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.

[†]Authors of equal contribution

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

Proceedings of the 4th Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 150–158
June 21, 2024 ©2024 Association for Computational Linguistics

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 handcrafted 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.

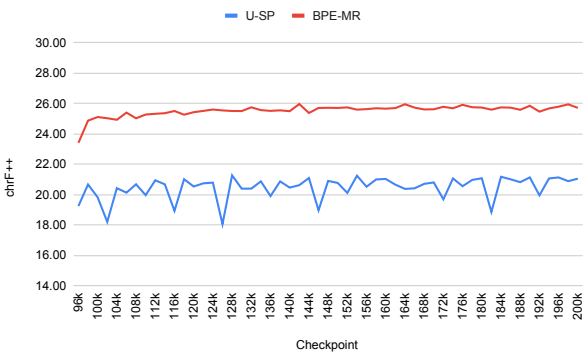

Figure 1: Development chrF++ scores (averaged over all 11 development sets) during phase 2 training of subword-level models.

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.

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.

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.