Moses and the Character-Based Random Babbling Baseline: CoAStaL at AmericasNLP 2021 Shared Task

Marcel Bollmann Rahul Aralikatte Héctor Ricardo Murrieta Bello
Daniel Hershcovich Miryam de Lhoneux Anders Søgaard
Department of Computer Science
University of Copenhagen
{marcel,rahul,dh,ml,soegaard}@di.ku.dk xhd160@alumni.ku.dk

Abstract

We evaluated a range of neural machine translation techniques developed specifically for low-resource scenarios. *Unsuccessfully.* In the end, we submitted two runs: (i) a standard phrase-based model, and (ii) a random babbling baseline using character trigrams. We found that it was surprisingly hard to beat (i), in spite of this model being, in theory, a bad fit for polysynthetic languages; and more interestingly, that (ii) was better than several of the submitted systems, highlighting *how* difficult low-resource machine translation for polysynthetic languages is.

1 Introduction

Shared tasks on machine translation are often conducted on large parallel training corpora: for example, the majority of datasets used in the WMT20 shared tasks have sentence pairs in the hundred thousands, often even millions (Barrault et al., 2020). In contrast, the AmericasNLP 2021 shared task (Mager et al., 2021) provided us with as little as 3,883 sentence pairs (for Ashaninka), and with the exception of Quechua (125k pairs), all languages had fewer than 30k sentence pairs. Additionally, many of these languages are polysynthetic, which is known to provide additional challenges for machine translation (Klavans et al., 2018; Mager et al., 2018b).

We initially focused our efforts on two areas: (i) obtaining more data, both parallel and monolingual (Sec. 2); and (ii) exploring a range of different neural machine translation techniques, particular those specifically developed for low-resource scenarios, to find a promising system to build on and tweak further. Unfortunately, we were wholly unsuccessful in the latter (Sec. 5). All neural models that we tried performed extremely poorly when compared to a standard statistical phrase-based model (Sec. 3.1). The overall low performance of all our models further prompted us to implement a "random babbling" baseline (Sec. 3.2): a model that outputs plausible-looking n-grams in the target language without any actual relation to the source sentences. This baseline, together with the phrase-based model, were the only two systems we ended up submitting. Our main findings are:

- It was surprisingly hard to beat a standard phrase-based model, as evidenced not only by our own failed attempts, but also by this system taking third place on three languages in the official evaluation (track 1).
- It is apparently challenging for many MT systems to even produce well-formed outputs in the target languages, as our random babbling baseline outperformed *at least* one other system on nine of the languages, and even took fifth place out of 12 on Ashaninka (track 2).

Language		Source(s)
AYM	Aymara	Prokopidis et al. (2016)
BZD	Bribri	Feldman and Coto-Solano (2020)
CNI	Asháninka	Ortega et al. (2020), Cushimariano Romano and Sebastián Q. (2008), Mihas (2011)
GN	Guaraní	Chiruzzo et al. (2020)
HCH	Wixarika	Mager et al. (2018a)
NAH	Nahuatl	Gutierrez-Vasques et al. (2016)
OTO	Hñähñu	Comunidad Elotl (2021)
QUY	Quechua	Agić and Vulić (2019)
SHP	Shipibo-Konibo	Galarreta et al. (2017)
TAR	Rarámuri	Brambila (1976)

Table 1: Languages in the shared task with sources of their training datasets

2 Data

We train models for all languages provided by the shared task, using their official training datasets (cf. Table 1). As the shared task allowed for using external datasets, we also tried to find more data sources to use for model training.

Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 248–254
June 11, 2021. ©2021 Association for Computational Linguistics

Parallel data We gathered parallel Spanish-to-target datasets for the following languages which should not overlap with the data provided by the shared task organizers: Aymara from JW300 (Agić and Vulić, 2019); Guarani from Tatoeba; and Nahuatl and Quechua from the Bible corpus by Christodouloupoulos and Steedman (2015). We note that for the Bible corpus, the Nahuatl portion is from a narrower dialectal region (NHG "Tetelcingo Nahuatl") than the data in the shared task, and it also covers a different variant of Quechua (QUW "Kichwa" vs. QUY "Ayacucho Quechua"), but we hoped that in this extremely low-resource scenario, this would still prove useful. All datasets were obtained from OPUS[1] (Tiedemann, 2012).

Monolingual data Wikipedias exist for Aymara, Guaraní, Nahuatl, and Quechua. We use WikiExtractor (Attardi, 2015) to obtain text data from their respective dumps,[2] then use a small set of regular expressions to clean them from XML tags and entities. This gives us between 28k and 100k lines of text per language.

We obtain further monolingual data from several online sources in PDF format. For Nahuatl and Hñähñu, we use a book provided by the Mexican government;[3] for Quechua, we use two books: *The Little Prince* (Saint-Exupéry, 2018) and Antonio Raimondi's *Once upon a time.. in Peru* (Villacorta, 2007). The Mexican government also publishes the series *Languages from Mexico* which contains books based on short stories in Nahuatl (Gustavo et al., 2007), Raramuri (Arvizu Castillo, 2002), Hñähñu (Mondragón et al., 2002b), and Wixárika (Mondragón et al., 2002a). Finally, we also use the Bible translated to Quechua, Guarani, and Aymara. We extract the text for all of these resources with the Google OCR API.[4]

3 Models

We first describe the two models we submitted: a standard phrase-based model (CoAStaL-1) and a random babbling baseline (CoAStaL-2). Other models that we experimented with but did not submit for evaluation are discussed later in Sec. 5.

3.1 Phrase-Based MT

We train a statistical phrase-based model with Moses (Koehn et al., 2007) using default settings, following the guidelines for training a baseline.[5] We do minimal preprocessing: we use the provided cleaning script and rely on plain whitespace tokenization, with the only exception that we also insert spaces around square brackets. The language model is trained with 5-grams instead of 3-grams, as this improved the results very slightly on the development sets. We train a separate model for each language and use the respective development set for tuning before translating the test set.

The models we submitted did, mistakenly, *not* make use of the additional parallel data we gathered (cf. Sec. 2). We evaluated the same system trained *with* this additional data after the deadline, but unfortunately did not observe an improvement; we present results for both variants in Sec. 4.

3.2 Random Babbling Baseline

Since we observed very low scores for all the models we tried, we wanted to compare with a baseline that generates text based only on (i) n-gram distributions in the target language, and (ii) lengths of the source sentences. We call this baseline *random babbling* because it is in no way conditioned on the actual words in the source sentences.

Concretely, we "train" our baseline by extracting and counting all *character trigrams* in the training file of the target language. Characters were chosen over words as the official evaluation metric of the shared task, chrF, is character-based. We also calculate the average *length ratio* of the sentence pairs in order to determine the desired length of our "translation" at test time. To generate output, we simply choose the top n most frequent character trigrams, with n chosen so that the desired sentence length is reached.[6]

Lastly, we perform a few tweaks to disguise this babbling as an actual translation: (i) we randomize the order of the chosen trigrams, (ii) reduce multiple consecutive whitespace characters to a single space, (iii) lowercase all characters that are not word-initial and uppercase the sentence-initial

[1] https://opus.nlpl.eu/

[2] https://dumps.wikimedia.org/

[3] https://www.gob.mx/inpi/documentos/libros-en-lenguas-indigenas

[4] https://cloud.google.com/vision/docs/pdf

[5] http://www.statmt.org/moses/?n=Moses.Baseline

[6] We also tried random baseline models with other n-gram lengths, sampling from the distribution (instead of always picking the most frequent items), and training a simple language model, but found nothing that significantly improved on this approach on the development set.

Set	System	Track	Languages									
			AYM	BZD	CNI	GN	HCH	NAH	OTO	QUY	SHP	TAR
Dev	CoAStaL-1: Phrase-based	1	.225	.213	.253	.235	.261	.204	.160	.276	.276	.174
	CoAStaL-2: Random	2	.178	.113	.214	.132	.195	.189	.094	.234	.182	.116
Test	Helsinki-2 (best)	1	.310	.213	.332	.376	.360	.301	.228	.394	.399	.258
	CoAStaL-1: Phrase-based	1	.191	.196	.265	.241	.257	.214	.184	.269	.297	.159
	+ extra data	1	.188	–	–	.242	–	.216	–	.250	–	–
	CoAStaL-2: Random	2	.168	.107	.212	.128	.191	.184	.101	.232	.173	.113
	Baseline	2	.157	.068	.102	.193	.126	.157	.054	.304	.121	.039

(a) chrF

			AYM	BZD	CNI	GN	HCH	NAH	OTO	QUY	SHP	TAR
Dev	CoAStaL-1: Phrase-based	1	2.57	3.83	2.79	2.59	6.81	2.33	1.44	1.73	3.70	1.26
	CoAStaL-2: Random	2	0.02	0.03	0.04	0.02	1.14	0.02	0.02	0.02	0.06	0.02
Test	Helsinki-2 (best)	1	2.80	5.18	6.09	8.92	15.67	3.25	5.59	5.38	10.49	3.56
	CoAStaL-1: Phrase-based	1	1.11	3.60	3.02	2.20	8.80	2.06	2.72	1.63	3.90	1.05
	+ extra data	1	1.07	–	–	2.24	–	2.06	–	1.24	–	–
	CoAStaL-2: Random	2	0.05	0.06	0.03	0.03	2.07	0.03	0.03	0.02	0.04	0.06
	Baseline	2	0.01	0.01	0.01	0.12	2.20	0.01	0.00	0.05	0.01	0.00

(b) BLEU

Table 2: Results for our submitted models on Dev and Test sets. All Test results are from the official evaluation except for the "Phrase-based + extra data" setting, which we evaluated after the deadline.

character, and (iv) if the sequence does not end in a punctuation mark but the Spanish source sentence did, we copy and add this punctuation character from the source side.

4 Results

Results of our models are shown in Table 2, both for our own evaluation on the development sets and for the official evaluation on the test sets (Ebrahimi et al., 2021).

Phrase-Based MT Our phrase-based model (Sec. 3.1) was ranked in track 1 of the shared task evaluation as it makes use of the development sets for tuning. Compared to the other systems evaluated in this track, we observe a solid average performance of our model—it usually ranks in the middle of the field, with the best placement being 3rd on Bribri, Hñähñu, and Shipibo-Konibo, and the worst ranking being 8th out of 11 on Guarani. In terms of chrF score, the model ranges between 0.159 (on Raramuri) and 0.297 (on Shipibo-Konibo), but we note that there is a noticeable gap to the best-performing system, Helsinki-2, which outperforms ours by about +0.09 chrF on average.

Random Babbling Our random babbling baseline (Sec. 3.2) did *not* make use of the development sets and was therefore ranked in track 2 of the official evaluation. Amazingly, it almost never ranks last and even takes 5th place out of 12 on Ashaninka. It also outperforms the official baseline on eight of the languages. In terms of BLEU score, on the other hand, this model usually scores close to zero. This is because we based it on character trigrams; if we wanted to optimize for BLEU, we could have chosen word-based babbling instead. Comparing across the tracks with our first, phrase-based system, we observe that the latter scores consistently better, which is reassuring.

4.1 Discussion

We intended our phrase-based Moses system more as a baseline for our experiments with different neural models than as an actual system submission. It was surprising to us how clearly this system outperformed our attempts at building a neural MT system, and that it already did so with its default configuration. In theory, whitespace tokenization should be a bad fit for polysynthetic languages, as a high degree of morphological complexity exacerbates the problem of data sparsity and rarely seen word forms. We experimented with different subword tokenization techniques in combination with Moses, but this always resulted in degraded performance on the development sets.

The random babbling baseline was motivated by two observations: (i) performance was extremely low for all models we tried, and (ii) outputs of the neural models frequently looked very unnatural, to the point that the models had not yet learned how to form plausible-looking sentences in the target languages. This is quite typical behavior for underfitted neural models. As an example, this is an output we observed when running the official baseline system on the development set for Raramuri:

(1) IN: *Realmente no me importa si tengo un lugar para vivir.*
GOLD: *Ke chibi iré mapure ke nirúlisaka kúmi ne betélima.*
PRED: *(2) (a) ké ne ga'rá ne ga'rá ne ga'rá ga'rá ne ga'rá ne ga'rá ne ga'rá ne ga'rá ne ga'rá ne ga'rá ne ga'rá ne ga'rá ne ga'rá ne ga'rá ne ga'rá ne ga'rá ne ga'rá ne ga'rá ne ga'rá ne ga'rá ne ga'rá ne ga'rá ne ga'rá [. . .]*

This prompted us to implement a baseline which, while having *no* relation to the actual input sentence, at least better resembles the typical distribution of character n-grams in the given language. Here is an example from the test set for Ashaninka with outputs from both our phrase-based (SYS-1) and random (SYS-2) model:

(2) IN: *Todavía estoy trabajando para este día.*
GOLD: *Irosatitatsi nantabeeti oka kitaiteriki.*
SYS-1: *Tekirata nosaikaki trabajando inchamenta itovantarori." día.*
SYS-2: *Iritsiri irotakntakanarishiantakiro aka.*

We can see that both system outputs bear very little resemblance to the gold translation or to each other. While Moses (SYS-1) copies a few Spanish words and includes implausibly placed punctuation marks, random babbling (SYS-2) produces output of similar length to the correct translation and overlaps with it in several observable character trigrams (e.g. *iro, tsi, ant*).

Obviously, the random babbling baseline is not meant as an actual suggestion for a translation system—it literally does not "translate" anything. However, as the official shared task evaluation and the examples above show, it can serve as a useful "sanity check" for situations where the performance of actual MT systems is so low that it is unclear whether they even acquired superficial knowledge of character distributions in the target language.

5 Things that did not work

Here we briefly describe other ideas that we pursued, but were unfortunately not successful with, so we did not submit any systems based on these techniques for evaluation.

Pre-trained Transformers Following Rothe et al. (2020), we use an auto-encoding transformer as the encoder and an auto-regressive transformer as the decoder of a sequence-to-sequence model. Out of the several configurations we experimented with, the best performance was observed when the encoder is pre-trained on the Spanish OSCAR corpus (Ortiz Suárez et al., 2020) and the decoders are pre-trained on language-specific monolingual corpora collected from the web (cf. Sec. 2) along with the target files of the training data. However, the results were not on-par with the simpler models; averaging over all languages, we observed a chrF score of 0.12 on the dev sets, compared to 0.23 with the phrase-based model (cf. Sec. 3.1). We postulate that the training data was just not enough to train the cross-attention weights between the encoder and decoders. Note that these weights need to be trained from scratch, as opposed to the other weights which are initialized from language modelling checkpoints.

Back-translation In an attempt to improve the transformer-based models, we used the shared task data to train similar transformer-based models in the reverse direction, i.e. *to* Spanish, in order to back-translate the monolingual corpora (cf. Sec. 2). This would give us automatically translated Spanish outputs to use as the source side for additional training data (Sennrich et al., 2016; Hoang et al., 2018). Since monolingual data in Spanish—which was used to pre-train the decoder's language model for this experiment—is abundant, we expected the machine-translated Spanish text to be of reasonably good quality. However, the models turned out to perform quite badly, with the resulting Spanish text being of very low quality and often very repetitive. We therefore decided to abandon this direction after preliminary experiments.

Character-Level NMT Since many of the languages in the shared task are polysynthetic, a character-level model might be better suited here, as it can better learn morphology (Belinkov et al., 2017). We train fully character-level models following Lee et al. (2017), which are based on combining convolutional and recurrent layers.[7] Finding a good hyperparameter configuration for this model proved very time-consuming; the best configuration we found modifies the original model by using half the number of units in the embedding layer and decoder layers (256 and 512, respectively). For Quechua, which we initially experimented on, this yielded a chrF score of 0.33 on the dev set vs. 0.27 with phrase-based MT, but we ran out of time to train models for the other languages. A post-hoc evaluation on the other languages failed to replicate this success, though. Potentially, the hyperparameter configuration is very sensitive to the language in question, or the amount of training data was not enough for the other languages (Quechua had by far the largest training set of all languages in the shared task).

Language Model Prior We train NMT models using a language model prior, following Baziotis et al. (2020). This method allows us to make use of the additional monolingual data we gathered (cf. Sec. 2) within a neural MT framework, and we hoped that this would help the model to produce valid words in the target languages, i.e., reduce the "babbling" effect we saw in outputs like Example (1) above. We focused our efforts on the LSTM-based models provided by the authors[8] rather than the transformer ones, since we believe that those should be easier to train in this extremely low-resource setting. Despite experimenting with different hyperparameters (including number and size of LSTM layers), we could not exceed an average 0.16 chrF on the dev sets (compared to 0.23 with the phrase-based model).

Graph Convolutional Encoders We experiment with graph convolutional encoders using the framework by Bastings et al. (2017). Thus, we train NMT systems that operate directly over graphs; in our case, syntactic annotations of the source sentences following the Universal Dependencies (UD) scheme (Nivre et al., 2020). We parsed the all the source sentences from training set provided by the task organizer with Stanza (Qi et al., 2020). We were initially motivated to follow this approach because UD annotation can provide extra information to the encoder to generate better translations, ideally with less data. Even though we tested several configurations, not even our best architecture—two layers of GCN encoder with 250 units, and LSTM decoder with 250 units, trained for 5 epochs, with a vocabulary of 5000 words in source and target—was able to outperform the random babbling system. We hypothesize that with this amount of examples, UD's external information is not sufficient to produce an efficient encoder.

6 Conclusion

The (relative) success of our random babbling baseline shows that many MT systems fail to reproduce even superficial characteristics of word formation and character distribution in the target languages; a result that was confirmed by our own failed attempts at training a competitive neural MT model.

Out of the neural models we tried, purely character-level MT was among the more promising ones. We speculate that in the Spanish-to-target setting, a model that combines a strong pre-trained Spanish encoder with a purely character-level decoder might be a promising direction for further experiments.

We also note that there are several language-specific resources, such as morphological segmentation tools,[9] that might be worth using. We focused our efforts here on finding a broadly applicable architecture without any language-specific components, but would be curious to see if including such components can yield significant improvements on individual languages.

Acknowledgements

Marcel Bollmann was funded from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No. 845995. Rahul Aralikatte and Anders Søgaard were funded by a Google Focused Research Award. Miryam de Lhoneux was funded by the Swedish Research Council (grant 2020-00437).

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.

Teresa Arvizu Castillo. 2002. *Relatos tarahumaras = Ki'á ra'ichaala rarámuli*. CNCA-Dirección Gen-

[7] We use our own reimplementation of the authors' code.

[8] `https://github.com/cbaziotis/lm-prior-for-nmt`

[9] e.g. Apertium for Guarani: `https://github.com/apertium/apertium-grn`

eral de Culturas Populares e Indígenas, Ciudad de Mexico.

Giusepppe Attardi. 2015. Wikiextractor. https://github.com/attardi/wikiextractor.

Loïc Barrault, Magdalena Biesialska, Ondřej Bojar, Marta R. Costa-jussà, Christian Federmann, Yvette Graham, Roman Grundkiewicz, Barry Haddow, Matthias Huck, Eric Joanis, Tom Kocmi, Philipp Koehn, Chi-kiu Lo, Nikola Ljubešić, Christof Monz, Makoto Morishita, Masaaki Nagata, Toshiaki Nakazawa, Santanu Pal, Matt Post, and Marcos Zampieri. 2020. Findings of the 2020 conference on machine translation (WMT20). In *Proceedings of the Fifth Conference on Machine Translation*, pages 1–55, Online. Association for Computational Linguistics.

Jasmijn Bastings, Ivan Titov, Wilker Aziz, Diego Marcheggiani, and Khalil Sima'an. 2017. Graph convolutional encoders for syntax-aware neural machine translation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1957–1967, Copenhagen, Denmark. Association for Computational Linguistics.

Christos Baziotis, Barry Haddow, and Alexandra Birch. 2020. Language model prior for low-resource neural machine translation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7622–7634, Online. Association for Computational Linguistics.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872, Vancouver, Canada. Association for Computational Linguistics.

David Brambila. 1976. *Diccionario Raramuri–Castellano (Tarahumara)*. Obra Nacional de la Buena Prensa, Mexico.

Luis Chiruzzo, Pedro Amarilla, Adolfo Ríos, and Gustavo Giménez Lugo. 2020. Development of a Guarani - Spanish parallel corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 2629–2633, Marseille, France. European Language Resources Association.

Christos Christodouloupoulos and Mark Steedman. 2015. A massively parallel corpus: the Bible in 100 languages. *Language Resources and Evaluation*, 49(2):375–395.

Comunidad Elotl. 2021. Tsunkua – corpus paralelo otomí-español.

Rubén Cushimariano Romano and Richer C. Sebastián Q. 2008. Ñaantsipeta asháninkaki birakochaki. Diccionario Asháninka-Castellano. Versión preliminar. http://www.lengamer.org/publicaciones/diccionarios/.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir, Gustavo A. Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando A. Coto Solano, Ngoc Thang Vu, and Katharina Kann. 2021. Americasnli: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages.

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.

Aguilar Gutiérrez Gustavo, Arellano Zamora Rogelio, Conde Reyes Magdaleno, Tepole Rivera Miguel Ángel, and Tzanahua Antonio. 2007. *Relatos nahuas = Nahua zazanilli*. CNCA-Dirección General de Culturas Populares e Indígenas, Ciudad de Mexico.

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).

Vu Cong Duy Hoang, Philipp Koehn, Gholamreza Haffari, and Trevor Cohn. 2018. Iterative back-translation for neural machine translation. In *Proceedings of the 2nd Workshop on Neural Machine Translation and Generation*, pages 18–24, Melbourne, Australia. Association for Computational Linguistics.

Judith Klavans, John Morgan, Stephen LaRocca, Jeffrey Micher, and Clare Voss. 2018. Challenges in speech recognition and translation of high-value low-density polysynthetic languages. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 2: User Track)*, pages 283–293, Boston, MA. Association for Machine Translation in the Americas.

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.

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.

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, Elisabeth Mager, Alfonso Medina-Urrea, Ivan Vladimir Meza Ruiz, and Katharina Kann. 2018b. Lost in translation: Analysis of information loss during machine translation between polysynthetic and fusional languages. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 73–83, Santa Fe, New Mexico, USA. 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, Anna Currey, Vishrav Chaudhary, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager, Ngoc Thang Vu, Graham Neubig, 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 NLP for Indigenous Languages of the Americas*, Online. Association for Computational Linguistics.

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

Lucila Mondragón, Jacqueline Tello, and Argelia Valdez. 2002a. *Relatos huicholes = Wixarika’ ’ïxatsikayari*. CNCA-Dirección General de Culturas Populares e Indígenas, Ciudad de Mexico.

Lucila Mondragón, Jacqueline Tello, and Argelia Valdez. 2002b. *Relatos otomíes. Nfini Hñähñu*. CNCA-Dirección General de Culturas Populares e Indígenas, Ciudad de Mexico.

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. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

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.

Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. 2020. A monolingual approach to contextualized word embeddings for mid-resource languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1703–1714, Online. Association for Computational Linguistics.

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

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Sascha Rothe, Shashi Narayan, and Aliaksei Severyn. 2020. Leveraging pre-trained checkpoints for sequence generation tasks. *Transactions of the Association for Computational Linguistics*, 8:264–280.

Antoine de Saint-Exupéry. 2018. *Quyllur Llaqtayuq Wawamanta*. Ediciones El Lector, Arequipa, Peru. Translated by Lydia Cornejo Endara & César Itier.

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016. Improving neural machine translation models with monolingual data. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 86–96, Berlin, Germany. Association for Computational Linguistics.

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).

Luis Felipe Villacorta. 2007. “Había una vez... El Perú de Antonio Raimondi”. Historia y alcances de un cuento para niños creado en el museo. *Illapa Mana Tukukuq*, (4):101–112.