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

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	$r_{agr \to es}$	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; $r_{agr \to es}$ = 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 $r_{agr \to es}$ are the lowest. On the other hand, the News by Ojo Público dataset has a $r_{agr \to es}$ 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		Test	
			BLEU	Chrf	BLEU	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