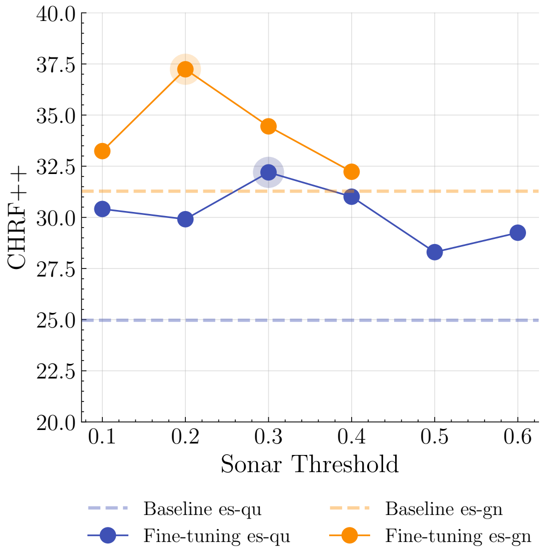


Figure 1: Performance evaluating on dev set. ChrF++ is in the vertical axis, and value for the threshold is in in the horizontal axis.

(en→gn and pt→gn) does not improve translation quality. However, our system trained only with Spanish to Guarani parallel data outperforms baseline models. We outperform Helsinki's baseline model by +2.74 ChrF++ points and Sheffield's baseline by +5.31 points.

Model size Table 2 compares the performance when increasing the model size. We compare the NLLB 1.3 Billion parameter model with the 3.3 Billion parameter one. When fine-tuning using the 3.3B model, we find that we do slightly worse (-1.35 ChrF++ points) than with the 1.3B model for the Spanish to Guarani direction, yet we gain +0.12 ChrF++ points for the Spanish to Quechua direction. These results indicate that increasing the model size does not yield superior performance. Mainly due to the small amount of data used for fine-tuning.

	Spanish → Quechua		**Spanish → Guarani**	
	ChrF++	**BLEU**	**ChrF++**	**BLEU**
NLLB-1.3B	36.27	3.77	**37.48**	**11.15**
NLLB-3.3B	**36.39**	**4.07**	36.13	10.20

Table 2: Experiments with model size.

LoRA Table 4 shows the results when fine-tuning using LoRA adapters. Notably, for the Spanish to Quechua direction, our system performs as well as the full fine-tuning with the NLLB 3.3B model

		Data			Spanish → Quechua		Spanish → Guarani	
		es	en	pt	ChrF++	BLEU	ChrF++	BLEU
Baseline	Helsinki				28.78	-	34.74	-
	Sheffield				30.22	-	32.17	-
	Google Translate				34.86	3.23	30.33	4.71
Inference	NLLB-1.3B				24.97	1.95	31.28	6.27
	NLLB-3.3B				26.84	1.64	32.03	6.31
Our	NLLB-1.3B	✓	✗	✗	32.20	3.28	**37.48**	**11.15**
	NLLB-1.3B	✓	✓	✗	32.71	3.18	33.94	7.89
	NLLB-1.3B	✓	✗	✓	32.40	3.35	33.92	7.88
	NLLB-1.3B	✓	✓	✓	31.59	2.94	34.01	8.26
Our	NLLB-1.3B + Synthetic	✓	✗	✗	35.12	3.55	32.50	6.08
	NLLB-1.3B + Synthetic	✓	✓	✓	**36.27**	**3.77**	32.48	6.00

Table 3: Evaluations computed on the development dataset. The symbol ✓ indicates that the parallel data in the corresponding direction was utilized for the fine-tuning whereas ✗ indicates that it was not used.

when fine-tuning with LoRA, even though it has 85.8% fewer trainable parameters. We perform just slightly worse when compared to the full fine-tuning in the Spanish to Guarani direction when using LoRA.

	Trainable Parameters	es→qu	es→gn
NLLB-3.3B	3,898,511,360	36.39	36.13
NLLB-1.3B	1,748,125,696	36.27	37.48
NLLB-3.3B + LoRA	553,648,128	36.40	35.24
NLLB-1.3B + LoRA	377,487,360	36.11	35.26

Table 4: Experiments with LoRA fine-tuning. ChrF++ scores are shown for Spanish to Quechua (es→qu) and Spanish to Guarani (es→gn) directions.

Inference experiments A grid search is performed to understand how the beam size and the repetition penalty term (β) parameters affect the translation quality. We plot the obtained results in Figure 2.

We observe that for the Spanish to Guarani direction, the penalty term has a significant impact on the translation quality as measured by the ChrF++ score. Specifically, we find that best ChrF++ results are obtained with a beam size of 10 and $\beta = 4$.

However, for the Spanish to Quechua direction, neither the repetition penalty term nor the beam size seem to have significant impact on the translation quality.

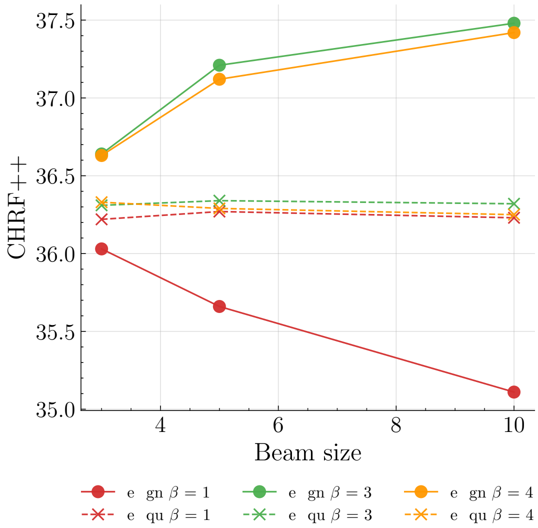


Figure 2: Performance evaluating on dev set grouped by repetition penalty. ChrF++ is in the vertical axis, and beam size is in the horizontal axis.

4.2 Test Set Results

In table 5 we show the results on the test sets. Our submitted systems ranked first in both directions. On Spanish to Quechua, we achieve 38.21 ChrF++ using LoRA adapters with the 1.3B model. We observe that our submitted systems significantly outperform baseline models by +4.2 points on Sheffield's baseline and by +9.4 points on Helsinki's baseline for this direction. On Spanish to Guarani direction we achieve 38.93 ChrF++ and our submitted systems outperform Helsinki's baseline model by up to +1.91 points and by up to

+3.17 points for Sheffield's baseline model.

Model	r_{qu}	es→qu	r_{gn}	es→gn
Baselines				
Helsinki		28.81		37.02
Sheffield		34.01		35.76
Full fine-tuning				
NLLB-3.3B	2	38.19	2	37.64
NLLB-1.3B	4	38.02	**1**	**38.93**
LoRA				
NLLB-3.3B + LoRA	3	38.10	3	37.63
NLLB-1.3B + LoRA	**1**	**38.21**	4	37.42

Table 5: Final test results. ChrF++ scores are shown for Spanish to Quechua (es→qu) and Spanish to Guarani (es→gn) directions. r_{qu} and r_{gn} denote the position in the final classification for Spanish to Quechua and Spanish to Guarani respectively.

5 Conclusions

In this paper we describe the Barcelona Supercomputing Center submission to the AmericasNLP 2024 shared task. We took part in the Spanish to Quechua and Spanish to Guarani tracks. We fine-tuned different versions of the NLLB-200 model. Our systems ranked in first place in both translation directions outperforming the provided baselines.

Our experiments show that increasing model size does not yield superior performance when data is scarce. Also, we show that training with multilingual data combined with synthetic data improves translation quality for the Spanish to Quechua direction. Finally, we show that by fine-tuning the models using LoRA, we can obtain a similar performance as a full parameter fine-tuning while training only 14.2% of the parameters.

Acknowledgments

This work has been supported by the Spanish project PID2021-123988OB-C33.

References

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.

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.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

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, Toronto, Canada. Association for Computational Linguistics.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoit Sagot. 2023. SONAR: sentence-level multimodal and language-agnostic representations.

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.

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.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Ilya Loshchilov and Frank Hutter. 2019. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

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.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Belgium, Brussels. Association for Computational Linguistics.

Gema Ramírez-Sánchez, Jaume Zaragoza-Bernabeu, Marta Bañón, and Sergio Ortiz-Rojas. 2020. Bifixer and bicleaner: two open-source tools to clean your parallel data. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 291–298, Lisboa, Portugal. European Association for Machine Translation.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Xianghui Sun, Yunjie Ji, Baochang Ma, and Xiangang Li. 2023. A comparative study between full-parameter and lora-based fine-tuning on chinese instruction data for instruction following large language model. *arXiv preprint arXiv:2304.08109*.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Rodolfo Zevallos, John Ortega, William Chen, Richard Castro, Núria Bel, Cesar Toshio, Renzo Venturas, Hilario Aradiel, and Nelsi Melgarejo. 2022. Introducing QuBERT: A large monolingual corpus and BERT model for Southern Quechua. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 1–13, Hybrid. Association for Computational Linguistics.

System Description of the NordicsAlps Submission to the AmericasNLP 2024 Machine Translation Shared Task

Joseph Attieh[1†], Zachary William Hopton[2 †], Yves Scherrer[1,3], Tanja Samardzic[2],

[1]Department of Digital Humanities, University of Helsinki, {first.last}@helsinki.fi
[2]Language and Space Lab, University of Zurich, {first.last}@uzh.ch
[3]Department of Informatics, University of Oslo, {first.last}@ifi.uio.no

Abstract

This paper presents the system description of the NordicsAlps team for the AmericasNLP 2024 Machine Translation Shared Task 1. We investigate the effect of tokenization on translation quality by exploring two different tokenization schemes: byte-level and redundancy-driven tokenization. We submitted three runs per language pair. The redundancy-driven tokenization ranked first among all submissions, scoring the highest average chrF2++, chrF, and BLEU metrics (averaged across all languages). These findings demonstrate the importance of carefully tailoring the tokenization strategies of machine translation systems, particularly in resource-constrained scenarios.

1 Introduction

The participation of the NordicsAlps team in the AmericasNLP 2024 Machine Translation Shared Task builds directly on the previous contributions by the Helsinki team. The main goal of the shared task, as in the previous editions, is to build machine translation (MT) systems capable of translating Spanish into eleven American languages. With limited training data, the MT solutions need to leverage cross-lingual transfer and data-efficient approaches to achieve a good level of performance on the translation tasks. Previous contributions of the Helsinki team performed cross-lingual transfer by pre-training a Spanish-English model, and transferring the knowledge learned to the language pairs of the task, i.e., Spanish-TARGET (any of the eleven indigenous target languages), by continued training. The previous Helsinki submissions primarily focused on increasing the data size by collecting additional sources and applying data augmentation techniques, but data efficiency was not directly addressed. Our submission builds on the previous findings and focuses on the data efficiency aspect of the challenge.

The core idea behind our proposal is that both cross-lingual transfer and data efficiency can be improved by optimizing the vocabulary size, which can be controlled by means of tokenization. Following the current understanding about the role of tokenization in machine translation (Section 2), we aim at small vocabularies (short tokens). We explore two options: (1) byte-level tokenization and (2) redundancy-driven subword-level tokenization, and compare them with the SentencePiece tokenization used in De Gibert et al. (2023). We submit three runs for each language pair. Among these runs, the redundancy-driven tokenization scheme gives the best scores on all language pairs. Furthermore, it ranks first among all submissions to the shared task in terms of average chrF++, chrF, and BLEU.

2 Related Work

2.1 Machine translation for indigenous languages of the Americas

As pointed out by Mager et al. (2018), despite the fact that there are millions of people in the Americas who identify as indigenous, there is a distinct lack of language technology for the hundreds of indigenous languages spoken in the Americas. Machine translation systems have the potential to aid in equality of access to information, educational technology, and language revitalization efforts for indigenous communities (Mager et al., 2018, 2023; Ebrahimi et al., 2023). However, building such systems for languages that are often relatively low-resourced presents a number of potential challenges, as delimited in a survey of the field by Haddow et al. (2022). These challenges can include the lack of reliable language identification tools to aid in data collection, a scarcity of parallel data sets, and non-standardized orthographies. Mager et al. (2018) also note that indigenous American languages are very typologically diverse, yet

[†]Authors of equal contribution

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 150–158
June 21, 2024

many are understudied from a linguistic standpoint compared to languages more commonly treated in NLP. This limits the opportunity to experiment with machine translation models informed by linguistic knowledge (i.e., via token annotations), which is an area that generally lacks study in low-resource machine translation settings according to Haddow et al. (2022).

Now in its fourth year, the AmericasNLP shared task has become a lively forum for progressing in machine translation for indigenous languages in the Americas. Previous submissions to the 2021 and 2023 shared tasks have taken a variety of creative steps to work around the challenges common in low-resourced language machine translation. Among other things, this has included experimenting with fine-tuning pre-trained machine translation models; data mining and filtering; exploiting monolingual language data to synthesize or back-translate more parallel data; multilingual translation models; knowledge distillation; in-context learning with GPT models; and model ensembling (Mager et al., 2021; Ebrahimi et al., 2023). Importantly, previous challenges have included qualitative analysis of some of the submitted translation systems. Indeed, other researchers have highlighted that community involvement is a key part of developing NLP tools that have a positive impact for indigenous communities and their languages (Mager et al., 2023; Zhang et al., 2022).

2.2 Subword segmentation in MT

With the introduction of subword tokenization to MT (Sennrich et al., 2016), the size of the vocabulary has become a hyper-parameter, which is most commonly set in an arbitrary way. For instance, the size of 32k is a frequent choice for multilingual MT at the moment. The vocabulary size can, in principle, be optimized for the task (Kudo, 2018), but this is hard to do in the framework of transfer learning because the vocabulary of pre-trained models is fixed and hard to map onto a different one for the end task. This is an important obstacle to improving cross-lingual transfer in general (Rust et al., 2021). Byte Pair Encoding (BPE) drop-out (Provilkov et al., 2020) is a popular general method of regularizing the vocabulary, which is suitable for transfer learning.

In search of a more principled approach to setting the vocabulary size, Mielke et al. (2019) find that the size of 0.4 $\times$ the initial (word-level) size results in the lowest negative log likelihood of a language model across multiple languages. The size of 32k appears the best when translating from German to English with a large training set. Otherwise, 2k seems to work best for varied data sizes and directions (Gowda and May, 2020). Defining linguistically motivated subword units is a criterion proposed by Ataman and Federico (2018). This method can help with a particular language (e.g. Turkish), but it depends on external language-specific knowledge. Using more linguistically driven algorithms is found to improve downstream performance on various tasks (Bostrom and Durrett, 2020; Park et al., 2021), but the improvements are surprisingly small and not very consistent. As a matter of fact, replacing standard BPE tokens with randomly selected ones gives almost the same MT scores (Saleva and Lignos, 2023).

Byte-level tokenization is an attempt to overcome the arbitrariness of the vocabulary size parameter and other limitations of subword tokenization (Shapiro and Duh, 2018). Instead of representing the text using subwords, the content is mapped to bytes using the Unicode Transformation Format 8-bit (UTF-8) encoding. This strategy removes the need for initial text processing by reducing all texts to a small vocabulary of only 256 byte types. This level of tokenization is similar to the character-level (bytes roughly encode Unicode characters), which looked promising with RNN models (Lee et al., 2017). However, later experiments yielded mixed results. Shaham and Levy (2021) trained models that operate on byte sequences, outperforming the subword-based models in bilingual translation. These findings were also confirmed in a many-to-one multilingual setup and for endangered languages (Zhang and Xu, 2022). On the other hand, Libovický et al. (2022) find that subword tokenization is still better. More generally, byte-level tokenization can improve the performance on various tasks in low-resource languages (Clark et al., 2022; Xue et al., 2022), but its use in high-resource settings is still questionable. Since the data sets in this shared tasks are relatively small, we explore the use of small byte- and subword-level vocabularies.

3 Data

Following De Gibert et al. (2023), we train multilingual one-to-many models that translate from Spanish to the eleven indigenous target languages and and include English as an additional high-resource target language.

Spanish–English Data We use Spanish–English parallel data from a subset of the sources mentioned in De Gibert et al. (2023): *Europarl*, *GlobalVoices*, *NewsCommentary*, *TedTalks*, and *Tatoeba* collected from OPUS (Tiedemann, 2012). In contrast to De Gibert et al. (2023) and due to time constraints, we do not include *Bibles* nor *OpenSubtitles*. Validation data for pre-training comes from the Spanish–English *WMT-News* corpus.

The Spanish–English parallel data underwent cleaning with OpusFilter (Aulamo et al., 2020), as described in De Gibert et al. (2023). Namely, this consisted in deduplication and a set of filters based on length difference, script identification and language identification.

Spanish–Indigenous Language Data Our models include all eleven indigenous American languages for which data was provided in the shared task: Asháninka (cni), Aymara (aym), Bribri (bzd), Chatino (ctp), Guarani (gn), Hñähñu (oto), Nahuatl (nah), Quechua (quy), Raramuri (tar), Shipibo-Konibo (shp), and Wixarika (hch).

We used all Spanish–indigenous language training and development data provided by the Shared Task organizers (Ortega et al., 2020; Cushimariano Romano and Sebastián Q., 2008; Mihas, 2011; Tiedemann, 2012; Feldman and Coto-Solano, 2020; Agić and Vulić, 2019; Montoya et al., 2019; Galarreta et al., 2017). Whenever available, we also included the *extra* and *synthetic* datasets provided by the Shared Task organizers (De Gibert et al., 2023).

The data used for this year's submissions differ from those described in De Gibert et al. (2023) in two crucial aspects. First, we did not include Bible data, since Bibles did not improve translation quality in earlier editions (Vázquez et al., 2021) and were not part of the organizer-provided datasets. Second, due to time constraints, we did not apply any filtering or cleaning to the parallel data.

No preprocessing has been applied to the byte-level models. Some general preprocessing was carried out on the Spanish–indigenous language data for the BPE-based models. This consisted in whitespace normalization, Unicode character normalization, and separation of punctuation from words. Separation of punctuation from words was done using the Moses tokenizer as well as hand-crafted rules to prevent tokenization at apostrophes that actually represented glottal stops. As documented in Vázquez et al. (2021), we also applied some spelling normalization scripts to the data for Wixarika and Raramuri.

Since all our models are multilingual models with several target languages, we include a target language tag at the beginning of the source sentence. We did not use the additional *variant* and *quality* tags proposed by De Gibert et al. (2023), and opted for simply relying on the target language for the tags.

3.1 Post-processing

The output produced by the MT models is post-processed by removing subword segmentation marks (if applicable), removing <unk> tokens, and detokenizing with the Moses detokenizer (with Spanish settings).

After inspecting the translations of the development sets, we also apply some language specific post-processing rules:

- For Aymara, Bribri and Raramuri, we normalize apostrophes and remove whitespaces surrounding them.
- For Guarani and Hñähñu, we apply the normalization functions of De Gibert et al. (2023)[1], complemented with some additional diacritic replacements.
- For Wixarika, we observed that the + sign was not properly detokenized; however, we could not find a simple post-processing routine to properly attach this symbol to preceding and/or following tokens.

4 Methods

Our subword-level settings follow previous model architectures and training regimes closely with a few updates. The main difference here is the tokenization. In the byte-level settings, we work with different architectures.

4.1 Subword-level Models

U-SP As a baseline, we segmented all data with the subword tokenizer provided by De Gibert et al. (2023). This tokenizer was trained jointly on all source and target languages with the Unigram model implemented in the SentencePiece toolkit, using a vocabulary size of 32k tokens.

[1]See `https://github.com/Helsinki-NLP/americasnlp2023-st/blob/main/create_opusfilter_config.py`

In preliminary experiments, we found that using joint or separate token embeddings did not make a significant impact, and neither did subword sampling. We report results on the model that most closely resembles the other subword setting (BPE-MR), namely with separate embeddings and without subword sampling.

BPE-MR The principle of BPE-MR is to use text redundancy as a criterion for the vocabulary size. We look for the vocabulary that approximately minimizes text redundancy (hence MR). This goal is inspired by connecting several observations from previous work.

The first relevant point is that, given a fixed vocabulary size, data compression efficiency of a tokenization algorithm has an impact on machine translation. That is, the tokenization that minimizes the length of the sentence gives the best BLEU score (Gallé, 2019). This finding is recently replicated by Zouhar et al. (2023) using Rényi entropy as the measure of compression efficiency. While these findings do not suggest a preferred vocabulary size, we note that the overall best scores are obtained with smaller vocabularies, in the range around 2k, already observed by Gowda and May (2020).

The second relevant point is that monolingual BPE models maximally compress a corpus after carrying out just 200–350 merges (Gutierrez-Vasques et al., 2021). Since each BPE merge adds exactly one new member to the vocabulary, the maximal compression happens with the vocabulary size of several hundreds (number of BPE merges + the set of characters). This compression is measured by information theoretic redundancy of a given corpus, and was shown to hold across a diverse sample of languages.

The third relevant point is that Shannon entropy converges to a similar value across different languages when the redundancy is maximized making different languages in some sense more similar (Gutierrez-Vasques et al., 2021). More compatible embedding spaces across languages coincide with identical vocabulary sizes (Maronikolakis et al., 2021), at least in alphabetic scripts, although the size itself does not seem to impact the performance on the zero-shot XNLI task.

We thus train a BPE subword tokenizer[2] to carry out 300 merges for each language. Note that this is far fewer merges than what is typical when using BPE for training subword tokenizers. For English and Spanish, the tokenizers were trained on the parallel Spanish-English training data. For the indigenous languages, we trained each tokenizer on the given language's training and development data, as well as the *extra* files where available. We did not use any provided synthetic data while training the tokenizers. Subword tokenization models for the indigenous languages were trained on preprocessed data. We then applied the trained subword tokenization models to their respective language's train, development, extra, and synthetic data, and added the tokenized extra and synthetic data to the train set.

For the indigenous languages, we experimented with an early stop criteria to determine exactly how many merges to train the tokenization models for. This consisted in iteratively training 350 tokenization models for each language to carry out 1 to 350 merges on the corpora. After applying each model, we determined the difference in the frequency of the vocabulary items merged by BPE at the given merge and the prior merge. Based off of previous unpublished experiments with smaller datasets, we stop training BPE models when seven models occurred where the difference in merged-item frequency was extremely low (i.e., -1 or 0). However, for all of the indigenous languages used here, the early stop criterion was never met in the first 350 merges. Therefore, we trained all models to carry out 300 merges, and will conduct further research on finding the ideal stopping point in the future.

Model Architecture and Training Regime All MT models use the Transformer architecture (Vaswani et al., 2017) with mostly the same hyperparameters as Model B of De Gibert et al. (2023).[3] The models are trained with OpenNMT-py 3.4.3 (Klein et al., 2020).

The training takes place in two phases. In phase 1, the model is trained on 89% of Spanish–English data and 1% of data coming from each of the eleven indigenous languages. In phase 2, the proportion of Spanish–English data is reduced to 50%, with the other half sampled to equal amounts from the eleven indigenous languages. We did not include a third phase of language-specific fine-tuning this year.

We train the first phase for 100k steps and pick the best intermediate savepoint according to the

[2] https://github.com/rsennrich/subword-nmt

[3] Notable differences include the use of separate source and target token embeddings, and of the ALiBi position encodings (Press et al., 2021).

Model	Source vocab.	Target vocab.
U-SP	21 511	25 949
BPE-MR	1 215	5 896
Byte-SESD, Byte-SEMD	256	256

Table 1: Vocabulary sizes of the different MT models.

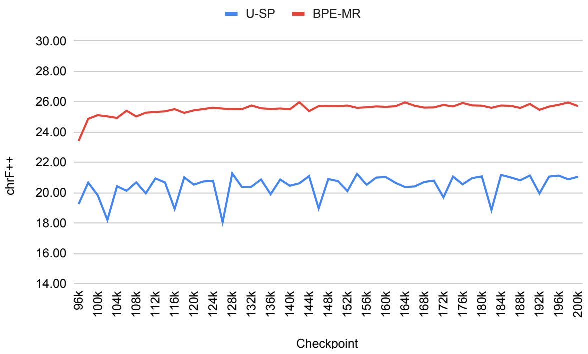


Figure 1: Development chrF++ scores (averaged over all 11 development sets) during phase 2 training of subword-level models.

English validation set. Depending on the model, this occurred after 96k or 100k steps. We initialize phase 2 with this savepoint and continue training until 200k steps, saving intermediate checkpoints every 2k steps. We then pick the most promising savepoint for each language based on the chrF++ score of the development set.

We train two models, one with a baseline SentencePiece tokenizer, and one with the proposed BPE-MR approach. They are described in detail below.

4.2 Byte-Level Models

For our byte-level models, we experiment with different architectures within a one-to-many setup. We define the following two variants: the first variant is the single encoder multiple decoder setup (**Byte-SEMD**) which involves one encoder for Spanish and one language-specific decoder for each target language. The second variant is a single encoder single decoder (**Byte-SESD**) setup comprising one encoder for Spanish and one decoder that is shared by all target languages. The model employs language tokens as a guide to generate text in the target language. We proceed with the same training regiment as before, by pre-training a model on English-Spanish data, and using the weights of the model to initialize the encoder and decoders in the proposed setups.

For all models, we use a total of 6 transformer layers for the encoder and 6 layers for the decoder with 8 attention heads, 512 hidden units and the feed-forward dimension of 2048. We follow the architecture of Shaham and Levy (2021) by replacing the dense trainable embedding matrix of the embeddingless models with a fixed one-hot encoding of the vocabulary. We use relative position encoding (Shaw et al., 2018) as the limit of the sequences supported by the framework is 5000 (lower than the largest byte sequence in the training data). We use the MAMMOTH toolkit (Mickus et al., 2024) as a basis for our implementation, since it is specifically designed for modular sequence-to-sequence model training, which allows to produce the different sharing patterns desired in this study. The models underwent training for 1.5 days, with an early stopping criterion in place. However, we observed that they were undertrained at the time of submission: the loss continued to decrease, and the early stopping mechanism had not yet been triggered. Consequently, we chose the most recent checkpoint for the submission. This issue rises due to the sequence length of such models that requires a larger batch size compared to the other variants as well as a longer training budget.

Language		Savepoint	Before	After
aym	Aymara	124k	33.35	33.42
bzd	Bribri	176k	23.01	22.99
cni	Asháninka	196k	24.48	
ctp	Chatino	200k	38.34	
gn	Guarani	194k	31.90	34.61
hch	Wikarika	142k	26.97	
nah	Nahuatl	152k	25.39	
oto	Hñähñu	130k	11.86	12.75
quy	Quechua	164k	31.76	
shp	Shipibo-Konibo	164k	27.51	
tar	Raramuri	142k	15.76	15.76

Table 2: Development set chrF++ scores of the BPE-MR model, before and after language-specific post-processing. No post-processing was applied to six languages. The table also shows the savepoints that yielded the reported scores. These savepoints were used for test set translation.

5 Results

The different tokenization strategies resulted in different vocabulary sizes of the MT models, as can be seen in Table 1.

5.1 Subword-level Model Evaluation

Figure 1 shows the evolution of the development chrF++ scores during the second phase of training.

Model	aym	bzd	cni	ctp	gn	hch	nah	oto	quy	shp	tar	Average
1 – BPE-MR	[2] 29.39	[4] 23.32	**[1] 23.20**	**[1] 37.38**	[5] 36.23	**[1] 27.64**	**[1] 22.87**	**[1] 12.98**	[11] 32.98	[2] 27.04	[5] 14.57	**[1] 26.15**
2 – Byte-SEMD	[8] 26.37	[8] 17.23	[9] 15.45	[2] 23.64	[10] 32.32	[9] 23.47	[8] 20.77	[7] 11.63	[14] 28.81	[10] 22.20	[9] 10.53	[8] 21.13
3 – Byte-SESD	[12] 15.77	[12] 12.24	[10] 15.23	[11] 12.96	[17] 14.80	[12] 15.97	[13] 14.57	[9] 11.22	[16] 25.15	[12] 21.28	[8] 12.63	[12] 15.62
Best competitor	**[1] 30.97**	**[1] 23.47**	[2] 22.98	[3] 20.70	**[1] 38.93**	[2] 26.46	[2] 21.71	[2] 12.63	**[1] 38.21**	**[1] 29.37**	**[1] 17.03**	[2] 23.32

Table 3: Official chrF++ scores on the test sets. Rankings are displayed in brackets.

We observe that the training curves are relatively flat, which suggests that phase 2 training can be limited to a few thousand steps without significant impact on translation performance.

The BPE-MR model clearly outperform the U-SP model. Moreover, the training scores of the U-SP model fluctuate much more. In particular, the U-SP model shows occasional language-specific "breakdowns", but recovers quickly from them. For example, the chrF++ scores for Guarani vary between 27.97 (100k), 3.66 (102k), and 28.17 (104k). We currently do not have an explanation why such breakdowns occur, and why they only occur for some of the languages.

On the basis of these observations, we decided not to submit the U-SP model. Table 2 shows the selected checkpoints per language and the corresponding development set chrF++ scores of the BPE-MR model. It also shows that language-specific post-processing (see Section 3.1) had a considerable impact on our Guarani and Hñähñu results.

5.2 Test results

We submitted three runs to the shared task: (1) BPE-MR, (2) Byte-SEMD, and (3) Byte-SESD. Table 3 reports the official results on the test set. Our BPE-MR submission was ranked first for 5 out of 11 languages and second for 2 additional languages. For Bribri, Asháninka, Hñähñu and Quechua, it was an extremely close competition: the first 6, 7, 6 and 4 submissions respectively are only within one chrF++ point. For all but Quechua, our BPE-MR submission is among these best submissions. In terms of average chrF++, chrF, and BLEU, the submitted BPE-MR model ranks first among all submissions to the shared task.

As mentioned previously, we notice that the byte-level models are undertrained at the time of the submission, due to the sequence length of such models that requires a larger batch size compared to the other variants, and a longer training budget.

6 Conclusions

This paper presents the NordicAlps submissions to the AmericasNLP 2024 machine translation shared task. Our contribution focuses on data efficiency, and in particular on optimizing subword-level tokenization. We trained four systems: a baseline system with a previously trained SentencePiece tokenizer (U-SP), a subword-level system based on the proposed minimized text redundancy BPE approach (BPE-MR), and two byte-level systems differing in their decoder architectures (Byte-SEMD with language-specific decoders and Byte-SESD with a single shared decoder). We did not submit the U-SP system.

The BPE-MR system reached the first rank in terms of average scores across all languages. It reached a top-five ranking for all languages except Quechua. The Byte-SEMD and Byte-SESD systems performed less well, but this is most likely due to undertraining.

Acknowledgments

This work was supported by the GreenNLP project, funded by the Research Council of Finland and the Swiss National Science Foundation Scientific Exchanges.

References

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.

Duygu Ataman and Marcello Federico. 2018. An evaluation of two vocabulary reduction methods for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 97–110, Boston, MA. Association for Machine Translation in the Americas.

Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann. 2020. OpusFilter: A configurable parallel corpus filtering toolbox. In *Proceedings of the 58th Annual*

Meeting of the Association for Computational Linguistics: System Demonstrations, pages 150–156, Online. Association for Computational Linguistics.

Kaj Bostrom and Greg Durrett. 2020. Byte pair encoding is suboptimal for language model pretraining. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 4617–4624, Online. Association for Computational Linguistics.

Jonathan H. Clark, Dan Garrette, Iulia Turc, and John Wieting. 2022. Canine: Pre-training an efficient tokenization-free encoder for language representation. *Transactions of the Association for Computational Linguistics*, 10:73–91.

Rubén Cushimariano Romano and Richer C. Sebastián Q. 2008. Ñaantsipeta asháninkaki birakochaki. diccionario asháninka-castellano. versión preliminar. `http://www.lengamer.org/publicaciones/diccionarios/`.

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, Toronto, Canada. Association for Computational Linguistics.

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.

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.

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, Varna, Bulgaria. INCOMA Ltd.

Matthias Gallé. 2019. Investigating the effectiveness of BPE: The power of shorter sequences. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1375–1381, Hong Kong, China. Association for Computational Linguistics.

Thamme Gowda and Jonathan May. 2020. Finding the optimal vocabulary size for neural machine translation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3955–3964, Online. Association for Computational Linguistics.

Ximena Gutierrez-Vasques, Christian Bentz, Olga Sozinova, and Tanja Samardzic. 2021. From characters to words: the turning point of BPE merges. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 3454–3468, Online. Association for Computational Linguistics.

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.

Guillaume Klein, François Hernandez, Vincent Nguyen, and Jean Senellart. 2020. The OpenNMT neural machine translation toolkit: 2020 edition. In *Proceedings of the 14th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 102–109, Virtual. Association for Machine Translation in the Americas.

Taku Kudo. 2018. Subword regularization: Improving neural network translation models with multiple subword candidates. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 66–75, Melbourne, Australia. Association for Computational Linguistics.

Jason Lee, Kyunghyun Cho, and Thomas Hofmann. 2017. Fully character-level neural machine translation without explicit segmentation. *Transactions of the Association for Computational Linguistics*, 5:365–378.

Jindřich Libovický, Helmut Schmid, and Alexander Fraser. 2022. Why don't people use character-level machine translation? In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2470–2485, Dublin, Ireland. Association for Computational Linguistics.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

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.

Antonis Maronikolakis, Philipp Dufter, and Hinrich Schütze. 2021. Wine is not v i n. on the compatibility of tokenizations across languages. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2382–2399, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Timothee Mickus, Stig-Arne Grönroos, Joseph Attieh, Michele Boggia, Ona De Gibert, Shaoxiong Ji, Niki Andreas Loppi, Alessandro Raganato, Raúl Vázquez, and Jörg Tiedemann. 2024. MAMMOTH: Massively multilingual modular open translation @ Helsinki. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations*, pages 127–136, St. Julians, Malta. Association for Computational Linguistics.

Sabrina J. Mielke, Ryan Cotterell, Kyle Gorman, Brian Roark, and Jason Eisner. 2019. What kind of language is hard to language-model? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4975–4989, Florence, Italy. 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.

John 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.

Hyunji Hayley Park, Katherine J. Zhang, Coleman Haley, Kenneth Steimel, Han Liu, and Lane Schwartz. 2021. Morphology matters: A multilingual language modeling analysis. *Transactions of the Association for Computational Linguistics*, 9:261–276.

Ofir Press, Noah A. Smith, and Mike Lewis. 2021. Train short, test long: Attention with linear biases enables input length extrapolation. *CoRR*, abs/2108.12409.

Ivan Provilkov, Dmitrii Emelianenko, and Elena Voita. 2020. BPE-dropout: Simple and effective subword regularization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1882–1892, Online. Association for Computational Linguistics.

Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3118–3135, Online. Association for Computational Linguistics.

Jonne Saleva and Constantine Lignos. 2023. What changes when you randomly choose BPE merge operations? not much. In *Proceedings of the Fourth Workshop on Insights from Negative Results in NLP*, pages 59–66, Dubrovnik, Croatia. Association for Computational Linguistics.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1715–1725, Berlin, Germany. Association for Computational Linguistics.

Uri Shaham and Omer Levy. 2021. Neural machine translation without embeddings. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 181–186, Online. Association for Computational Linguistics.

Pamela Shapiro and Kevin Duh. 2018. BPE and CharCNNs for translation of morphology: A cross-lingual comparison and analysis. *Preprint*, arXiv:1809.01301.

Peter Shaw, Jakob Uszkoreit, and Ashish Vaswani. 2018. Self-attention with relative position representations. *Preprint*, arXiv:1803.02155.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

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.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Mengjiao Zhang and Jia Xu. 2022. Byte-based multilingual NMT for endangered languages. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4407–4417, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Shiyue Zhang, Ben Frey, and Mohit Bansal. 2022. How can NLP help revitalize endangered languages? a case study and roadmap for the Cherokee language. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1529–1541, Dublin, Ireland. Association for Computational Linguistics.

Vilém Zouhar, Clara Meister, Juan Gastaldi, Li Du, Mrinmaya Sachan, and Ryan Cotterell. 2023. Tokenization and the noiseless channel. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5184–5207, Toronto, Canada. Association for Computational Linguistics.

On the Robustness of Neural Models for Full Sentence Transformation

Michael Ginn, Ali Marashian, Bhargav Shandilya, Claire Benét Post, Enora Rice, Juan Vásquez, Marie McGregor, Matt Buchholz, Mans Hulden, Alexis Palmer
University of Colorado, Depts. of Linguistics and Computer Science
first.last@colorado.edu,
{benet.post,juan-vasquez-1}@colorado.edu

Abstract

This paper describes the LECS LAB submission to the AmericasNLP 2024 Shared Task on the Creation of Educational Materials for Indigenous Languages (Chiruzzo et al., 2024). The task requires transforming a base sentence with regards to one or more linguistic properties (such as negation or tense). We observe that this task shares many similarities with the well-studied task of word-level morphological inflection, and we explore whether the findings from inflection research are applicable to this task. In particular, we experiment with a number of augmentation strategies, finding that they can significantly benefit performance, but that not all augmented data is necessarily beneficial. Furthermore, we find that our character-level neural models show high variability with regards to performance on unseen data, and may not be the best choice when training data is limited.

1 Introduction

Morphological inflection is an NLP task with a rich history of rule-based, statistical, and neural methods (Clark 2002; Durrett and DeNero 2013; Nicolai et al. 2015; Cotterell et al. 2016; Faruqui et al. 2016; Wu et al. 2021; inter alia). Typically, systems must predict an inflected form of a word (such as *"cats"*) given a lemma form (*"cat"*) and an inflectional change (plural).

In the AmericasNLP 2024 Shared Task on Creation of Educational Materials for Indigenous Languages (Chiruzzo et al., 2024), systems must convert a base sentence into a target sentence by changing one or more linguistic properties (example in Table 1). Generally, this transformation involves inserting or deleting helper words, modifying the inflection of words in the source sentence, or both, and we observe many similarities (and some differences) between this task and the morphological inflection task.

Source	Ko'one'ex ich kool
Change	PERSON:1_PL
Target	Ko'ox ich kool

Table 1: Example from the Yukatek Maya training data.

In approaching this task, we apply lessons from research on inflection models. The shared task poses particular difficulties due to the limited amount of available training data. To alleviate this issue, we utilize sequence-to-sequence (seq2seq) neural models and explore various techniques, focusing in particular on exploring various **data augmentation** strategies. We present results for all three task languages: Bribri, Yukatek Maya,[1] and Guaraní. Our code is available on GitHub.[2]

2 Background

In 2021, the first edition of the workshop (and shared task) on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP) was proposed. For this edition, the task of machine translation was presented to the participants. The goal of this shared task was to learn machine translation models for ten indigenous languages. The participants were given ten sets of language pairs: Quechua–Spanish, Wixarika–Spanish, Shipibo-Konibo–Spanish, Asháninka–Spanish, Raramuri–Spanish, Nahuatl–Spanish, Otomí– Spanish, Aymara–Spanish, Guarani–Spanish, and Bribri–Spanish (Mager et al., 2021). For the 2022 edition, the participants were asked to present novel speech-to-text translation systems for Bribri–Spanish, Guaraní–Spanish, Kotiria–Portuguese, Wa'ikhana–Portuguese, and Quechua–Spanish (Ebrahimi et al., 2022). Finally, in 2023, the task was machine translation for the

[1]This language is referred to by task organizers (and many speakers) simply as 'Maya' - we also use this shorter form.
[2]https://github.com/lecs-lab/americasnlp2024

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 159–173
June 21, 2024 ©2024 Association for Computational Linguistics

ten pairs mentioned above, plus a new language pair, Chatino-Spanish (Ebrahimi et al., 2023).

3 Related Work

Many of our strategies are inspired by research in morphological inflection. Morphological (re)inflection is the task of predicting an inflected form given a lemma or wordform and one or more target morphological features, and has been studied extensively through several shared tasks (Cotterell et al., 2016, 2017, 2018; Vylomova et al., 2020; Pimentel et al., 2021; Kodner et al., 2022; Goldman et al., 2023).

Morphological inflection has been studied with neural models such as RNNs (Kann and Schütze, 2016), convolutional neural networks (Östling, 2016), variational autoencoders (Zhou and Neubig, 2017), and transformers (Wu et al., 2021).

Data augmentation has been proposed as a strategy to address the challenges of training neural models on inflection tasks, particularly with limited data. Approaches have included creating artificial examples that copy the inputs directly to the outputs (Kann and Schütze, 2017; Bergmanis et al., 2017; Liu and Hulden, 2022; Yang et al., 2022), creating synthetic examples using morphological analyzers (Nicolai et al., 2017), and editing substrings using various methods to identify candidate stems (Silfverberg et al., 2017; Anastasopoulos and Neubig, 2019).

4 Models

We explore a number of models, including small sequence-to-sequence models, pretrained multilingual models, and large language models. For most models, input for a given instance consists of the source sentence plus the expected set of linguistic changes (e.g. PERSON:1_PL in Table 1).

4.1 Character-level neural models

We compare several different small character-level sequence-to-sequence models, using the `Yoyodyne` library for implementation.[3]

LSTM. We use a standard encoder-decoder LSTM with cross-attention. LSTMs have proven effective at inflection tasks (Cotterell et al., 2018), outperforming transformers under certain conditions (Wu et al., 2021). The expected linguistic changes are concatenated with the source sentence.

Transformer. Wu et al. (2021) also finds that in many cases, the transformer can outperform recurrent networks at character-level tasks. Thus, we also compare with an encoder-decoder transformer. Linguistic changes are treated as in the LSTM.

Pointer-generator. For tasks such as summarization (and the current task!) where the output sequences may share many tokens with the input sequence, the pointer-generator mechanism (See et al., 2017) has proven effective. The mechanism is a modification of an encoder-decoder architecture that introduces a pointing mechanism, where the model can copy a token from the input sequence rather than generating a novel token. Unlike the prior models, linguistic changes are encoded and attended to separately, so that they cannot be "pointed to" by the pointer-generator mechanism. We explore both LSTMs and Transformers with pointer-generator mechanisms.

We performed a hyperparameter search to determine the optimal hyperparameters for both the attentive-LSTM and pointer generator. The results of our search, our final hyperparameters, are given in Table 2. The full hyperparameter space we explored is reported in Appendix A. We train all models on a NVIDIA A100 GPU, with Adam optimization, a linear scheduler, a learning rate of 0.001, and a dropout of 0.2. We also explored using a larger architecture with the parameters described in Yang et al. (2022), however, we find these models nearly always underperform by a wide margin.[4]

4.2 Pretrained multilingual models

Transfer learning is a common strategy used to overcome limited data in lower-resource languages. To this end, we utilize mBART (Liu et al., 2020), which has shown a promising capability of generalization in the case of unseen languages (Liu et al., 2021). The desired linguistic change is appended to the source sentence, separated by the model separation token.

4.3 Large language models

Large language models (LLMs) generally struggle on rare, low-resource languages that are not well-represented in their training corpora (Robinson et al., 2023; Ahuja et al., 2023). However,

[3] `https://github.com/CUNY-CL/yoyodyne`

[4] Results are given in Appendix B. We observe that the larger models tend to overfit the training data, with much higher validation loss than their smaller counterparts.

Model	Language	Hyperparameters					
		Batch Size	Embedding Size	Hid. Size	Attn Heads	Enc. Layers	Dec. Layers
LSTM	Bribri	32	512	448	1	1	1
	Maya	32	256	896	1	2	1
	Guaraní	16	256	1152	1	1	1
PG	Bribri	32	256	1280	2	1	1
	Maya	64	448	1728	1	1	1
	Guaraní	16	192	1152	1	1	1

Table 2: Hyperparameters for LSTM and Pointer Generator models for three languages

LLMs may be able to achieve better performance on these languages through **in-context learning** (also known as *few-shot prompting*), where a small number of examples for a novel task are provided in the prompt at inference time (Brown et al., 2020). With ever-increasing context lengths, LLMs have even been able to learn completely novel languages using comprehensive linguistic resources provided in the context (Tanzer et al., 2024).

We utilize the ChatGPT API and the GPT-4 model to study in-context learning for our sentence transformation task (OpenAI et al., 2024). Since the provided training splits are very small, we provide the entire training set as context in our prompts. We also experiment with attempting to provide a more focused, relevant context, by filtering training examples to only those that have a linguistic change in common with the test sentences.

We utilize the `gpt-4-0125-preview` model, with temperature of 0 and a fixed random seed of 430. Full details about our prompting strategy are provided in Appendix C. As making an API call for every unique test example is fairly expensive, we prompt the model to make predictions on chunks consisting of multiple examples. We experiment with chunks of 20 and 80 examples.

Split	# examples Bribri	Guaraní	Maya
Train	309	178	594
Dev	212	79	149
SENTENCE COPYING	331	226	749
TRANSITIVE TRANSFORM.	3392	195	1671
STEM PERTURB.	200	200	200
CONCATENATION	500	500	500
EMBEDDINGS	300	250	-

Table 3: The number of examples in the train and dev split (top) and the number of artificial examples created by each augmentation strategy (bottom).

5 Data Augmentation

In very low-resource settings, data augmentation can be highly effective at improving output quality and performance. We employ a number of strategies for augmentation. Table 3 summarizes the training splits and number of artificial examples created by each strategy. Examples of each augmentation strategy appear in Table 10 (Appendix D).

Sentence copying (COPY). A major challenge in this task is that the sentences in the evaluation set include lemmas and words which are not present in the training set. To address this, we use a variation of the *lemma copying* technique described in Liu and Hulden (2022); Yang et al. (2022), which we designate *sentence copying*.

In this technique, we create additional training examples where the source and target sentence are identical and the *Change* field is blank. We create examples for every source sentence and target sentence in the training set (COPY_{tr}). We also experiment with creating examples for every source sentence in the dataset being used for evaluation, and add these to the former to create COPY_{all}. This technique, a form of domain adaptation, provides the model with a bias towards copying and aids the decoder in producing coherent sentences in the language. COPY_{all} was not an allowable strategy for our final shared task submission, but we include the results here for comparison.

External sentence copying (COPY_{ext}). As external resources are valid for the shared task, we can extend the coverage provided by the sentence copying technique by using data from outside the provided datasets, similar to the approach used in Kann and Schütze (2017). We find existing unlabeled text corpora in the languages and create additional COPY rows for every sentence.

For Maya, we extract transcriptions from the

ELAN[5] (Sloetjes and Wittenburg, 2008) data in the Yucatec Maya DoReCo dataset (Skopeteas, 2022). We discard non-utterance transcriptions (such as pauses) but keep the same segmentation as the original transcription (which may not be grammatically complete sentences). For Bribri, we leverage the dataset provided by the AmericasNLP 2024 Shared Task 1;[6] we also use the provided orthographic conversion tool.[7] Finally, for Guaraní, we use a portion of the CC-100 corpus (Conneau et al., 2020).

All datasets were sanity-checked to ensure they used orthographies comparable to the training data for a given language, but no comprehensive analysis was performed for orthographic alignment. We also filter the datasets by excluding utterances which are significantly longer than those in the shared task training or dev sets.[8]

Transitive transformations (TRANS). In the standard inflection task, inputs are lemmas and outputs are inflected word forms. In this task, however, the inputs are grammatical sentences (as there is no clear equivalent for a lemma form of a sentence) and have non-null linguistic features already.

For example, there are instances in the datasets which transform a sentence to carry second person inflection. Presumably, the source sentence in these instances is either first or third person; the linguistic features of the source sentences are not specified. If there is *also* an instance in the dataset where the same source sentence is transformed to carry third person inflection, then we know there is a relationship between the two target sentences (in addition to their relationships to the common source sentence).

In these cases, we can create an additional example that takes one of the target sentences as input and produces the other target sentence, using the linguistic change from the latter instance (and vice versa). We can use this strategy for any pair of examples where the source sentence is identical and the linguistic change of the latter sentence replaces all of the feature values of the former. We describe this strategy as *transitive transformations*.

Stem perturbation (PER). We follow the insights of Silfverberg et al. (2017) and Anastasopoulos and Neubig (2019), which seek to replace stems with random character sequences from the language. Different approaches have been used to identify stems: Silfverberg et al. (2017) uses the longest common substring, while Anastasopoulos and Neubig (2019) uses character alignment to select substrings that are aligned between the lemma and inflected form.

We use an alternate strategy based on edit trees. Starting with a source sentence, we randomly change one or two characters (via deletion, or via insertion of or replacement with a random character from the domain character set); if the edit trees which could be applied to the original source can be also applied to the altered sentence, the latter is considered valid and added to the pool of possible augmentations. We repeat this process ten times per original source sentence (with each altered sentence serving as the new 'source' sentence), then randomly sample from the pool of possible augmentations for training.

Concatenation (CON). For this strategy, we select sentence pairs that have exactly the same set of linguistic transformations. We then produce a new training example by concatenating the two source sentences to be the new source, and concatenating the two target sentences to be the new target output.

Embedding-based augmentation. A more structured approach to augmentation is to replace words with their synonyms whenever possible while keeping the sentence structure and type of transformation constant. To find synonyms in our vocabulary, we first train language-specific static embeddings over external datasets for Guaraní and Bribri. For this purpose, we simply use the data provided as part of the first shared task of AmericasNLP 2024.

Deviating from our previous character-based approach, we use byte-pair encoding to tokenize our data. We then train a word2vec model and use these vectors as subword representations. Words that are not inflected in the training data[9] are replaced with a randomly sampled word from its top 3 most similar words in the embedding space. This allows us to create duplicates of both source and target sentences with minimal, targeted alteration

[5] https://archive.mpi.nl/tla/elan

[6] https://turing.iimas.unam.mx/americasnlp/2024_st_1.html

[7] https://github.com/AmericasNLP/americasnlp2024/

[8] Arbitrarily defined, per language, as 1.5 times the max length in characters of a sentence in the training or dev set.

[9] After byte-pair encoding, we create a list of standalone tokens and use them as candidates for synonym replacement. Our BPE encoder uses underscores to denote that a token is inflected or acts as an inflection. We assume that these standalone tokens that frequently appear without underscores can be replaced with a synonym.

to the semantic and morpho-syntactic content of the data.

6 Results and Discussion

6.1 Evaluation

We report results on the evaluation split provided for the shared task. Models are evaluated with per-sentence accuracy, BLEU score (Papineni et al., 2002), and CHRF score (Popović, 2015).

6.2 Models

We compare the various architectures described in section 4 and report results in Table 4.

Character-level neural models. Our character-level models strongly outperform the baselines on Maya, are competitve on Bribri, and underperform on Guaraní. Within the character-level architectures, the LSTM models perform best in nearly all cases. For the smaller datasets (which have roughly 200-300 training examples), the standard LSTM model achieves the best performance, while on Maya ($\sim$ 600 examples) the pointer-generator LSTM outperforms. This may indicate that the pointer-generator model needs a certain amount of training data to effectively utilize the pointing mechanism and outperform a standard LSTM, and only the Maya dataset meets that threshold.

For Guaraní, all of the sequence-to-sequence models perform very poorly. Qualitative analysis of the results shows that the models struggle to repeat back valid sentences in the language at all.

Pretrained multilingual models. mBART achieves our second best performance on Maya (second to pointer-generator LSTM), and the results for Guaraní and Bribri are also competitive with those of ChatGPT models. Unlike the character-level models, mBART tokenizes the source into subwords; hinting at the possible advantages of using subwords and the information they could carry from the model being pretrained on other languages.

Large language models. The ChatGPT-based approach achieves competitive performance, providing evidence that the model is able to capture some patterns correctly through in-context learning. The approach outperforms all other models on Guaraní (the language with the least training data), demonstrating that the LLM is able to leverage its vast training knowledge as a strong prior on the task at hand, and to make robust generalizations from the available data.

We observe minimal differences based on the chunk size, except for Maya where the smaller chunk size performs significantly better. The system using smart retrieval (SR) is able to achieve close performance for Guaraní and Maya, but underperforms on Bribri; SR is potentially a viable way to reduce prompt size and thereby cost.

LLMs offer a promising approach to building NLP systems for under-resourced languages, particular when using in-context learning for rare languages, as here. However, the high cost of inference, lack of control (due to the closed-source nature of the models), and privacy concerns are major considerations for practical usage in an endangered language context.

6.3 Data augmentation

Based on the results of the previous section, we select the LSTM and pointer-generator LSTM for our experiments with various augmentation strategies. Noting that the three shared task metrics do not always align in their assessment of best-performing model, we primarily focus on chrF, as accuracy and BLEU score tend to have high variability.[10]

We present results for models trained using each of the data augmentation strategies in Figure 1. The copying strategies tend to be the strongest, followed by the stem perturbation strategy. The other strategies show mixed results, and in some cases underperform the baseline.

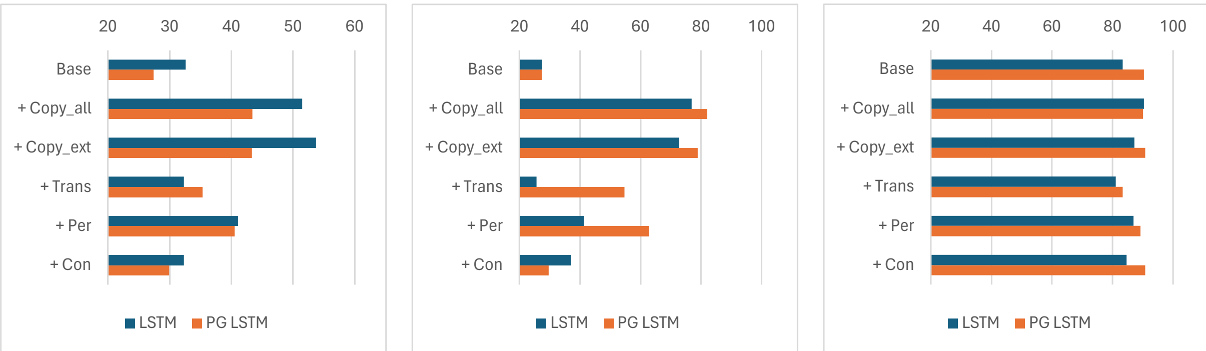
Sentence copying. We focus on a number of variations and combinations of the copy strategy and report results in Figure 2, finding that all of our strategies generally improve over the baseline. Unsurprisingly, the models trained on data including the source sentences of the evaluation set outperformed those without by an average of 14.46 chrF points. This strategy, in which the model is retrained before running inference and the target outputs are neither required nor exposed, provides clear benefits in this highly low-resource scenario.

The COPY_{ext} strategies show mixed results, sometimes matching or outperforming the COPY_{all} strategies (as in Bribri) but sometimes underperforming (as in Maya, LSTM). Combining strategies shows mixed results, and we suspect that after

[10]We observe these metrics jump wildly during training. Furthermore, having even a single incorrect output character can affect the accuracy and BLEU metrics significantly.

Architecture	Bribri Acc.	Bribri BLEU	Bribri chrF	Guaraní Acc.	Guaraní BLEU	Guaraní chrF	Maya Acc.	Maya BLEU	Maya chrF
Naive Copy	0.00	10.59	38.42	0.00	23.33	71.47	0.00	33.67	69.15
Edit Trees	5.66	20.35	45.56	22.78	34.99	77.14	26.17	52.38	78.72
LSTM	0	**19.73**	**32.57**	0	**1.95**	27.43	40.94	61.24	83.33
PG-LSTM	0	17.38	27.36	0	1.64	27.34	**51.68**	**75.51**	**90.37**
TRANSFORMER	0	13.29	29.17	0	1.27	**27.90**	16.11	42.33	70.33
PG-TRANSFORMER	0	7.9	23.09	0	0.64	22.16	10.74	36.45	64.74
MBART	5.66	40.13	60.43	**32.91**	35.12	77.62	50.34	74.12	88.70
ChatGPT									
chunksize = 20	12.26	43.43	**63.31**	**32.91**	45.63	79.21	**48.99**	74.46	**89.54**
chunksize = 80	**12.74**	**43.87**	62.39	**32.91**	**48.70**	80.32	32.89	51.36	69.84
chunksize = 1, SR	6.13	39.42	57.67	30.38	45.55	**81.80**	48.32	**74.50**	88.47

Table 4: Results for different models on development data, with no data augmentation. We **bold** the best results overall and the best results within each section. PG = pointer-generator. SR = smart retrieval.



(a) Bribri (b) Guaraní (c) Maya

Figure 1: chrF results for various data augmentation strategies.

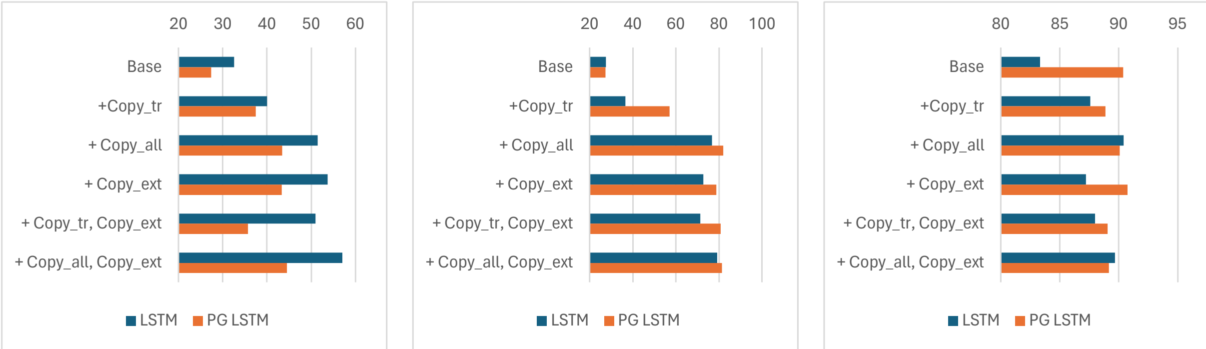
a certain number of synthetic examples, the utility of this strategy declines.

Combined strategies. Finally, we experiment with combinations of augmentation strategies, directly concatenating the synthetic datasets, with results in Figure 3. We observe mixed results—for Guaraní and Maya, none of the combined strategies show significant improvements over individual strategies, and in some cases performance degrades somewhat. We do see improvements in Bribri with the combined COPY$_{all}$ + PER strategy and the COPY$_{all}$ + PER + CON strategy over any of the individual strategies. Broadly, we find that synthetic data of this sort can only help up to a certain amount, and creating more synthetic data does not necessarily continue to improve performance.

7 Shared Task Submission

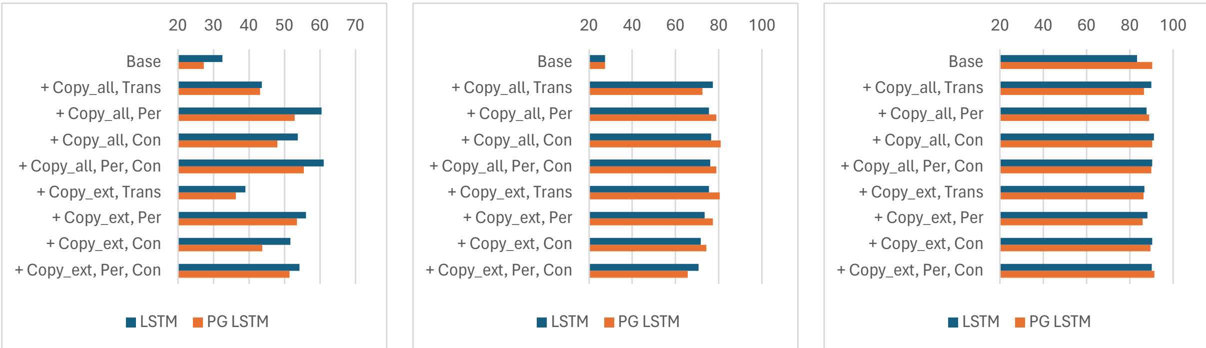
We selected a number of systems for final submission to the shared task, based on our evaluation results. We use the ChatGPT system with a chunk size of 20, the MBART system, and several of the augmented character-level neural systems. We aim to select a diverse set of augmented systems, so we select the COPY$_{ext}$, COPY$_{tr}$ + COPY$_{ext}$, and COPY$_{ext}$ + PER systems for the LSTM model and the COPY$_{ext}$, COPY$_{ext}$+TRANS, and COPY$_{ext}$+PER+CON systems for the pointer-generator model.

We train final models using the training data and specified synthetic dataset. We perform hyperparameter search and select the optimal model architecture for each language and model, which we report in appendix A. We train models for 1000 epochs, selecting the best model according to vali-



(a) Bribri (b) Guaraní (c) Maya

Figure 2: chrF results for strategies incorporating sentence copying using various sources. COPY$_{tr}$ uses only the training data. COPY$_{all}$ uses training data and source sentences from the evaluation data. COPY$_{ext}$ uses sentences from external corpora.



(a) Bribri (b) Guaraní (c) Maya

Figure 3: chrF results for combinations of strategies.

dation accuracy.

We report results from the covered test set in Table 5. Disappointingly, we observe significant performance discrepancies from our dev set results, with only the ChatGPT-based system maintaining similar scores. We propose three possible factors that could have caused this.

First, all of the datasets involved are quite small, and it is possible that through random variability, the test set was meaningfully different in distribution from the evaluation set. Neural models can be vulnerable to distributional shift, particularly when training data is scarce (Linzen, 2020), which may explain why the non-neural baseline model fared better.

We briefly investigate whether this is the case by examining the types of linguistic changes in each data split. Specifically, for each desired linguistic change in the evaluation and test datasets (which might include multiple changes from a single example), we compute the number of times that change occurs in the training dataset, and average over all changes. This gives us a rough estimate of how common the linguistic changes are in the model's training data.

We report these results in Table 6. We find that for Bribri and Guaraní, the distribution is very similar between the dev and test sets, while for Maya, the test set contains changes that are far more rare (-23.6 points) on average. As Maya was the language where we observed the greatest discrepancy in performance, this could be a contributing factor, and represents an important consideration for neural models.

The other potential contributing factor is that due to the small datasets and difficult nature of the task, the performance of our models was highly variable. For augmentation strategies such as synonym replacement, the base assumption that synonyms are even present in a dataset of this size might not be accurate. During training, we often observed dev accuracy curves that swung wildly, sometimes jumping up or down by 10 points in a single epoch. Furthermore, since we performed a large number

#	Architecture	Bribri Acc.	Bribri BLEU	Bribri chrF	Guaraní Acc.	Guaraní BLEU	Guaraní chrF	Maya Acc.	Maya BLEU	Maya chrF
	Baseline (Edit Trees)	8.75	22.11	52.73	14.84	25.03	76.10	25.81	53.69	80.23
1	ChatGPT	12.08	36.95	66.75	30.77	45.18	82.33	51.61	76.82	90.29
	LSTM models									
2	+COPY$_{ext}$	3.96	16.45	47.74	7.69	17.80	70.54	19.35	57.60	78.29
3	+COPY$_{ext}$ + COPY$_{tr}$	5.00	19.77	48.26	9.34	13.15	67.20	18.71	50.21	76.19
4	+COPY$_{ext}$ + PER	4.17	16.34	51.81	8.24	15.34	66.82	16.77	59.19	79.34
	PG models									
5	+COPY$_{ext}$	0.62	13.52	34.75	7.69	20.74	71.18	30.32	60.14	79.70
6	+COPY$_{ext}$ + TRANS	0.21	7.73	31.29	11.81	17.55	69.13	15.81	43.75	71.75
7	+COPY$_{ext}$ + PER + CON	5.21	27.72	56.81	12.09	22.54	71.85	34.84	69.18	85.89
8	MBART	0.83	9.90	36.47	3.30	13.84	61.46	35.16	68.11	86.04
9	EMBEDDING AUG. + LSTM	0.83	7.91	47.76	0.55	3.80	56.21	-	-	-

Table 5: Test results for our submitted models.

Language	Dev	Test
Bribri	71.9	77.5
Guaraní	12.8	12.1
Maya	71.4	47.8

Table 6: Average frequency in the *training* data of each linguistic change observed in the dev and test set.

of experiments and selected our final models using the same evaluation set, we may have unintentionally overfit to the specific evaluation set and chosen systems that did not generalize well to the new data. In the future, this could be avoided by using many-fold cross-validation to select models rather than a single dev set.[11]

Finally, we saw significant performance benefits to including sentence copying in Figure 2, and we employed this in all of our submitted character-level systems. However, this strategy is most beneficial when it includes the sentences and lemmas that appear in the data being evaluated. It is possible that our external corpora happened to contain more overlap with the dev set examples than those in the test set, which could significantly impact performance. We suspect the strategy of retraining including the test inputs as synthetic examples could alleviate this.

Overall, these results serve as a cautionary example of the risks of selecting final systems based on limited evaluation metrics in extremely low-resource scenarios.

[11]We considered this, but it was ultimately too resource-intensive for the number of experiments we wished to run.

8 Conclusion

We describe our systems for the 2024 Americas-NLP Shared Task on the Creation of Educational Materials for Indigenous Languages, which include LLM-based systems, character-level neural networks, and finetuned multilingual models. We observe potential benefits from augmentation strategies for character-level models, particularly the *sentence copying* strategy, which helps a model adapt to new examples.

However, we find that nearly all of our systems, with the exception of the LLM system, do not generalize well to the covered test set, resulting in poor performance on the shared task. These results reaffirm the difficulty of training robust neural models in low-resource scenarios and the importance of thorough validation.

Acknowledgements

Thanks to Adam Wiemerslage for assistance with Yoyodyne. This work utilized the Blanca condo computing resource at the University of Colorado Boulder. Blanca is jointly funded by computing users and the University of Colorado Boulder.

Portions of this work were supported by the National Science Foundation under Grant No. 2149404, “CAREER: From One Language to Another”. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the National Science Foundation.

References

Kabir Ahuja, Harshita Diddee, Rishav Hada, Millicent Ochieng, Krithika Ramesh, Prachi Jain, Akshay Nambi, Tanuja Ganu, Sameer Segal, Mohamed Ahmed, Kalika Bali, and Sunayana Sitaram. 2023. MEGA: Multilingual evaluation of generative AI. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4232–4267, Singapore. Association for Computational Linguistics.

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

Toms Bergmanis, Katharina Kann, Hinrich Schütze, and Sharon Goldwater. 2017. Training data augmentation for low-resource morphological inflection. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 31–39, Vancouver. Association for Computational Linguistics.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Luis Chiruzzo, Pavel Denisov, Samuel Canul Yah, Lorena Hau Ucán, Marvin Agüero-Torales, Aldo Alvarez, Silvia Fernandez Sabido, Alejandro Molina Villegas, Abteen Ebrahimi, Robert Pugh, Arturo Oncevay, Shruti Rijhwani, Rolando Coto-Solano, Katharina von der Wense, and Manuel Mager. 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)*. Association for Computational Linguistics.

Alexander Clark. 2002. Memory-based learning of morphology with stochastic transducers. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 513–520.

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

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

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sandra Kübler, David Yarowsky, Jason Eisner, and Mans Hulden. 2017. CoNLL-SIGMORPHON 2017 shared task: Universal morphological reinflection in 52 languages. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 1–30, Vancouver. Association for Computational Linguistics.

Ryan Cotterell, Christo Kirov, John Sylak-Glassman, David Yarowsky, Jason Eisner, and Mans Hulden. 2016. The SIGMORPHON 2016 shared Task—Morphological reinflection. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 10–22, Berlin, Germany. Association for Computational Linguistics.

Greg Durrett and John DeNero. 2013. Supervised learning of complete morphological paradigms. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1185–1195, Atlanta, Georgia. Association for Computational Linguistics.

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.

Abteen Ebrahimi, Manuel Mager, Adam Wiemerslage, Pavel Denisov, Arturo Oncevay, Danni Liu, Sai Koneru, Enes Yavuz Ugan, Zhaolin Li, Jan Niehues, Monica Romero, Ivan G Torre, Tanel Alumäe, Jiaming Kong, Sergey Polezhaev, Yury Belousov, Wei-Rui Chen, Peter Sullivan, Ife Adebara, Bashar Talafha, Alcides Alcoba Inciarte, Muhammad Abdul-Mageed, Luis Chiruzzo, Rolando Coto-Solano, Hilaria Cruz, Sofía Flores-Solórzano, Aldo Andrés Alvarez López, Ivan Meza-Ruiz, John E. Ortega, Alexis Palmer, Rodolfo Joel Zevallos Salazar, Kristine Stenzel, Thang Vu, and Katharina Kann. 2022. Findings of the second americasnlp competition on speech-to-text translation. In *Proceedings of the NeurIPS 2022 Competitions Track*, volume 220 of *Proceedings of Machine Learning Research*, pages 217–232. PMLR.

Manaal Faruqui, Yulia Tsvetkov, Graham Neubig, and Chris Dyer. 2016. Morphological inflection generation using character sequence to sequence learning. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 634–643, San Diego, California. Association for Computational Linguistics.

Omer Goldman, Khuyagbaatar Batsuren, Salam Khalifa, Aryaman Arora, Garrett Nicolai, Reut Tsarfaty, and Ekaterina Vylomova. 2023. SIGMORPHON–UniMorph 2023 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 117–125, Toronto, Canada. Association for Computational Linguistics.

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

Katharina Kann and Hinrich Schütze. 2017. Unlabeled data for morphological generation with character-based sequence-to-sequence models. In *Proceedings of the First Workshop on Subword and Character Level Models in NLP*, pages 76–81, Copenhagen, Denmark. Association for Computational Linguistics.

Jordan Kodner, Salam Khalifa, Khuyagbaatar Batsuren, Hossep Dolatian, Ryan Cotterell, Faruk Akkus, Antonios Anastasopoulos, Taras Andrushko, Aryaman Arora, Nona Atanalov, Gábor Bella, Elena Budianskaya, Yustinus Ghanggo Ate, Omer Goldman, David Guriel, Simon Guriel, Silvia Guriel-Agiashvili, Witold Kieraś, Andrew Krizhanovsky, Natalia Krizhanovsky, Igor Marchenko, Magdalena Markowska, Polina Mashkovtseva, Maria Nepomniashchaya, Daria Rodionova, Karina Scheifer, Alexandra Sorova, Anastasia Yemelina, Jeremiah Young, and Ekaterina Vylomova. 2022. SIGMORPHON–UniMorph 2022 shared task 0: Generalization and typologically diverse morphological inflection. In *Proceedings of the 19th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 176–203, Seattle, Washington. Association for Computational Linguistics.

Tal Linzen. 2020. How can we accelerate progress towards human-like linguistic generalization? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5210–5217, Online. Association for Computational Linguistics.

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

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

Zihan Liu, Genta Indra Winata, and Pascale Fung. 2021. Continual mixed-language pre-training for extremely low-resource neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2706–2718, Online. 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.

Garrett Nicolai, Colin Cherry, and Grzegorz Kondrak. 2015. Inflection generation as discriminative string transduction. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 922–931, Denver, Colorado. Association for Computational Linguistics.

Garrett Nicolai, Bradley Hauer, Mohammad Motallebi, Saeed Najafi, and Grzegorz Kondrak. 2017. If you can't beat them, join them: the University of Alberta system description. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 79–84, Vancouver. Association for Computational Linguistics.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Robert Östling. 2016. Morphological reinflection with convolutional neural networks. In *Proceedings of the 14th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 23–26, Berlin, Germany. 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.

Tiago Pimentel, Maria Ryskina, Sabrina J. Mielke, Shijie Wu, Eleanor Chodroff, Brian Leonard, Garrett Nicolai, Yustinus Ghanggo Ate, Salam Khalifa, Nizar Habash, Charbel El-Khaissi, Omer Goldman, Michael Gasser, William Lane, Matt Coler, Arturo Oncevay, Jaime Rafael Montoya Samame, Gema Celeste Silva Villegas, Adam Ek, Jean-Philippe Bernardy, Andrey Shcherbakov, Aziyana Bayyr-ool, Karina Sheifer, Sofya Ganieva, Matvey Plugaryov, Elena Klyachko, Ali Salehi, Andrew Krizhanovsky, Natalia Krizhanovsky, Clara Vania, Sardana Ivanova, Aelita Salchak, Christopher Straughn, Zoey Liu, Jonathan North Washington, Duygu Ataman, Witold Kieraś, Marcin Woliński, Totok Suhardijanto, Niklas Stoehr, Zahroh Nuriah, Shyam Ratan, Francis M. Tyers, Edoardo M. Ponti, Grant Aiton, Richard J. Hatcher, Emily Prud'hommeaux, Ritesh Kumar, Mans Hulden, Botond Barta, Dorina Lakatos, Gábor Szolnok, Judit Ács, Mohit Raj, David Yarowsky, Ryan Cotterell, Ben Ambridge, and Ekaterina Vylomova. 2021. SIGMORPHON 2021 shared task on morphological reinflection: Generalization across languages. In *Proceedings of the 18th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 229–259, Online. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Nathaniel Robinson, Perez Ogayo, David R. Mortensen, and Graham Neubig. 2023. ChatGPT MT: Competitive for high- (but not low-) resource languages. In *Proceedings of the Eighth Conference on Machine Translation*, pages 392–418, Singapore. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

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

Stavros Skopeteas. 2022. Yucatec maya doreco dataset. In Frank Seifart, Ludger Paschen, and Matthew Stave, editors, *Language Documentation Reference Corpus (DoReCo) 1.2*. Leibniz-Zentrum Allgemeine Sprachwissenschaft & laboratoire Dynamique Du Langage (UMR5596, CNRS & Université Lyon 2), Berlin & Lyon.

Han Sloetjes and Peter Wittenburg. 2008. Annotation by category: ELAN and ISO DCR. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. In *The Twelfth International Conference on Learning Representations*.

Ekaterina Vylomova, Jennifer White, Elizabeth Salesky, Sabrina J. Mielke, Shijie Wu, Edoardo Maria Ponti, Rowan Hall Maudslay, Ran Zmigrod, Josef Valvoda, Svetlana Toldova, Francis Tyers, Elena Klyachko, Ilya Yegorov, Natalia Krizhanovsky, Paula Czarnowska, Irene Nikkarinen, Andrew Krizhanovsky, Tiago Pimentel, Lucas Torroba Hennigen, Christo Kirov, Garrett Nicolai, Adina Williams, Antonios Anastasopoulos, Hilaria Cruz, Eleanor Chodroff, Ryan Cotterell, Miikka Silfverberg, and Mans Hulden. 2020. SIGMORPHON 2020 shared task 0: Typologically diverse morphological inflection. In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 1–39, Online. Association for Computational Linguistics.

Shijie Wu, Ryan Cotterell, and Mans Hulden. 2021. Applying the transformer to character-level transduction. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1901–1907, Online. Association for Computational Linguistics.

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

Chunting Zhou and Graham Neubig. 2017. Morphological inflection generation with multi-space variational encoder-decoders. In *Proceedings of the CoNLL SIGMORPHON 2017 Shared Task: Universal Morphological Reinflection*, pages 58–65, Vancouver. Association for Computational Linguistics.

A Hyperparameter Search Space

We performed a hyperparameter search for the attentive-LSTM and pointer-generator models using the **sentence copying** data augmentation strategy. We used random search with the goal of maximizing validation accuracy. We report the search space we considered in Table 7.

Hyperparameter	Distribution	Values
Batch Size	categorical	16, 32, 64
Embedding Size	q_uniform	128 to 1024; q=64
Hidden Size	q_uniform	128 to 2048; q=64
Attention Heads	values	1, 2
Encoder Layers	values	1, 2, 3
Decoder Layers	values	1, 2, 3

Table 7: Hyperparameter Search Space

B Larger Architectures

For thoroughness, we also compare architectures using the architecture size described in Yang et al. (2022). We report these results in Table 8.

Except for the transformer models, these larger models well underperform their smaller counterparts, in many cases overfitting the training data and completely failing to generalize. The transformer models perform more robustly, and seem to benefit from deeper and larger architectures.

C LLM Prompting

We attempted two different prompting strategies for our Chat-GPT implementation.

In the first strategy, we used a full-context approach, using the entire language's training split as the context. We tried these two different chunk size settings, calling the API with chunks of 20 or 80 test sentences at a time.

In the second strategy, we tried a smart-retrieval approach with a chunk size of one to only provide relevant examples as context. Relevant examples were those with the same changes as the test sentences within the language's training split.

In Table 9, an example of the prompt we provided using the smart-retrieval approach for a sentence in Bribri is shown. Note that this prompt provides just one training instance; in our experiments we provided multiple instances per prompt.

D Augmentation Examples

We provide examples of the rows created by each augmentation strategy in Table 10

	Bribri			Guarani			Maya		
Architecture	Acc.	BLEU	chrF	Acc.	BLEU	chrF	Acc.	BLEU	chrF
LSTM	0	9.44	26.21	0	0.59	18.38	0	5.53	27.13
PG-LSTM	0	8.45	25.54	0	0.85	18.32	24.16	49.66	**76.77**
TRANSFORMER	0	**18.19**	**32.93**	0	1.42	**29.96**	**27.52**	**53.14**	74.18
PG-TRANSFORMER	0	0	0.26	0	0	0.33	0	0	1.61

Table 8: Results for different architectures, using larger model sizes of Yang et al. (2022). PG = pointer-generator.

Prompt

Below is an example of a sentence in Bribri, the linguistic change, and the target sentence after applying the change.

ID:	Bribri0303
Source:	Ye' shka'
Change:	TYPE:NEG, TENSE:PRF_PROG
Target:	Ye' kë̀ ku̱'bak shkṓk

Below is a similar example, where the source sentence and linguistic change are given, and the output sentence is not known. For this example, please output only the id and target sentence values, as in:

ID:	Some ID
Target:	Sentence after applying the change

Do not output any additional text, and do not output the Source or Change fields. This is very important, take your time and do not mess up or I will lose my job.

Example Input:

ID:	Bribri0367
Source:	Pûs kapë'wa̱
Change:	TYPE:NEG, TENSE:PRF_PROG
Target:	*

Model Response:

ID:	Bribri0367
Target:	Pûs kë̀ ku̱'bakapë'wa̱

Table 9: Example prompt given while LLM prompting.

Strategy		Source	Change	Target
COPY	(original)	Ko po ojupi	TENSE:FUT_SIM	Ko po ojupíta
	(augmented)	Ko po ojupi	NOCHANGE	Ko po ojupi
COPY$_{ext}$	(original)	-	-	-
	(augmented)	Nde ruvichápe	NOCHANGE	Nde ruvichápe
TRANS	(originals)	Che rasy Che rasy	PERSON:2_SI PERSON:1_PL_EXC	Nde nderasy Ore rorasy
	(augmented)	Nde nderasy	PERSON:1_PL_EXC	Ore rorasy
PER	(original)	Ha'e oguapy	PERSON:3_PL	Hikuái oguapy
	(augmented)	Ha'e ocguapy	PERSON:3_PL	Hikuái ocguapy
CON	(originals)	Nde nderejapói Apurahéi kuri	PERSON:3_PL PERSON:3_PL	Ha'ekuéra ndojapói Ha'ekuéra opurahéikuri
	(augmented)	Nde nderejapói apurahéi kuri	PERSON:3_PL	Ha'ekuéra ndojapói ha'ekuéra opurahéikuri
EMBED	(original)	Mombe'ukuéra omboty kuri pende arete	ASPECT:IPFV	Mombe'ukuéra omboty kuri hína pende arete
	(augmented)	Sombezlkuéra omboty-kuri pende arete	ASPECT:IPFV	ombeãrkuéra omboty kurir hína pende arete

Table 10: Example applications of our augmentation strategies. All examples are Guaraní.

The unreasonable effectiveness of large language models for low-resource clause-level morphology: In-context generalization or prior exposure?

Coleman Haley
School of Informatics
University of Edinburgh
Edinburgh, United Kingdom
coleman.haley@ed.ac.uk

Abstract

This paper describes the submission of Team "Giving it a Shot" to the AmericasNLP 2024 Shared Task on Creation of Educational Materials for Indigenous Languages. We use a simple few-shot prompting approach with several state of the art large language models, achieving competitive performance on the shared task, with our best system placing third overall. We perform a preliminary analysis to determine to what degree the performance of our model is due to prior exposure to the task languages, finding that generally our performance is better explained as being derived from in-context learning capabilities.

1 Introduction

This paper describes the submission of Team "Giving it a Shot" to the AmericasNLP 2024 Shared Task on Creation of Educational Materials for Indigenous Languages (Chiruzzo et al., 2024). This task covers three indigenous languages of the Americas: Yucatec Maya (yua), Guaraní (grn), and Bribri (bzd). The task is similar to the clause-level reinflection task described by Goldman and Tsarfaty (2022) and explored in a 2022 MRL shared task (Goldman et al., 2022). However, it is more challenging in a number of ways. The first is structural: the present shared task provides an input sentence, and what values of features should be changed, while the previous task provided all feature values present in the input and what they should be changed to. As such, the features describing the source must be learned latently. Other challenges come from differences in the languages covered: all three languages in this shared task are relatively low-resource, and correspondingly the training data in the shared task is also very limited (595 training examples at most).

However, as a morphological/morphosyntactic task[1], the input-output functions are relatively simple compared to many common tasks in NLP, being in all likelihood context-free or even regular (Karttunen and Beesley, 2005; Pullum and Gazdar, 1982; Roark and Sproat, 2001). Increasingly in NLP, even computationally complex tasks such as sentiment analysis are being framed as few-shot tasks for large language models (LLMs), with impressive results being obtained by presenting a few examples to a language model and allowing it to perform next-token prediction (Wang et al., 2024; Wei et al., 2022; Brown et al., 2020). The ability of such paradigms to improve performance over raw language model probabilities has been termed in-context learning; however, this term has been the subject of controversy, as it is not learning in the traditional machine learning sense, nor is it clear exactly how much information is being extracted from the context. For example, in the "in-context learning" of sentiment analysis, much of the relation between a sentence and a sentiment label is presumably latent in the pre-trained weights, and the examples serve moreso to "extract" that information from the model, enabling better generalization than could be expected from the information in the provided in-context examples alone.

This setting therefore represents an interesting case: if few-shot prompting works well here, will it be due to prior language exposure, or an ability to generalize simple functions from limited data? To explore this question, we create three simple few-shot prompting-based systems, based on two closed-source LLMs (GPT-3.5 and GPT-4) and one openly available model (Command R+), finding they perform competitively on the shared task. We permute the characters in the dataset to preserve the problem stucture while ablating language information, finding some evidence that the models primarily generalize in-context data, rather than using prior language exposure.

[1]Note that, as in prior work on clause-level morphology, the functions involved sometimes operate at the clause level

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 174–178

```
Here’s some examples.
Source,Change,Target
Táan a bin koonol tu k’íiwikil koonol,TYPE:NEG,Ma’ táan a bin koonol tu k’íiwikil koonoli’
Táan u bin koonol tu k’íiwikil koonol,TYPE:NEG,Leti’e’ ma’ táan u bin koonol tu k’íiwikil koonoli’
Jach k’a’abéet in bin tu k’íiwikil koonol,TYPE:NEG,Ma’ jach k’a’abéet in bin tu k’íiwikil koonoli’
Táan a bine’ex ich kool,TYPE:NEG,Ma’ táan a bine’ex ich kooli’
Teche’ ka bin xíimbal tu yotoch,TYPE:NEG,Teche’ ma’ ta bin xíimbal tu yotochi’
...
Now fill in the third column:
Te’exe’ táan a bine’ex koonol tu k’íiwikil koonol,TYPE:NEG,Te’exe’ ma’ táan a bine’ex koonol tu k’íiwikil koonoli’
```

Figure 1: Sample prompt (examples abbreviated). The real output of GPT-4 is shown in **bold**. The same prompt format is used for all systems and languages.

2 Method

We treat the task as a simple few-shot prompting problem, using no external data. We consider three models: two closed-source (`gpt-4-0125` and `gpt-3.5-turbo-0125` from OpenAI) and one open-source (Command R+ from Cohere). Our prompt is minimal, de-emphasising problem specific factors. It simply presents relevant examples from the training data in a CSV format, then asks the model to complete the third column of a test item CSV row. We use no additional data besides the provided training set, and perform no fine-tuning. A sample prompt is shown in Figure 1. In contrast to prior work showing the utility of expert prompting (Xu et al., 2023), describing the task domain (Zhang et al., 2024), and tipping (Salinas and Morstatter, 2024), preliminary evidence showed limited effects of any of these techniques when augmenting our prompt format. Indeed, treating the problem as a simple CSV completion task seems to have triggered interesting behavior in all 3 models: almost without exception, the first line contained either just the predicted target, or all three completed columns separated by commas. Indeed, even on our worst-performing model and language pair, Bribri using the Cohere model, only 4/480 examples in the test set are miss-parsed by these heuristics (i.e., yielding something other than the model’s prediction), in contrast to prior works where large language models typically place their completions unpredictably in unstructured text, causing parsing errors.

While the format of the prompt is simple, some heuristics are required to best make use of the provided training data and compute costs. Typically, only examples of the requested change are shown. In cases where more than 10 examples of a particular change occur in the training data, the training data exhibiting this change is sorted according to the sum of BLEU (Papineni et al., 2002) and chrF (Popović, 2015) with the source of the test item as a reference, and the 10 examples with the highest score used in the prompt. Further, when the specific change occurs fewer than 3 times in the training data (as is often the case in Bribri), we back off to similar changes: we break the queried change into the component feature changes, and take up to 3 instances of each component change. If a feature change does not occur on its own in the training data, we add one example containing the feature change which maxmizes the sum of BLEU and chrF between its source and the target source. Finally, we add up to 8 examples which contain some of the component changes, again chosen by their source BLEU+chrF similarity.

We use temperature 0.1 for the OpenAI models and temperature 0.3 for Cohere models. Preliminary evidence suggested that lower temperatures aided consistency.

3 Results

Our results on the test set are shown in full in Table 1. All of the models improve over the provided baseline for at least one of the languages. Command R+ struggles the most with the task, scoring below the baseline for Bribri and Guaraní (though improving substantially in terms of BLEU and chrF for the former). All systems improve dramatically over the baseline for Maya, which had the most provided training examples (595), with few requiring backoff. The systems performed competitively in the shared task, with

(e.g. the addition of a particle to express negation)

System	Bribri			Maya			Guaraní		
	Acc.	BLEU	chrF	Acc.	BLEU	chrF	Acc.	BLEU	chrF
`command-r-plus`	*7.08*	31.68	62.45	49.03	73.09	88.54	*9.34*	*22.64*	*73.40*
`gpt-3.5-turbo-0125`	11.67	33.80	65.51	50.97	75.09	89.76	18.13	31.94	79.36
`gpt-4-0125`	**17.71**	**39.48**	**69.28**	**53.87**	**78.54**	**91.66**	**25.00**	**40.55**	**81.71**
Baseline (edit trees)	8.75	22.11	52.73	25.81	53.69	80.23	14.84	25.03	76.10

Table 1: Performance of our three submissions for each language compared to the provided baseline. Our simple GPT-4-based system is our best across all metrics (shown in **bold**), placing third overall in the shared task. Scores below the baseline are shown in *italics*.

Command R+ placing 8th, `gpt-3.5-turbo-0125` placing 6th, and `gpt-4-0125`, placing third overall and coming in first for Maya (tying the second place system in accuracy but out-performing in terms of the secondary metrics). Overall, these results indicate that even very simple approaches using large language models can be useful for low-resource morpho-syntactic tasks, when training data is limited. However, choice of model remains important–despite the fact that Command R+ both out-performs `gpt-3.5-turbo` on MMLU and ranks higher on the LMSys Chatbot Arena[2], it substantially under-performs on both Bribri and Guaraní.

4 In-context generalization or prior exposure?

While our results suggest that large language models *can* solve complex clause-level reinflection tasks for some indigenous languages, it is unclear *what drives* this behavior. One hypothesis is that it is driven largely by prior exposure to these languages. The Glot500 dataset, which attempts to collate large amounts of data for low-resource languages for language model pretraining, contains 610,052 Maya sentences; 87,568 Guaraní sentences, and none for Bribri (ImaniGooghari et al., 2023). Attempts to develop a large corpus for Bribri have so far maxed out at just around 100,000 tokens, even with manual gathering of data from books not on the internet (Coto-Solano, 2022). This lines up relatively neatly with our results, with Maya > Guaraní > Bribri.

Another possibility is that the model is primarily generalizing the patterns of in-context examples it is provided. Support for this account is provided from the observation that the same pattern of Maya > Guaraní > Bribri is evident in the baseline, which has no prior language exposure; suggesting that the inherent difficulty of the task may vary between the languages/their datasets. As such, our primary results alone are ambiguous between these two hypotheses.

To differentiate these hypotheses, we develop a simple test involving *permuting* the alphabet for each language, such that most characters are mapped to other characters. This should provide a problem of an equivalent difficulty to the original, but which has a very different distribution over tokens, which should limit the degree to which the model uses information from prior exposure to the languages. To ensure the difficulty characteristics of the problem are preserved, letters with diacritics are permuted analogously to their counterparts without diacritics. This is due to the observation that a positive quality of our systems is their tendency to generalize patterns that apply to one set of diacritics on a letter to different diacritics on that letter. As an example, here is a real Maya sentence from the dataset followed by its permuted counterpart:

(1) Teche’ ka bin xíimbal tu najil *Original*

(2) Kitsi’ pe dun cúumder ko neyur *Permuted*

As the structure of the shared task prevents us from evaluating on a permuted test set, we present results on the development set for this experiment, shown in Table 2. We note that the results should be interpreted in light of the fact that there is substantial variability (on the order of ≈ 5 percentage points) from run-to-run. Ideally, we would run this experiment repeatedly to compute confidence intervals, but resource constraints prevent this.

Overall, our results suggest that our model performance is mostly a result of generaliz-

[2]https://leaderboard.lmsys.org

System	Data	Bribri Acc.	Bribri BLEU	Bribri chrF	Maya Acc.	Maya BLEU	Maya chrF	Guaraní Acc.	Guaraní BLEU	Guaraní chrF
`command-r-plus`	orig.	**10.85**	**40.61**	**55.93**	**44.96**	**72.96**	**88.55**	**22.78**	35.42	**76.09**
	perm.	5.66	34.47	55.22	43.62	72.77	88.14	21.52	**41.95**	75.22
`gpt-3.5-turbo-0125`	orig.	4.72	37.23	57.52	**51.68**	76.20	**90.16**	**34.18**	44.95	**82.07**
	perm.	**9.91**	**37.72**	**58.22**	**51.68**	**76.62**	89.97	32.91	**49.17**	81.83
`gpt-4-0125`	orig.	15.57	41.35	62.88	53.02	75.32	**91.05**	**39.24**	**52.68**	83.06
	perm.	**18.87**	**42.15**	**63.94**	**55.03**	**76.98**	90.86	31.65	46.90	**83.10**

Table 2: We isolate the role of in-context generalization for our models using a permuted version of the development set. For each model, we show in **bold** whether performance is better on the permuted variant of the development set (lower), or the original development set (upper). Generally, systems perform similarly on both datasets, suggesting performance is primarily derived from the in-context examples.

ing in-context information, rather than applying language-level knowledge. For Maya, all 3 models retain their level of performance on the permuted test set. For Bribri, we see a moderate decrease in performance for Command R+, but an *increase* in performance for GPT-3.5 and GPT-4. This suggests an effect on the (e.g. distributional) properties of subword tokens on in-context generalization behaviours. On the other hand, Guaraní performance clearly degrades for GPT-4, suggesting either a subword issue as in the case of Bribri, or some amount of prior knowledge of the language from pre-training or instruction tuning being recruited. Taken together, though, these results suggest that in-context learning in these models is able to generalize a small set of examples in a linguistically plausible way, even in the absence of prior exposure to the language of the stimuli.

5 Conclusion

We present a simple few-shot learning setup for the AmericasNLP 2024 Shared Task on the Creation of Educational Materials for Indigenous Languages, applied to three state-of-the-art large language models. We find even simple few-shot prompting techniques are able to beat the baseline, with our best system (GPT-4) placing third in the shared task. We investigate the extent to which the performance of our approach is due to a model's prior exposure to the language, by using a character-permuted version of the development set to maintain the problem structure while ablating the language information. We find from this preliminary evidence that the performance of these systems is driven more by in-context learning capabilities than prior exposure to these low-resource indigenous languages. We also find preliminary evidence of performance sensitivity to subwords, as we find that sometimes the model performs *better* on the permuted language than the original language.

One question not addressed here is the cause of the relative performance of the models on each of the three languages. The differences in performance mirror the performance of the baseline, suggesting that in some sense perhaps e.g. the Maya data is simpler or the training data is more informative than for some of the other languages. However, future work could characterize this further, investigating what kind of data sparsity these systems can generalize over and what kinds of functions they are better or worse at generalizing. For example, anecdotally for Bribri we found the systems struggled to generalize morphophonological stem changes (e.g., sú + ök should be sawök, but the model produces súök).

Acknowledgments

This work was in part supported by the UKRI Centre for Doctoral Training in Natural Language Processing, funded by the UKRI (grant EP/S022481/1) and the University of Edinburgh, School of Informatics and School of Philosophy, Psychology Language Sciences. The author would like to thank Paul J.W. Schauenburg, Verna Dankers, Sharon Goldwater, and Edoardo Ponti for their comments on this work. Any errors are my own.

References

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Luis Chiruzzo, Pavel Denisov, Samuel Canul Yah, Lorena Hau Ucán, Marvin Agüero-Torales, Aldo Alvarez, Silvia Fernandez Sabido, Alejandro Molina Villegas, Abteen Ebrahimi, Robert Pugh, Arturo Oncevay, Shruti Rijhwani, Rolando Coto-Solano, Katharina von der Wense, and Manuel Mager. 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)*. Association for Computational Linguistics.

Rolando Coto-Solano. 2022. Evaluating word embeddings in extremely under-resourced languages: A case study in Bribri. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4455–4467, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Omer Goldman, Francesco Tinner, Hila Gonen, Benjamin Muller, Victoria Basmov, Shadrack Kirimi, Lydia Nishimwe, Benoît Sagot, Djamé Seddah, Reut Tsarfaty, and Duygu Ataman. 2022. The MRL 2022 shared task on multilingual clause-level morphology. In *Proceedings of the 2nd Workshop on Multilingual Representation Learning (MRL)*, pages 134–146, Abu Dhabi, United Arab Emirates (Hybrid). Association for Computational Linguistics.

Omer Goldman and Reut Tsarfaty. 2022. Morphology without borders: Clause-level morphology. *Transactions of the Association for Computational Linguistics*, 10:1455–1472.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Lauri Karttunen and Kenneth R. Beesley. 2005. Twenty-five years of finite-state morphology. In *Inquiries into Words, Constraints and Contexts. Festschrift for Kimmo Koskenniemi on his 60th Birthday*. CSLI Publications, Stanford, CA, USA.

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, PA, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Geoffrey K. Pullum and Gerald Gazdar. 1982. Natural languages and context-free languages. *Linguistics and Philosophy*, 4(4):471–504.

Brian Roark and Richard Sproat. 2001. The Formal Characterization of Morphological Operations. In *Computational Approaches to Morphology and Syntax*. Oxford University Press.

Abel Salinas and Fred Morstatter. 2024. The butterfly effect of altering prompts: How small changes and jailbreaks affect large language model performance. *Preprint*, arXiv:2401.03729.

Zengzhi Wang, Qiming Xie, Yi Feng, Zixiang Ding, Zinong Yang, and Rui Xia. 2024. Is ChatGPT a good sentiment analyzer? A preliminary study. *Preprint*, arXiv:2304.04339.

Jason Wei, Maarten Bosma, Vincent Y. Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V. Le. 2022. Finetuned language models are zero-shot learners. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Benfeng Xu, An Yang, Junyang Lin, Quan Wang, Chang Zhou, Yongdong Zhang, and Zhendong Mao. 2023. Expertprompting: Instructing large language models to be distinguished experts. *Preprint*, arXiv:2305.14688.

Kexun Zhang, Yee Man Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. Hire a linguist!: Learning endangered languages with in-context linguistic descriptions. *Preprint*, arXiv:2402.18025.

A Comparison of Fine-Tuning and In-Context Learning for Clause-Level Morphosyntactic Alternation

Jim Su[*], Justin Ho[*], George Aaron Broadwell, Sarah Moeller, Bonnie J. Dorr
University of Florida
{jimsu, justinho, broadwell, smoeller, bonniejdorr}@ufl.edu

Abstract

This paper presents our submission to the AmericasNLP 2024 Shared Task on the Creation of Educational Materials for Indigenous Languages. We frame this task as one of morphological inflection generation, treating each sentence as a single word. We investigate and compare two distinct approaches: fine-tuning neural encoder-decoder models such as NLLB-200, and in-context learning with proprietary large language models (LLMs). Our findings demonstrate that for this task, no one approach is perfect. Anthropic's Claude 3 Opus, when supplied with grammatical description entries, achieves the highest performance on Bribri among the evaluated models. This outcome corroborates and extends previous research exploring the efficacy of in-context learning in low-resource settings. For Maya, fine-tuning NLLB-200-3.3B using StemCorrupt augmented data yielded the best performance.

1 Introduction

The AmericasNLP 2024 Shared Task on the Creation of Educational Materials for Indigenous Languages (Chiruzzo et al., 2024) focuses on the challenge of clause-level morphosyntactic alternation for low-resource indigenous languages of the Americas. The objective of this task is to develop a system capable of applying a set of grammatical attributes to a given source sentence, thereby generating a target sentence with the desired changes. The motivation behind this task lies in the potential for such systems to aid in the preservation and revitalization of endangered languages (Anastasopoulos and Neubig, 2019).

This task involves three indigenous languages of the Americas: Bribri, Guaraní, and Maya. Two examples are provided below:

Example 1.1. *Bribri*
Source sentence: Ye'shka' ("I walked")
Attributes: TYPE:NEG (negative polarity)
Target sentence: Ye'kë shkàn̲e ("I didn't walk")

Example 1.2. *Maya*
Source sentence: Táan in xíimbal tu jáal já' ("I'm walking on the beach")
Attributes: TYPE:NEG (negative polarity)
Target sentence: Ma' táan in xíimbal tu jáal ja'i' ("I'm not walking on the beach")

We frame this task as one of morphological inflection generation, treating each sentence as a single word. Our objective is thus twofold: to develop a system that performs sentence-level morphological inflection for low-resource indigenous languages of the Americas, and to provide insight into what techniques are effective for future practitioners who attempt this task. In pursuit of this goal, we compare the performance of two distinct approaches: fine-tuning pre-trained transformer models and leveraging LLMs through in-context learning. By evaluating these two approaches, we aim to contribute to the understanding of effective strategies for addressing the unique challenges posed by low-resource languages in tasks such as morphosyntactic alternation.

2 Background

This task is unique, as previous literature has explored morphological inflection generation on the word level rather than on the sentence level (Nicolai et al., 2023). Further, this task is challenging for two reasons:

1. Data scarcity: low-resource indigenous languages, by definition, have limited available data for training and evaluating machine learning models (Liu and Dorr, 2024). The scarcity of parallel corpora, annotated texts, and linguistic resources poses significant obstacles in developing robust morphological inflection

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 179–187
June 21, 2024

systems (Moeller, 2021). This scarcity is compounded by the novel nature of this task as prior literature is scarce.

2. Unusual linguistic properties: Indigenous languages of the Americas exhibit a wide range of linguistic properties that diverge from those of well-studied languages like English or Spanish. These languages often feature intricate morphological, phonological, and orthographic systems (Dagostino et al., 2024). They may be polysynthetic and adhere to irregular morphological paradigms. Such linguistic properties can make it challenging to model computationally, especially in the context of limited training data.

Prior results have demonstrated the effectiveness of transformer-based models (Vaswani et al., 2023) on the word-level inflection task (Anastasopoulos and Neubig, 2019). Building upon this success, we evaluate the performance of fine-tuned transformer models on the sentence-level task, exploring their ability to capture and generate morphological inflections in context.

In an effort to extend the available data, we search for external sentence-level parallel corpora aligned with the task format. However, our search yields no suitable resources. While it may be possible to preprocess and adapt data from other formats in a separate pre-training stage, this approach is complex and may require a significant time investment for developing custom preprocessing pipelines, which is not possible in our study, given the short time-frame of this shared task.

To address the challenge of limited data resources, we opt for data augmentation using StemCorrupt (Anastasopoulos and Neubig, 2019), generating synthetic instances based on the existing data. StemCorrupt is a data augmentation technique created for generating additional instances for the word-level inflection task. The use of StemCorrupt is motivated by the availability of pre-existing code and the relative simplicity of this technique, which allow us to quickly run data augmentation and focus our efforts on other aspects of the task.

The limited supervised data challenge also prompts us to explore the use of proprietary large language models (LLMs). These models have the capability to process long context windows of arbitrary text as input and do not require fine-tuning, making them a promising alternative for low-resource settings where extensive task-specific data is unavailable.

Recent advancements have demonstrated that by scaling training compute and corpus size, LLMs may excel in tasks for they which they are not explicitly trained (Wei et al., 2022). Studies exploring the use in-context learning with LLMs on low-resource machine translation have shown promising results (Tanzer et al., 2024). More recent work in the area suggests that when paired with appropriate language resources, LLMs can even surpass human baselines in translation quality (Reid et al., 2024). These findings highlight the potential of in-context learning in LLMs for addressing the challenges posed by low-resource languages and the importance of incorporating relevant linguistic knowledge to maximize their performance.

3 Data

For fine-tuning, we use the provided training dataset[1] and an augmented dataset that we create by applying StemCorrupt to the provided training dataset. For in-context learning, we experiment with the inclusion of a grammatical description in the prompt. Previous work investigating the use of proprietary LLMs on low resource languages has shown that, when combined with grammatical descriptions, these models obtain strong performance on tasks such as machine translation (Tanzer et al., 2024). We hypothesize that using grammatical descriptions in an in-context learning setting can improve performance on this task as well.

3.1 Training Data

The training set provided by the organizers contains 1199 training instances. These instances consist of 594 Maya instances, 427 Bribri instances, and 178 Guarani instances. This dataset is somewhat imbalanced, with Guarani comprising only 14.8% of all training instances.

Each instance contains a set of *change tags*, i.e., morphosyntactic attributes that act as functors (such as TYPE:NEG to indicate negation of sentence polarity). Across all languages, there are 77 unique change tags. These follow a long-tailed distribution: some tags are shared across languages while others are unique to a particular language. Refer to the Appendix A for an exhaustive distribution of change tags.

[1] https://github.com/AmericasNLP/americasnlp2024

3.2 Data Augmentation

We perform data augmentation to generate 1000 additional instances for each language. Prior literature has demonstrated the efficacy of StemCorrupt for improving the performance of language models on word-level inflection tasks (Samir and Silfverberg, 2023). We explore the effect of StemCorrupt on this task at the sentence level.

3.3 Utilizing Grammatical Descriptions

We encounter two challenges to using published grammatical descriptions of these three languages:

1. Grammatical descriptions are difficult to find, and the orthographies used in them vary. Many of the resources that do exist were published in the 1960-80s or earlier and are only accessible online as PDF images of printed text or in digital formats that do not translate easily into correct Unicode characters.[2] We narrow our search to resources that use English as the analysis language which limit our choices further since other descriptions seem to be available in Spanish. Finally, we search for grammatical descriptions with interlinear glossed text (e.g. Umaña et al. (1998)) in order to provide information similar to the change tokens provided in the shared task data.

2. The length of data passed into an LLM is limited by its context window, establishing a hard limit on how much data (in particular, excerpts from the published resources) can be passed into the model. Even within this hard limit, particularly long input sequences can degrade performance (Li et al., 2024).

We employ the following grammatical descriptions, focusing on passages that contained interlinear glossed texts:

1. Bribri - Dickeman-Datz (1985) and Jara (1995)

2. Guarani - Estigarribia (2020)

We are unable to find a suitable grammatical description for the Yucatec Maya language that matched the orthography used in this task.

[2]For example, the scans of these typewritten Peace Corps language learning lessons: https://www.livelingua.com/project/peace-corps/guarani or this image of a Bribri grammatical description: http://journals.uvic.ca/index.php/WPLC/article/view/5054/1954

3.4 Data Processing

Since curated data is provided by the shared task organizers, minimal preprocessing is required. The Bribri data needs some additional preparation. For Maya and Guarani, no preprocessing is done.

For the Bribri language, training instances are provided in both the data/ and pilotdata/ directories. We concatenate the training sets and development sets across data/ and pilotdata/.

The Bribri data/ directory contains the straight apostrophe (') character while the pilotdata/ directory contains the right single quotation mark (’). We replace each instance of the right single quotation mark in the Bribri pilot training data with the straight apostrophe.

4 Experiments

We perform four experiments and compare the results:

1. Fine-tuning the pre-trained encoder-decoder models

2. Fine-tuning the pre-trained encoder-decoder models with data augmentation

3. In-context learning on proprietary LLMs

4. In-context learning on proprietary LLMs with a grammatical description

4.1 Experiment Setup

We apply two classes of experimental setups: fine-tuning and in-context learning. Fine-tuning adapts a pre-trained model to predict the target column one instance at a time. In-context learning includes the full training set and instances from the validation set in the prompt of an LLM, predicting multiple targets per inference run.

Both setups have strengths and weaknesses. In-context learning is constrained by a fixed context window but can work on arbitrary forms of task information such as grammatical descriptions. In contrast, fine-tuning allows the model's parameters to be updated on an arbitrarily large training dataset but requires task-specific parallel data that is challenging to find for low-resource languages.

4.1.1 Fine-Tuning

For each training instance, we concatenate the source sentence with the change tags. A separator token is used to delimit the end of the source sentence and the start of the change token. A model

Model	Bribri	Guarani	Maya
Baseline			
(Kann and Schütze, 2016)	5.66	**22.78**	26.17
BART Family			
BART-Large	7.11	2.53	44.96
MBART-50	12.89	0.00	9.39
T5-FLAN Family			
FLAN-T5-XL	1.33	0.00	2.01
NLLB-200 Family			
NLLB-200-distilled-600M	19.55	21.51	49.66
NLLB-200-distilled-600M (+ StemCorrupt)	20.00	16.45	**58.39**
NLLB-200-3.3B	24.88	16.45	53.02
NLLB-200-3.3B (+ StemCorrupt)	28.44	21.51	52.35
In-Context Learning			
Claude 3 Opus	30.53	18.99	54.36
Claude 3 Opus (+ grammatical description)	**36.73**	17.72	N/A
Gemini 1.5 Pro	8.41	N/A	44.97
Gemini 1.5 Pro (+ grammatical description)	12.21	N/A	N/A

Table 1: Dev set accuracy score for all fine-tuned models. Bold means best performing model for that language. It is worth noting that for Maya, we are not able to find grammatical descriptions that matched the orthography of the task. As for Gemini 1.5 Pro, we suspect there may be an issue with the tokenizer for Guarani as the model would generate few predictions before failing.

is trained for each language as opposed to creating a single multi-lingual inflection model, since we find the former results in better performance over the latter. We run our experiments on a single A100 GPU using a batch size of 64. We follow the same evaluation scheme proposed by the organizers using accuracy, chrF, and BLEU (Popović, 2015; Papineni et al., 2002).

4.1.2 In-Context Learning

For each in-context learning experiment, the LLM is provided the following:

1. The training set, with IDs replaced by the row number
2. The development set with changes removed
3. A relevant prompt (refer to the Appendix A for exact prompts used)

4.2 Fine-Tuning Pre-Trained Encoder-Decoder Models

We fine-tune a variety of encoder-decoder model families. Different variants of BART are used such as mBART to evaluate the effect of multi-lingual pre-training on this task (Lewis et al., 2019; Liu et al., 2020). The FLAN-T5 series of models are also evaluated as these models incorporate a unique pre-training process that is promising in terms of boosting model performance (Chung et al., 2022). The last family of models examined is the NLLB-200 family of models for their strong performance on low-resource translation (Team et al., 2022). We experiment with the 600M and 3.3B parameter version of each model. Although the NLLB-200 also includes a Mixture of Experts (MoE) model that may outperform the other versions, this model is not investigated due to its prohibitive size (54B parameters, which exceeds the memory capacity of an A100 GPU).

4.3 Fine-Tuning Pre-Trained Encoder-Decoder Models with Data Augmentation

StemCorrupt is used to generate 1000 instances for each language. Only NLLB-200 3.3B is trained using the augmented StemCorrupt data as this is the best performing model found during fine-tuning on non-augmented data.

4.4 In-Context Learning on Proprietary LLMs

We evaluate two proprietary LLMs:

1. Gemini 1.5 Pro. This model is selected for its long context window and strong performance

Model	Window Size	Strategy
Gemini 1.5 Pro	1 Million	All
Claude 3 Opus	200k	Relevant

Table 2: Context window size for each model and document strategy used.

on in-context low-resource machine translation (Reid et al., 2024)

2. Claude 3 Opus. This model is selected as the current state of the art in proprietary LLMs (Anthropic, 2024).

We briefly evaluate GPT-4 Turbo but encounter significant challenges (OpenAI et al., 2024). The model produces outputs of unacceptably low quality, rendering them effectively unusable. Additionally, GPT-4 Turbo proves unstable, consistently failing to fully process the full test dataset.

4.5 In-Context Learning on Proprietary LLMs with a Grammatical Description

Each LLM evaluated by our team is constrained by a different context window length, which affects the strategy used for passing in the grammatical description. Our team relies on two strategies: passing all grammar resources to the model and passing the most relevant grammar resources to the model. The most relevant grammar resource for each language is determined by hand, with the most frequent change tokens for each language guiding this search. The selected section for each language describes the language's morphology, verbal agreement system, and syntax of various sentence types including affirmative statements, negation, and questions.

5 Results

5.1 Dev Set Results

Table 1 shows the results of our fine-tuned models and in-context learning experiments on the dev set. For all languages except for Guarani, we are able to exceed the baseline performance significantly. For Bribri, in-context learning combined with a grammatical description is the highest performing technique with an accuracy of 36.73% over the baseline of 5.66%. For Maya, fine-tuned NLLB-200-distilled-600M with StemCorrupt augmented data is the best technique with an accuracy of 58.39% over the baseline of 26.17%.

Model	Bribri	Guarani	Maya
Baseline	8.75	**14.84**	25.81
Submission 1			
NLLB-200-3.3B	9.79	0.00	**37.42**
Submission 2			
Claude 3 Opus	**26.88**	0.00	33.23

Table 3: Test set accuracy score for all fine-tuned models. Bold means best performing model for that language.

There are unique trends that can be observed from the results of our system runs on the dev set. As anticipated, Guarani proves challenging to improve upon due to particularly limited data. Even when data augmentation techniques are applied, results of neural techniques are still below that of the statistical-backed baseline. This result reaffirms the findings of prior literature in terms of the weaknesses of neural techniques under sparse data conditions. Furthermore, this result hints at morphological or linguistic complexities in Guarani that make this task challenging.

Comparing fine-tuning and in-context learning, no technique was optimal across all languages. This result affirms two ideas: fine-tuning models is still relevant in the age of LLMs, and LLMs empowered with language resources are a viable approach for this task. For Bribri, fine-tuned models–even with data-augmentation–are not able to match the best performing in-context learning LLM.

5.2 Test Set Results

Table 3 details the test results for our submissions. Our model's performance on the test set exhibits an unexpected discrepancy compared to its performance on the dev set. Both of our best systems for the dev set underperform significantly when evaluated on the test set. Compared to the dev set, the accuracy on the test set is 20% lower for Maya and 10% lower for Bribri. This significant drop in performance warrants further investigation to identify potential causes, such as differences in domain, style, or linguistic properties between the dev and test sets.

Despite this unexpected discrepancy, it is worth noting that our team achieved the second best system submission for Bribri. Without access to the target column of the test set, the exact reason remains unclear. With such a limited number of training instances, both the in-context learning and fine-tuned model may not have enough examples to generalize to different data distributions. Due to weak performance of our system runs applied to the dev set (and our misunderstanding that only above-baseline runs are submissible), our team has no submitted Guarani results for the test set.

6 Future Work

Much future work remains. Our search for language resources reveals a wide variety of language resources of varying types and orthographies. A future area of research is an exploration of the effect of different orthographies on the LLM performance in an in-context learning setting.

Additionally, a significant advantage of in-context learning is the reduction of restrictions on data types that can be utilized by the model. Exploring the effect of different resource types, such as dictionaries and learning worksheets, would be valuable. A historic bottleneck for the translation or inflection of low-resource languages has been data, specifically gold-standard data that adheres to a specialized format. By leveraging in-context learning, the variety of usable data is greatly increased and can offer opportunities for further exploration.

StemCorrupt has shown promise for sentence-level inflection despite initially being developed for word-level inflection. Exploring the feasibility of extending this technique to other languages is a worthwhile future endeavor.

7 Conclusion

In this paper, we present the systems submitted by our team for the AmericasNLP 2024 Shared Task on the Creation of Educational Materials for Indigenous Languages. We find that while LLMs–through in-context learning–exhibit impressive capabilities, fine-tuning still has a role to play in the modern NLP space. Moreover, we reaffirm the results of prior literature regarding the promise of LLMs when applied to low-resource languages using in-context learning. Additional work must be done to explore the abilities of such systems, but initial results point to promising potential for the task of morphosyntactic alternation. Our work also extends prior literature on StemCorrupt and demonstrates potential applications for the technique on sentence-level inflection generation.

Limitations

The main limitation of our work is selecting only grammatical descriptions published in English. More grammatical descriptions are available in Spanish.

References

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

Anthropic. 2024. The claude 3 model family: Opus, sonnet, haiku.

Luis Chiruzzo, Pavel Denisov, Samuel Canul Yah, Lorena Hau Ucán, Marvin Agüero-Torales, Aldo Alvarez, Silvia Fernandez Sabido, Alejandro Molina Villegas, Abteen Ebrahimi, Robert Pugh, Arturo Oncevay, Shruti Rijhwani, Rolando Coto-Solano, Katharina von der Wense, and Manuel Mager. 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)*. Association for Computational Linguistics.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *Preprint*, arXiv:2210.11416.

Carmen Dagostino, Marianne Mithun, and Keren Rice, editors. 2024. *The Languages and Linguistics of Indigenous North America*. De Gruyter Mouton, Berlin, Boston.

Margaret Dickeman-Datz. 1985. Transitivity in indefinite voice in bribri. *International journal of American linguistics*, 51(4):388–390.

Bruno Estigarribia. 2020. *A Grammar of Paraguayan Guarani*. UCL Press.

Carla Victoria Jara. 1995. *Text and context of the Suwo': Bribri oral tradition*. Louisiana State University and Agricultural & Mechanical College.

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.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *Preprint*, arXiv:1910.13461.

Tianle Li, Ge Zhang, Quy Duc Do, Xiang Yue, and Wenhu Chen. 2024. Long-context llms struggle with long in-context learning. *Preprint*, arXiv:2404.02060.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Preprint*, arXiv:2001.08210.

Zoey Liu and Bonnie J. Dorr. 2024. The Effect of Data Partitioning Strategy on Model Generalizability: A Case Study of Morphological Segmentation. In *Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Sarah Moeller. 2021. Computational morphology for language documentation and description. *Colorado Research in Linguistics*, 25.

Garrett Nicolai, Eleanor Chodroff, Frederic Mailhot, and Çağrı Çöltekin, editors. 2023. *Proceedings of the 20th SIGMORPHON workshop on Computational Research in Phonetics, Phonology, and Morphology*. Association for Computational Linguistics, Toronto, Canada.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

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 on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Machel Reid, Nikolay Savinov, Denis Teplyashin, Dmitry Lepikhin, Timothy Lillicrap, Jean baptiste Alayrac, Radu Soricut, Angeliki Lazaridou, Orhan Firat, Julian Schrittwieser, Ioannis Antonoglou, Rohan Anil, Sebastian Borgeaud, Andrew Dai, Katie Millican, Ethan Dyer, Mia Glaese, Thibault Sottiaux, Benjamin Lee, Fabio Viola, Malcolm Reynolds, and Yuanzhong Xu. 2024. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *Preprint*, arXiv:2403.05530.

Farhan Samir and Miikka Silfverberg. 2023. Understanding compositional data augmentation in typologically diverse morphological inflection. *Preprint*, arXiv:2305.13658.

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. *Preprint*, arXiv:2309.16575.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation. *Preprint*, arXiv:2207.04672.

A.C. Umaña, F.E. Figueroa, and F.P. Mora. 1998. *Curso básico de bribri*. Editorial de la Universidad de Costa Rica.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2023. Attention is all you need. *Preprint*, arXiv:1706.03762.

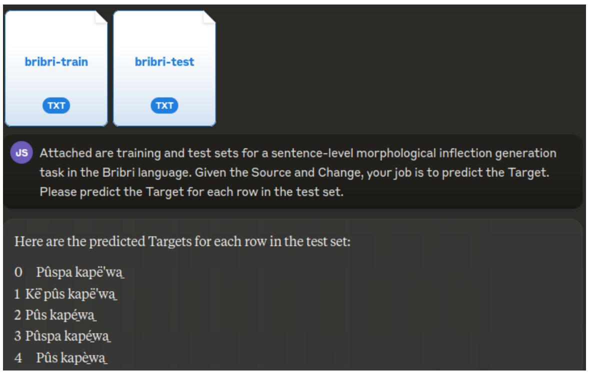
Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, Ed H. Chi, Tatsunori Hashimoto, Oriol Vinyals, Percy Liang, Jeff Dean, and William Fedus. 2022. Emergent abilities of large language models. *Preprint*, arXiv:2206.07682.

A Appendix

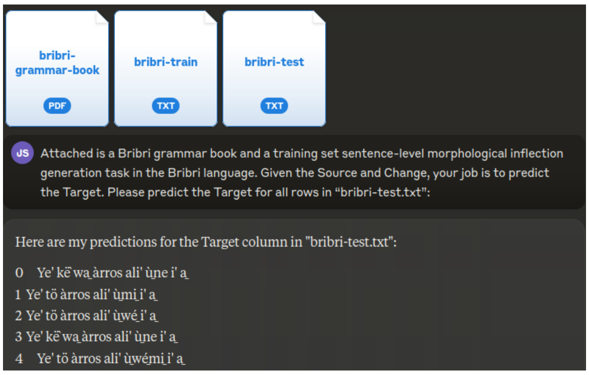
A.1 Proprietary LLM Prompts

The same prompt is used for all LLMs tested. The below screenshots are taken from Anthropic's claude.ai interface.

A.1.1 Without grammatical description



A.1.2 With grammatical description



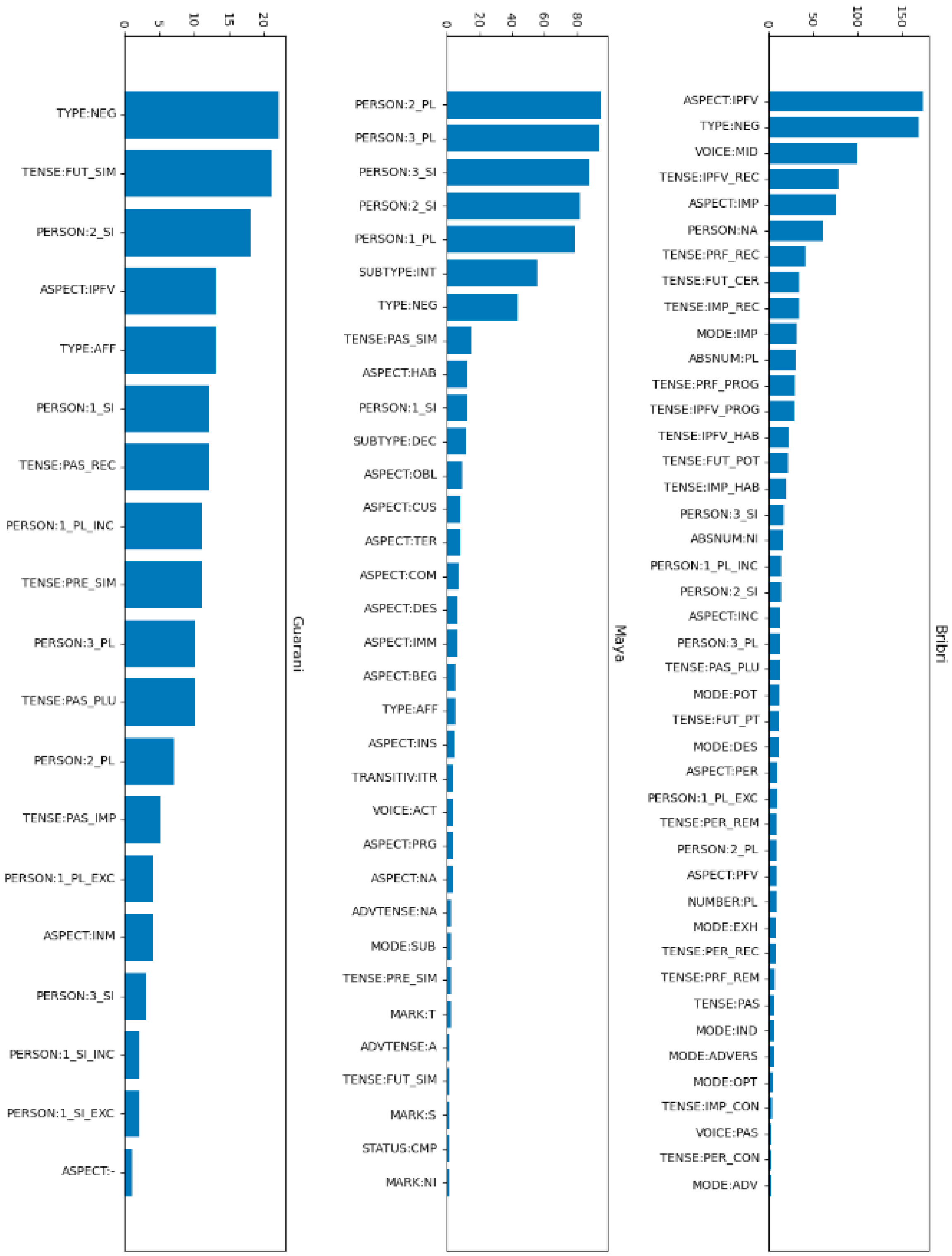


Figure 1: Distribution of change tags for each language.

Experiments in Mamba Sequence Modeling and NLLB-200 Fine-Tuning for Low Resource Multilingual Machine Translation

Dan DeGenaro*
Department of Linguistics,
Georgetown University
37th and O Sts NW, Washington, D.C.
drd92@georgetown.edu

Tom Lupicki*
Department of Computer Science,
Georgetown University
37th and O Sts NW, Washington, D.C.
tml89@georgetown.edu

Abstract

This paper presents DC_DMV's submission to the AmericasNLP 2024 Shared Task 1: Machine Translation Systems for Indigenous Languages. Our submission consists of two multilingual approaches to building machine translation systems from Spanish to eleven Indigenous languages: fine-tuning the 600M distilled variant of NLLB-200, and an experiment in training from scratch a neural network using the Mamba State Space Modeling architecture. We achieve the best results on the test set for a total of 4 of the language pairs between two checkpoints by fine-tuning NLLB-200, and outperform the baseline score on the test set for 2 languages.

1 Introduction

The 2024 AmericasNLP Shared Task on machine translation (MT) for Indigenous languages consists of developing an MT system (or systems) for the purpose of translating Spanish to 11 Indigenous languages of the Americas: Aymara (aym), Bribri (bzd), Asháninka (cni), Chatino (ctp), Guaraní (gn), Wixarika (hch), Nahuatl (nah), Hñähñu/Otomí (oto), Quechua (quy), Shipibo-Konibo (shp), and Rarámuri (tar). We take two approaches in parallel, namely finetuning NLLB-200 (Team et al., 2022) and training a Mamba architecture-based neural network (Gu and Dao, 2023) from scratch.[1]

2 Data

2.1 Data Sources

We utilize data from a number of sources, namely the training and development sets provided by the task organizers, data gathered as part of last year's HelsinkiNLP submission (De Gibert et al., 2023), parallel data from Tatoeba[2] released under a CC-BY 2.0 FR., and pivot translations generated from non-Spanish-to-target language parallel data from the Tatoeba Translation Challenge (Tiedemann, 2020). We include additional data to try to compensate for the sparseness of data available in the target languages more generally.

Organizer-provided Data Training and development data for the 11 target languages included in the shared task were released by task organizers[3]. The provided data includes data explicitly denoted as the training set, supplemental translation data from Spanish, and supplemental translation data from English. An overview of the organizer-provided data we used can be found in Table 1.

HelsinkiNLP Data collected for the 2023 HelsinkiNLP submission to the shared task (De Gibert et al., 2023) was also provided by the task organizers. This data is sourced from the OPUS corpus collection (Tiedemann, 2012), the FLORES-200 corpus (Team et al., 2022), the JHU Bible corpus (McCarthy et al., 2020), and various other texts spanning legal, educational, and news domains.

Tatoeba Translation Challenge Spanish-to-target-language parallel data is available from the Tatoeba website[2] for Guarani, Nahuatl, and Quechua.

Pivot Translations The Tatoeba Translation Challenge (Tiedemann, 2020) provides non-Spanish parallel data for Guarani, Nahuatl, and Quechua. We utilize machine translation systems to construct additional parallel language data. Data in English, Esperanto, French, German, Hebrew, Japanese, Macedonian, Polish, Russian, and Ukrainian was translated using bilingual Opus-MT

*Both authors contributed equally to this work.
[1]Code for both of our models is available here: https://github.com/tomlup/americasnlp-2024-st1-dc_dmv

[2]Tatoeba website.
[3]AmericasNLP 2024 Shared Task GitHub

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 188–194
June 21, 2024

Target Language	Data Source(s)
aym	Global Voices (Tiedemann, 2012)
bzd	(Feldman and Coto-Solano, 2020)
cni	AshanikaMT (Ortega et al., 2020; Cushimariano Romano and Sebastián Q., 2008; Mihas, 2011)
ctp	`https://scholarworks.iu.edu/dspace/handle/2022/21028`
gn	(Chiruzzo et al., 2020)
hch	(Mager et al., 2018)
nah	Axolotl (Gutierrez-Vasques et al., 2016)
oto	`https://tsunkua.elotl.mx/about/`
quy	JW300 (Agić and Vulić, 2019), Global Voices (Tiedemann, 2012)
shp	(Montoya et al., 2019), (Galarreta et al., 2017), `https://www.sil.org/resources/archives/30143`
tar	(Brambila, 1976)

Table 1: Sources of data provided by task organizers.

systems (Tiedemann and Thottingal, 2020). Data in Chinese, Javanese, and Portuguese was translated into Spanish using NLLB-200 (Team et al., 2022). Additionally, English-Indigenous language data that was provided as supplemental data by task organizers were also translated using Opus-MT. We make use of pivot translations only in the Mamba model.

2.2 Data Organization

For the purposes of training, we organize our collected data into three stages. Stage 1 includes all synthetic parallel texts created by means of pivot translation and synthetic data provided by task organizers. Stage 2 includes the supplemental data sourced from the 2023 HelsinkiNLP submission, as well as other Spanish-source supplemental data provided by task organizers. Stage 3 includes training data provided by the shared task organizers.

2.3 Duplicate Filtering

After all training data was organized into stages, all data for each target language was then filtered to remove duplicates using OpusFilter (Aulamo et al., 2020). The pipeline for filtering was as follows: All duplicates within Stage 3 data were removed. Then, all duplicates within Stage 2 and overlap with Stage 3 were removed from Stage 2. Finally, all duplicates within Stage 1 and any overlap with Stage 2 and Stage 3 were removed from Stage 1. The total number of training examples from each stage is shown in Table 2.

Language	Stage 1	Stage 2	Stage 3
aym	16,338	17,679	6,453
bzd	0	0	7,303
cni	13,018	0	3,860
ctp	2,762	2,246	357
gn	617,894	42,184	14,500
hch	505	2,628	6,587
nah	9,279	2,493	15,450
oto	0	9,012	4,531
quy	64,337	16,112	119,471
shp	23,125	16,719	14,511
tar	0	2,254	14,658
Total	747,258	110,787	207,681

Table 2: Overview of data organization by number of examples.

3 Methods

3.1 Finetuning NLLB-200

Our first method involves fine-tuning the NLLB-200 model (Team et al., 2022). We use the distilled 600M parameter variant, and leave all parameters trainable. We motivate this decision as follows. Given that we are tokenizing previously unseen languages using an already-trained tokenizer, the distribution and linear ordering of tokens in our fine-tuning data will differ vastly from the distribution and linear ordering in the languages previously seen by the model. As such, it is sensible to retrain the entire model, including the embeddings, to model this very different distribution. To that end, we introduce additional language tokens for the eight target languages in the shared task not already represented in the model (all except for Aymara, Guarani, and Quechua), which are randomly initialized.

We finetune on padded mini-batches of size 4 with a maximum sequence length of 384, in which all 4 training examples in a given batch have the same target language. However, batches from all 11 target languages are shuffled together. We optimize using AdamW, with a learning rate of $1 \cdot 10^{-5}$ and a weight decay of $1 \cdot 10^{-4}$.

With regard to training stages, we do not use the Stage 1 data to fine-tune NLLB. The number of epochs through each stage for each of our fine-tuned NLLB models are presented in Table 3.

The generation process for producing translations for evaluation uses a maximum sequence length of 384 and beam search with 4 beams and early stopping.

3.2 Mamba State Space Model

Our second method involves training a neural network using repeating multiple Mamba architecture layers and a language model head. We submit results for a model containing 3 Mamba layers and a final linear layer with 256 dimensions, and a vocabulary of 16,000 subword tokens trained on all data using SentencePiece (Kudo and Richardson, 2018) using a unigram language model algorithm (Kudo, 2018).

For the purposes of training our Mamba model, we modify our training data by appending a target language token to the beginning of each source sentence. We additionally append a start of sentence token and end of sentence token to the start and end of each sentence, respectively.

We train our model on padded mini-batches of size 128 with a maximum sequence length of 512. Each mini-batch contains shuffled data taken from all languages and all data used for training during an epoch. We optimize the model using AdamW using a learning rate of $1 \cdot 10^{-3}$ and a weight decay of $1 \cdot 10^{-4}$. The model is trained for 5 epochs through all data (Stage 1, Stage 2, and Stage 3), followed by an additional 25 epochs on combined Stage 2 and Stage 3 data. We motivate our decision to include Stage 1 data only in early training by our belief that our synthetic pivot translations are noisier than original Spanish-source translation data, but find it important to train our model on a wide range of data early on. In this regard, we view our later stages of training on Stage 2 and Stage 3 data as tuning our model on higher quality data.

4 Results

We present our results in Tables 5 and 6, alongside results for the two baseline systems. The reported scores are calculated using the chrF++ metric (Popović, 2017), as stipulated by the shared task.

Our NLLB+FT(v2) model beats both baseline systems on the development set for Aymara and Quechua, and both baseline systems on the test set for Quechua and Rarámuri. Additionally, several of our models beat at least one baseline system on the development set for Bribri, Nahuatl, Quechua, and Shipibo-Konibo.

Of all submissions this year, our NLLB+FT(v2) model achieves the best result for Aymara, Shipibo-Konibo, and Rarámuri, and our NLLB+FT(v4) model achieves the best result for Bribri, as evaluated on the test set. Our NLLB+FT(v2) and NLLB+FT(v4) models achieve average chrF++ scores across all languages of 22.17 and 23.32 respectively, with NLLB+FT(v4) representing the second best overall submission.

Interestingly, while our models did not achieve the best result on the test set for Asháninka, Hñähñu, and Quechua as measured by the official metric, at least one of our NLLB+FT models outperformed the best submission in BLEU score (Post, 2018). We report these scores in Table 4.

Our Mamba model shows poor performance at the stage in training at time of submission. However, we believe much of this to be due to undertraining given that our model is trained from scratch. With this in mind, we believe continued training may lead to success of our Mamba model, and plan to continue experiments with this architecture.

5 Conclusion

In this paper, we presented our submission to the AmericasNLP 2024 shared task on machine translation systems for Indigenous languages. Our submissions included six versions of a fine-tuned 600M parameter distilled variant of NLLB-200, and one Mamba-based model trained from scratch. We trained all of our models on multilingual data to translate from Spanish to 11 target Indigenous languages. We achieve the best chrF++ scores on 4 languages with our fine-tuned NLLB-200 models, improving upon the baseline systems for two languages and setting a new highest score for Rarámuri. Additionally, we find our Mamba-based

Version	# Epochs Stage 2	# Epochs Stage 3	# Epochs Addl. Stage 2	# Epochs Addl. Stage 3
v1	3	10	0	10
v2	3	10	3	0
v3	3	10	3	3
v4	3	10	3	4
v5	3	10	0	6
v6	3	10	3	8

Table 3: Our six fine-tuned NLLB submissions differ solely in the number of epochs through each fine-tuning stage. All models were trained for 3 epochs on the Stage 2 data (# Epochs Stage 2), followed by 10 epochs on the stage 3 data (# Epochs Stage 3). We then experiment with training the models on the Stage 2 data again (# Epochs Addtl. Stage 2), on the Stage 3 data again (# Epochs Addtl. Stage 2), or both. The order in which this process occurs is laid out left-to-right in the table. For instance, NLLB+FT(v6) was trained, in order, for 3 epochs through Stage 2, followed by 10 epochs through Stage 3, followed by 3 more epochs through stage 2, and finally 8 epochs through Stage 3.

Language	v1	v2	v3	v4	v5	v6
cni	**3.56**	3.52*	3.56*	3.51*	3.41*	3.49*
oto	1.55*	1.46*	**1.66**	1.49*	1.52*	1.36
quy	4.01	**5.41**	4.13	4.32	3.91	4.05

Table 4: BLEU scores for our six NLLB+FT submissions for the languages on which we achieve a higher BLEU score than the winning submission. The highest score for each language is bolded. All other results that achieve a higher BLEU score than the submission with the highest chrF++ score for that language are denoted with an asterisk.

model to perform poorly given its training, but plan to continue training and experimentation with this architecture.

Limitations

Due to dialectal and orthographic variation of the Indigenous languages included in this shared task, it is unclear how our systems would perform on language data that spans such variation not represented in the task data. For example, the provided data for Quechua belongs to the Quechua Ayacucho variant of the Southern Quechua dialect[4]. It is unclear how performance would vary for different varieties of Quechua.

Ethics Statement

To our knowledge, our work on this project adheres to the principles set forth in Schwartz, 2022.

Acknowledgements

We would like to acknowledge Dr. Kenton Murray from Johns Hopkins University for guiding us as we undertook this project.

[4]Datasets Information

Language	Helsinki	Sheffield	NLLB+FT(v1)	NLLB+FT(v2)	NLLB+FT(v3)	NLLB+FT(v4)	NLLB+FT(v5)	NLLB+FT(v6)	Mamba
aym	32.63	34.28	30.83	**34.38**	31.95	31.90	31.12	31.56	9.98
bzd	22.65	**25.03**	22.82*	19.18	23.59*	22.98*	23.36*	23.25*	5.01
cni	25.68	**26.34**	23.30	20.08	24.05	23.42	23.22	23.69	11.12
ctp	30.06	**37.33**	16.73	8.25	16.17	15.83	16.41	16.29	2.90
gn	**34.74**	32.17	29.28	31.67	30.08	30.10	29.18	29.99	9.25
hch	**27.98**	**27.98**	26.16	20.25	25.65	26.16	25.90	25.92	8.56
nah	22.78	**25.58**	23.90*	19.26	22.96*	23.50*	23.56*	23.79*	11.11
oto	13.10	12.69	12.17	11.20	12.11	12.33	12.18	12.12	4.05
quy	28.78	30.22	32.49*	**35.46**	32.64*	33.13*	32.54*	32.61*	10.78
shp	**30.59**	28.39	24.39	29.94*	26.35	26.70	25.48	25.37	10.70
tar	**17.58**	16.91	14.75	17.46	15.56	15.28	14.79	15.11	7.34

Table 5: Comparison of chrF++ scores for our models versus this year's baseline scores for the development set. The top score on each language is bolded. We denote with an asterisk any other result that beats one or both baselines. The differences in the NLLB versions are specified in Table 3.

Language	Helsinki	Sheffield	NLLB+FT(v1)	NLLB+FT(v2)	NLLB+FT(v3)	NLLB+FT(v4)	NLLB+FT(v5)	NLLB+FT(v6)	Mamba
aym	29.36	**31.84**	27.36	30.97*	28.12	28.32	27.09	27.48	8.69
bzd	23.47	**25.58**	23.19	19.60	23.32	23.47	23.41	23.15	4.72
cni	**24.92**	24.76	22.44	19.89	22.87	22.46	22.53	22.98	9.81
ctp	29.84	**37.05**	16.52	8.06	16.17	15.78	16.11	16.04	0.91
gn	**37.02**	35.76	31.58	33.31	32.58	32.44	31.22	31.66	8.91
hch	**28.67**	28.28	26.46	19.56	25.60	26.23	25.97	25.66	7.12
nah	22.78	**23.28**	21.63	18.52	21.07	21.44	21.43	21.41	10.46
oto	**13.32**	12.87	12.63	11.50	12.42	12.42	12.40	12.20	3.63
quy	28.81	34.01	33.91*	**36.02**	34.08*	34.29*	33.94*	33.56*	11.42
shp	**30.21**	30.06	22.25	29.37	23.84	24.74	23.59	23.05	9.67
tar	16.98	16.25	14.39	**17.03**	15.42	14.92	14.51	14.57	5.29

Table 6: Comparison of chrF++ scores for our models versus this year's baseline scores for the test set. The top score on each language is bolded. We denote with an asterisk any other result that beats one or both baselines. The differences in the NLLB versions are specified in Table 3.

References

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.

Mikko Aulamo, Sami Virpioja, and Jörg Tiedemann. 2020. OpusFilter: A configurable parallel corpus filtering toolbox. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 150–156, Online. Association for Computational Linguistics.

David Brambila. 1976. Diccionario raramuri – castellano (tarahumara).

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.

Rubén Cushimariano Romano and Richer C. Sebastián Q. 2008. Ñaantsipeta asháninkaki birakochaki. diccionario asháninka-castellano. versión preliminar. `http://www.lengamer.org/publicaciones/diccionarios/`.

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, Toronto, Canada. Association for Computational Linguistics.

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.

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, Varna, Bulgaria. INCOMA Ltd.

Albert Gu and Tri Dao. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces. *arXiv preprint*. ArXiv:2312.00752 [cs].

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).

Taku Kudo. 2018. Subword Regularization: Improving Neural Network Translation Models with Multiple Subword Candidates. *arXiv preprint*. ArXiv:1804.10959 [cs].

Taku Kudo and John Richardson. 2018. SentencePiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 66–71, Brussels, Belgium. Association for Computational Linguistics.

Manuel Mager, Carrillo Dionico, and Ivan Meza. 2018. The wixarika-spanish parallel corpus.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.

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.

John 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.

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.

Matt Post. 2018. A call for clarity in reporting BLEU scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium. Association for Computational Linguistics.

Lane Schwartz. 2022. Primum Non Nocere: Before working with Indigenous data, the ACL must confront ongoing colonialism. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages

724–731, Dublin, Ireland. Association for Computational Linguistics.

NLLB Team, Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No Language Left Behind: Scaling Human-Centered Machine Translation. *arXiv preprint*. ArXiv:2207.04672 [cs].

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in OPUS. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2214–2218, Istanbul, Turkey. European Language Resources Association (ELRA).

Jörg Tiedemann. 2020. The Tatoeba Translation Challenge – Realistic Data Sets for Low Resource and Multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Jörg Tiedemann and Santhosh Thottingal. 2020. OPUS-MT – Building open translation services for the World. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 479–480, Lisboa, Portugal. European Association for Machine Translation.

JGU Mainz's Submission to the AmericasNLP 2024 Shared Task on the Creation of Educational Materials for Indigenous Languages

Minh Duc Bui
Johannes Gutenberg University Mainz
minhducbui@uni-mainz.de

Katharina von der Wense
Johannes Gutenberg University Mainz
University of Colorado Boulder
k.vonderwense@uni-mainz.de

Abstract

In this paper, we present the four systems developed by the Meenzer team from JGU for the AmericasNLP 2024 shared task on the creation of educational materials for Indigenous languages. The task involves accurately applying specific grammatical modifications to given source sentences across three low-resource Indigenous languages: Bribri, Guarani, and Maya. We train two types of model architectures: finetuning a sequence-to-sequence pointer-generator LSTM and finetuning the Mixtral 8x7B model by incorporating in-context examples into the training phase. System 1, an ensemble combining finetuned LSTMs, finetuned Mixtral models, and GPT-4, achieves the best performance on Guarani. Meanwhile, system 4, another ensemble consisting solely of fine-tuned Mixtral models, outperforms all other teams on Maya and secures the second place overall. Additionally, we conduct an ablation study to understand the performance of our system 4.[1]

1 Introduction

Natural language processing (NLP) serves as a valuable educational tool for facilitating the learning of (endangered) languages. One effective method for generating learning material involves a system automatically transforming sentences based on specific properties. Subsequently, language learners are tasked with replicating the transformation, thus reinforcing their understanding of the language structure. The AmericasNLP 2024 shared task on the creation of educational materials for Indigenous languages (ST 2) (Chiruzzo et al., 2024) focuses on creating such material for three low-resource Indigenous languages: Bribri, Guarani, and Maya. Participants are tasked with applying a specific grammatical property to a given source sentence and producing the accurate modification.

[1]The code is available at https://github.com/MinhDucBui/SharedTaskAmericasNLP2024.

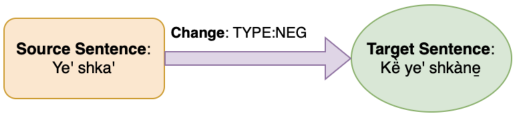


Figure 1: A Bribri sample from the shared task.

Our systems (which we submitted under the name "Meenzer Team") are ensembles composed of a range of models: finetuned character-level pointer-generator LSTMs (See et al., 2017), finetuned Mixtral 8x7B large language models (LLMs) (Jiang et al., 2024) via training on in-context examples, and GPT-4 (OpenAI, 2023). The main metric of the shared task is accuracy. We outperform all teams on Guarani by employing an ensemble across all models. Additionally, our ensemble of finetuned Mixtral models achieves the highest performance on Maya and reaches the second place overall.

The remainder of this paper is organized as follows: Section 2 details the task at hand and introduces the provided data. Following that, Section 3 dives into the details of our four system submissions. Section 4 presents the outcomes observed on both the development and test sets of the shared task. Lastly, an ablation study on our best performing system is provided in Section 5.

2 Task and Data

2.1 Task

In the context of this shared task, a source sentence is accompanied by a designated change feature, which the system is tasked with applying, see Figure 1. These features include modifications related to grammar, such as negation, and each sample can entail multiple concatenated grammatical alterations. While the shared task bears resemblance to morphological inflection shared tasks (Cotterell et al., 2016), where the goal is to modify a single word, our scenario necessitates adjustments to the

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 195–200
June 21, 2024

	Train	Dev	Test
Bribri	310	213	481
Guarani	179	80	365
Maya	595	150	311

Table 1: Dataset sizes for each language and split.

entire sentence to accurately represent a specified property.

2.2 Data

The dataset encompasses three Indigenous languages: Bribri, Guarani, and Maya.[2] For each language, a training and a development set are provided. Additionally, the input side of the test set is given and used to submit predictions for the shared task's final evaluation. Within the training set, Bribri comprises 28 unique features, resulting in 135 distinctive combinations; Guarani encompasses 19 unique features, forming 21 combinations; and Maya has 33 unique features, yielding 52 combinations. A summary of the sample distribution per language and split is presented in Table 1.

3 Meenzer Team's System

Our systems consist of ensembles comprising various models, including finetuned character-level pointer-generator LSTMs, finetuned Mixtral 8x7B LLMs utilizing in-context finetuning, and GPT-4.

3.1 Pointer-Generator LSTM

Our first model group is a character-level sequence-to-sequence LSTM architecture, featuring an LSTM encoder and decoder equipped with an attention mechanism, alongside a pointer-generator (Bahdanau et al., 2015; See et al., 2017). The pointer-generator allows the LSTM to both copy words through pointing and generate characters from a predefined vocabulary (Vinyals et al., 2015).

In contrast to the typical sequence-to-sequence LSTM models, we use a separate LSTM encoder to encode the provided change features. For a detailed explanation of the sequence-to-sequence LSTM, we refer to Bahdanau et al. (2015). Furthermore, we deploy a pointer generator with a character-level vocabulary: At timestep t, given the attention distribution a^t over the characters in the source sequence, the decoder state s_t and the context vector h_t^*, the generation probability $p_{\text{gen}} \in [0, 1]$ is determined as:

$$p_{\text{gen}} = \sigma(w_{h^*}^T h_t^* + w_s^T s_t + w_x^T x + b_{ptr})$$

where vectors $w_{h^*}^t, w_s, w_x$ and the scalar b_{ptr} are all learnable parameters, while σ represents the sigmoid function. The probability p_{gen} serves as a soft switch, enabling the model to decide whether to generate a character from the vocabulary or to copy a character from the source sequence by sampling from the attention distribution a_t:

$$P(w) = p_{\text{gen}} P_{\text{vocab}}(w) + (1 - p_{\text{gen}}) \sum_{i:w_i=w} a_i^t,$$

where $P_{\text{vocab}}(w)$ represents the probability distribution across all characters in the vocabulary, while $P(w)$ additionally adds all characters present in the source sequence.

Training We adopt a two-step training approach for our model: Initially, we train a model on the combined training sets of all three languages for 100 epochs, incorporating early stopping. Additionally, we employ hyperparameter tuning through 100 trials; see Appendix A.1. Subsequently, in preparation for our ensemble approach, we select the top 10 models and conduct further finetuning on each model using the dataset of the target language. This process is repeated independently for all three languages. Each change feature is assigned a distinct feature token, and we include language tags for each individual dataset, treating them as a change feature.

[2] https://github.com/AmericasNLP/americasnlp2024/tree/master/ST2_EducationalMaterials/data

<USER>: This is [LANGUAGE]. I will give you a source sentence and a grammar change and you have to output the correct change!
<Assistant>: Ok!
<USER>: Source Sentence: [SOURCE$_1$] / Grammar Change: [CHANGE$_1$]
<Assistant>: [TARGET$_1$]
<USER>: Source Sentence: [SOURCE$_{\text{sample}}$] / Grammar Change: [CHANGE$_{\text{sample}}$]
<Assistant>: [TARGET$_{\text{sample}}$]

Figure 2: An example of a 1-shot prompt for a sample, with [LANGUAGE] being replaced by the specific language under consideration. During training, we predict and compute the loss based on the [TARGET$_{\text{sample}}$] sequence. However, during testing, [TARGET$_{\text{sample}}$] is left blank and must be predicted.

	Bribri			Guarani			Maya			Avg.
	Acc.	BLEU	ChrF	Acc.	BLEU	ChrF	Acc.	BLEU	ChrF	**Acc.**
Dev Set										
`(1) LSTMs+Mixtrals+GPT4s`	**30.19**	**51.96**	**67.60**	**53.16**	**61.98**	**88.53**	**70.46**	**85.14**	**93.75**	**51.27**
`(2) LSTMs+Mixtrals`	**30.19**	**51.96**	**67.60**	49.36	58.09	86.33	**70.46**	**85.14**	**93.75**	50.00
`(3) LSTMs`	24.10	50.30	61.47	41.77	43.28	77.65	70.47	85.13	93.59	45.45
`(4) Mixtrals`	22.17	47.28	66.80	44.30	54.78	84.60	61.74	80.67	91.60	42.74
Test Set										
`(1) LSTMs+Mixtrals+GPT4s`	17.50	44.20	70.09	**34.62**	**49.60**	**84.93**	38.39	66.81	83.70	30.17
`(2) LSTMs+Mixtrals`	17.50	44.20	70.09	23.08	35.95	79.71	38.39	66.81	83.70	26.32
`(3) LSTMs`	8.54	32.50	61.24	12.64	20.01	71.61	27.74	58.59	79.29	16.31
`(4) Mixtrals`	**19.38**	**46.93**	**73.02**	23.90	36.94	79.48	**53.87**	**77.68**	**90.94**	**32.38**

Table 2: Our results on the development set (upper part) and the official results on the test set (lower part).

3.2 Mixtral 8x7B (Instruct)

Our second model is the Mixtral 8x7B (Instruct),[3] a LLM finetuned on instructional data (Jiang et al., 2024).

Architecture The Mixtral 8x7B model is a sparse mixture of experts language model (Shazeer et al., 2017), employing the same decoder-only transformer architecture as Mistral 7B (Jiang et al., 2023). However, it distinguishes itself by having each layer composed of 8 feedforward blocks, referred to as *experts*. At every token and layer, a router network selects two experts, which may vary at each timestep, to process the current state and combines their outputs. Consequently, while each token theoretically has access to 47B parameters, only 13B active parameters are utilized during inference. We leverage the instruction-tuned version.

Training We employ, what Li et al. (2023) call, supervised in-context learning (SICL), which differs itself from conventional in-context learning (ICL) by integrating in-context examples directly into the training phase (Min et al., 2022; Chen et al., 2022). We concatenate the task instruction, labeled in-context examples, and the target sequence to predict. Subsequently, we finetune the model to predict the target sequence, see Figure 2 for an example. In contrast, ICL generate predictions without adjusting model parameters.

To enhance both training and inference efficiency, we implement 4-bit quantization with LoRA (Dettmers et al., 2023). We train multiple LoRA adapters by varying the number of examples per prompt (k) and the number of epochs (m). Specifically, we experiment with $k = 5, 10, 20$ and $m = 10, 20$, resulting in a total of 6 models per language. Each LoRA adapter, applied onto the query and value projection matrices in the self-attention module, possesses a rank of 8. For each sample, examples are selected based on their overlap with the same or similar changes, with the top-k most similar examples chosen. Additionally, the order of the top-k examples is randomized for each epoch. We employ a learning rate of 1e-4 alongside a cosine learning rate scheduler, with a weight decay of 0.1.

[3]Model taken from `https://huggingface.co/mistralai/Mixtral-8x7B-Instruct-v0.1`

3.3 GPT-4

In addition to Mixtral 8x7B, we incorporate GPT-4 using ICL. GPT-4, another LLM, is configured with $k = 20$ examples. We maintain consistency in example selection and prompt style with Mixtral 8x7B (Instruct). Specifically, we leverage the *gpt-4-turbo-2024-04-09* version of GPT-4.

3.4 Ensembling Strategy

Our four final systems consist of different ensembles constructed from the previously mentioned models, leveraging majority voting to reach a final decision, with the best-performing model on the development set breaking ties. To introduce more diversity for the LLMs, we generate two inference prompts: While one prompt organizes the top-k examples in ascending order, the other arranges them in descending order. Consequently, for each language, we have 10 LSTM, 12 Mixtral, and 2 GPT-4 predictions. For each system, we choose the best combination of models by evaluating their performance on the development set.

System 1 This system incorporates predictions from the LSTM, Mixtral 8x7B, and GPT-4 models. It is denoted by `(1) LSTMs+Mixtrals+GPT4s`.

System 2 This system comprises predictions from the LSTM and Mixtral 8x7B models, labeled as `(2) LSTMs+Mixtrals`.

System 3 This system solely relies on predictions from the LSTM models, identified as `(3) LSTMs`.

System 4 This system only considers the Mixtral models and is denoted by `(4) Mixtrals`.

4 Results

The primary metric for evaluating the shared task performance is accuracy (acc.), supplemented by BLEU (Papineni et al., 2002) and ChrF (Popović, 2015) as additional metrics. We present the results for the development set and test set in Table 2.

4.1 Development Set Results

The ensemble of all models demonstrates the highest performance, achieving an average accuracy of 51.24 and attaining the top scores across all languages. Notably, the only difference between `(1) LSTMs+Mixtrals+GPT4s` and `(2) LSTMs+Mixtrals` is in the Guarani language, where the addition of ChatGPT improves performance. When considering only LSTM models, we still achieve an average accuracy of 45.45, compared to 42.74 for Mixtral models.

4.2 Test Set Results

On the test set, we observe a significant difference from the reported development set results. The Mixtral ensemble performs best, achieving an accuracy of 32.38, approximately 10 points lower than its development set performance. Surprisingly, the LSTM ensemble performs notably worse, with an average accuracy of only 16.31. This decline in performance cascades through all other ensembles incorporating LSTM models: `(2) LSTMs+Mixtrals` achieves an average accuracy of 26.32, while `(1) LSTMs+Mixtrals+GPT4s` reaches an average of 30.17.

Nevertheless, our `(1)` system achieves the highest performance on Guarani among all shared task systems, while `(4) Mixtrals` attains the highest accuracy on Maya (tied with another team). Overall, our `(4) Mixtrals` system secures second place among all systems based on average accuracy.

Development & Test Set Discrepancy The LSTMs, constructed at the character-level and trained from scratch with a limited training set, might encounter numerous unknown characters. Analyzing the case-sensitive character overlap between the language specific training, development, and test sets reveals a substantial disparity. For instance, in the case of Bribri, we observe that, while 21% of samples in the development set contain unseen characters, this figure rises to 65.4% in the test set. Similarly, for Guarani, the proportion increases from 11.4% in the development set to 22.3% in the test set. Conversely, for Maya, while there are no unseen characters in the development set, they account for 15.5% of samples in the test set.

	Bribri	Guarani	Maya	**Avg.**
Ensemble vs. (Best) Single Model				
Mixtral (Single)	17.45	40.50	57.71	38.55
Mixtrals (Ensemble)	**22.17**	**44.30**	**61.74**	**42.74**
ICL vs. SICL				
Mixtral (ICL)	7.08	18.99	35.57	20.55
Mixtral (SICL)	**14.15**	**36.7**	**57.71**	**36.19**
Random Prompt Order				
Mixtral (Fix)	8.49	35.44	54.36	32.76
Mixtral (Random)	**14.15**	**36.70**	**57.71**	**36.19**

Table 3: Ablation study on the development set for `(4) Mixtrals`, our best system.

5 Ablation Study

In this section, we conduct a brief ablation study on our best-performing system, `(4) Mixtrals`. The results on the development set are presented in Table 3.

Ensemble vs. (Best) Single Model We demonstrate that assembling the Mixtral models into an ensemble boost performance by approximately 4.19 average accuracy points compared to the single best Mixtral model.

ICL vs. SICL For this and the following comparison, we fix the number of examples to $k = 20$ and epochs to $m = 10$. We observe that ICL, which does not adjust parameters, demonstrates an average accuracy of only 20.55, a notable 15.64 lower than SICL.

Random Order per Epoch: Finally, we investigate the impact of randomly varying the order of the k examples in the prompt per epoch on performance. We find that maintaining a fixed order (consistent during inference) leads to decreased performance across all languages, with an average accuracy decrease of 3.43.

6 Conclusion

We presented the systems of the Meenzer Team by JGU for the AmericasNLP 2024 shared task on the creation of educational resources. We trained character-level pointer-generator LSTMs as well as Mixtral 8x7B models finetuned through SICL. In addition, we used GPT-4 models via in-context learning. We secured second place with an ensemble of the finetuned Mixtral 8x7B models and reached the highest accuracy on Maya. Additionally, we achieved the highest performance on Guarani using an ensemble of LSTM, Mixtral, and GPT-4 models.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Yanda Chen, Ruiqi Zhong, Sheng Zha, George Karypis, and He He. 2022. Meta-learning via language model in-context tuning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 719–730, Dublin, Ireland. Association for Computational Linguistics.

Luis Chiruzzo, Pavel Denisov, Samuel Canul Yah, Lorena Hau Ucán, Marvin Agüero-Torales, Aldo Alvarez, Silvia Fernandez Sabido, Alejandro Molina Villegas, Abteen Ebrahimi, Robert Pugh, Arturo Oncevay, Shruti Rijhwani, Rolando Coto-Solano, Katharina von der Wense, and Manuel Mager. 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)*. Association for Computational Linguistics.

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

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms. *ArXiv*, abs/2305.14314.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

Chengzu Li, Han Zhou, Goran Glavaš, Anna Korhonen, and Ivan Vulić. 2023. On task performance and model calibration with supervised and self-ensembled in-context learning. *Preprint*, arXiv:2312.13772.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.

OpenAI. 2023. Gpt-4 technical report. *ArXiv*, abs/2303.08774.

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.

Maja Popović. 2015. chrF: character n-gram F-score for automatic MT evaluation. In *Proceedings of the Tenth Workshop on Statistical Machine Translation*, pages 392–395, Lisbon, Portugal. Association for Computational Linguistics.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *Preprint*, arXiv:1701.06538.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. In *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc.

	Hyperparameter	values
Optimization	Batch size	{2, 4, ..., 128}
	Learning rate	[$1e^{-5}$, 0.01]
	β_1	[.8, .999]
	β_2	[.98, .999]
	Label smoothing	[0, .2]
	Scheduler	{reduceonplateau, warmupinvsqrt, (none)}
	Warmup samples*	{0, 10, ..., 1000}
	Factor*	[.1, .9]
	Min. learning rate*	[$1e^{-7}$, .001]
	Learning rate patience*	{1, 2, ..., 5}
Architectural	Embedding Size	{16, 32, ..., 512}
	Hidden layer size	{64, 128, ..., 2048}
	Encoder & Decoder layers	{1, 2}
	Feature Attention heads	{1, 2}
	Dropout	[0, .5]

Table 4: LSTM hyperparameter space. Continuous distributions are denoted by intervals [...], while discrete ones show step sizes 1, 2, ..., max. We uniformly sample from these, except for the learning rate, which follows a log uniform distribution. Hyperparameters and the distributions we sample from. * marks conditional hyperparameters, relevant only with chosen schedulers.

A Appendix

A.1 Hyperparameter Grid

We report in Table 4 the hyperparameter grid for our LSTMs.

Applying Linguistic Expertise to LLMs for Educational Material Development in Indigenous Languages

Justin Vasselli and **Arturo Martínez Peguero** and **Junehwan Sung** and **Taro Watanabe**
{vasselli.justin_ray.vk4, martinez_peguero.arturo.ma3, sung.junehwan.sl9, taro}@is.naist.jp

Abstract

This paper presents our approach to the AmericasNLP 2024 Shared Task 2 as the JAJ (/dæz/) team. The task aimed at creating educational materials for indigenous languages, and we focused on Maya and Bribri. Given the unique linguistic features and challenges of these languages, and the limited size of the training datasets, we developed a hybrid methodology combining rule-based NLP methods with prompt-based techniques. This approach leverages the meta-linguistic capabilities of large language models, enabling us to blend broad, language-agnostic processing with customized solutions. Our approach lays a foundational framework that can be expanded to other indigenous languages languages in future work.

1 Introduction

In recent years, there has been a surge of interest in developing natural language processing (NLP) technologies for low-resource languages. This is motivated by the desire to preserve cultural heritage and foster linguistic diversity.

Among 7,000 living languages on Earth, 42% of languages are in danger of disappearing, according to Ethnologue[1]. In the Americas, there are approximately 1,200 languages (Hammarström et al., 2024), and about 70% of them are classified as endangered or already extinct by the United Nations Educational, Scientific and Cultural Organization (UNESCO) (Moseley and Alexandre, 2010). The endangered languages are losing their vitality, and intergenerational transmission is barely occurring or not occurring at all, which suggests these languages are at a high risk of becoming extinct in the near future, along with the the cultural legacy residing within. While many communities are engaged in revitalization endeavors, such as developing educational resources for language preservation, these initiatives are confronted with significant challenges in terms of cost and time investment. NLP can be one way of ameliorating the situation.

The AmericasNLP 2024 Shared Task 2 is specifically aimed at creating educational materials for indigenous languages in the Americas, such as Maya, Guarani, and Bribri (Chiruzzo et al., 2024). These languages each have their own unique linguistic features and challenges, offering a rich opportunity for the application of innovative NLP techniques.

In this shared task, we are given a dataset with base sentences and the grammatical changes expected to be applied to each base sentence. Based on this information, we will train a system that can transform the base sentences according to the indicated changes.

Base sentence: *Ye' shka'* (Bribri for "I walked")
Expected change: Polarity: Negative
Target sentence: *Ye' kë shkane* (Bribri for "I didn't walk")

We participated in this task with the goal of developing a method that effectively handles the linguistic complexities of these languages. Our hybrid methodology combines rule-based NLP methods with prompt-based techniques, making use of the meta-linguistic capabilities of large language models (LLMs). This approach enabled us to blend broad, language-agnostic processing with customized solutions that address each language's unique needs.

Maya largely does not conjugate verbs based on grammatical tense; instead, it uses aspectual marking to express time-related information. The abundance of apparently simple examples with only a single change, presented an interesting challenge. Rather than using all the available examples, we experimented with selection methods to find the most syntactically similar example sentences.

Bribri, on the other hand, had a much more com-

[1] https://www.ethnologue.com/insights/how-many-languages-endangered/

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 201–208
June 21, 2024

plicated system of verb conjugation. Using more examples was not as effective as a separate system to handle the intricate verb morphology. This aspect of Bribri posed a substantial challenge, further complicated by variations in notation and spelling observed across different educational resources.

While our system addresses specific linguistic features of Maya and Bribri, we encountered time constraints that limited our ability to apply similar dedicated modifications to Guarani. Consequently, Guarani was processed using the base system, without the benefit of language-specific adaptations. Recognizing the unique linguistic structure and cultural significance of Guarani, we are committed to expanding our system in the future to include tailored strategies that cater specifically to its needs, enhancing the model's overall effectiveness.

2 Prior Work

Our approach is inspired by previous work on Rosetta Stone puzzles (Bozhanov and Derzhanski, 2013). These puzzles are carefully designed to encourage solvers to deduce implicit linguistic features from a low-resource language, relying solely on a limited set of bilingual sentence pairs, one of which is a low-resource language with uncommon linguistic features. Leveraging the inferred linguistic patterns, solvers are tasked with translating new sentences in both directions, which is in essence a few-shot translation task.

Şahin et al. (2020) explored the resolution of Rosetta Stone problems using a range of algorithms, from random word selection to transformer-based language models. Sung et al. (2024) conducted similar experiments, specifically focusing on assessing the metalinguistic awareness of pre-trained language models. Their findings not only underscored the limitations of various algorithms in tasks requiring linguistic inference skills but also outlined directions for advancing machines' comprehension of human language.

Recent advancements in the application of LLMs to Rosetta Stone puzzles have shown encouraging outcomes, demonstrating the potential of LLMs to extract and apply linguistic knowledge from limited data (Vamvas, 2022; Chi et al., 2024). This research supports the feasibility of our approach, suggesting that LLMs can adeptly adapt to the intricacies of low-resource languages by leveraging their inherent capabilities in pattern recognition and language comprehension.

Another relevant research area is the Machine Translation from One Book (MTOB) task (Tanzer et al., 2024). In this task, models are trained to translate between English and Kalamang using only a single Kalamang textbook. LLMs seem to be breaking into this space as well, as shown by Gemini 1.5l's state-of-the-art performance on this task. The MTOB task is particularly pertinent to our work because it demonstrates the potential of machine learning models to operate effectively even with severely limited datasets, mirroring the conditions we faced in the shared task (Pichai and Hassabis, 2024).

3 Method

3.1 Preparing the Examples

Due to the limited dataset size, ensuring the quality and consistency of data was paramount. We implemented several preprocessing steps to optimize the training data:

Duplicate Removal and Capitalization: We removed duplicate entries where the source and change tags were identical and standardized capitalization across the dataset to ensure uniformity. This reduced the number of example sentences in Maya from 594 to 584, but didn't affect Bribri or Guarani.

Tag Collapsing: Through detailed analysis of the training data, we identified and merged change tags that consistently appeared together, reducing redundancy. For example for changes that included both ASPECT and TENSE in Maya, we removed the TENSE tags as they did not appear to change the Target sentence.

Sub-step Expansion: We split some complex changes into simpler, sequential sub-steps. For instance, a change tagged as `TYPE:NEG, SUBTYPE:INT` was decomposed into two stages: first applying `TYPE:NEG` to reach an intermediary form, followed by `SUBTYPE:INT` to achieve the final state. This added 12 new cases to Maya and 6 new cases to Bribri.

Backward Labeling: When feasible, we generated additional training cases by labeling from the target back to the source, effectively doubling the data for those instances. The total number of examples used were 1199 for Maya (up from 594), 329 for Bribri (up from 309), and the 178 Guarani examples remained unchanged.

Independence Analysis: By consulting language-specific textbooks and analyzing the training data, we identified which changes were independent and which were interdependent, enabling more precise modeling of language rules.

3.2 Base System

Our base system is designed to leverage the capabilities of large language models by prompting them with relevant example cases. The system's operation during inference takes the same form, regardless of language, with some language specific modifications at key points.

Initially we simply included all available examples with the same change as the test case. However, we quickly found that not all test case changes were represented in the training data, particularly those with compound changes. To address this, we implemented a language-specific strategy for decomposing and processing the changes:

Maya: Changes that commonly occurred together (e.g., `ASPECT:BEG, TENSE:PAS_SIM`) were collapsed into a single step (`ASPECT:BEG`). The remaining changes were then split and processed sequentially in the order: `STATUS` → `PERSON` → `ASPECT` → `TYPE` → `SUBTYPE`.

Bribri: Changes that typically co-occurred (e.g., `ABSNUM:PL, PERSON:3_PL`) were combined into a single change (`PERSON:3_PL`). `PERSON` is then processed separately from the other changes.

Guarani: Changes are applied consecutively in the case of compound changes.

When a test case has a compound change, it is passed to the language specific function that determines which changes should be applied and if they can be applied together or sequentially. In the case of sequential changes, the system will reprompt the model with the modified output from the previous step until all specified changes are implemented. For instance, a Bribri case with the change `TENSE:PRF_REC, ASPECT:PFV, ABSNUM:PL, PERSON:3_PL` will undergo two rounds of processing, each focusing on one specific change. First the tense will be changed to recent perfect, then the result of that will be changed to 3rd person plural.

3.3 POS Tagging

A key component of our system is the application of custom, simplified part of speech (POS) taggers tailored to each target language. These taggers are primarily dictionary-based and are used to supplement the example sentences being passed to the LLM by explaining better the grammatical role of the words of the provided examples. The POS tagger for Maya focuses predominantly on function words, as these play a crucial role in understanding the grammatical structure of sentences. We didn't use a full dictionary for Maya, but made sure to have coverage of aspect markers such as *táan* and pronouns such as *in* or *teen* (Bolles and Bolles, 1996). Additionally our tagger is designed to recognize and handle known suffixes such as *e'ex*. For Bribri, we developed a POS tagger using a comprehensive dictionary of Bribri words from Professor Haakon S. Krohn's website[2] (Krohn, 2023).

3.4 The Prompt

The prompt was adapted from the one Vamvas (2022) used for the Rosetta Stone Puzzles. It integrates the retrieved examples with part of speech tags, as well as some language specific hints.

Here is a full example from Maya:

> System message:
> You are a helpful assistant with a strong background in linguistics.
>
> User message:
> This is a linguistic puzzle. Below are example sentences in a foreign language and sets of changes to apply to them.The examples are followed by the problem sentence and desired change.Your task is to look closely at the example sentences and to change the sentence correctly.
>
> Example 1:
> Sentence: Tin kaxtik a najil
> ((Tin, PRON:TA:1_S) (kaxtik, ::) (a, PRON:A:2_X) (najil, ::))
> Change(s): PERSON:3_PL, ASPECT:PRG
> Answer: Táan u kaxtiko'ob a naajil
> ((Táan, PART:ASP:) (u, PRON:A:3_X) (kaxtiko'ob, :kaxtik:3_P) (a, PRON:A:2_X) (naajil, ::))
>
> *Example 2-5*
>
> Here is the problem. Answer first, then explain your reasoning.
> Sentence: Tin wilaj u sáastal tin kaajal ((Tin, PRON:TA:1_S) (wilaj, ::) (u, PRON:A:3_X) (sáastal, ::) (tin, PRON:TA:1_S) (kaajal, ::))
> Change(s): PERSON:3_PL
> Hint:

3.5 Generation and Postprocessing

We tested generation on the development set using `gpt-3.5-turbo-0125`,

[2] https://www.haakonkrohn.com/bribri/index.html

`gpt-4-0125-preview` (OpenAI et al., 2024), and `Mixtral-8x7B-Instruct-v0.1` (Jiang et al., 2024). For the GPT models, we used temperature of 0. For Mixtral we used a greedy search.

While the generation can be stopped after "`\n\n`", we found that not setting the stop allowed the model to generate an explanation of the change. This was useful for interpretability, and we were able to spot some issues during development this way.

For postprocessing, the preamble (Answer:) and POS tags are stripped from the first line. Mixtral is less consistent with the answers, so we wrote some rule based processing to find potential answers in quotes.

Here is an example output from GPT-4:

> Answer: Tu yilo'ob u sáastal tu kaajalo'ob ((Tu, PRON:TA:3_X) (yilo'ob, :yil:3_P) (u, PRON:A:3_X) (sáastal, ::) (tu, PRON:TA:3_X) (kaajalo'ob, :kaajal:3_P))
>
> Explanation:
> The task involves changing the sentence from first person singular to third person plural. To do this, we follow the patterns observed in the examples.
> 1. **Pronoun Change**: In all examples, the first person singular pronoun "Tin" changes to the third person plural pronoun. In Examples 2 and 5, "Tin" changes to "Tu" for third person plural. Therefore, both instances of "Tin" in the problem sentence are changed to "Tu".
> 2. **Verb and Action Word Changes**: In the examples, verbs and action words that are in first person singular form are changed to their third person plural forms by adding an "o'ob" suffix or changing the suffix to match third person plural conventions. This is seen in Example 5 with "wilaj" changing to "yilo'ob" and "kaajali"' changing to "kaajalo'obi"'. Since "kaajali"' is not present in the problem sentence, but "kaajal" is, we apply a similar transformation to "kaajal", changing it to "kaajalo'ob". The word "wilaj" similarly changes to "yilo'ob" following the pattern from Example 5.
> 3. **Maintaining Unchanged Parts**: Elements that do not pertain to person, such as "u sáastal" (indicating possession or association with "sáasta"), remain unchanged in the transformation, as seen across all examples.
> By applying these rules derived from the examples, we correctly transform the problem sentence to third person plural as requested.

3.6 Maya Adaptations

Initially, our system used all available examples for each change, but we refined our example selection process for Maya, where certain changes could return over 100 relevant examples.

We conducted experiments on the development set to find the optimal number of examples. These experiments varied the number of examples from one to ten per test case. The results indicated that selecting five examples struck the best balance between performance and cost-efficiency.

System	Accuracy	BLEU	ChrF
5 examples	20.81	50.72	78.60
+ selected	47.56	72.39	89.36
+ POS Tags	**56.38**	**78.26**	**91.33**
All matches	**56.38**	77.63	91.30

Table 1: Different types of example retrieval on the Maya development set

The selection of these five examples was based primarily on syntactic relevance rather than semantic similarity or word form matching. By comparing the POS tagged source sentences of the examples with that of the input, we were able to identify examples that shared similar syntactic structures. This approach ensured that the examples used were not structurally analogous, enhancing the effectiveness of the model's performance.

For example if the input sentence is *Ma' ta jutaj a najili'*, an example sentence *Ta manaj a najil* overlaps significantly in word form, but the first sentence is a negative sentence, while the second is affirmative. This means they will react to the next change slightly differently. Instead, we match *Ma' ta jutaj a najili'* with *Ma' tin wa'alik u k'aaba'* which is structurally similar, despite less character-level overlap.

This method reduced computational costs by approximately 75% across the full test set, achieving comparable results to using the full example set, as shown in Table 1. Using five carefully selected examples with POS tags not only matches the performance of using all examples on the development set, but also achieves substantial cost savings, further justifying our approach.

3.7 Bribri Adaptations

Managing the complexity of Bribri verb conjugation, especially for irregular verbs, necessitated innovative methods to enhance translation accuracy. An initial experiment used an oracle verb conjugation hint, which leveraged correct verb forms from target sentences in the development set, to evaluate potential performance enhancements. This experiment led to a substantial increase in accuracy from 15% to 65%, signifying the crucial role of accurate verb conjugation in model performance.

Motivated by the success of the oracle hint, we

System	Accuracy	BLEU	ChrF
Examples	14.62	43.69	64.61
+ VERB form hint	46.70	64.84	79.85
+ MODE hints	**47.17**	**67.01**	**80.75**

Table 2: Ablation results for the hint categories on the Bribri development set

Data	Accuracy	BLEU	ChrF
Maya dev	56.38	78.26	91.33
Maya test	54.17	71.72	82.78
Baseline	25.81	53.69	80.23
Bribri dev	47.17	67.01	80.75
Bribri test	53.55	78.41	91.53
Baseline	8.75	22.11	52.73
Guarani dev	41.77	55.81	86.12
Guarani test	36.81	48.29	84.12
Baseline	14.84	25.03	76.10

Table 3: The results of the submitted system on the development and test sets.

developed a rule-based verb conjugator, employing a comprehensive database of verb conjugations sourced from Jara Murillo (2018)[3]. In this system, verbs identified by our POS tagger are looked up in the database to ascertain the correct conjugational form based on the grammatical context. For instance, in the sentence *Ye' tö i kötwa* with changes `TYPE:NEG, TENSE:FUT_CER, ASPECT:IPFV`, the verb *kötwa* is located by the POS tagger and looked up in the verb conjugation database. It's found to be the perfect remote form of *ujtökwa*. The conjugator transforms the verb into *ujtèpawa* for the negative certain future tense. This transformation is then included as a hint in the model's prompt:"The correct form of *kötwa* is likely *ujtèpawa*."

These enhancements, particularly the integration of verb conjugation hints, have markedly improved our system's performance, as detailed in the ablation study results in Table 2. While the rule-based conjugator does not replicate the perfect accuracy of the oracle hint, it significantly contributes to the overall effectiveness of the system in managing Bribri's complex verb conjugations.

3.8 Additional hints

Drawing from the Machine Translation from One Book (MTOB) task, we supplemented our approach with grammatical hints taken from textbooks. For each language, we incorporated short summaries of grammatical rules related to the changes from language textbooks. While time constraints limited the coverage of all possible changes, the preliminary results from these hints were promising and represent a straightforward avenue for further improvement.

4 Results

Our systems for Maya and Bribri improved on the baselines provided by the task organizers by considerable amounts. In particular, our system's Bribri accuracy was over six times higher than the edit-tree based baseline. This is likely due to the challenges of complex verb conjugation using an edit-tree approach. The complete results can be seen in Table 3.

As indicated in Tables 1, Maya became more resource efficient with example selection and POS tagging. Bribri performance saw the largest boost from verb hints and moderate improvements from mode hints, as shown in Table 2.

Our Bribri system was the best performing in the competition. Our Maya system came a close third. Due to this, and despite regrettably not submitting Guarani results, our contribution was ranked first overall. We submitted Guarani after the deadline, and report the results of all three languages on the development and test sets compared to the baseline in Table 3.

5 Discussion

We submitted our results using GPT-4 for system prompting due to its superior performance on the development set. Table 4 details the performance across different LLMs, noting that while Mixtral scored more competitively with GPT 3.5 for Maya, it was outperformed by both GPT models in the other languages.

5.1 Error Analysis

5.1.1 Maya

The errors in Maya predominantly stem from inconsistencies in the example data and the complex syntactic structures that require deeper linguistic insights beyond mere pattern matching. For instance, the development set case *Te'exe' ti' kajakbale'ex tu yotoche'* (`TYPE:NEG`) changes to *Ma' kajakbale'ex tu wotochi'*, contrasting with a similar training example *Leti'obe' ti' u taalo'obi'* (`TYPE:NEG`) that

[3]https://www.lenguabribri.com/gramática-de-la-lengua-bribri

Lang	System	Acc.	BLEU	ChrF
Maya	Mixtral	44.97	69.19	83.52
	GPT-3.5	42.28	67.84	86.04
	GPT-4	**56.38**	**78.26**	**91.33**
Bribri	Mixtral	34.43	42.86	72.06
	GPT-3.5	40.57	61.15	77.04
	GPT-4	**47.17**	**67.01**	**80.75**
Guarani	Mixtral	12.66	20.95	69.84
	GPT-3.5	36.71	51.38	83.35
	GPT-4	**41.77**	**55.81**	**86.12**

Table 4: The results on the development set for the different LLMs.

becomes *Leti'obe' ma' ti' u taalo'obi'*, where *Leti'obe* and *ti'* are retained. These inconsistencies, including the absence of examples for changes like *tu yotoche'* to *tu wotochi'*, contribute to over 50% of the errors.

Additionally, positioning of *wáaj* in interrogative sentences varied without clear rules, leading to misplacement about 25% of the time. These idiosyncratic cases highlight the need for more robust language-specific rules in our system.

The remaining errors involved rule misapplication, such as overuse of the suffix *-o'ob*, and spelling mistakes like failing to correctly modify *tu wotoch* to *tu yotoch*.

5.1.2 Bribri

For Bribri, verb conjugation continues to be a major issue, accounting for 57% of the errors. These range from minor issues like incorrect accent placement (e.g., *sùr* instead of *sú̀r*) to significant errors such as incorrect verb forms (e.g., *kötwa* instead of *ujtèkèulur*). Another 19% of errors were due to omissions, where words present in the reference were missing in the hypothesis (e.g., *Ppö* instead of *I ppö*).

Less frequently, errors involved extraneous words in the hypothesis, making up 9% of the total errors (e.g., *Ye' wa stsa'* instead of *Ye' stsa'*). Incorrect pronoun use accounted for 8% of the errors, and the remaining 6% were due to words appearing out of order (e.g., *Kë ie' stsö* instead of *Ie' kë stsö*).

5.1.3 Guarani

For Guarani, although no language-specific optimizations were implemented, the error analysis indicates that the majority of the issues are related to verb conjugation. Specifically, incorrect verb forms account for 75% of the errors observed. This suggests that developing a system similar to the rule-based verb conjugator used for Bribri, which provides hints based on accurate verb conjugation, could be highly beneficial in improving the accuracy for Guarani. Implementing such a system could significantly reduce errors and enhance the model's overall performance for this language.

6 Conclusion

Supplementing the capabilities of LLMs, such as GPT-4, by incorporating simple rule-based natural language processing techniques, our approach to the AmericasNLP 2024 Shared Task 2 has laid a foundational framework that can be expanded in future work to include other low-resource languages, contributing to the creation of educational materials for indigenous languages.

Particularly effective were the custom verb conjugation hints for Bribri, which markedly improved accuracy. This approach underscores the ongoing need for focused linguistic tools tailored to the specific structural complexities of each language.

Throughout the project, we encountered several challenges, including inconsistencies in the training data and the complex nature of indigenous language structures that often deviate significantly from those of more widely studied languages. These issues underscore the importance of developing tailored NLP tools that can adapt to the idiosyncrasies of any given language.

Looking ahead, we aim to extend our methodology to include Guarani more comprehensively, enhance our rule-based systems for better accuracy, and further explore the potential of LLMs in processing linguistically diverse and low-resource languages. This work not only contributes to the field of computational linguistics by providing valuable insights into the treatment of indigenous languages but also plays a crucial role in the preservation and revitalization of these vital cultural heritages.

By continuing to refine our approaches and expand our linguistic coverage, we hope to contribute to a more inclusive and equitable representation of languages in the digital age, ensuring that technology serves as a bridge rather than a barrier in the education and preservation of linguistic diversity.

Acknowledgments

A special thank you to Professor Carla Victoria Jara Murillo and Professor Haakon S. Krohn who generously allowed us to use and repackage their

Bribri data for use in this project.

References

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

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

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

Luis Chiruzzo, Pavel Denisov, Samuel Canul Yah, Lorena Hau Ucán, Marvin Agüero-Torales, Aldo Alvarez, Silvia Fernandez Sabido, Alejandro Molina Villegas, Abteen Ebrahimi, Robert Pugh, Arturo Oncevay, Shruti Rijhwani, Rolando Coto-Solano, Katharina von der Wense, and Manuel Mager. 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)*. Association for Computational Linguistics.

Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank. 2024. glottolog/glottolog: Glottolog database 5.0.

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

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *Preprint*, arXiv:2401.04088.

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

Christopher Moseley and Nicholas Alexandre. 2010. *Atlas of the world's languages in Danger*, 3rd edition. Memory of peoples. United Nations Educational, Scientific and Cultural Organization (UNESCO).

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav

Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Sundar Pichai and Demis Hassabis. 2024. Our next-generation model: Gemini 1.5.

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

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

Garrett Tanzer, Mirac Suzgun, Eline Visser, Dan Jurafsky, and Luke Melas-Kyriazi. 2024. A benchmark for learning to translate a new language from one grammar book. *Preprint*, arXiv:2309.16575.

Jannis Vamvas. 2022. Translation puzzles are in-context learning tasks.

Exploring Very Low-Resource Translation with LLMs: The University of Edinburgh's Submission to AmericasNLP 2024 Translation Task

Vivek Iyer Bhavitvya Malik* Wenhao Zhu* Pavel Stepachev*
Pinzhen Chen Barry Haddow Alexandra Birch
School of Informatics, University of Edinburgh
vivek.iyer@ed.ac.uk

Abstract

This paper describes the University of Edinburgh's submission to the AmericasNLP 2024 shared task on the translation of Spanish into 11 indigenous American languages. We explore the ability of multilingual Large Language Models (LLMs) to model low-resource languages by continued pre-training with LoRA, and conduct instruction fine-tuning using a variety of datasets, demonstrating that this improves LLM performance. Furthermore, we demonstrate the efficacy of checkpoint averaging alongside decoding techniques like beam search and sampling, resulting in further improvements. We participate in all 11 translation directions. Our models are released here: https://tinyurl.com/edi-amnlp24

1 Introduction

We participated in AmericasNLP 2024's shared task on machine translation (MT). It requires participants to translate from Spanish to 11 indigenous American languages: Aymara (aym), Bribri (bzd), Ashaninka (cni), Chatino (ctp), Guarani (gn), Huichol (hch), Nahuatl (nhe), Otomi (ote), Quechua (quy), Shipibo-Konibo (shp) and Tarahumara (tar). We adopted multilingual large language models (LLMs) and our workflow consists of data curation, continued pre-training, instruction tuning, and several decoding techniques. We submitted to all 11 translation directions.

We study and report the feasibility of using LLMs for very low-resource machine translation tasks. LLMs have recently been the focus of recent research interest, and in machine translation, they have demonstrated competitive or better performance against traditional neural MT systems in high-resource languages (Hendy et al., 2023; Robinson et al., 2023; Iyer et al., 2023; Alves et al., 2024). Nonetheless, research has shown that these models struggle in low-resource settings if used off-the-shelf (Robinson et al., 2023), and there has been limited exploration of adapting LLMs to extremely low-resource MT. Existing approaches rely on massively multilingual dictionaries (Lu et al., 2023) or a series of complex grammatical and linguistic tools (Zhang et al., 2024). Despite their effectiveness, a pitfall of these approaches is that it can be hard to scale them up to build multilingual, low-resource LLMs. Moreover, it is unclear how the (scarce) monolingual and parallel data available for these languages can be effectively utilised, and how recent developments in MT of high-resource languages (Xu et al., 2024; Alves et al., 2024) scale to very low-resource settings.

This work attempts to take a step towards answering these questions. We build multilingual LLMs for these indigenous American languages by fine-tuning Llama-2 7B (Touvron et al., 2023), Mistral 7B (Jiang et al., 2023) and MaLA-500 (Lin et al., 2024). We explore continued pre-training with LoRA on various monolingual and parallel data sources. We then conduct instruction tuning using a variety of tasks and language pairs, and show this contributes to performance improvements in MT. We end this work by demonstrating how familiar techniques such as checkpoint averaging, beam search, and sampling help boost LLM performance for low-resourced translation as well.

2 Data

2.1 Monolingual data

We summarize statistics of the monolingual data used in our experiments in Table 1. We curate this data from various sources:

MADLAD-400 (Kudugunta et al., 2024): This is a manually audited general domain dataset sourced from Common Crawl, spanning 419 languages. Given this corpus has many dialects among

*denotes equal contribution

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 209–220
June 21, 2024

the American languages of interest, we create a dictionary[1] mapping each language to the ISO 639-3 codes of all its dialects, and download all of them. We remark on various strategies we tried for handling dialects in Section 3.1. We sample 150000 sentences from the English and Spanish splits to maintain comparable data quantities.

Glot500 (ImaniGooghari et al., 2023): This dataset belongs to multiple domains, covers 500 languages and spans multiple licences. We downloaded the publicly available version of this dataset from Hugging Face, for the languages of interest to us, and concatenated the train, dev, and test splits for these languages. We handled dialects similar to the MADLAD-400 corpus.

Wikipedia: We download Wikipedia dumps for the languages of interest and parse them with WikiExtractor (Attardi, 2015) for downstream use.

Helsinki'23 datasets (De Gibert et al., 2023): We reuse the monolingual data extracted by the winning team from the AmericasNLP 2023 Shared Task, University of Helsinki. We separate out the Bibles, UDHR, Wikipedia, and Miscellaneous (Misc) domains.

OCR data: In the pursuit of additional data, we utilized alternative external resources. We manually extracted[2] various text resources (summarised in Table 9 and classified them into groups and languages. The extracted files were converted to PDF format. Each page of the file was transformed into PNG format and upscaled to a resolution of 600 DPI. Our approach employed **ocrmac**[3](based on the Apple Vision Framework) for OCR. The methodology focused solely on bounding box text spans, without the application of sentence or paragraph restoration. We summarize statistics of the OCR data in Tables 3, 8, 9.

2.2 Instruction Tuning data

Inspired from Alves et al. (2024), we try to make our instruction tuning dataset as diverse as possible, and observe that multi-task instruction tuning yields performance gains on the singular task of Machine Translation as well. We summarize the statistics of our instruction tuning dataset in Table 4, and detail our sources as follows:

Aya (Singh et al., 2024): We use the Cohere Aya Dataset for the English, Portuguese and Spanish languages which consist of about 3.8K, 3.8K and 9K instructions respectively. The Aya Dataset consists of freshly created human annotations to existing prompts, as well as re-annotations by humans of machine-generated prompt completions. Given that this dataset relies strongly on human annotation, we include it in our instruction tuning dataset - even though the languages provided are not the indigenous American languages we are interested in. We could not find any data for these American languages in the Aya project.

MT Data: We use the official datasets provided by the organizers (`official`), the NLLB and the FLORES 200 corpora (Costa-jussà et al., 2022), the Helsinki'23 OPUS parallel corpora (De Gibert et al., 2023) as well as our own extraction of the OPUS dataset (Tiedemann, 2009) – from which we were able to extract more languages and pairs than the original Helsinki collection. For the NLLB corpus, which is sorted in decreasing order of scores indicating translation quality, we sample sentences from the top to ensure the highest quality sentences are chosen for instruction tuning. Finally, as far as possible, we try to ensure uniform sampling across all these languages and corpora to prevent imbalance.

Cross-lingual QA: We also generate synthetic cross-lingual instruction data using a powerful open-source LLM, Mixtral-8x7B-Instruct (Jiang et al., 2024), for data augmentation. Our generation process is illustrated in Figure 1. Given a translation pair (X, Y), where X is from a high-resource language and Y is from a low-resource language, we follow the prompt of Köksal et al. (2024) and ask Mixtral to generate a question Q based on X. As X and Y are semantically equivalent, Y is now used as the answer to the question Q. Finally, we add an instruction at the end of the prompt to generate in the target language. This is, thus, similar to a cross-lingual QA task - where the question is in a high-resource language, but the answer is in the indigenous American language and the LLM is instructed to generate its response in the latter. In this way, we use (Q, Y) as synthetic cross-lingual instruction data.

During training, we convert all our instruction-

[1] `https://tinyurl.com/uedin-dialectsdict`

[2] We are not speakers of any indigenous languages in this shared task.

[3] `https://github.com/straussmaximilian/ocrmac` v0.1.6 with parameters: recognition_level="accurate", language_preference=["es-ES", "en-US", "ru-RU", "fr-FR", "de-DE"]

Language	Total	MADLAD 400	GLOT 500	Wikipedia	Helsinki'23 (Bibles)	Helsinki'23 (Misc)	Helsinki'23 (UDHR)	Helsinki'23 (Wikipedia)	*OCR (multilingual)*†
Aymara (ay)	779835	58572	355229	19272	61182	0	120	16081	269379
Bribri (bzd)	41123	0	0	0	7659	0	0	0	33464
Asháninka (cni)	74964	0	0	0	0	0	0	0	74964
Chatino (ctp)	113415	0	0	0	23764	0	0	0	89651
Guarani (gn)	531478	98351	97470	39546	7849	0	102	39593	248567
Huichol (hch)	68411	0	0	0	7936	373	0	0	60102
Nahuatl (nhe)	547187	84647	23615	0	70988	0	91	8641	359205
Otomi (oto)	284988	131139	7991	0	7943	443	156	0	137316
Quechuan (quy)	986947	113640	168189	62777	61131	0	277	58073	522860
Shipibo-Konibo (shp)	32326	4897	0	0	16025	0	122	0	11282
Tarahumara (tar)	63438	0	0	0	7894	0	0	0	55544
Total	3384364	491246	652494	121595	272371	816	868	122388	1862334

Table 1: Monolingual dataset used for continued pre-training, in terms of number of sentences, for the indigenous American languages. †OCR data is inherently multilingual, with significant amounts of English and/or Spanish, so the data per language is likely overestimated.

Corpus	English	Spanish
MADLAD 400	150000	150000
Wikipedia	100000	100000
Helsinki'23 (Bibles)	148060	487006
Helsinki'23 (UDHR)	0	120
Total	398060	737126

Table 2: Monolingual dataset used for continued pre-training, in terms of number of sentences, for high-resourced languages (English, Spanish) we use as replay data to prevent catastrophic forgetting.

tuning datasets to the Alpaca format.

3 Approach

To adapt LLMs for the task of translating indigenous American languages, we follow the 2-stage training paradigm proposed in related work (Xu et al., 2024; Alves et al., 2024) and explore its effectiveness for low-resource languages.

3.1 Stage 1: Continued Pre-training with LoRA

In order to "teach" our LLMs the indigenous American languages, we first fine-tuned LLMs with monolingual data for each of these languages. Given these low-resource languages are out-of-distribution from the original pre-training data, we also included replay data from two high-resource languages (English and Spanish) to prevent catastrophic forgetting (Ibrahim et al., 2024). For each American language, given that there were often several (distinctive) dialects, we found that the easiest setting, i.e., to concatenate all of them together, performed very similarly to more careful dialect separation techniques. Inspired by Nguyen et al. (2023), who filtered data from various domains into quality buckets, we segregated our data based on dialects - we assigned the test/dev set dialects to "higher-quality" buckets, and the rest to lower quality. We then tried out a variety of approaches in our preliminary experiments that involved pre-training on various buckets at various stages, but none of these settings performed significantly better[4] than our earlier baseline that concatenated all dialects. Our conclusion here was that these LLMs are only just beginning to learn to model these very low-resourced languages, and cannot separate between dialects at this stage.

For efficiency reasons, we opted for low-rank (LoRA) adaptation (Hu et al., 2021), rather than full-fine tuning. We attached rank 8 LoRA adapters to query and value matrices, following Hu et al. (2021), and also fine-tuned input and output (LM head) embeddings – which we empirically observed to yield significant gains in validation performance. We used average cross-entropy loss σ on the official development set as our validation metric, which we computed as the weighted average of average perplexity on high-resource languages (English and Spanish) and that of the indigenous American languages:

$$\sigma = 0.9 \cdot \sigma_{\text{avg}}^{\{En,Es\}} + 0.1 \cdot \sigma_{\text{avg}}^{\{American\}}$$

where $\sigma_{\text{avg}}^{\{En,Es\}}$ and $\sigma_{\text{avg}}^{\{American\}}$ are the average perplexities on English and Spanish, as well as the indigenous American languages respectively.

We explored adaptation of four LLMs: Llama-2 7B (Touvron et al., 2023), MaLA-500 (Lin et al.,

[4]from a validation loss perspective

Source	Files	Characters
Grammar/Education Book	156 (52.2%)	39,971,932 (46.6%)
Scientific Paper	58 (19.4%)	9,880,833 (11.5%)
Dictionary	55 (18.4%)	28,579,012 (33.3%)
Book	16 (5.4%)	3,360,407 (3.9%)
Other	14 (4.7%)	4,009,128 (4.7%)
Total	299	85,801,312

Table 3: Summary statistics of the OCR data grouped by **source**. We exclude whitespaces while counting **characters**. Percentages of the total are displayed in parentheses.

Task(s)	Dataset	Languages	Instruction Count
Human-annotated Prompt Completions	Aya Dataset	{es, pt, en}	16795
Cross-lingual QA	Synthetic	{es} → All	82538
Machine Translation	Official	{es} → All	76511
	NLLB	{en} → {aym, gn}	13276
	FLORES 200	{es, en, pt} → {aym, gn, quy}	18081
	Helsinki'23	{es} → {gn, hch, nhe, quy, shp}	27976
	OPUS	{es, en, pt} → {aym, cni, gn, nhe, quy}	112681

Table 4: Datasets used for instruction tuning. Languages are denoted by their ISO 639 codes.

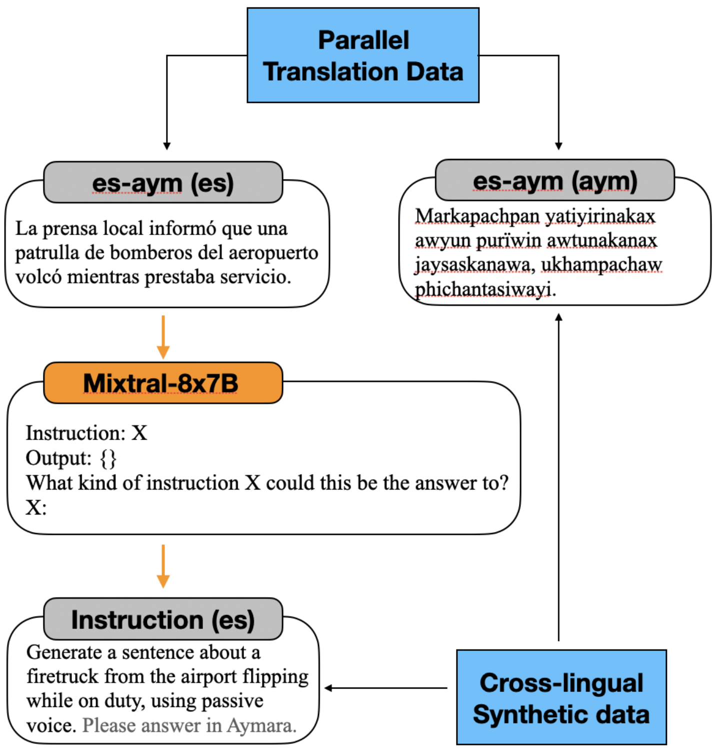


Figure 1: Illustration of our designed process of generating cross-lingual synthetic instruction data.

2024), Mistral 7B (Jiang et al., 2023) and Mistral 7B v0.2[5] for this task. We chose Llama-2 and Mistral since they are the most widely used general-purpose models while MaLA-500, which is the Llama-2 model scaled to 500 languages using LoRA adapters, could potentially enable better cross-lingual transfer.

To examine in greater detail the role of parallel data for continued pre-training under low-resource settings, we trained primarily 2 sets of models, dubbed `v1` and `v2`. `v1` used a concatenation of all available monolingual data[6], while the `v2` models integrated not only monolingual data from `v1`, but also the parallel corpora. Inspired from related work, we explored 3 techniques of leveraging this parallel data: i) `v2.0`: considering the target side of es-X bitext as additional monolingual data, and using the same for pre-training, ii) `v2.1`: following Alves et al. (2024), concatenating[7] the source and target sentences of a certain percentage of sentences (25%, in our experiments[8]), while the rest is used for its target-side data, and iii) `v2.2`: 'interleaving' concatenated Es-X and X-Es parallel text, closely following Guo et al. (2024), and fine-tuning with the same after pre-training on exclusively monolingual data (i.e. `v1` models in our case). For our best-performing model, Mistral 7B, we found `v2.2` baselines overfit and lead to divergence of validation loss, as a result we discard these models.

Given that validation loss cannot be compared fairly across models with different tokenizers, and may not correlate well with downstream MT performance (Iyer et al., 2023), a key challenge we faced was our inability to reliably estimate downstream MT performance after stage 1 pre-training. We, thus, resorted to instruction-tuning all our top-performing models and directly evaluated downstream MT quality– similar to related works (Xu et al., 2024; Alves et al., 2024).

3.2 Stage 2: Instruction Tuning

For instruction-tuning, we continue fine-tuning the stage 1 LoRA adapters on our curated multi-task dataset (Table 4). We fine-tune both input and output embeddings, along with the LoRA adapters, since we observe that this leads to marginal improvements in MT quality. We show these results in Table 5, along with ablations showing how each dataset contributes to improving our overall average performance.

4 Experiments

4.1 Experimental Settings

Stage 1: We used temperature sampling with $\tau = 80$ to ensure uniform data distribution across the relatively higher-resourced (English, Spanish, Quechua, Aymara, Guarani) and the other lower-resourced languages in this setup – since our objective in this work was to build a multilingual LLM that generalizes well to all the languages in this task. However, given the temperature is quite high, and low-resource languages might thus be oversampled excessively, we used a 'clipping factor' of 10 to ensure oversampling does not exceed 10x the original data size.

We conducted continued pre-training of our models using Hugging Face PEFT (Mangrulkar et al., 2022) with the DeepSpeed ZeRO3 configuration (Rajbhandari et al., 2020) on 2 A100-80GB GPUs. We used LoRA adapters on the query and value matrices of rank 4, alpha 8, and dropout 0.1. We used a batch size of 3 per GPU and 16 gradient accumulation steps. We used a learning rate of 2e-5 and a cosine scheduler. We did not use warm-up since we also provided replay data, and empirically found this to be a better choice for validation performance. We saved and evaluated every 100 steps, with a patience value of 5 for early stopping and average evaluation loss as the validation metric. We pre-trained our models for 1 epoch only, due to the enormous training costs.

Stage 2: For instruction tuning, we used the LLaMa-Factory (Zheng et al., 2024) library – which is an easy-to-use package for instruction tuning, built on top of Hugging Face libraries. We continued to tune the LoRA adapters from Stage 1 for 4 epochs using tf32 floating point precision. We used a learning rate of 1e-4, with a cosine scheduler and warm-up ratio of 3%. We used a batch size of 8 per GPU and 16 gradient accumulation steps.

Decoding: We used LLaMa-Factory for decoding on the test set. We used the following default parameters for sampling: a sampling temperature

[5] `https://models.mistralcdn.com/mistral-7b-v0-2/mistral-7B-v0.2.tar`

[6] except the OCR data, which we were only able to obtain for `v2` pre-training

[7] During concatenation, we prepend the language code `L` before each sentence `X`, like so `[L]: X`. Source and target sentences are then joined with the newline character `\n`.

[8] We observed higher percentages (like 75%) decreased validation perplexity more significantly.

No.	Base Model	Tuned Part	Data	Avg. chrF++
1	Llama-2-7B	LoRA	Parallel	7.09
2	Llama-2-7B	LoRA	Parallel+Aya	8.11
3	Mistral-7B	LoRA	Parallel	9.54
4	Mistral-7B	LoRA	Parallel+Aya	9.85
5	Llama-2-7B-Stage1	LoRA	Parallel+Aya	15.17
6	Llama-2-7B-Stage1	LoRA+Emb	Parallel+Aya	15.20
7	Mistral-7B-Stage1	LoRA+Emb	Parallel+Aya	16.24
8	Mistral-7B-Stage1-v1	LoRA+Emb	Parallel+Aya+Syn	16.81
9	MaLA-7B-Stage1	LoRA+Emb	Parallel+Aya+Syn	17.41
10	Mistral-7B-Stage1-v2	LoRA+Emb	Parallel+Aya+Syn	17.32

Table 5: chrF++ scores on the AmericasNLP24 development set using greedy decoding.

of 0.95, top-p sampling with p=0.7 and top-k sampling with k=50. We used beam search with a beam size of 10, repetition and length penalty of 1.0. We used a batch size of 16 and set the maximum number of new tokens for generation to 512.

4.2 Instruction Tuning Experiments

We report our empirical experiment results in Table 5 and introduce our main findings below.

Continued pre-training is crucial. As evident from the instruction-tuning experiments performed on two raw LLMs, i.e. Llama-2-7B & Mistral-7B, and their corresponding stage 1 variants (Llama-2-7B-Stage1 & Mistral-7B-Stage1), we can see that the pre-trained stage 1 models outperform raw instruction-tuned models by a large margin – indicating that LLMs benefit significantly from in-domain monolingual data, even if it is scarce compared to usual high-resourced setups.

However, these gains can potentially suffer from limited returns over time. For the Stage 1 `v2.0` models, which have been trained on 2.5M sentences (78M tokens) more, and obtained a gain in stage 1 validation loss of almost -1.0 point, the corresponding gains in downstream performance (chrF++) was not as significant. Further research is required to verify and analyse the findings from these preliminary experiments.

The general purpose Aya instruction dataset boosts MT performance. This was a surprising finding that showed that even though: a) the language of the generation is not an American indigenous language, and b) the task is not Machine Translation, general-purpose instruction data do not focus on the translation task - we still found significant gains in MT performance. This is likely because this data helps the LLM to reason and follow instructions better.

Adding cross-lingual synthetic instruction data also helps Another interesting exploration in our work is the usage of cross-lingual synthetic instruction data (Section 2.2). While we observe that the quality of the synthetic is not perfect and contains some degree of noise, it does improve the system's translation quality on average. Preliminary experiments also suggested that substituting this with higher quality (but less quantity) data end up performing worse, suggesting that LLMs likely do not know how to generate in these low-resource languages and more data, even if synthetic, can help.

Fine-tuned Mistral usually outperforms Llama-2 Mistral 7B, which has been shown to consistently outperform Llama 13B (Jiang et al., 2023), seems to be more effective in low-resource settings as well. It consistently beats the latter by significant margins. Hence, we choose Mistral as our primary LLM and decide to improve on the same for our final models.

4.3 Checkpoint Averaging

Inspired by (Gao et al., 2022), we use a straightforward low computational approach to boost the performance of our instruction-tuned LLMs. We selected the last 4 model checkpoints from the same run and averaged the model (LoRA) parameters to obtain a better model. Checkpoint averaging is relatively cheaper and does not require storing and querying multiple models at test time. Additionally, we explore all 10 combinations of the last 4 model checkpoints, combining them in triplets and pairs. However, the most significant improvement was observed when averaging the last 4 models checkpoints.

We perform decoding using default parameters of LLaMa-Factory— a sampling temperature of 0.95, top-p and top-k sampling with p=0.7 and

#	Checkpoint	Avg. chrF++ score per model		
		Mistral-7B-v1	MaLA-500	Mistral-7B-v2
(a)	Final checkpoint (step=8151)	19.05	19.18	19.34
(b)	Checkpoint 8000	19.42	19.20	19.16
(c)	Checkpoint 7500	19.18	19.34	18.82
(d)	Checkpoint 7000	19.27	19.08	19.14
(e)	AVG(a,b,c,d)	**20.29**	**19.94**	**20.07**

Table 6: Checkpoint averaging with different models on AmericasNLP development set using default generation parameters of LLaMa-Factory.

k=50 respectively, beam size 1, length and repetition penalty of 1.0 and maximum number of new tokens for generation 512. In Table 6, it's evident that the model with averaged checkpoints consistently outperforms the others. We believe the reason behind its superior performance is that checkpoint averaging acts as a form of regularization.

During the training process, it is possible for a few layers of the model to start over-fitting after certain steps, leading to a degradation in performance if training continues. However, by averaging later checkpoints with the initial ones from earlier in the training process, the effects of over-fitting can be mitigated. This combination helps to regularize the model, preventing it from over-fitting to the training data while still leveraging the useful information learned during the later stages of training.

For future work, we will explore two approaches: a) combining last k checkpoints instead of last 4 during model averaging. b) Weighted averaging of checkpoints, where checkpoints with better performance on the development set receive higher weights. Our hypothesis is that these methods could improve model performance over the current unweighted averaging of the last 4 checkpoints.

4.4 Final Test Set Results

The final systems we submit to the shared task are, therefore (all model IDs are from Table 6 and are open-sourced at `https://tinyurl.com/edi-amnlp24`):

- System 1: Checkpoint `e` i.e. average of checkpoints `a`, `b`, `c` and `d`, for `Mistral-7B-v1`
- System 2: Checkpoint `e` i.e. average of checkpoints `a`, `b`, `c` and `d`, for `MALA-7B-stage2`
- System 3: Average of checkpoints `a`, `c` and `d` for `Mistral-7b-stage2-v2`

For final inference, we use a beam size of 10 expecting a performance boost. Other decoding parameters remained the same. We show our final results on the AmericasNLP 2024 test sets in Table 7. We observe that while our models do not outperform the best systems, the gap is relatively lower for lower resourced languages like Huichol, Nahuatl and Otomi. While this does align with our stated goal of building a general purpose LLM for the languages in this task, as part of future research, we shall explore how we can model better across the other pairs too and increase our competitiveness.

Ethical Considerations

None of the authors of this paper speak any indigenous American languages in this shared task. We rely on the language-labelled datasets suggested by the task organizers and from other reputable sources. We actively sought data manual inspection using Google Translate.

Acknowledgments

This work has received funding from UK Research and Innovation under the UK government's Horizon Europe funding guarantee [grant numbers 10039436 and 10052546].

The computations described in this research were performed using the Baskerville Tier 2 HPC service (https://www.baskerville.ac.uk/). Baskerville was funded by the EPSRC and UKRI through the World Class Labs scheme (EP/T022221/1) and the Digital Research Infrastructure programme (EP/W032244/1) and is operated by Advanced Research Computing at the University of Birmingham.

Language	Metrics	Best system 1	Best system 2	UEdin Submission 1	UEdin Submission 2	UEdin Submission 3
aym	BLEU	3.49	3.23	1.14	1.06	1.13
	chrF++	30.97	29.39	21.77	21.37	21.89
bzd	BLEU	4.84	4.56	2.21	1.89	1.75
	chrF++	23.47	23.41	16.54	16.32	15.56
cni	BLEU	2.41	3.49	0.41	0.37	0.43
	chrF++	23.20	22.98	14.82	13.68	14.50
ctp	BLEU	13.44	4.65	3.35	4.30	3.38
	chrF++	37.38	23.64	17.66	20.70	17.57
gn	BLEU	12.04	11.28	3.38	1.78	3.21
	chrF++	38.93	37.64	29.20	24.61	29.13
hch	BLEU	10.08	9.62	9.87	7.03	9.60
	chrF++	27.64	26.46	24.41	22.03	24.37
nah	BLEU	2.30	1.09	0.48	0.37	0.44
	chrF++	22.87	21.71	18.12	17.21	18.98
oto	BLEU	1.42	1.55	0.43	0.21	0.44
	chrF++	12.98	12.63	8.91	7.81	9.19
quy	BLEU	4.85	4.83	1.32	0.94	1.31
	chrF++	38.21	38.19	25.23	22.77	25.04
shp	BLEU	4.45	4.14	1.34	1.56	1.55
	chrF++	29.37	27.04	22.04	22.43	22.86
tar	BLEU	0.92	1.01	0.11	0.11	0.15
	chrF++	17.03	15.42	9.65	9.49	9.48

Table 7: AmericasNLP 2024 test set results. We show the performances of the top 2 best systems from each language, as well as each of the 3 systems we submit. Languages are denoted by their ISO 639 codes.

References

Duarte M Alves, José Pombal, Nuno M Guerreiro, Pedro H Martins, João Alves, Amin Farajian, Ben Peters, Ricardo Rei, Patrick Fernandes, Sweta Agrawal, et al. 2024. Tower: An open multilingual large language model for translation-related tasks. *arXiv preprint arXiv:2402.17733*.

Giusepppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

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, Toronto, Canada. Association for Computational Linguistics.

Yingbo Gao, Christian Herold, Zijian Yang, and Hermann Ney. 2022. Revisiting checkpoint averaging for neural machine translation. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 188–196, Online only. Association for Computational Linguistics.

Jiaxin Guo, Hao Yang, Zongyao Li, Daimeng Wei, Hengchao Shang, and Xiaoyu Chen. 2024. A novel paradigm boosting translation capabilities of large language models. *arXiv preprint arXiv:2403.11430*.

Amr Hendy, Mohamed Abdelrehim, Amr Sharaf, Vikas Raunak, Mohamed Gabr, Hitokazu Matsushita, Young Jin Kim, Mohamed Afify, and Hany Hassan Awadalla. 2023. How good are gpt models at machine translation? a comprehensive evaluation. *arXiv preprint arXiv:2302.09210*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Adam Ibrahim, Benjamin Thérien, Kshitij Gupta, Mats L Richter, Quentin Anthony, Timothée Lesort, Eugene Belilovsky, and Irina Rish. 2024. Simple and scalable strategies to continually pre-train large language models. *arXiv preprint arXiv:2403.08763*.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Vivek Iyer, Pinzhen Chen, and Alexandra Birch. 2023. Towards effective disambiguation for machine translation with large language models. *arXiv preprint arXiv:2309.11668*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, Lélio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Théophile Gervet, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2024. Mixtral of experts. *arXiv preprint arXiv:2401.04088*.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2024. Madlad-400: A multilingual and document-level large audited dataset. *Advances in Neural Information Processing Systems*, 36.

Abdullatif Köksal, Timo Schick, Anna Korhonen, and Hinrich Schütze. 2024. Longform: Effective instruction tuning with reverse instructions. *Preprint*, arXiv:2304.08460.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André FT Martins, and Hinrich Schütze. 2024. Mala-500: Massive language adaptation of large language models. *arXiv preprint arXiv:2401.13303*.

Hongyuan Lu, Haoyang Huang, Dongdong Zhang, Haoran Yang, Wai Lam, and Furu Wei. 2023. Chain-of-dictionary prompting elicits translation in large language models. *arXiv preprint arXiv:2305.06575*.

Sourab Mangrulkar, Sylvain Gugger, Lysandre Debut, Younes Belkada, Sayak Paul, and Benjamin Bossan. 2022. Peft: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft.

Xuan-Phi Nguyen, Wenxuan Zhang, Xin Li, Mahani Aljunied, Qingyu Tan, Liying Cheng, Guanzheng Chen, Yue Deng, Sen Yang, Chaoqun Liu, et al. 2023. Seallms–large language models for southeast asia. *arXiv preprint arXiv:2312.00738*.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. 2020. Zero: Memory optimizations toward training trillion parameter models. In *SC20:*

International Conference for High Performance Computing, Networking, Storage and Analysis, pages 1–16. IEEE.

Nathaniel R Robinson, Perez Ogayo, David R Mortensen, and Graham Neubig. 2023. Chatgpt mt: Competitive for high-(but not low-) resource languages. *arXiv preprint arXiv:2309.07423*.

Shivalika Singh, Freddie Vargus, Daniel Dsouza, Börje F Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura OMahony, et al. 2024. Aya dataset: An open-access collection for multilingual instruction tuning. *arXiv preprint arXiv:2402.06619*.

Jörg Tiedemann. 2009. News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing*, volume 5, pages 237–248.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*.

Haoran Xu, Young Jin Kim, Amr Sharaf, and Hany Hassan Awadalla. 2024. A paradigm shift in machine translation: Boosting translation performance of large language models. In *The Twelfth International Conference on Learning Representations*.

Kexun Zhang, Yee Man Choi, Zhenqiao Song, Taiqi He, William Yang Wang, and Lei Li. 2024. Hire a linguist!: Learning endangered languages with in-context linguistic descriptions. *arXiv preprint arXiv:2402.18025*.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. *arXiv preprint arXiv:2403.13372*.

Appendix

Combinations of languages	**Source type**	**Files**	**Characters**
Aymara	Mono	8	682,766
English/Asháninka	Mixed	2	1,605,073
English/Aymara	Mixed	9	2,945,037
English/Chatino	Mixed	8	2,708,631
English/Guaraní	Mixed	12	2,773,253
English/Hñähñu	Mixed	5	2,181,855
English/Nahuatl	Mixed	24	8,950,757
English/Quechua	Mixed	7	1,429,763
English/Spanish/Aymara	Mixed	1	246,850
English/Spanish/Quechua	Mixed	3	953,120
English/Spanish/Rarámuri	Mixed	2	1,250,289
English/Wixarika	Mixed	1	544,090
French/Aymara	Mixed	1	52,022
French/Bribri	Mixed	1	1,099,198
French/Hñähñu	Mixed	1	111,296
French/Quechua	Mixed	1	194,163
French/Rarámuri	Mixed	1	23,418
German/Guaraní	Mixed	1	178,220
German/Quechua	Mixed	2	1,361,053
Nahuatl	Mono	2	224,394
Quechua	Mono	10	492,504
Russian/Guaraní	Mixed	1	51,939
Russian/Nahuatl	Mixed	1	75,205
Russian/Quechua	Mixed	2	193,794
Spanish/Asháninka	Mixed	9	2,133,942
Spanish/Asháninka/Quechua	Mixed	1	65,046
Spanish/Aymara	Mixed	45	9,546,160
Spanish/Aymara/Nahuatl/Quechua	Mixed	1	208,828
Spanish/Bribri	Mixed	4	801,911
Spanish/Chatino	Mixed	3	1,162,349
Spanish/Guaraní	Mixed	20	8,101,890
Spanish/Hñähñu	Mixed	10	3,059,227
Spanish/Nahuatl	Mixed	17	6,171,836
Spanish/Quechua	Mixed	67	19,830,467
Spanish/Rarámuri	Mixed	6	1,311,759
Spanish/Shipibo-Konibo	Mixed	4	461,930
Spanish/Wixarika	Mixed	5	1,478,789
Wixarika	Mono	1	1,138,488

Table 8: Summary statistics of the OCR data, grouped by **Combinations of languages**. **Characters** counted without whitespaces.

Source	Low-resource languages	Source type	Files	Characters
Book	Nahuatl	Mono	1	195,009
	Quechua	Mixed	8	1,727,827
		Mono	6	299,083
	Wixarika	Mono	1	1,138,488
Dictionary	Asháninka	Mixed	3	783,665
	Aymara	Mixed	15	4,792,382
	Chatino	Mixed	2	1,012,744
	Guaraní	Mixed	8	5,509,379
	Nahuatl	Mixed	5	3,424,235
	Quechua	Mixed	19	12,354,240
	Rarámuri	Mixed	3	702,367
Grammar/Education Book	Asháninka	Mixed	5	2,279,964
	Aymara	Mixed	25	6,212,691
		Mono	8	682,766
	Bribri	Mixed	3	714,131
	Chatino	Mixed	1	149,605
	Guaraní	Mixed	16	4,585,622
	Hñähñu	Mixed	13	4,441,870
	Nahuatl	Mixed	24	9,877,127
		Mono	1	29,385
	Quechua	Mixed	47	9,072,258
		Mono	5	247,344
	Rarámuri	Mixed	3	1,146,458
	Shipibo-Konibo	Mixed	3	314,443
	Wixarika	Mixed	2	218,268
Other	Aymara	Mixed	4	1,136,545
	Hñähñu	Mixed	1	95,944
	Nahuatl	Mixed	5	1,461,840
	Rarámuri	Mixed	1	54,278
	Wixarika	Mixed	3	1,260,521
Scientific Paper	Asháninka	Mixed	3	675,386
	Asháninka/Quechua	Mixed	1	65,046
	Aymara	Mixed	12	648,451
	Aymara/Nahuatl/Quechua	Mixed	1	208,828
	Bribri	Mixed	2	1,186,978
	Chatino	Mixed	8	2,708,631
	Guaraní	Mixed	10	1,010,301
	Hñähñu	Mixed	2	814,564
	Nahuatl	Mixed	8	434,596
	Quechua	Mixed	7	754,112
	Rarámuri	Mixed	2	682,363
	Shipibo-Konibo	Mixed	1	147,487
	Wixarika	Mixed	1	544,090

Table 9: Summary statistics of the OCR data, grouped by **Source** and **Low-resource languages**. **Characters** counted without whitespaces.

The role of morphosyntactic similarity in generating related sentences

Michael Hammond
Dept. of Linguistics
U. of Arizona
Tucson, Arizona 85721, USA
hammond@arizona.edu

Abstract

In this paper we describe our work on Task 2: AmericasNLP 2024 Shared Task on the Creation of Educational Materials for Indigenous Languages. We tried three approaches, but only the third approach yielded improvement over the baseline system. The first system was a fairly generic transformer model. The second system was our own implementation of the edit tree approach from the baseline system. Our final attempt was a version of the baseline system where if no transformation succeeded, we applied transformations from similar morphosyntactic relations. We describe all three here, but, in the end, we only submitted the third system.

1 Introduction

The nature of the task was to transform one sentence into another in three languages based on a specification of the morphosyntactic differences between the input and output sentences (Chiruzzo et al., 2024). The three languages are Bribri, Guarani, and Maya. We give sample data in Table 1.

Glosses or translations were not provided. In addition, we do not know what morphosyntactic features might be appropriate for the input sentence.

We were provided with definitions for the different tags. In the first example in Table 1 we convert the absolutive argument to a plural. In the second example, we convert to an affirmative. Finally, in the third case, we switch one of the arguments to a first person plural subject.

We tried three different approaches: a simple transformer, our own implementation of edit trees in terms of a single regular transduction, and using related morphosyntactic tags when possible. In the following sections, we describe these three attempts.

2 Transformer

The first model we tried was a simple transformer (Vaswani et al., 2017).[1] We first concatenate the input and morphological tags to serve as input.

These are fed into the encoder which first creates embeddings, applies drop-out, and feeds these to a GRU layer (Cho et al., 2014).

The output and hidden weights are then fed to an attention-based decoder with two layers of GRUs and a simple linear layer. Attention was Bahdanau (Bahdanau et al., 2014).

Batch size varied, but was typically around 32. The dimensions for all hidden layers was either 512 or 1024 for different runs. Dropout for the encoder was set at 0.1. Loss was negative log likelihood and the Adam optimizer was used. We tried a variety of different configurations, but best performance was at 600 epochs with 512 hidden nodes at all layers. See Table 2 for performance with dev data.

The data are extremely limited and this surely impaired performance. Our sense is that simply concatenating the input and morphosyntactic tags was also not the best choice.

3 Edit trees as a single transduction

The baseline system for the task is based on the notion of *edit trees*. The basic idea is to build a tree representation of changes that the input must undergo to be converted to the output (Chrupała, 2008).

Chrupała gives the edit tree in Figure 1 for the Polish word pair *najtrudniejszy* ‘hardest’ and *trudny* ‘hard’. The basic logic is that we identify the largest shared span, in this case characters 3 through 6. To the left of that, we replace *naj* with

[1] All of our code can be obtained at https://github.com/hammondm/americasnlp24task2. Our specific transformer architecture is an adaptation from https://pytorch.org/tutorials/intermediate/seq2seq_translation_tutorial.html.

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 221–223
June 21, 2024

Language	Input	Features	Target
Bribri	Pûs kapë’wa̲	ABSNUM:PL	Pûs kapë’ulur
Guarani	Ore ndorombyai kuri	TYPE:AFF	Ore rombyai kuri
Maya	Janalnajen tu k’íiwikil koonol	PERSON:1_PL	Janalnajo’on tu k’íiwikil koonol

Table 1: Example data

Language	Accuracy	BLEU	ChrF
Bribri	0.00	3.14	12.51
Guarani	0.00	0.29	4.56
Maya	0.67	14.17	40.08

Table 2: Performance with transformer model on dev data

Language	Accuracy	BLEU	ChrF
Bribri	3.30	12.93	39.14
Guarani	0.00	22.19	72.63
Maya	1.34	30.68	71.95

Table 3: Performance with our implementation of edit trees on dev data

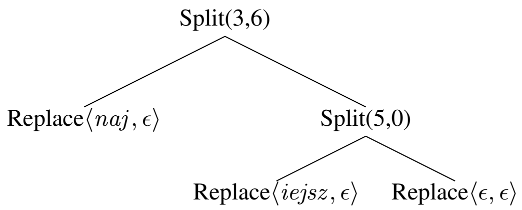


Figure 1: Example edit tree (Chrupała, 2008, p.127)

ϵ. To the right, we repeat the process and identify the longest shared span: *y*. To the left of this, we replace *iejsz* with ϵ. To the right, we do nothing: Replace(ϵ,ϵ).

Formally, Chrupała defines a function lcs from two strings $(\Sigma^* \times \Sigma^*)$, specifically $w_{1\ldots n}$ and $w'_{1\ldots m}$, to four natural numbers $(\mathbb{N} \times \mathbb{N} \times \mathbb{N} \times \mathbb{N})$, (i, j, k, l), representing indices into the strings where the shared string is indexed as $w_{i\ldots n-j} = w'_{k\ldots m-l}$.

There is then a function $split$ which maps from a string and indices to three strings $(\Sigma^* \times \mathbb{N} \times \mathbb{N}) \rightarrow (\Sigma^* \times \Sigma^* \times \Sigma^*)$, taking a string $w_{1\ldots n}$ and indices i and j and returning the triple $(w_{1\ldots i}, w_{i+1\ldots n-j}, w_{n-j+1\ldots n})$.

An edit tree is then either a Replace node with two strings or it is a Split node with two indices and two daughter nodes that are themselves edit trees.

An edit tree is built with the function et which is defined as follows with respect to strings w and w'. If there is no lcs span, then the tree is simply Replace(w, w').

Otherwise it is defined as:

$$\begin{aligned}&\text{Split}(i_w, j_w),\\&et(w_{\textit{prefix}}, w'_{\textit{prefix}}),\\&et(w_{\textit{suffix}}, w'_{\textit{suffix}})\end{aligned} \tag{1}$$

In our implementation, these operations are mapped to a single regular transduction with back-references. For convenience, we use python for this.

First, Split nodes are represented as a list of three elements: the left daughter, the two indices, and the right daughter. Replace nodes are represented as a pair of strings. The tree in Figure 1 would be represented as:

$$[(naj, \epsilon), (3, 6), [(iejsz, \epsilon), (5, 0), (\epsilon, \epsilon)]] \tag{2}$$

We traverse the tree from left to right converting each node into a pair of strings. All string pairs map to themselves. Pairs of indices i, j are translated into maps from `.{1,k}`, where $k = j - i$, to the next available backreference. In the case of the tree in Figure 1, we have the translation `naj.{1,3}iejsz.{1,1}` mapping to `\1\2`. This mapping is executed in our code using the python `re.sub` function.

This approach approximates the baseline system, but does not perform as well. See Table 3.

4 Morphosyntactic similarity

Our final model, and the one we submitted, was an addition to the baseline system.

The baseline system records the edit trees that are associated with specific morphosyntactic tag combinations along with the relative frequency of each tree.

At inference stage, one selects the edit trees associated with the tags for the test item, sorts these by how frequently they’re used in training, and try them one by one starting from the most frequent. If a rule succeeds, the output of that rule is returned.

Language	Accuracy	BLEU	ChrF
Bribri	9.38	17.13	55.07
Guarani	25.81	50.36	79.46
Maya	14.84	22.55	73.18

Table 4: Performance with morphosyntactic tag similarity on test data

If no rule succeeds or the specific tag combination is not in the training data, then the input is returned as the output.

Our adaptation here was to add additional options to the list. If the procedure above produced no distinct output form, we then applied additional rules. These additional rules were generated from the full list of rules, sorted by how similar the tag sequence is to the test item tag sequence. If the above procedure results in no change, we then turn to these rules, going through them one by one. This procedure is terminated the same way as the baseline system: when a rule produces a change, that is the final output and further rules are not considered. If no rule produces a distinct output, the input itself is returned.

This change results in a modest overall improvement in the baseline as seen in Table 4.

5 Conclusion

To summarize, we tried three different techniques: transformer, edit trees as transductions, and exploiting morphosyntactic similarity in selecting edit trees.

As implemented, the transformer performed the worst. Systems built on edit operations seem to perform much better in these character-to-character mapping domains, so this is not really a surprise.

Translating edit trees into regular transductions did not reach the level of the baseline, but is not an unreasonable approach to pursue further. Edit operations are clearly useful. The question is what is the scope of those operations. Are they separate operations as in edit trees as originally developed or can some amalgam of those operations be more successful.

Finally, using morphosyntactic similarity is successful and this is thus clearly an approach worth pursuing further.

One very obvious way to go further is to build a model of the morphosyntactic structure of the input. We do not know what the words mean, but perhaps we can get some mileage toward identifying parts of speech from the meanings of the tags. With this in hand, we could exploit that part of speech information in our edit trees.

This last approach is purely speculative, but it seems like a fairly obvious way to go (in hindsight!) given the nature of the task.

Acknowledgments

Thanks to Diane Ohala for useful discussion. All errors are my own.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Luis Chiruzzo, Pavel Denisov, Samuel Canul Yah, Lorena Hau Ucán, Marvin Agüero-Torales, Aldo Alvarez, Silvia Fernandez Sabido, Alejandro Molina Villegas, Abteen Ebrahimi, Robert Pugh, Arturo Oncevay, Shruti Rijhwani, Rolando Coto-Solano, Katharina von der Wense, and Manuel Mager. 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)*. Association for Computational Linguistics.

Kyunghyun Cho, Bart Van Merriënboer, Caglar Gulcehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078*.

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

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.

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

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

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

Abstract

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

1 Introduction

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

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

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

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

2 Related Work

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

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

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

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

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

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

3 Task Description

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

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

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

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

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

Exercise 1. Transform the following sentence to first-person plural:

```
J-jaan _____ najil
a) béet u
b) o'ob tu
c) o tin
d) o'ob janal
```

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

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

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

The task was evaluated in terms of exact accuracy (fraction of times the system output matched the expected output), and also two classic metrics for generative tasks: BLEU (Papineni et al., 2002) and chrF (Popović, 2015). The main metric for the task was exact accuracy.

4 Dataset

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

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

Table 1: Size of the dataset.

4.1 Bribri

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

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

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

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

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

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

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

4.2 Maya

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

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

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

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

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

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

[1]https://duolingo.com/

2016).

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

4.3 Guarani

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

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

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

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

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

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

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

5 Approaches and Results

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

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

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

5.1 Baseline

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

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

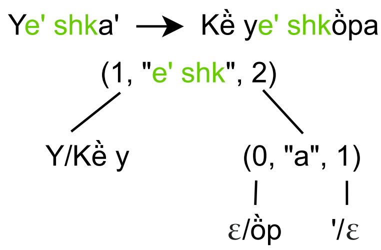


Figure 1: Edit tree for the training sample `Bribri0315`, from *Ye' shka'* 'I walked' to *Kë̀ ye' shkö̀pa* 'I won't walk'. The root node is a match node with the match "e' shk", prefix length 1 and suffix length 2. Its prefix node is a substitution node that replaces "Y" with "Kë̀ y". Its suffix node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

5.2 JAJ (/dʒæz/)

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

5.3 Meenzer Team

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

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

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

5.4 Giving it a Shot

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

5.5 LECS Lab

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

The team develops a variation of the *lemma copying* technique (Liu and Hulden, 2022; Yang et al., 2022), which they name *sentence copying*. The idea is to create additional training samples by copying same sentence as both source and target with an empty change field. All LSTM systems except of system 9 use the external sentence copying for data augmentation, where the copied sentences are taken from external datasets, namely transcriptions from the Yucatec Maya DoReCo dataset (Skopeteas, 2022) for Maya, Guarani portion of the CC-100 dataset (Conneau et al., 2020) for Guarani, and Bribri portion of the AmericasNLP2024 Shared Task 1 data for Bribri. System 3 additionally performs the sentence copying with all sentences from the training data.

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

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

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

System 9 works with byte-pair encoding sub-

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

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

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

5.6 UF_NLP

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

5.7 Arizona Linguistics

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

5.8 Task Results

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

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

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

6 Conclusions

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

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

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

Acknowledgements

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

References

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

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

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

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

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

Minh Duc Bui and Katharina von der Wense. 2024. Jgu mainz's submission to the americasnlp 2024 shared task on the creation of educational materials for indigenous languages. In *Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*. Association for Computational Linguistics.

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

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

Stavros Skopeteas. 2022. Yucatec Maya DoReCo dataset.

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

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

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

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

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

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

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

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

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

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

Findings of the AmericasNLP 2024 Shared Task on Machine Translation into Indigenous Languages

Abteen Ebrahimi[◇*] **Ona de Gibert**[♠*] **Raúl Vázquez**[♠*] **Rolando Coto-Solano**[Ω*]
Pavel Denisov[▽] **Robert Pugh**[ψ] **Manuel Mager**[♣] **Arturo Oncevay**[♡]
Luis Chiruzzo[Φ] **Katharina von der Wense**[◇♯] **Shruti Rijhwani**[~]
[◇]University of Colorado Boulder [♠]University of Helsinki [Ω]Dartmouth College
[▽]University of Stuttgart [ψ]Indiana University [♣]Amazon AWS AI
[♡]Pontificia Universidad Católica del Perú [Φ]Universidad de la República, Uruguay
[♯]Johannes Gutenberg University Mainz [~]Google Deepmind
abteen.ebrahimi@colorado.edu

Abstract

This paper presents the findings of the third iteration of the AmericasNLP Shared Task on Machine Translation. This year's competition features eleven Indigenous languages found across North, Central, and South America. A total of six teams participate with a total of 157 submissions across all languages and models. Two baselines – the Sheffield and Helsinki systems from 2023 – are provided and represent hard-to-beat starting points for the competition. In addition to the baselines, teams are given access to a new repository of training data which consists of data collected by teams in prior shared tasks. Using ChrF++ as the main competition metric, we see improvements over the baseline for 4 languages: Chatino, Guarani, Quechua, and Rarámuri, with performance increases over the best baseline of 4.2 ChrF++. In this work, we present a summary of the submitted systems, results, and a human evaluation of system outputs for Bribri, which consists of both (1) a rating of meaning and fluency and (2) a qualitative error analysis of outputs from the best submitted system.

1 Introduction

Though the field of natural language processing (NLP) has seen a steep increase in interest and impressive performance improvements over the past decade, a large performance gap still remains between a handful of so-called "high-resource," mostly colonial, languages and the remaining majority of the world's languages (Blasi et al., 2022). The Indigenous languages of the Americas exemplify this reality, representing nearly 15% of the world's linguistic diversity (Eberhard et al., 2024) and yet, until recently, receiving little attention in NLP research.

Language	Family	Train	Extra	Syn.	Dev.
Asháninka (cni)	Arawak	3,883	-	13,195	883
Aymara (aym)	Aymaran	6,531	24,331	16,750	996
Bribri (bzd)	Chibchan	7,508	-	-	996
Chatino (ctp)	Oto-Manguean	357	2,246	-	499
Guarani (gn)	Tupi-Guarani	26,032	42,186	40,516	995
Nahuatl (nah)	Uto-Aztecan	16,145	2,493	9,222	672
Otomí (oto)	Oto-Manguean	4,889	9,012	-	599
Quechua (quy)	Quechuan	125,008	6,469	60,399	996
Rarámuri (tar)	Uto-Aztecan	14,720	2,254	-	995
Shipibo-Konibo (shp)	Panoan	14,592	16,721	23,595	996
Wixarika (hch)	Uto-Aztecan	8,966	2,653	511	994

Table 1: Languages of the shared task, their ISO codes, language families, and dataset statistics.

The AmericasNLP Shared Task on Machine Translation (MT), now in its third iteration (2021, 2023, and 2024), is focused on pushing the performance of MT on this group of languages through two main avenues: by applying modeling and architectural advancements, and through the creation of new linguistic resources which support the training and evaluation of these systems.

This year's shared task continues to focus on the eleven Indigenous languages from the last competition. While this year's competition does not feature new data for evaluation, competitors are given access to a new repository of training data which extends the original set of parallel examples with additional data collected by teams in prior years. This repository represents the first step in creating a new living source of data which can grow through contributions from teams participating in future iterations of the shared task. This year's competition also features two baselines: the University of Sheffield (Gow-Smith and Villegas, 2023) and University of Helsinki (De Gibert et al., 2023) systems which each achieved the best performance for a subset of languages in 2023 (Ebrahimi et al., 2023). These baselines are strong and hard-to-beat; across 157 submissions from 6 different teams, we see improvements for only 4 of the 11

[*] Equal contribution.

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 236–246
June 21, 2024

languages: Chatino, Guarani, Quechua, and Rarámuri. As two of these four languages are the relatively highest-resourced, this finding may indicate that we are approaching a plateau in performance gains achievable purely through modeling and architectural approaches; therefore, a focus on collecting additional training data may yield the most future improvements.

The paper is structured as follows. In Section 2, we provide a brief overview of the data and languages provided by the organizers at the beginning of the competition. Section 3 contains summary descriptions of the approaches used by each team. Section 4 discusses the results of the competition. In Section 5, we conduct a human evaluation of system outputs for Bribri. In the first part of this evaluation, we follow the prior shared tasks in quantitatively rating a sample of outputs on two axes: meaning and fluency. For the second part, we conduct a qualitative error analysis, comparing baseline systems to the best submitted system. In Section 6, we conclude with a brief discussion of future directions in improving MT quality for Indigenous languages of the Americas.

2 Data and Languages

The shared task features 11 Indigenous languages of the Americas. The language direction we are interested in is from Spanish into the low-resource language.

We use the AmericasNLP 2021 data for development and evaluation. It consists of a multi-way parallel dataset of the Spanish XNLI test set into 10 languages of the Americas (Asháninka, Aymara, Bribri, Guarani, Nahuatl, Otomí, Quechua, Rarámuri, Shipibo-Konibo, and Wixarika). The task also includes Chatino, for which the data comes from Mexican court proceedings. Chatino was introduced as a surprise language in last year's edition (Ebrahimi et al., 2023). For an in-depth review of development and evaluation data, please refer to Ebrahimi et al. (2022) and Mager et al. (2021).

For training data, besides the data used in previous editions, this year we include the data collected by De Gibert et al. (2023) as part of their Helsinki-NLP submission. This consists of `extra` data, made up of different sources listed in their system description paper, as well as `syn`, which refers to synthetic data obtained through backtranslation. Table 1 provides an overview of the languages, their linguistic families, and the total number of parallel sentences with Spanish. While there is no new data for Bribri, this year's data sizes increased considerably for Shipibo-Konibo, Aymara, Quechua and Guarani, with over 40k added sentences (although the majority comes from backtranslations). The test data for all languages consists of 1,003 sentences, except for Chatino, which contains 1,000 sentences.

We publicly release the training and development data in our Github repostitory.[1]

3 Metrics

For evaluation, we use the automatic metric ChrF++ (Popović, 2017) as implemented in SACREBLEU (Post, 2018). It is an overlap-based metric at the character-level, which is adequate for our task since most languages are morphologically rich.

While teams are not required to submit a system for all languages, the final score for each submission (ChrF++ column in Table 3) is calculated by taking an average over all eleven languages; if there is no model output for a given language, the score is taken as 0.

4 Baselines and Submitted Systems

In this section, we describe the 2024 baseline systems and each team's approach. We present a summary of all approaches in Table 2.

4.1 Baselines

This year, we consider two different baselines, based on the strongest submissions of the previous edition of our shared task, shown to be competitive among each other. The overall winning team in the previous edition was Sheffield (Gow-Smith and Villegas, 2023). They exploited the knowledge from different distilled versions of NLLB (Costa-jussà et al., 2022), a large pretrained model. We use their Submission 3, which chooses a single checkpoint with best average ChrF across all languages.

We also include Helsinki-NLP's Submission 6 (De Gibert et al., 2023), given that it outperforms the previous system on several language pairs. Their winning model is a multilingual one-to-many system, pretrained on Spanish–English data.

4.2 Submitted Systems

BSC The BSC team submitted systems for two languages: Quechua and Guarani, and followed the

[1] https://github.com/AmericasNLP/americasnlp2024/tree/master/ST1_MachineTranslation

Team	Models	Data	Overview
BSC Gilabert et al. (2024)	• NLLB-200 (1.3B)	• Length-based data filtering • Train set deduplication • Embedding-based sentence similarity	• Multilingual and bilingual fine-tuning of NLLB-200 • Low-Rank Adaptation (LoRA; 15% trainable params.) and full finetuning achieves
NordicAlps Attieh et al. (2024)	• From-scratch transformer encoder–decoder models	Various tokenizations: • Byte-level BPE • SentencePiece • Redundancy-driven tokenization	2 stage training: • First focus on Spanish-English data • Second, reduce Spanish-English to 50% with the other 50% sampled to equal amounts from the 11 TGT languages
DC_DMV DeGenaro and Lupicki (2024)	• NLLB-200 (600M, distil.) • State-space model (Mamba) from scratch	• Partition data into three stages, with deduplication	• Fully fine-tune a distilled NLLB-200 model using two data stages • Train a 3-layered Mamba network from scratch followed by a language model head using three data stages
Edinburgh Iyer et al. (2024)	• Llama-2 (7B) • Mistral (7B) • MaLA-500	• Collect additional data through OCR • Grammar and Education books, Scientific Papers, Dictionaries, and Books as sources	• Fine-tune LLama-2, Mistral and MaLA-500 models using a 2-stage training • LoRA fine-tuning with monolingual data first, then continue with instruction tuning • Regularize outputs using model averaging of the 4 last checkpoints

Table 2: Summary overview of each team's approach.

prior year's baseline approach of finetuning NLLB-200. In addition to the data provided by the organizers, the team collected new data from multiple sources, including the Monolingual-Quechua-IIC dataset (Zevallos et al., 2022), Flores-200 (Team et al., 2022), and other online datasets.[2] After collection, the data is cleaned in a multi-step process to remove duplicates and filter sentences. In the first step, sentences with more than 150 tokens and sentence pairs with a length ratio greater than 3 are removed. Next, various libraries are used to further clean the data, including Bifixer (Ramırez-Sanchez and Zaragoza-Bernabeu, 2020) and NLPDedup.[3] Finally, an embedding-based approach is used to calculate similarities between the source and targets side of a sentence pair; similarity scores are used with various thresholds to determine the final training examples.

NLLB is finetuned separately for each target language, and parallel sentences between each target and English, Portuguese, and Spanish are used. Two model sizes are considered: the 3.3B and 1.3B parameter version. Interestingly, the larger model only shows improvements for Quechua while performance decreases for Guarani; this relationship depends on the finetuning method used. Increasing the similarity score threshold offers better performance up to a point, after which performance begins to decrease, likely due to the greatly reduced amount of available data for finetuning. Overall, the best performance is found by using NLLB 1.3B with full finetuning for Guarani, improving over the prior best model by 1.91 ChrF++. For Quechua, NLLB 1.3B + LoRA (Hu et al., 2021) finetuning improves over the prior best score by 4.2 ChrF++. For these two languages, both systems achieved the highest performance across all submitted systems in this year's shared task.

NordicAlps The NordicAlps team submitted systems for all eleven languages of the shared task, building on the Helsinki system (De Gibert et al., 2023) from the 2023 shared task. The final models are one-to-many, trained to output translations in any of the competition languages as well as English. Target language tags are used to specify the output language. Data used is similar to the prior year's system, but this year's submission does not include Bible data. Preprocessing steps include whitespace normalization, Unicode normalization, and punctuation tokenization; these steps were implemented using the Moses tokenizer as well as through handwritten rules. The models do not make use of additional meta-data tags describing the language variant or quality on the input side. Of the three submitted systems, the main difference lies in the tokenization: a traditional byte-level BPE tokenization, SentencePiece tokenization, and `BPE-MR` tokenization, which consists of a BPE subword tokenizer trained using only 300 merges. `BPE-MR` tokenization is motivated by prior work on text compression through tokenization, and the finding that monolingual text can be compressed optimally using a small number of merge operations. Model training is carried out in stages, with the first stage covering a high-resource language pair (Spanish–English), and the second stage introducing more Indigenous language pairs (up to 50% of the examples used for training). Of the three sub-

[2] https://huggingface.co/datasets/somosnlp-hackathon-2022/spanish-to-quechua

[3] https://github.com/saattrupdan/NLPDedup

Rank	Team	Ver.	Count	Tot. BLEU	Tot. ChrF	Tot. ChrF++	Avg. BLEU	Avg. ChrF	Avg. ChrF++	BLEU	ChrF	**ChrF++**
1	NordicsAlps	1	11	55.48	321.69	287.60	5.04	29.24	26.15	5.04	29.24	26.15
2	DC_DMV	4	11	40.14	288.67	256.51	3.65	26.24	23.32	3.65	26.24	23.32
3	DC_DMV	3	11	39.63	287.64	255.49	3.60	26.15	23.23	3.60	26.15	23.23
4	DC_DMV	1	11	38.97	284.66	252.38	3.54	25.88	22.94	3.54	25.88	22.94
5	DC_DMV	5	11	37.84	284.26	252.21	3.44	25.84	22.93	3.44	25.84	22.93
6	DC_DMV	6	11	37.95	284.04	251.77	3.45	25.82	22.89	3.45	25.82	22.89
7	DC_DMV	2	11	34.15	272.59	243.83	3.10	24.78	22.17	3.10	24.78	22.17
8	NordicsAlps	2	11	27.28	265.46	232.41	2.48	24.13	21.13	2.48	24.13	21.13
9	UEdin	3	11	23.41	236.36	208.56	2.13	21.49	18.96	2.13	21.49	18.96
10	UEdin	1	11	24.04	235.42	208.34	2.19	21.40	18.94	2.19	21.40	18.94
11	UEdin	2	11	19.62	224.50	198.44	1.78	20.41	18.04	1.78	20.41	18.04
12	NordicsAlps	3	11	18.03	195.03	171.81	1.64	17.73	15.62	1.64	17.73	15.62
13	Z-AGI_Labs	1	4	8.35	103.03	87.32	2.09	25.76	21.83	0.76	9.37	7.94
14	DC_DMV	9	11	2.08	96.67	83.69	0.19	8.79	7.61	0.19	8.79	7.61
15	BSC	3	2	16.48	85.68	76.95	8.24	42.84	38.47	1.50	7.79	7.00
16	BSC	4	2	16.10	84.56	75.83	8.05	42.28	37.91	1.46	7.69	6.89
17	BSC	2	2	16.09	84.56	75.73	8.04	42.28	37.86	1.46	7.69	6.88
18	BSC	1	2	15.89	84.42	75.63	7.95	42.21	37.82	1.44	7.67	6.88
19	BSC	5	1	11.53	38.37	35.73	11.53	38.37	35.73	1.05	3.49	3.25
20	ND-NAIST	1	1	2.60	38.51	32.88	2.60	38.51	32.88	0.24	3.50	2.99

Table 3: Main ranking of all submitted systems. Count denotes the number of languages a particular system was submitted for, with the Avg.* columns showing the average metric score across submitted systems. The final three columns represent the average over all 11 languages of the shared task, with ChrF++ being used to calculate the overall ranking.

missions, the model using `BPE-MR` tokenization offered the best performance and achieved the best result for 5 of the shared task languages, and 2nd for 2 other languages.

DC_DMV The DC_DMV team submitted a system for each of the eleven languages, and followed two main approaches: finetuning a single version of the distilled 600m version of NLLB-200 for all the languages, as well as using a state-space model based on the Mamba architecture (Gu and Dao, 2023). Similar to the BSC team, duplicate examples are filtered, and the data is split into mutually exclusive stages. Stage 1 contains the largest set of data with over 700k examples, while Stages 2 and 3 have 100k and 200k examples, respectively. For the NLLB approach, the model is fully finetuned using data from the latter two stages, and the various submitted systems following this approach differ in the amount of training done using data from each stage. For the Mamba approach, a model is trained from scratch using all available data. While this approach did not yield strong results, likely due to the lack of pretraining, an NLLB-based submission achieved the best result across all submitted systems for Aymara, Shipibo-Konibo, and Rarámuri, while a different NLLB model achieved the best results for Bribri.

University of Edinburgh The University of Edinburgh participated with three system submissions for each of the eleven languages. These are the best performing systems in a series of experiments where the authors explore finetuning three well-known open-source LLMs: Llama-2 7B (Touvron et al., 2023), Mistral 7B (Jiang et al., 2023) and MaLA-500 (Lin et al., 2024). The finetuning consists of a two-stage training process employing Low-Rank Adaptation (LoRA) (Hu et al., 2021) and instruction tuning. In a nutshell, the first stage consists of finetuning LoRA adapters by continued pretraining on the LLM monolingual data, to adapt the models to specific linguistic features of each of the target languages. This setup includes using diverse data sets such as MADLAD-400 (Kudugunta et al., 2023) and Glot500 (ImaniGooghari et al., 2023). The second stage focuses on instruction tuning where models are finetuned using a combination of human-annotated and synthetic cross-lingual data, which helps improve the models' efficiency in real translation tasks. Furthermore, the authors explore n-last checkpoint averaging, with different beam search, and sampling setups to boost model performance at inference time.

5 Results

The overall ranking for the shared task can be found in Table 3, and the best per-language performance for each team can be found in Table 4. The full results for all submissions and teams can be found in Table 6.

LANG. / TEAM	AYM	BZD	CNI	CTP	GN	HCH	NAH	OTO	QUY	SHP	TAR
Baseline - Helsinki	29.36	23.47	**24.92**	29.84	37.02	**28.67**	22.78	**13.32**	28.81	**30.21**	16.98
Baseline - Sheffield	**31.84**	**25.58**	24.76	37.05	35.76	28.28	**23.28**	12.87	34.01	30.06	16.25
BSC	-	-	-	-	<u>**38.93**</u>	-	-	-	<u>**38.21**</u>	-	-
DC_DMV	<u>30.97</u>	<u>23.47</u>	22.98	16.52	33.31	26.46	21.63	12.63	36.02	<u>29.37</u>	<u>**17.03**</u>
ND-NAIST	-	-	-	-	-	-	-	-	32.88	-	-
NordicsAlps	29.39	23.32	<u>23.20</u>	<u>**37.38**</u>	36.23	<u>27.64</u>	<u>22.87</u>	<u>12.98</u>	32.98	27.04	14.57
UEdin	21.89	16.54	14.82	20.70	29.20	24.41	18.98	9.19	25.23	22.86	9.65
Z-AGI_Labs	11.89	-	22.65	-	-	-	21.71	-	31.07	-	-

Table 4: The best CHRF++ scores 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.

The first place in the shared task, across all eleven language pairs, is awarded to the NordicAlps team (Submission 1). Their overall score significantly surpasses those of the second and third place teams, DC_DMV and UEdin, respectively. Notably, only three of the six teams submit entries for all eleven languages.

NordicAlps secures the top performance on five language pairs (Spanish to Asháninka, Chatino, Wizarika, Nahuatl, and Otomí), although they only exceed the baseline for Chatino. Similarly, the second-ranked team, DC_DMV, leads for four language pairs (Spanish to Aymara, Bribri, Shipibo-Konibo, and Rarámuri) but surpasses the baselines solely for Rarámuri. These results highlight the importance of meticulous pipeline design for data preprocessing and segmentation, as implemented by NordicAlps and the use of large multilingual models (NLLB) for finetuning, as employed by DC_DMV, for achieving robust results across most language pairs.

Finally, the BSC team, which participates for only two language pairs, Spanish to Guarani and Quechua, achieves the highest performance on both, surpassing the established baselines. Their strategic focus on finetuning a large multilingual model (NLLB) and gathering new data for these languages is key to their success.

6 Human Evaluation

Following prior AmericasNLP shared tasks (Mager et al., 2021; Ebrahimi et al., 2023), we also conduct a human evaluation of system outputs, focusing on Bribri.

6.1 Quantitative Analysis

As the test set has remained consistent across these competitions, we extend the prior evaluation using the best performing system from this year's shared task: Submission 4 by DC_DMV (DeGenaro and Lupicki, 2024). We consider the same 50 test inputs as in the prior analysis for this experiment, and a speaker of Bribri rates the system output on two axes: meaning and fluency. We consider a 5-point scale for evaluation, with a score of 5 being the best, and present results in Figure 1.

Similar to the pattern shown by the automatic metrics, we see a decrease in the perceived quality of translations from the best 2024 system as compared to the baseline (Gow-Smith and Villegas, 2023); i.e., scores suffer more, with a larger proportion being rated with a score of 1. For both metrics, scores of 5 are non-existent, showing a decrease in top-end performance as well. To further gain insights into the errors, we qualitatively look at the system outputs from the best 2024 system.

6.2 Qualitative Analysis

Table 5 shows examples of Bribri sentences translated by the best performing submission, organized by their score for meaning. The sentence with a score of 4 is readable and the original meaning is understandable, but there are parts that are not quite correct. In this example, "*Yes, you know she was great*", the hypothesis is very good, but it has at least one spelling mistake (**ujche̱n* instead of *ujché̱n* for "*it's known*"), and the word '*good*', *bua'*, is missing the intensifier {-ë} that it would need in order to become *bua'ë* '*great*'. In the case of the

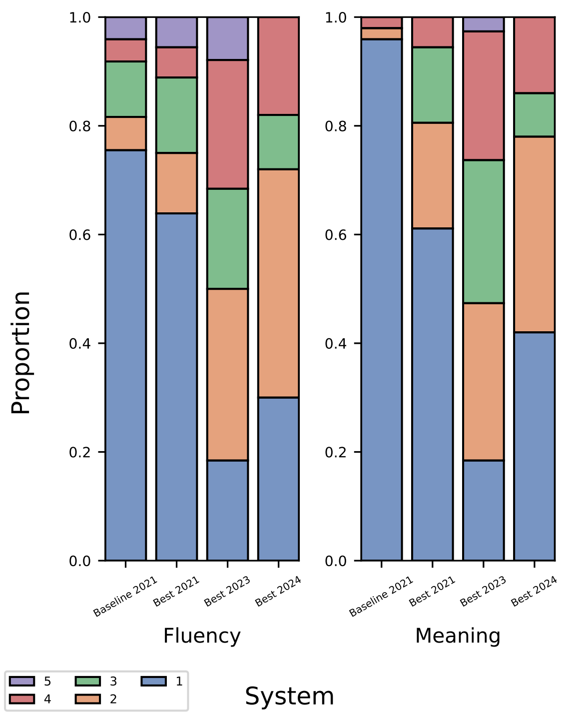


Figure 1: Results of the human evaluation for Bribri. Figure presents the proportion of evaluated example for each rating classification, with 1 representing the lowest quality and 5 representing the best. Values for the Baseline 2021, Best 2021, and Best 2023 systems are taken from (Ebrahimi et al., 2023).

sentence with the score 3, the words in the hypothesis still allow for an understanding of the meaning, but there are more mistakes. The example "*Hm, afterwards we moved to a new house*", has at least one spelling mistake: **pâali* instead of one of the other documented spellings of ‘*new*’, for example *pàali* or *páli*. More importantly, it has a reflexive pronoun *e'* which does not belong in the sentence, and the verb is missing the plural marker {-yal} in the verb *mìneyal* ‘went.PL’.

The remaining hypotheses from Table 5 have more significant issues in their meanings. The example for meaning score 2, "*I spoke to Ramona again*", has some words correct, but there are errors and entire components missing. The translation is missing the postposition *ta* ‘*with*’, which would be necessary to link the oblique argument ‘*Ramona*’ to be verb *ujté* ‘spoke’. It is possible that the system hallucinated the word *tamalé*, which resembles the word *tamáli* ‘*cuajiniquil fruit, Inga spuria*’ because the word starts with the same letters as the postposition *ta*. But, in doing so, the system changed the meaning of the translation. A factor that might contribute to the hallucination is that there is an iterative morpheme, {-male}, which can mean ‘*again*’ when it is attached to verbs (e.g. *ie démale* ‘*he came again*’ (Constenla et al., 2004, 119)). Unfortunately this morpheme is only found attached to verbs, not to postpositions,[4] and this makes the system hypothesis more difficult to understand.

Finally, the example for meaning score 1 can be translated, in its gold-standard version, as "*I am finishing with my project for next week*". The hypothesis produced by the system can be translated approximately as "*I am working*[sic], *finish, other*[sic] *weapon*". The verbs in the Bribri version are not connected properly, and the meaning of ‘*week*’ is not present in the translation. Moreover, the system hallucinated the word *mókö̈l* ‘*weapon, rifle*’, and it used the wrong numerical classifier to describe the rifle, **ië̀k* ‘*another* [round] *one*’, when it should have used the classifier for long objects (e.g. *rifles*): *ië̀töm*. These errors combined make it so that the meaning of the original sentence cannot be inferred from the system's translation.

In summary, while we have made considerable progress as a community in the translation of Indigenous languages of the Americas, there is still much work ahead of us, both in terms of data collection and algorithm development.

7 Future Directions

In this section, we briefly discuss several possible future directions for the AmericasNLP shared task, given the results from the current as well as prior competitions.

Evaluation Data One bottleneck in the advancement of language technologies for low-resource, and particularly Indigenous, languages is the availability of evaluation data. High quality, gold standard data in target low-resource languages supports many important roles in the NLP research pipeline. First, and most importantly, it is the single resource which is necessary for experimentation; without held out data for evaluation, there cannot be any idea of how well a system performs for a given language. Second, the domain and source of data is important, as, over time, models are created to perform best on the data they are evaluated on. Particularly for low-resource languages, where there may not be great diversity in available data, it becomes vital to consider what data is used for evalu-

[4]There is a Bribri iterative morpheme, {-ne}, which can be attached to adverbs and verbs, but it has not been observed with postpositions either.

MS	Bribri	English translation
4	Tő, be’ é̱n a̱ i̱à̱na̱ tö ie’ dör bua’ë.	Yes, you know she was great.
	Tő, be’ wa̱ i ujche̱n tö ie’ bák bua’.	Yes, you know[sic] that she was good.
3	Hum, ukö̀ki̱ sa’ mì̱ne̱yal ù pá̱li̱ a̱.	Hm, afterwards we moved to a new house.
	Um, e’ ukö̀ki̱ sa’ e̱’ mì̱ne̱ ù pâ̱a̱li̱ a̱.	Hm, after that we went us to a new[sic] house.
2	Ramona ta̱ ye’ ujté skàne̱	I spoke to Ramona again.
	Ye’ ujté Ramona ta̱ma̱lé̱.	I spoke, Ramona, [cuajiniquil] fruit [sic].
1	Ye’ tso’ ka̱nè̱ ma̱ú̱k è̱we̱wa semana ië̀t wa.	I am finishing with my project for next week.
	Ye’ tso’ kanè̱balök è̱nu̱k mó̱köl ië̀k.	I am working[sic]. Finish. Other[sic] weapon.

Table 5: Examples of Bribri sentences for each of the meaning scores (MS), accompanied by their translations in English. The first sentence is the gold standard, and the second sentence is the hypothesis by the best performing system.

ation. Future shared tasks should strive to continue creating new evaluation sets, both for currently supported languages (in order to increase diversity) as well as for new languages. Evaluation sets which contain data which is relevant to speakers and contain minimal biases increases the chances that good performance on the evaluation set is correlated with good real-world performance.

Additional Training Data This iteration of the shared task marks the first where performance did not increase for the majority of languages in the shared task. Of the four languages which did see improvements, two are relatively high-resource and have recently been included in large pretrained models (Costa-jussà et al., 2022). As such, additional data for training likely plays a large role in improving the performance for these languages. While teams continue to find new digital data for training, other non-digital sources may need to be considered for future systems.

Language Identification One of the main bottlenecks for gathering additional data is that every process of collecting resources from online sources starts with a good language identifier. Investing efforts into developing a language identification system for the shared task languages could boost the collection of additional training data.

New Language Pairs The performance of low-resource language pairs in multilingual MT models can benefit from incorporating additional data from other language pairs. Furthermore, our goal is to expand the scope of our shared task in future editions to include more underserved languages of the Americas. To achieve this, we plan to engage more researchers who have developed and published resources for the Indigenous languages of the Americas, both at our workshop and in other venues.

8 Conclusion

In this work, we present the results of the AmericasNLP 2024 Shared Task on Machine Translation. Overall, 6 teams participated in the shared task, and submitted a combined 157 submissions across all eleven supported languages. Prior to the start of the competition, the organizers provided two strong baselines and a training data set which includes data collected from prior submissions. While there were improvements for four languages in this year's shared task, the majority of languages did not see any performance gains over the baselines, which were the strongest systems from 2023.

Acknowledgments

We would like to thank all teams for participating in the shared task. Additionally, we would like to thank the University of Sheffield team for providing the necessary documentation for running the baseline system.

References

Joseph Attieh, Zachary William Hopton, Yves Scherrer, and Tanja Samardzic. 2024. System description of the nordicsalps submission to the americasnlp 2024 machine translation shared task. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, Mexico City, Mexico. Association for Computational Linguistics.

Damian Blasi, Antonios Anastasopoulos, and Graham Neubig. 2022. Systematic inequalities in language technology performance across the world's languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5486–5505, Dublin, Ireland. Association for Computational Linguistics.

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

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. 2022. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*.

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.

Dan DeGenaro and Tom Lupicki. 2024. Experiments in mamba sequence modeling and nllb-200 fine-tuning for low resource multilingual machine translation. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, Mexico City, Mexico. Association for Computational Linguistics.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2024. *Ethnologue: Languages of the world.*, 25 edition. SIL International, Dallas, Texas.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios Gonzales, Ivan Meza-Ruiz, et al. 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.

Javier García Gilabert, Aleix Sant Savall Carlos Escolano, Francesca De Luca Fornaciari, Audrey Mash, and Maite Melero. 2024. Bsc submission to the americasnlp 2024 shared task. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, Mexico City, Mexico. Association for Computational Linguistics.

Edward Gow-Smith and Danae Sánchez Villegas. 2023. Sheffield's Submission to the AmericasNLP Shared Task on Machine Translation into Indigenous Languages.

Albert Gu and Tri Dao. 2023. Mamba: Linear-Time Sequence Modeling with Selective State Spaces.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-Rank Adaptation of Large Language Models. In *International Conference on Learning Representations*.

Ayyoob ImaniGooghari, Peiqin Lin, Amir Hossein Kargaran, Silvia Severini, Masoud Jalili Sabet, Nora Kassner, Chunlan Ma, Helmut Schmid, André Martins, François Yvon, and Hinrich Schütze. 2023. Glot500: Scaling Multilingual Corpora and Language Models to 500 Languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1082–1117, Toronto, Canada. Association for Computational Linguistics.

Vivek Iyer, Bhavitvya Malik, Wenhao Zhu, Pavel Stepachev, Pinzhen Chen, Barry Haddow, and Alexandra Birch. 2024. Exploring very low-resource translation with llms: The university of edinburgh's submission to americasnlp 2024 translation task. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, Mexico City, Mexico. Association for Computational Linguistics.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7B.

Sneha Kudugunta, Isaac Caswell, Biao Zhang, Xavier Garcia, Christopher A. Choquette-Choo, Katherine Lee, Derrick Xin, Aditya Kusupati, Romi Stella, Ankur Bapna, and Orhan Firat. 2023. MADLAD-400: A Multilingual And Document-Level Large Audited Dataset.

Peiqin Lin, Shaoxiong Ji, Jörg Tiedemann, André F. T. Martins, and Hinrich Schütze. 2024. MaLA-500: Massive Language Adaptation of Large Language Models.

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.

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.

Matt Post. 2018. A call for clarity in reporting bleu scores. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, page 186. Association for Computational Linguistics.

Gema Ramırez-Sanchez and Jaume Zaragoza-Bernabeu. 2020. Bifixer and Bicleaner: Two open-source tools to clean your parallel data.

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

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, Brian Fuller, Cynthia Gao, Vedanuj Goswami, Naman Goyal, Anthony Hartshorn, Saghar Hosseini, Rui Hou, Hakan Inan, Marcin Kardas, Viktor Kerkez, Madian Khabsa, Isabel Kloumann, Artem Korenev, Punit Singh Koura, Marie-Anne Lachaux, Thibaut Lavril, Jenya Lee, Diana Liskovich, Yinghai Lu, Yuning Mao, Xavier Martinet, Todor Mihaylov, Pushkar Mishra, Igor Molybog, Yixin Nie, Andrew Poulton, Jeremy Reizenstein, Rashi Rungta, Kalyan Saladi, Alan Schelten, Ruan Silva, Eric Michael Smith, Ranjan Subramanian, Xiaoqing Ellen Tan, Binh Tang, Ross Taylor, Adina Williams, Jian Xiang Kuan, Puxin Xu, Zheng Yan, Iliyan Zarov, Yuchen Zhang, Angela Fan, Melanie Kambadur, Sharan Narang, Aurelien Rodriguez, Robert Stojnic, Sergey Edunov, and Thomas Scialom. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models.

Rodolfo Zevallos, John Ortega, William Chen, Richard Castro, Núria Bel, Cesar Toshio, Renzo Venturas, Hilario Aradiel, and Nelsi Melgarejo. 2022. Introducing QuBERT: A Large Monolingual Corpus and BERT Model for Southern Quechua. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 1–13, Hybrid. Association for Computational Linguistics.

Appendix

Lang.	Team	Ver.	BLEU	ChrF	ChrF++
aym	DC_DMV	2	3.49	35.43	30.97
aym	NordicsAlps	1	3.23	33.46	29.39
aym	DC_DMV	4	2.74	32.36	28.32
aym	DC_DMV	3	2.52	32.24	28.12
aym	DC_DMV	6	2.24	31.43	27.48
aym	DC_DMV	1	2.27	31.26	27.36
aym	DC_DMV	5	2.37	30.91	27.09
aym	NordicsAlps	2	1.99	30.37	26.37
aym	UEdin	3	1.13	25.14	21.89
aym	UEdin	1	1.14	24.94	21.77
aym	UEdin	2	1.06	24.56	21.37
aym	NordicsAlps	3	1.10	17.55	15.77
aym	Z-AGI_Labs	1	0.74	13.30	11.89
aym	DC_DMV	9	0.15	9.51	8.69
bzd	DC_DMV	4	4.84	22.23	23.47
bzd	DC_DMV	5	4.56	22.15	23.41
bzd	DC_DMV	3	4.63	22.02	23.32
bzd	NordicsAlps	1	5.00	22.27	23.32
bzd	DC_DMV	1	4.68	21.97	23.19
bzd	DC_DMV	6	4.75	21.99	23.15
bzd	DC_DMV	2	3.44	18.11	19.60
bzd	NordicsAlps	2	1.72	15.98	17.23
bzd	UEdin	1	2.21	15.43	16.54
bzd	UEdin	2	1.89	15.17	16.32
bzd	UEdin	3	1.75	14.53	15.56
bzd	NordicsAlps	3	1.39	13.17	12.24
bzd	DC_DMV	9	0.09	4.36	4.72
cni	NordicsAlps	1	2.41	27.76	23.20
cni	DC_DMV	6	3.49	26.15	22.98
cni	DC_DMV	3	3.56	26.05	22.87
cni	Z-AGI_Labs	1	3.22	26.75	22.65
cni	DC_DMV	5	3.41	25.63	22.53
cni	DC_DMV	4	3.51	25.53	22.46
cni	DC_DMV	1	3.56	25.48	22.44
cni	DC_DMV	2	3.52	22.13	19.89
cni	NordicsAlps	2	0.06	20.13	15.45
cni	NordicsAlps	3	1.68	17.30	15.23
cni	UEdin	1	0.41	17.54	14.82
cni	UEdin	3	0.43	17.08	14.50
cni	UEdin	2	0.37	16.26	13.68
cni	DC_DMV	9	0.14	11.83	9.81
ctp	NordicsAlps	1	13.44	40.37	37.38
ctp	NordicsAlps	2	4.65	26.61	23.64
ctp	UEdin	2	4.30	23.01	20.70
ctp	UEdin	1	3.35	19.50	17.66
ctp	UEdin	3	3.38	19.50	17.57
ctp	DC_DMV	1	1.73	20.58	16.52
ctp	DC_DMV	3	1.68	20.18	16.17
ctp	DC_DMV	5	1.68	20.06	16.11
ctp	DC_DMV	6	1.75	19.90	16.04
ctp	DC_DMV	4	1.74	19.59	15.78
ctp	NordicsAlps	3	1.78	14.97	12.96
ctp	DC_DMV	2	0.96	9.72	8.06
ctp	DC_DMV	9	0.00	3.38	2.62
gn	BSC	3	12.04	41.81	38.93
gn	BSC	4	11.28	40.66	37.64
gn	BSC	2	11.37	40.69	37.63
gn	BSC	1	11.04	40.38	37.42
gn	NordicsAlps	1	8.82	39.36	36.23
gn	BSC	5	11.53	38.37	35.73
gn	DC_DMV	2	5.46	36.78	33.31
gn	DC_DMV	3	6.30	35.72	32.58
gn	DC_DMV	4	6.42	35.51	32.44
gn	NordicsAlps	2	6.81	35.23	32.32
gn	DC_DMV	6	5.82	34.69	31.66
gn	DC_DMV	1	5.97	34.66	31.58
gn	DC_DMV	5	5.66	34.18	31.22
gn	UEdin	1	3.38	32.22	29.20
gn	UEdin	3	3.21	32.31	29.13
gn	UEdin	2	1.78	27.61	24.61
gn	NordicsAlps	3	1.60	16.11	14.80
gn	DC_DMV	9	0.32	10.10	8.91
hch	NordicsAlps	1	10.08	31.13	27.64
hch	DC_DMV	1	9.62	29.83	26.46
hch	DC_DMV	4	8.51	29.54	26.23
hch	DC_DMV	5	8.64	29.21	25.97
hch	DC_DMV	6	8.83	28.95	25.66
hch	DC_DMV	3	8.85	28.75	25.60
hch	UEdin	1	9.87	27.40	24.41
hch	UEdin	3	9.60	27.50	24.37
hch	NordicsAlps	2	6.46	26.92	23.47
hch	UEdin	2	7.03	24.51	22.03
hch	DC_DMV	2	3.29	22.36	19.56
hch	NordicsAlps	3	1.35	18.43	15.97
hch	DC_DMV	9	0.49	8.10	7.12
nah	NordicsAlps	1	2.30	26.91	22.87
nah	Z-AGI_Labs	1	1.09	26.29	21.71
nah	DC_DMV	1	1.79	25.58	21.63
nah	DC_DMV	4	1.73	25.41	21.44
nah	DC_DMV	5	1.86	25.35	21.43
nah	DC_DMV	6	1.78	25.24	21.41
nah	DC_DMV	3	1.85	24.84	21.07
nah	NordicsAlps	2	1.52	24.84	20.77
nah	UEdin	3	0.44	22.86	18.98
nah	DC_DMV	2	1.75	21.69	18.52
nah	UEdin	1	0.48	21.75	18.12
nah	UEdin	2	0.37	20.78	17.21
nah	NordicsAlps	3	1.64	17.08	14.57
nah	DC_DMV	9	0.12	13.14	10.46
oto	NordicsAlps	1	1.42	14.95	12.98
oto	DC_DMV	1	1.55	14.61	12.63
oto	DC_DMV	3	1.66	14.30	12.42
oto	DC_DMV	4	1.50	14.34	12.42
oto	DC_DMV	5	1.52	14.29	12.40
oto	DC_DMV	6	1.36	14.14	12.20
oto	NordicsAlps	2	0.20	13.80	11.63
oto	DC_DMV	2	1.46	13.05	11.50
oto	NordicsAlps	3	1.41	13.14	11.22
oto	UEdin	3	0.44	10.87	9.19
oto	UEdin	1	0.43	10.56	8.91
oto	UEdin	2	0.21	9.32	7.81
oto	DC_DMV	9	0.04	4.39	3.63
quy	BSC	1	4.85	44.04	38.21
quy	BSC	4	4.83	43.91	38.19
quy	BSC	2	4.72	43.87	38.10
quy	BSC	3	4.44	43.86	38.02
quy	DC_DMV	2	5.41	41.43	36.02
quy	DC_DMV	4	4.32	39.67	34.29
quy	DC_DMV	3	4.13	39.49	34.08
quy	DC_DMV	5	3.91	39.33	33.94
quy	DC_DMV	1	4.01	39.24	33.91
quy	DC_DMV	6	4.05	38.95	33.56
quy	NordicsAlps	1	4.08	37.92	32.98
quy	ND-NAIST	1	2.60	38.51	32.88
quy	Z-AGI_Labs	1	3.29	36.69	31.07

Lang.	Team	Ver.	BLEU	ChrF	ChrF++
quy	NordicsAlps	2	2.65	33.36	28.81
quy	UEdin	1	1.32	29.54	25.23
quy	NordicsAlps	3	2.77	28.99	25.15
quy	UEdin	3	1.31	29.37	25.04
quy	UEdin	2	0.94	26.69	22.77
quy	DC_DMV	9	0.40	13.08	11.42
shp	DC_DMV	2	4.45	32.95	29.37
shp	NordicsAlps	1	4.14	30.55	27.04
shp	DC_DMV	4	3.90	27.77	24.74
shp	DC_DMV	3	3.44	26.86	23.84
shp	DC_DMV	5	3.17	26.58	23.59
shp	DC_DMV	6	3.07	25.91	23.05
shp	UEdin	3	1.55	25.90	22.86
shp	UEdin	2	1.56	25.52	22.43
shp	DC_DMV	1	2.95	25.04	22.25
shp	NordicsAlps	2	1.09	25.68	22.20
shp	UEdin	1	1.34	25.08	22.04
shp	NordicsAlps	3	2.60	23.83	21.28
shp	DC_DMV	9	0.27	11.13	9.67
tar	DC_DMV	2	0.92	18.94	17.03
tar	DC_DMV	3	1.01	17.20	15.42
tar	DC_DMV	4	0.93	16.72	14.92
tar	DC_DMV	6	0.81	16.69	14.57
tar	NordicsAlps	1	0.55	17.03	14.57
tar	DC_DMV	5	1.04	16.58	14.51
tar	DC_DMV	1	0.86	16.41	14.39
tar	NordicsAlps	3	0.73	14.49	12.63
tar	NordicsAlps	2	0.12	12.54	10.53
tar	UEdin	1	0.11	11.46	9.65
tar	UEdin	2	0.11	11.07	9.49
tar	UEdin	3	0.15	11.32	9.48
tar	DC_DMV	9	0.07	7.65	6.64

Table 6: Full results of the shared task.

Author Index