Enriching Wayúunaiki–Spanish Neural Machine Translation with Linguistic Information

Nora Graichen,[1] **Josef van Genabith,**[1,2] **Cristina España-Bonet**[2]
[1]Saarland University, Saarland Informatics Campus, Germany
[2]German Research Center for Artificial Intelligence (DFKI GmbH)
`graichen@coli.uni-saarland.de`
`{cristinae, Josef.Van_Genabith}@dfki.de`

Abstract

We present the first neural machine translation system for the low-resource language pair Wayúunaiki–Spanish and explore strategies to inject linguistic knowledge into the model to improve translation quality. We explore a wide range of methods and combine complementary approaches. Results indicate that incorporating linguistic information through linguistically motivated subword segmentation, factored models, and pretrained embeddings helps the system to generate improved translations, with the segmentation contributing most. In order to evaluate translation quality in a general domain and go beyond the available religious domain data, we gather and make publicly available a new test set and supplementary material. Although translation quality as measured with automatic metrics is low, we hope these resources will facilitate and support further research on Wayúunaiki.

1 Introduction

Due to a lack of data (text or speech data), languages are digitally divided between high and low-resourced (LRL) (Bender, 2019). Actually, the low-resource scenario has been identified as one of the main challenges in the field of Natural Language Processing (NLP) (Koehn and Knowles, 2017). At the same time, research conducted and presented at major conferences often focuses on a few highly resourced languages, languages with similar characteristics, or a handful of well-studied languages (Joshi et al., 2020). Fortunately, research in low-resource settings and with LRLs is slowely becoming quite popular in the NLP community, with a steadily growing body of work for the low-resource scenario (Wang et al., 2021). This does not imply that the division between low- and high-resourced NLP scenarios has been overcome. In fact, there are many open challenges for research on and with LRLs.

The majority of the world's 7000 languages are understudied and underresourced (Joshi et al., 2020), due to the lack of research and resources. LRLs face a lack of data quality and quantity, NLP tools, and engagement with native speakers of that language, which, if overcome, can support the conservation and preservation of those languages and their culture, preserving cultural and linguistic diversity.

In this work, we aim at fostering research for Wayúunaiki by providing data and pretrained Neural Machine Translation (NMT) models. We present the first Wayúunaiki–Spanish NMT system, and explore different approaches to inject linguistic knowledge to improve translation quality. We aim at assisting the Wayúu community, whose language is emerging from an endangered situation according to Ethnologue.[1] Even though the Wayúu people are the most numerous indigenous people in Colombia (Departamento Administrativo Nacional de Estadística, 2021), Wayúunaiki is vulnerable, i.e. the language is spoken by children but only in certain, restricted domains, for instance at home. Our research hypothesis in this work is that the injection of linguistic knowledge will increase the translation quality for the language pair Wayúunaki–Spanish. We enrich the data to represent implicit linguistic information (e.g., linguistically motivated subword segmentation, annotating POS tag factors, and pretrained embeddings) as, if insufficient amounts of training data is available, linguistic information may help the model identify patterns present in the text, which may alleviate the data sparsity problem. We build on and extend previous work on NMT for LRLs by Sennrich and Haddow (2016) and Chen and Fazio (2021). We combine complementary approaches to maximize improvements. We find that while linguistically motivated subword segmentation helps, factored models and pretrained embeddings lead

[1]`https://www.ethnologue.com/language/guc/`

Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP), pages 67–83
July 14, 2023

to a performance degradation due to data sparsity and low quality annotations. While the results of this work do not provide good quality translation models yet, we expect to contribute to the development of NMT systems for LRLs and to inspire further research. We integrate our best-performing system for Wayúunaiki to Spanish into the document translation interface *TransIns*[2] (Steffen and van Genabith, 2021) for public use. Our collected supplementary material, the new general domain test data set, as well as code are also publicly available.[3]

2 Related Work

Various ways of incorporating linguistic knowledge into NMT systems have been explored. These include the addition of (linguistic) factors (e.g., Sennrich and Haddow (2016), España-Bonet and van Genabith (2018), Manzanares, 2020), or using different subword segmentation techniques (e.g., Sennrich et al. (2016), Kudo and Richardson (2018) Grönroos et al., 2014) with the aim of improving translation quality. Improvements are possible, especially in LRL scenarios (e.g., Sennrich and Zhang, 2019), morphologically rich languages (e.g., Ortega et al., 2020), and for out-of-domain texts (e.g., Chen and Fazio, 2021).

Subword segmentation is essential in NMT since it eases the out-of-vocabulary (OOV) problem and allows training smaller models (Mielke et al., 2021). Subword units offer a representation, that builds a bridge between word and character-level, based on the statistical properties of the text. A good choice of subword units will offer a good balance between the vocabulary size, the size of the model and therefore the decoding efficiency.

Data-driven, unsupervised subword segmentation is a statistically-informed process that incorporates implicit linguistic knowledge present in the text, like statistical patterns that present regularities of encountered word forms. This approach is limited to the data used during training the segmentation model, such that text variations (e.g., inconsistent orthography or out-of domain context) might result in segmentation variations and over-segmentation (Amrhein and Sennrich, 2021).

The *Byte-Pair-Encoding* (BPE) algorithm (Gage, 1994) is a widely used, unsupervised approach for subword segmentation. BPE merges the most frequent pairs of characters in a corpus to create a new subunit, and repeats the process until the desired number of merge operations are performed. With BPE, common words form a single unit while rare words are split into subunits. The first application in MT by Sennrich et al. (2016) lead to a strong improvement in performance. Further approaches include SentencePiece (Kudo and Richardson, 2018), a tokeniser that implements both BPE and *unigram language model* (LM) (Kudo, 2018). In Kudo (2018) subword segmentation is combined with a regularization method, offering a robust alternative to the deterministic BPE. For the segmentation technique by Kudo (2018), an initial subword set is pruned, according to the contribution of each subword to the unigram LM (Mielke et al., 2021). Another alternative for creating more segmentation variety in the training data is the regularization method particularly for BPE called BPE-dropout (Provilkov et al., 2020).

Semi-supervised segmentation techniques incorporate and exploit linguistically labeled training data to guide the segmentation process. Linguistic annotation can help to learn the correct segmentation rules, especially in low quantity and quality data scenarios (Chen and Fazio, 2021).

The semi-supervised segmentation technique, *Prefix-Root-Postfix-Encoding* (PRPE) by Zuters et al. (2018) is a morphologically guided algorithm, that incorporates linguistic knowledge without requiring any morphological rules. Nonetheless, a list of affixes is essential during the construction of the segmenter. In comparison to the BPE algorithm, subwords that include positional information of a word are extracted in form of prefixes, roots, and postfixes. This subword segmentation algorithm has been shown to improve translation quality, measured with BLEU, in comparison to other systems, in which unsupervised algorithms were applied (Chen and Fazio, 2021). The algorithm is not thought to be used as a morphological segmentation tool, even though it produces text that resembles morphologically segmented text. Moreover, it avoids over-segmentation by sometimes only partially performing the morphological splitting with the motivation that too many subwords would reduce the translation quality (Zuters et al., 2018).

FlatCat (Grönroos et al., 2014) is a variant of the toolkit Morfessor (Smit et al., 2014) for statistical morphological segmentation which can be

[2] https://transins.dfki.de

[3] https://github.com/norgrai/wayuunaiki

applied in an unsupervised or semi-supervised manner. The system consists of a category-based hidden Markov model (HMM) and a flat lexicon structure for morphological segmentation. The states of the HMM are the morph categories (prefix, stem, suffix, and non-morphs, with the last category catching subwords that are not proper morphemes but segments of a longer morph). Morfessor FlatCat is best suited for semi-supervised training where some morphological splitting guidelines are given; in fully unsupervised training with no annotations over-segmentation or under-segmentation will probably occur (Grönroos et al., 2014). Zuters et al. (2018), in their comparison between PRPE and Morfessor FlatCat, acknowledge previous, small improvements using Morfessor for inflected languages in statistical MT, but these improvements are not reproduced in their experiments.

Sennrich and Haddow (2016) were one of the first to introduce **linguistic factors** like lemmas, part-of-speech (POS) tags, dependency labels, and morphological features as factors into an NMT model.[4] The additional linguistic information is coupled with each subword by concatenating or averaging the embeddings. As their main objective was reducing data sparsity, they tested the factored architecture on high and LRL pairs, obtaining significant translation improvements in BLEU for the model with all factors included, for both high and low resource scenarios. In their experiments, the best results with only one factor were achieved with a POS tag or lemma factor in a RNN encoder–decoder architecture with attention for English to German translation. Similar performance for lemma factors was observed by Armengol-Estapé et al. (2021) with the Transformer architecture (Vaswani et al., 2017). By adding a lemma factor to the subwords, different inflections of a words are linked to the same representation. By introducing POS tags, it is possible to discriminate between different word categories, that share the same surface word.

Word embeddings capture both semantic knowledge (Mikolov et al., 2013; Brunila and LaViolette, 2022) and, to a lower extent, syntactic knowledge (Mikolov et al., 2013; Andreas and Klein, 2014). Syntax is more evident in embeddings when the training data is scarce (Andreas and Klein, 2014). Qi et al. (2018) showed that leveraging pretrained word embeddings can lead to significant improvements for certain LRL pairs. However, Qi et al. (2018) use of pretrained embeddings by Bojanowski et al. (2017) limits the scope of the comparison, since only a few Indigenous languages, such as Quechua, have access to such rich representations or have sufficient data available for training them.

According to Fernandez et al. (2013), there were very few projects that involve the development of a translator for Indigenous languages in Colombia such as Wayúunaiki. At the same time Llerena García (2013) presented the reasons and need for a "Software traductor de español a lengua wayuu" (*Spanish to Wayúu language translator software*). Unfortunately, to the best of our knowledge, even now, 10 years after Fernandez et al. (2013) and Llerena García (2013), there still exists no publicly accessible translation system, that supports the Wayúu community.

3 Language Description

Wayúunaiki is the native language spoken by a minority (compared to Spanish) in the Wayúu community, located in the Caribbean region, connecting Colombia and Venezuela. More than half a million people of this bi-national community speak this LRL. The Wayúu community is the most numerous indigenous community in Colombia (Departamento Administrativo Nacional de Estadística, 2021). There are 380,460 Wayúus in Colombia[5] and about 415,500 Wayúus in Venezuela (INE, 2012).

Wayúunaiki belongs linguistically to the Arawak languages. This language family flourished among ancient, indigenous nations in South America and consists of polysynthetic, mainly head-marking languages with different degrees of agglutination (Méndez-Rivera, 2020). Spanish, the high-resourced language spoken in the same countries, is a fusional, inflected language with a flexible syntactic order. The preferred pattern is subject + verb + object (SVO), while Wayúunaiki has a VSO order. Both languages have their own phonological system and do not share the same alphabet: Spanish has 22 consonants and 5 vowels in its phonological repertoire, while Wayúunaiki has 16 consonant and

[4]Linguistic information was earlier introduced by Alexandrescu and Kirchhoff (2006) in a neural NLP model.

[5]According to the latest census information: the *Censo Nacional de Población y Vivienda* (CNPV) was conducted in 2018 by the National Administrative Department of Statistics (DANE).

data set	# of samples	tokens		TTR	
		esp	guc	esp	guc
train	41499	776k	591k	0.029	0.048
development	1001	18.7k	14.0k	0.175	0.220
in-domain test set	1001	18.7k	14.2k	0.181	0.219
Total	43501	814k	620k	0.028	0.047
additional data:					
out-of-domain test	1107	15.1k	10.6k	0.203	0.360

Table 1: Description of the bitext data sets: number of samples, words, and type-token-ratio (TTR) for the Wayúunaiki (guc) and Spanish (esp) data set from the Tatoeba MT Challenge with our partitions, and the additional, manually collected data.

12 vowel phonemes —6 vowel pairs of long and short ones (Viloria Rodríguez et al., 2022). An inconsistent writing system for the Wayúu language, due to the two main "official" orthographic systems, in combination with a very small amount of written material in Wayúunaiki, make the orthographic situation challenging (Álvarez, 2017).[6]

4 Data Collection and Preprocessing

Parallel corpora. We use the only online parallel corpus for Wayúunaiki and Spanish available in the Tatoeba MT Challenge, version v2021-08-07 (Tiedemann, 2020). The bitext is a subpart of the no longer available JW300, a parallel corpus from Agić and Vulić (2019) with religious-themed data, addressing a wider range of topics including bible psalms.[7] The Wayúunaiki part of the bitext follows the official writing norm ALIV (Alfabeto de Lenguas Indígenas de Venezuela, *alphabet of indigenous languages of Venezuela*). The corpus consists of ∼43k sentence pairs, which we divided into a train, development, and test set. Table 1 gives a summary of the parallel corpora utilized.

The usage of highly domain-specific (here religious) data limits the translation quality in other domains and when used for other domains introduces a strong ideological, and gender-related bias, given the biblical content: gender pronouns and person names do not appear in the data with a balanced frequency,[8] nor do they share a similar word context, regarding activities or occupations (Storks et al., 2019). Furthermore, we asked two native Wayúunaiki speakers to perform a revision of random Wayúu sentences in the Tatoeba corpora. The revision showed the low quality of the resource. Some sentences are not direct translations and miss important information. In the example below, the personal name (Margaret) is absent in the Wayúunaiki sentence (a), but given in the official translation (b). According to bilingual Spanish and Wayúunaiki speakers, the correct translation would be (c).

(a) Sü'lakajaaka pireewa sümaa saatsa aainjuushi süka keesü nayaalu'u na süikeyuukana süka shiain nekaajüin ma'in.

(b) Margaret trajo la comida y la puso en el centro de la mesa, donde estaban todos sentados.
Margaret brought the food and put it in the center of the table, where everyone was sitting.

(c) Nos cocinaron fideos en salsa con queso porque es la comida que comen ellos.
They cooked us noodles in sauce with cheese because that's the food they eat.

In order to create a general domain parallel data set and assess the generalizability of the translation systems, we collected data from Spanish–Wayúunaiki dictionaries and illustrative grammar booklets for non-Wayúunaiki speakers to learn the language. Table 2 shows the number of samples and sources we used to build the general domain test set.

source	# of samples	parallel sentences
Lozano R. and Mejía V. (2007)	402	yes & aligned text
Álvarez (2016)	211	yes
Álvarez (2011)	425 69	yes aligned text
Total:	1107	

Table 2: Description of out-of-domain data set, collected bitext for Spanish–Wayúunaiki.

Monolingual corpora Table 3 lists the details of the monolingual data we collected. We extracted Wayúunaiki text from the translation[9] of

source	language	# of samples, tokens		language unit
de Saint-Exupéry et al. (2016)	guc	1933	19.5k	sentence
David M. Captain (2005)	guc	3177	3.2k	word
Total:	5.1k units			
WikiDump (Wikipedia, 2020)	esp	29.02M	597M	sentence

Table 3: Description of monolingual data in Wayúunaiki (guc) and Spanish (esp).

[6]Since 1984, the official *Alfabeto de Lenguas Indígenas de Venezuela*, the alphabet of indigenous languages of Venezuela has been the norm in Colombia and Venezuela, but the system of Miguel Ángel Jusayú is being utilized alongside.

[7]The web-crawled data stems from the website jw.org of a religious society, covering many low-resource languages. Aside from the Bible, the Jehovah's Witnesses provide magazines, books, and other multi-media content.

[8]For instance, the female pronoun *ella* occurs less than one-fourth of the times the male pronoun *él* occurs.

[9]https://www.academia.edu/37583043/Pürinsipechonkai

the book *The Little Prince* by Antoine de Saint-Exupery. This corpus is used as monolingual data, since it does not align at sentence level with the Spanish version. We also extract from a bilingual Spanish–Wayúunaiki dictionary (David M. Captain, 2005) entries in Wayúunaiki, which we used, one token per line, as additional data. The Wayúu data follows the the official writing norm ALIV. For Spanish, we use a subset of 10M sentences from the Spanish Wikipedia dump from May 2020 (Wikipedia, 2020) extracted with *WikiTailor* (España-Bonet et al., 2023). Notice the data asymmetry between Wayúunaiki and Spanish. While we obtain 5000 sentences in Wayúunaiki, the Spanish Wikipedia alone has almost 30M sentences. This reflects the typical data imbalance between high- and low-resourced languages.

The monolingual corpus is used in our work combined with the monolingual parts of the parallel corpus to train word embeddings.

Supplementary Material Some of our experiments require supplementary information in the form of linguistic annotations, or dictionaries. We extracted morphological analyses of verb conjugations in Wayúunaiki from the work of Álvarez (2017) to guide the semi-supervised training of the segmentation models (Prefix-Root-Postfix-Encoding and FlatCat). For this, the morph categories prefix, stem, and suffix were manually annotated. An example file is listed in Appendix A and we make all files available online.[10] We perform a similar morphological annotation with Spanish samples taken from lecture slides from Doctor Lluís Simarro Lacabra (2014), an educational institution.

Preprocessing We split the monolingual text into sentences and tokens using the *nltk* tokenizer. Since there is no tokenizer for Wayúunaiki, we use regular expressions (RE). The character ’ in Wayúunaiki, which in the Latin alphabet represents the glottal stop consonant [ʔ] known as "saltillo", *little skip*, had to be stripped from additional white spaces. For simplification, all possible saltillos (‘ ′ ’ ´) were mapped to the ' character in the parallel data sets. Likewise, quotations (» « “ ”) were normalized to ". Bible verses number references were detected with REs and removed. Enumerations with brackets, numbers with punctuation at the beginning of the sentence, and URLs were also removed. We train a truecaser with Moses scripts (Koehn et al., 2007) for each language on the parallel data and applied them to all data sets accordingly.

5 NMT Systems

All our models are based on a transformer architecture (Vaswani et al., 2017) and developed with Marian v1.11.0 (Junczys-Dowmunt et al., 2018).

5.1 Baseline System

We perform a wide hyperparameter search on a transformer following van Biljon et al. (2020) (see Appendix B for the parameters, the ranges we explore and the best configuration). With the gained insights from the random search, we chose the configuration of the most promising model, a small transformer model with 3 encoder, 3 decoder layers, 4 heads and hidden layers with a size of 1024, and use it in all systems.

We train a baseline system on unsegmented data without (**BASE**) and with (**BASE+EMB**) pretrained embeddings. The embeddings for each language are trained independently with *fastText* (Bojanowski et al., 2017) on the preprocessed, unsegmented monolingual text, using the continuous skip-gram model (Mikolov et al., 2010). In our experiments, the model achieved the best results with embeddings that have a dimension of 256.

5.2 Subword Segmentation Techniques

We investigate different subword segmentation algorithms and apply them separately for each language: BPE without (**SUBW-bpe**) and with applied dropout (**SUBW-dp**), a unigram LM (**SUBW-uni**) for segmentation, PRPE (**SUBW-prpe**), and Morfessor FlatCat (**SUBW-fc**).

For SUBW-bpe, we explore both the impact of separate and joint vocabulary, and of different vocabulary sizes, using the *subword-nmt* toolkit (Sennrich et al., 2016). The chosen merge operations range from 100 to 15000 merges. According to the results (detailed numbers in Appendix C), we use for SUBW-bpe with 4k merge operations with separate vocabularies if not stated otherwise.

Reported models with pretrained embeddings (SUBW-bpe+EMB) are trained with *fastText* like the ones for the baseline but with segmented monolingual text.

[10] https://github.com/norgrai/wayuunaiki

5.3 Factored Models

We investigate factored models, where POS tag information is injected. Since an NLP tool for POS tagging or lemmatization in Wayúunaiki is not available, we adapt Spanish–Wayúunaiki dictionaries into linguistic knowledge-based vocabularies: Wayúu vocabulary entries were annotated with the Spanish translation and POS tag to represent implicit linguistic information. We use a bilingual dictionary from the Apertium (Forcada et al., 2011) GitHub[11] and an illustrated dictionary from David M. Captain (2005). We match their different POS tag annotations for Wayúu with the POS tag categories of the *FreeLing* analyzer (Padró and Stanilovsky, 2012) for Spanish.[12]

Approximately 40% of the Wayúu training data could be annotated in this way, mostly due to annotation of the closed class "punctuation" with makes up about 15% of the tokens. The high number of unclassified words is mainly due to the lack of a lemmatizer: only dictionary entries can be looked up automatically, so most tokens with inflectional and derivational variation cannot be matched with their corresponding POS tag. This stands in stark contrast to the annotation with *FreeLing* for Spanish, where much more fine-grained classes were used and every word is assigned a POS tag.

5.4 Evaluation

For the automatic evaluation, we use SacreBLEU (Post, 2018) to calculate BLEU[13] (Papineni et al., 2002) and chrF2++[14] (Popović, 2015). As semantic metric we use BLEURT[15] (Sellam et al., 2020) and for all cases, we estimate 95% confidence intervals via bootstrap resampling (Koehn et al., 2003) with 1000 samples.

Since the surface-based n-gram scoring methods can strongly restrict the expressiveness of agglutinative languages like Wayúunaiki, we also include example model translations for a qualitative manual comparison.

[11] https://github.com/apertium/apertium-guc-spa

[12] See the detailed resulting alignments among languages and the percentage of categories in our training data in Appendix A.

[13] BLEU|nrefs:1|bs:1000|seed:12345|case:mixed|eff:no |tok:13a|smooth:exp|version:2.3.1

[14] chrF2++|nrefs:1|bs:1000|seed:12345|case:mixed|eff:yes |nc:6|nw:0|space:no|version:2.3.1

[15] BLEURT v0.0.2 using checkpoint BLEURT-20

6 Results and Discussion

We report the translation scores for Wayúunaiki to Spanish in Tables 4 (religious domain) and 5 (general domain) for each method with the best system per metric boldfaced. In Table 6 we report translation results for Spanish to Wayúunaiki for the most representative systems (the best segmentation approach together with a factored and a pretrained embeddings model).

Model Architecture. van Biljon et al. (2020) demonstrated improvements for translating English text into agglutinative LRLs with a transformer by halving the model's depth to 3 encoder and 3 decoder layers. We obtain the same conclusion from the hyperparameter search for translating from and into Wayúunaiki. Our BASE model is also a small transformer with 3 encoder and 3 decoder layers but Wayúunaiki–Spanish turns out to be a challenging language pair with baseline translation quality close to zero.

Pretrained embeddings alone do not significantly improve the results (BASE+EMB, SUBW-bpe+EMB), although they have been shown to provide a better representation of less frequent concepts in LRLs (Haddow et al., 2022). Qi et al. (2018) showed that pretrained embeddings seem to be effective for not-too-distant translation pairs. This may well be the reason for our lack of improvement, Wayúunaiki and Spanish are very distant, but we conjecture that the most important problem we face is the lack of sufficient data to train Wayúu embeddings: the monolingual Wayúu corpus we use is almost equivalent to the size of the parallel corpus. Still the results of Qi et al. (2018) indicate that pretrained embeddings seem to introduce semantic and syntactic information of words improving translations even for distant translation pairs: systems are able to capture overall basic language characteristics and generate more grammatically well-formed sentences. Qi et al. (2018) indicate that for very little but sufficient training data, that allows training the system, using pretrained word embeddings from (Bojanowski et al., 2017) are most effective. Their usage of pretrained embeddings by Bojanowski et al. (2017) make comparison with our results very difficult, as such embeddings are trained on billions of tokens.

Notice that our BASE systems trained on unsegmented data are well below any subword segmentation we apply. This contradicts the conclusions for Quechua-Spanish in Chen and Fazio (2021):

model guc-esp	BLEU	chrF2	BLEURT
BASE	0.5 ± 0.2	6.0 ± 0.3	0.17 ± 0.01
BASE+EMB	0.7 ± 0.2	11.8 ± 0.4	0.094 ± 0.007
SUBW-bpe	4.2 ± 0.7	20.5 ± 0.8	0.21 ± 0.01
SUBW-dp	3.1 ± 0.5	16.7 ± 0.8	**0.22 ± 0.01**
SUBW-uni	3.3 ± 0.6	**22.0 ± 0.7**	0.20 ± 0.01
SUBW-prpe	1.0 ± 0.3	7.0 ± 0.3	0.15 ± 0.01
SUBW-fc	**4.5 ± 0.8**	21.0 ± 0.8	0.21 ± 0.01
SUBW-bpe+			
+FACT	1.0 ± 0.2	8.9 ± 0.4	0.127 ± 0.006
+EMB	0.6 ± 0.2	7.9 ± 0.3	0.090 ± 0.005
+FACT+EMB	0.8 ± 0.2	13.6 ± 0.4	0.115 ± 0.007

Table 4: Automatic evaluation scores of the **Wayúunaiki to Spanish** translations with the religious **in-domain** test set.

model guc-esp	BLEU	chrF2	BLEURT
BASE	0.08 ± 0.04	4.8 ± 0.3	0.106 ± 0.006
BASE+EMB	0.06 ± 0.03	8.8 ± 0.6	0.048 ± 0.004
SUBW-bpe	**0.20 ± 0.10**	13.2 ± 0.9	0.075 ± 0.006
SUBW-dp	0.14 ± 0.08	8.8 ± 0.7	**0.132 ± 0.006**
SUBW-uni	0.16 ± 0.08	13.8 ± 0.9	0.070 ± 0.005
SUBW-prpe	0.11 ± 0.08	4.5 ± 0.3	0.104 ± 0.006
SUBW-fc	0.12 ± 0.03	**14.0 ± 0.8**	0.067 ± 0.005
SUBW-bpe+			
+FACT	0.07 ± 0.02	6.5 ± 0.5	0.082 ± 0.004
+EMB	0.07 ± 0.03	6.8 ± 0.6	0.067 ± 0.004
+FACT+EMB	0.03 ± 0.01	9.6 ± 0.6	0.059 ± 0.005

Table 5: Automatic evaluation scores of the **Wayúunaiki to Spanish** translations with the **general domain** test set.

in an out-of-domain evaluation their model outperformed all of their systems trained with different segmentation methods (e.g., BPE, unigram LM, PRPE).

Segmentation technique. Although all segmentation methods yield a statistically significant improvement over the baseline, the scores both on the general and in-domain test set emphasize that models do not provide good or even reasonable quality translation yet. Notice also that no single model outperforms other models in all automatic evaluation metrics.

While the results show some potential of Morfessor Flatcat to be used as a segmentation technique,[16] the need to tune additional parameters (perplexity threshold and weight) make the approach more complex and provide no statistically significant improvements with respect to the most straightforward SUBW-bpe. We therefore use SUBW-bpe in our factored models.

The unigram LM subword segmentation method of SentencePiece, used in many NLP systems (Mielke et al., 2021), offers a non-deterministic alternative, though with the SUBW-uni model for the first time in our experiments we observe subwords that are ungrammatical. For instance, the verbs gobernar (Eng: rule) in the reference (2) and the translation (4), which has an incorrect duplication of the character "r":

(1) mapa, kettaapa tü miit juya Nuluwataainjachikalü o'u, nüle'ejireerü tü aluwataayakat nümüin chi nüshikai .

(2) después de gobernar como rey por mil años , le devolverá el reino a su padre .
and after ruling as king for a thousand years, he will return the kingdom to his father

(3) finalmente , cuando llegue el día de su vida , comenzó a goberna**rr**se con él .
finally, when the day of his life came, he began to govern himself with it .

Observed word repetitions and hallucinations in SUBW-uni or SUBW-dp suggest that the training is still not optimized. The following examples are common translation outputs (they appear several times with diffent and unrelated source sentences) for general domain inputs unrelated to the Bible:

(a) la biblia dice : " el nombre de Jehová
the bible says : " the name of Jehovah

(b) Jesús dijo : " tú , tú , tú ,
Jesus said: " you, you, you,

Fu et al. (2020) argue that the repetition problem is the expression of human language itself: words that produce high probabilities tend to be chosen as the subsequent word again, constructing prediction loops, which result in repetitions. We observe single-word repetitions; however, word pair repetitions are more common, exemplified with "tú" and "," in (b).

Similar to the findings of Raunak et al. (2021), we encounter fluent but "detached", and non-grammatical translation outputs with repetitive structure of hallucinations. The investigation of Lee et al. (2018) on hallucinations with a medium-sized corpus (4.5M training sentences) let them conclude that the noisy and finite characteristics of the data sets are the source for the phenomenon. They propose data augmentation as the

[16] Zuters et al. (2018) introduced a method of segmentation post-processing to control the effective vocabulary size and support an open vocabulary: they performed the Morfessor subword segmentation in an unsupervised fashion on the data on which they applied additionally the BPE algorithm. We tried out this approach but could not achieve comparable results to the reported SUBW-fc.

model esp-guc	BLEU	chrF2	BLEURT
religious domain:			
SUBW-bpe+	1.2 ± 0.3	13.9 ± 0.4	0.239 ± 0.007
+FACT	0.7 ± 0.2	10.7 ± 0.4	0.240 ± 0.008
+EMB	0.5 ± 0.1	17.1 ± 0.6	0.255 ± 0.008
+FACT+EMB	0.7 ± 0.2	19.3 ± 0.6	0.252 ± 0.008
general domain:			
SUBW-bpe+	0.10 ± 0.06	11.3 ± 0.5	0.205 ± 0.007
+FACT	0.06 ± 0.01	9.9 ± 0.5	0.212 ± 0.007
+EMB	0.01 ± 0.01	9.3 ± 0.7	0.232 ± 0.007
+FACT+EMB	0.02 ± 0.00	13.0 ± 0.8	0.228 ± 0.005

Table 6: Automatic evaluation scores of the **Spanish to Wayúunaiki** translations with the religious **in-domain test** set (top rows) and the **general domain** test set (bottom rows).

most promising approach for preventing hallucinations. Still, their techniques require knowledge of hallucinations and exhaustive filtering of the training data. Similar conclusions are made by Raunak et al. (2021); furthermore, they emphasize that invalid or misaligned sentence pairs that do not provide accurate translations should be removed.

Although the overall scores are very low, we find that introduced linguistic knowledge in the shape of linguistically inspired morphs helps the system to better accomplish the translation task. Yet, the segmentation has to be carried out invariably: one possible explanation for the qualitatively lower translations of the models with applied BPE Dropout or the SentencePiece unigram LM is the statistical noise introduced in the segmentation process, being both non-deterministic segmentations contrary to the BPE algorithm.

Linguistic Factors and Embeddings. The performance of the +FACT methods is worse than the original SUBW-bpe. The same happens when adding pretrained word embeddings (+EMB). The introduced linguistic information in the shape of POS tags, pretrained embeddings, and the combination of both does not help to overcome the difficulties of this LRL translation pair. The main reason is the low coverage for Wayúunaiki, both in the amount of data to train the embeddings and therefore their quality, and in POS annotations as explained in Section 5.3.

It is generally acknowledged that introducing linguistic factors coupled with a word or its subwords improves translation quality only to a modest extent (Sennrich and Haddow, 2016). Hence, for language pairs in a high resource setting, it is not advisable to invest time and effort in a factored NMT approach (Casas et al., 2021). Still, in an LRL setting that possibly involves morphologically rich languages, the data sparsity problem can be eased by converting the plain parallel text into a factored representation on the source side.

Translation quality should not be evaluated only automatically though, as low scores are difficult to compare and different metrics show different trends (see their correlations in Appendix C). No single model outperforms all of the others in Table 4 measured across all three metrics. Although none of the proposed models achieved a higher BLEU score than SUBW-bpe for translating into Wayúunaiki in Table 6, the chrF2 score indicates improvements (± 3.2), which we verified by manually examining example translations, e.g., (2) and (3).

(1) **input**: hablémosle sin prisas .
let's talk to him without haste .

(2) **SUBW-bpe+EMB**: püküja **nümüin** tü alatakat **nümüin** .
. .] .] *tell those who cut for him . . .] .]*

(3) **SUBW-bpe**: shia süka tü kee'ireekat paa'in .
this is what you want .

(4) **reference**:
nnojoishii ashapajaainjanain waya waashajaapa **nümaa** .

7 Conclusion and Future Work

In this work we applied various unsupervised and semisupervised subword segmentation methods to enrich the data used to train a transformer-based NMT model with linguistic information. Additionally, we extended the architecture of the standard SUBW-bpe model by adding linguistic information in the form of POS tag factors and/or supplying the system with pretrained embeddings. In line with previous research on Indigenous LRL pairs that include Spanish, we observed that the addition of subword information is crucial to improve translation quality (e.g., Ortega et al. (2020), Mager et al. (2021), Chen and Fazio, 2021). In particular, the Indigenous languages of America, which are mostly characterized by a rich morphology, and part of agglutinative and polysynthetic languages, benefit from approaches that consider the LRL's morphology and apply subword segmentation techniques that are suitable for the language pair. In contrast, we did not achieve any improvement with factors and pretrained embeddings. The lack of resources, in terms of data and annotation coverage, is the likely cause for the low performance of these models.

Our next steps are focused on investigating the effectiveness of injecting linguistic knowledge for the Wayúu language by exploring datasets without repetitive sequences and less sparse and noisy annotations. To do this, more sophisticated approaches to obtain implicit linguistic knowledge from LRL text, such as introducing linguistic information also on the target side in the form of POS-tag or lemma factors are possible.

Problems related to the lack of resources for factored training could in principle be overcome by applying a linguistically inspired subword segmentation technique, for instance, Morfessor's FlatCat. By splitting a word into its subwords, chances of determining the stem are higher, if the segmentation into subwords representing stems and suffixes is both accurate and consistent. Given the stem, the word can be annotated with its POS tag from the linguistic knowledge-based vocabulary. We note that this is limited to languages without infixation and would work only for words without assimilatory processes between affixes and stem. Still, it presents a possible approach to obtain labeled data.

Besides enriching the data with linguistic information, our observations on word repetitions and hallucinations indicate that additional cleaning, filtering of unaligned source and target translations, and orthographic normalization could significantly enhance data quality and hence translation performance.

We believe that injecting linguistic information, especially for LRL pairs can alleviate the data sparsity problem and aid the models with the annotation of implicit linguistic knowledge present in the data. By enriching the data to represent such information present in the text (e.g., annotating POS tags), a model can better identify patterns inherent in the data. Still, choosing between the different approaches and techniques requires taking into account the nature of the LRL pair and the available resources, particularly supported NMT tools and data sets.

Limitations

In this work we explored transfer learning approaches only by using pretrained word embeddings. Transfer learning should be explored further. Some of the segmentation methods have their own hyperparameters which are usually obtained for high-resourced languages and might be suboptimal in our case. These hyperparameters should be systematically explored. Finally, token-free pretrained models fine-tuned on our data should be investigated.

It is costly and difficult to acquire human translations, due to the limited number of speakers and exclusive LRL communities; moreover, the fact that we are not Wayúunaiki speakers limited our qualitative assessment.

Acknowledgements

Thanks to two supportive natives of the Wayúu community, we were able to analyze our translation results beyond automatic evaluation scores. Both bilingual speakers are non-professional interpreters and translators that helped voluntarily. Adolfo y señora Gladys: ¡Muchas gracias por su ayuda con las traducciones, interés y confianza en el proyecto! We thank Jörg Steffen for the integration of the Wayúunaiki–Spanish system in TransIns.

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.

Andrei Alexandrescu and Katrin Kirchhoff. 2006. Factored neural language models. In *Proceedings of the Human Language Technology Conference of the NAACL, Companion Volume: Short Papers*, pages 1–4, New York City, USA. Association for Computational Linguistics.

Chantal Amrhein and Rico Sennrich. 2021. How suitable are subword segmentation strategies for translating non-concatenative morphology? In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 689–705, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Jacob Andreas and Dan Klein. 2014. How much do word embeddings encode about syntax? In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 822–827, Baltimore, Maryland. Association for Computational Linguistics.

Jordi Armengol-Estapé, Marta R. Costa-jussà, and Carlos Escolano. 2021. Enriching the transformer with linguistic factors for low-resource machine translation. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 73–78, Held Online. INCOMA Ltd.

Emily Bender. 2019. The #benderrule: On naming the languages we study and why it matters. *The Gradient*.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching Word Vectors with Subword Information. volume 5, pages 135–146.

Mikael Brunila and Jack LaViolette. 2022. What company do words keep? revisiting the distributional semantics of J.R. firth & zellig Harris. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4403–4417, Seattle, United States. Association for Computational Linguistics.

Noe Casas, Jose A. R. Fonollosa, and Marta R. Costa-jussà. 2021. Sparsely factored neural machine translation. *CoRR*, abs/2102.08934.

William Chen and Brett Fazio. 2021. Morphologically-guided segmentation for translation of agglutinative low-resource languages. In *Proceedings of the 4th Workshop on Technologies for MT of Low Resource Languages (LoResMT2021)*, pages 20–31, Virtual. Association for Machine Translation in the Americas.

Linda B. Captain David M. Captain. 2005. *Diccionario Básico Ilustrado, wayuunaiki-español español-wayuunaiki*. Editorial Fundación para el Desarrollo de los Pueblos Marginados, Editorial Buena Semilla.

A. de Saint-Exupéry, José Álvarez, and Jean-Marc Probst Foundation. 2016. *The Little Prince*. The Jean-Marc Probst Foundation.

DANE Departamento Administrativo Nacional de Estadística. 2021. *Información sociodemográfica del pueblo Wayúu*. Number 2805-6345 in Informes de Estadística Sociodemográfica Aplicada. DANE Colombia.

Shuoyang Ding, Adithya Renduchintala, and Kevin Duh. 2019. A call for prudent choice of subword merge operations in neural machine translation. In *Proceedings of Machine Translation Summit XVII: Research Track*, pages 204–213, Dublin, Ireland. European Association for Machine Translation.

L'Institut d'Educació Secundària Doctor Lluís Simarro Lacabra. 2014. Análisis morfológico de la palabra. lengua castellana y literatura. Lecture slides http://Análisis-morfológico-de-la-palabra.pdf.

Cristina España-Bonet, Alberto Barrón-Cedeño, and Lluís Màrquez. 2023. Tailoring and Evaluating the Wikipedia for in-Domain Comparable Corpora Extraction. *Knowledge and Information Systems*, pages 1365–1397.

Cristina España-Bonet and Josef van Genabith. 2018. Multilingual Semantic Networks for Data-driven Interlingua Seq2Seq Systems. In *Proceedings of the LREC 2018 MLP-MomenT Workshop*, pages 8–13, Miyazaki, Japan.

Dayana Fernandez, Jose Atencia, Ornela Gamboa, and Óscar Bedoya. 2013. Design and implementation of an "Web API" for the automatic translation Colombia's language pairs: Spanish-Wayuunaiki case. pages 1–9.

Mikel L. Forcada, Mireia Ginestí-Rosell, Jacob Nordfalk, Jimmy O'Regan, Sergio Ortiz Rojas, Juan Antonio Pérez-Ortiz, Felipe Sánchez-Martínez, Gema Ramírez-Sánchez, and Francis M. Tyers. 2011. Apertium: a free/open-source platform for rule-based machine translation. *Machine Translation*, 25:127–144.

Zihao Fu, Wai Lam, Anthony Man-Cho So, and Bei Shi. 2020. A theoretical analysis of the repetition problem in text generation. In *AAAI Conference on Artificial Intelligence*.

Philip Gage. 1994. A new algorithm for data compression. *The C Users Journal archive*, 12:23–38.

Stig-Arne Grönroos, Sami Virpioja, Peter Smit, and Mikko Kurimo. 2014. Morfessor FlatCat: An HMM-based method for unsupervised and semi-supervised learning of morphology. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers*, pages 1177–1185, Dublin, Ireland. Dublin City University and Association for Computational Linguistics.

Barry Haddow, Rachel Bawden, Antonio Valerio Miceli Barone, Jindřich Helcl, and Alexandra Birch. 2022. Survey of low-resource machine translation. volume 48, pages 673–732, Cambridge, MA. MIT Press.

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Marcin Junczys-Dowmunt, Roman Grundkiewicz, Tomasz Dwojak, Hieu Hoang, Kenneth Heafield, Tom Neckermann, Frank Seide, Ulrich Germann, Alham Fikri Aji, Nikolay Bogoychev, André F. T. Martins, and Alexandra Birch. 2018. Marian: Fast neural machine translation in C++. In *Proceedings of ACL 2018, System Demonstrations*, pages 116–121, Melbourne, Australia. Association for Computational Linguistics.

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.

Philipp Koehn and Rebecca Knowles. 2017. Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver. Association for Computational Linguistics.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proceedings of the 2003 Human Language Technology Conference of the North American Chapter of the Association for Computational Linguistics*, pages 127–133.

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.

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.

Katherine Lee, Orhan Firat, Ashish Agarwal, Clara Fannjiang, and David Sussillo. 2018. Hallucinations in neural machine translation. In *Interpretability and Robustness in Audio, Speech, and Language (IRASL) Workshop, Conference on Neural Information Processing Systems (NeurIPS)*, Montreal, Canada.

Ernesto Llerena García. 2013. Software traductor de español a lengua wayuu. pages 353–356.

Jorge Lozano R. and Julián David' Mejía V. 2007. Wayuunkeera - cartilla trilingüe & cuaderno de actividades, wayuunaiki español english. Universidad Libre.

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.

Noé Casas Manzanares. 2020. *Injection of linguistic knowledge into neural text generation models*. Ph.D. thesis, Universitat Politècnica de Catalunya.

Sabrina Mielke, Zaid Alyafeai, Elizabeth Salesky, Colin Raffel, Manan Dey, Matthias Gallé, Arun Raja, Chenglei Si, Wilson Lee, Benoît Sagot, and Samson Tan. 2021. Between words and characters: A Brief History of Open-Vocabulary Modeling and Tokenization in NLP. *CoRR*, abs/2112.10508.

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

Tomas Mikolov, Martin Karafiát, Lukas Burget, Jan Cernockỳ, and Sanjeev Khudanpur. 2010. Recurrent neural network based language model. In *Interspeech*, volume 2, pages 1045–1048. Makuhari.

Nelson J. Méndez-Rivera. 2020. *Linguistic outcomes of the Wayuunaiki-Spanish Language contact situation*. Ph.D. thesis, University of Ottawa.

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.

Lluís Padró and Evgeny Stanilovsky. 2012. FreeLing 3.0: Towards wider multilinguality. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 2473–2479, Istanbul, Turkey. European Language Resources Association (ELRA).

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, page 311–318, USA. Association for Computational Linguistics.

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

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.

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.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 529–535, New Orleans, Louisiana. Association for Computational Linguistics.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183, Online. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. BLEURT: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892, Online. Association for Computational Linguistics.

Rico Sennrich and Barry Haddow. 2016. Linguistic input features improve neural machine translation. In *Proceedings of the First Conference on Machine Translation: Volume 1, Research Papers*, pages 83–91, Berlin, Germany. 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.

Rico Sennrich and Biao Zhang. 2019. Revisiting low-resource neural machine translation: A case study. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 211–221, Florence, Italy. Association for Computational Linguistics.

Peter Smit, Sami Virpioja, Stig-Arne Grönroos, and Mikko Kurimo. 2014. Morfessor 2.0: Toolkit for statistical morphological segmentation. Technical report, Gothenburg, Sweden.

Jörg Steffen and Josef van Genabith. 2021. TransIns: Document translation with markup reinsertion. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 28–34, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Shane Storks, Qiaozi Gao, and Joyce Yue Chai. 2019. Recent Advances in Natural Language Inference: A Survey of Benchmarks, Resources, and Approaches. *CoRR*, abs/1904.01172.

Jörg Tiedemann. 2020. The tatoeba translation challenge – realistic data sets for low resource and multilingual MT. In *Proceedings of the Fifth Conference on Machine Translation*, pages 1174–1182, Online. Association for Computational Linguistics.

Elan van Biljon, Arnu Pretorius, and Julia Kreutzer. 2020. On optimal transformer depth for low-resource language translation. *CoRR*, abs/2004.04418.

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 *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Leonel Viloria Rodríguez, Johan Yacomelo, Rudecindo González, and Daniela Segura. 2022. Pronombres en wayuunaiki y español; una mirada contrastiva. *Íkala, Revista de Lenguaje y Cultura*, 27.

Rui Wang, Xu Tan, Renqian Luo, Tao Qin, and Tie-Yan Liu. 2021. A survey on low-resource neural machine translation. pages 4636–4643.

Inc. Wikimedia Foundation Wikipedia. 2020. In *Wikimedia downloads.* [link].

Jānis Zuters, Gus Strazds, and Kārlis Immers. 2018. *Semi-automatic Quasi-morphological Word Segmentation for Neural Machine Translation: 13th International Baltic Conference, DBIS 2018, Trakai, Lithuania, July 1-4, 2018, Proceedings*, pages 289–301.

José Álvarez. 2011. Pütchimaajatü komputatoorachiki wayuunaikiru'usu. Diccionario de computación en wayuunaiki. Microsoft Venezuela.

José Álvarez. 2016. La conjugación del verbo en la lengua wayuu. Instituto Caro Y Cuervo.

José Álvarez. 2017. *Manual de la lengua wayuu, Karalouta atüjaaya saa'u wayuunaikikuwa'ipa*. Organización Indígena de La Guajira Yanama.

A Supplementary Material Annotation

A.1 Morph Categories

We manually annotate the morph categories prefix, stem, and suffix of 26 words in Wayúunaiki and 91 in Spanish for the Morfessor Flatcat approach. To perform Prefix-Root-Postfix-Encoding, we created two heuristics that contain the common suffixes, prefixes and endings for the Wayúu and Spanish languages. The example below shows 10 words annotated for Wayúunaiki.

Listing 1 Example annotations for Wayúunaiki used for semi-supervision in the Morfessor Flatcat (Grönroos et al., 2014) system. Morph categories are indicated by PRE (prefix), STM (stem), and SUF (suffix).

```
aya'lajaa a/PRE ya'laja/STM a/SUF
aya'lajeewaa a/PRE ya'laja/STM ee/SUF a/SUF
aya'lajiraa a/PRE ya'laja/STM ira/SUF a/SUF
aya'lajünaa a/PRE ya'laja/STM na/SUF a/SUF
apütüshi a/PRE pütü/STM shi/SUF
apütüichi a/PRE pütü/STM i/SUF chi/SUF
apütüeechi a/PRE pütü/STM ee/SUF chi/SUF
apütüinjachi a/PRE pütü/STM inja/SUF chi/SUF
apütüshijachi a/PRE pütü/STM shi/SUF ja/SUF chi/SUF
apütüichipa a/PRE pütü/STM i/SUF chi/SUF pa/SUF
```

A.2 POS Tagset Alignment

We summarize our alignment between the POS tags of the different sources in Wayúunaiki and the POS tag categories of the *FreeLing* analyzer for Spanish in Table 7. Due to different categorizations of some determiners, we replaced entries that were referring to the determiners as either adverb or pronoun in David M. Captain (2005) and mapped them uniformly to the POS tag D. About 80 references to another surface form of the same word were looked up and matched with their corresponding POS tag.

Spanish		Wayúunaiki	
class	abbr.	class	abbr.
adjective	**A**	(1)(2) adjetivo	adj.
conjunction	**C**	(1) conjunción	conj.
determiner	**D**	(3) determinante	det
punctuation	**F**	puntuación	punct.
pronoun	**P**	(1) pronombre	pron.
adverb	**R**	(1) adverbio	adv.
adposition	**S**	(1) posposición	posp.
		(2) Postposición	post.
verb	**V**	(1) verbo transitivo	v.t.
		(1) verbo intransitivo	v.i.
		(2) verbos	vblex
noun	**N**	(1) nombre	n
		(2) Alineable	ali.
		(2) Inalineable	ina.
interjection	**I**	(1) interjección	interj.
		(2) Interjeccion	ij

Table 7: Description of Tagset for Spanish (left): POS classes with the category and the abbreviation used. Alignment with the Wayúunaiki data (right): (1) refers to the dictionary in David M. Captain (2005), (2) Forcada et al. (2011), and (3) the manually extracted, closed classes in Lozano R. and Mejía V. (2007).

A.3 POS Tags Distribution

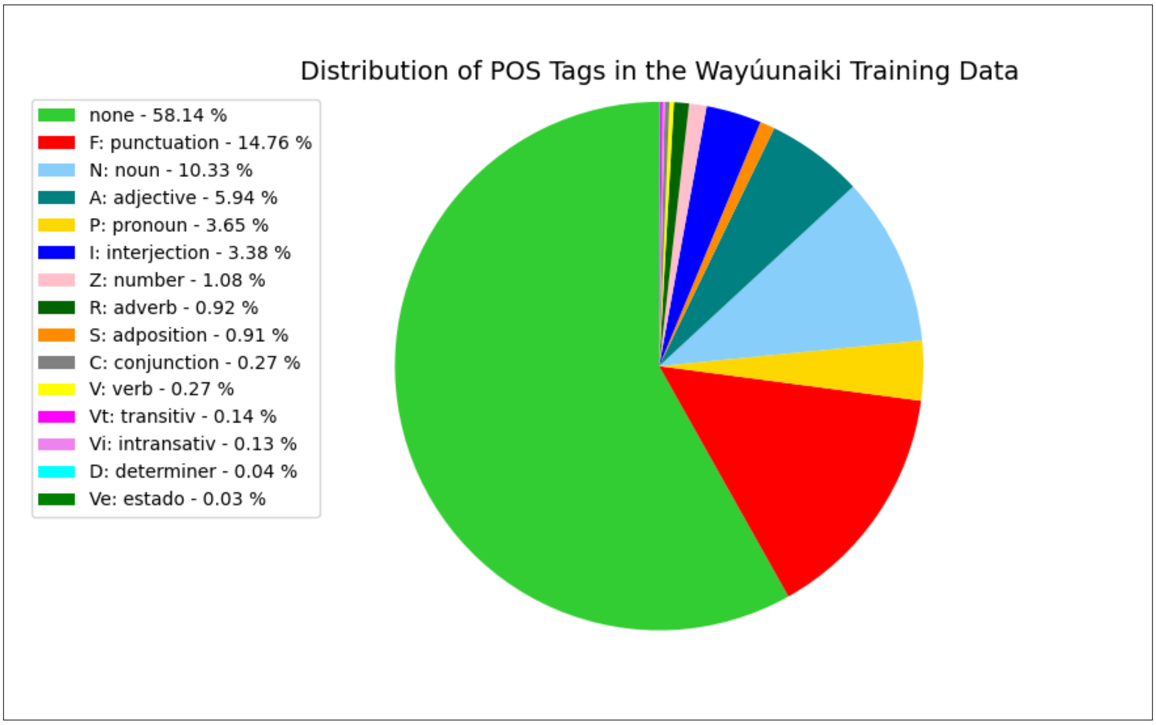

Figure 1: POS tags of the Wayúu training data, which we annotated based on linguistic knowledge-based vocabularies.

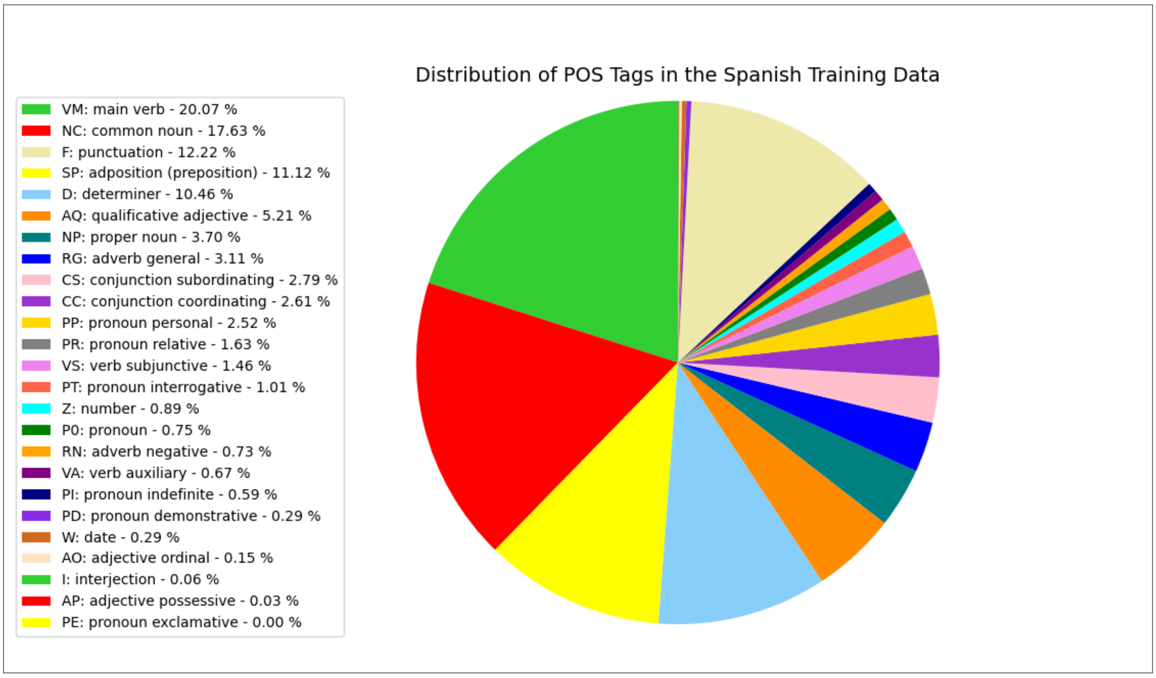

Figure 2: POS tags of the Spanish training data, annotated with *FreeLing* (Padró and Stanilovsky, 2012). We summarized the subclasses of determiner (D), numbers (Z), and punctuation (F) for representation purposes only.

B NMT Hyperparameter Exploration

Building upon findings from van Biljon et al. (2020), we explore different hyperparameters which are specially relevant in the LR scenario. Table 8 summarizes the hyperparameter space explored. Table 9 shows the best configuration that is used for the baseline system (BASE). Finally, we show the segmentation-related hyperparameters used for the segmented-based models (SUBW-*) in Table 10.

Hyperparameter	Values
# attention heads:	2, 4, 8
# of encoder/decoder layers:	2, 3, 4
embedding size:	256, 512, 1024
tied embeddings:	True, False
learning-rate:	1e-3, 1e-4 3e-4, 5e-4
warm-up steps:	1000, 4000
adam optimizer beta:	0.98, 0.999
label-smoothing:	0, 0.1, 0.2
layer-normalization:	True, False
train-position-embeddings:	True, False
exponential-smoothing:	0, 0.0001
clip-norm:	0, 1, 5
seeds:	0, 42, 1111

Table 8: Hyperparameters explored (as required by Marian software) with the corresponding values considered.

```
type: transformer
hidden layer size: 1024
embedding size: 256
tied embeddings: False
decoder depth: 3
encoder depth: 3
transformer heads: 4
transformer-dim-ffn: 1024
transformer-postprocess: da
transformer-preprocess: n
dropout - transformer: 0.3
        - ffn: 0.25
        - attention: 0
clip-norm: False
exponential-smoothing: 0
layer normalization: False
label smoothing: 0.1
learning-rate (lr): 3e-4
    lr-warmup: 1000
    lr-decay-inv-sqrt: 4000
optimizer (betas): adam (0.9, 0.999,1e-9)
seed: 42
early stopping patience: 15
beam size: 5
mini-batch-words: 1000
max-sentence length: 100
```

Table 9: Network configuration for the baseline **BASE**. Operation: d=dropout, a=add, n=normalize. As in Table 8, the parameters are those used by Marian.

```
(0) subword_nmt/learn_bpe.py
    bpe_operations: 4000
    separate vocabulary setting

(1) subword_nmt/apply_bpe.py
    dropout: 0.05

(2) sentencepiece-options:
    vocab size: 4000
    character coverage: 0.9998
    sentencepiece-alphas: 0 0

(3) segmentation:
    prefix rate: 32
    suffix rate: 500
    postfix rate (esp): 180
    postfix rate (guc): 500
    vocab size: 5000
    model training:
    dim-vocabs 4000 4000

(4) segmentation:
    perplexity (esp): 200
    perplexity (guc): 15
    α: 0.1
    β: 1.0
```

Table 10: Additional configuration for (0) **SUBW-bpe**, (1) **SUBW-dp**, (2) **SUBW-uni**, (3) **SUBW-prpe**, (4) **SUBW-fc**.

C Systems Evaluation

C.1 Translation Quality vs Vocabulary Size

The size of the vocabulary is very important in low resourced settings. We therefore perform a deep exploration of the merge operations in our SUBW-bpe system. Figure 3 shows translation quality with the three metrics (BLEU, chrF and BLEURT) varying the merge operations between 100 and 15000 per language.

Similarly to Ding et al. (2019), we find performance drops with increasing merge operations, confirming made findings, that in low-resource settings fewer merge operations, hence smaller vocabulary sizes seem to be appropriate (Mielke et al., 2021). Interestingly, we note a strong decline in performance for merge operations greater than 2k and smaller than 4k merges, Figure 3. Since the merge-depending vocabulary size influences the final amount of parameters, we suppose that for 2k or 4k, an optimal setting for the SUBW-bpe architecture is encountered.

1

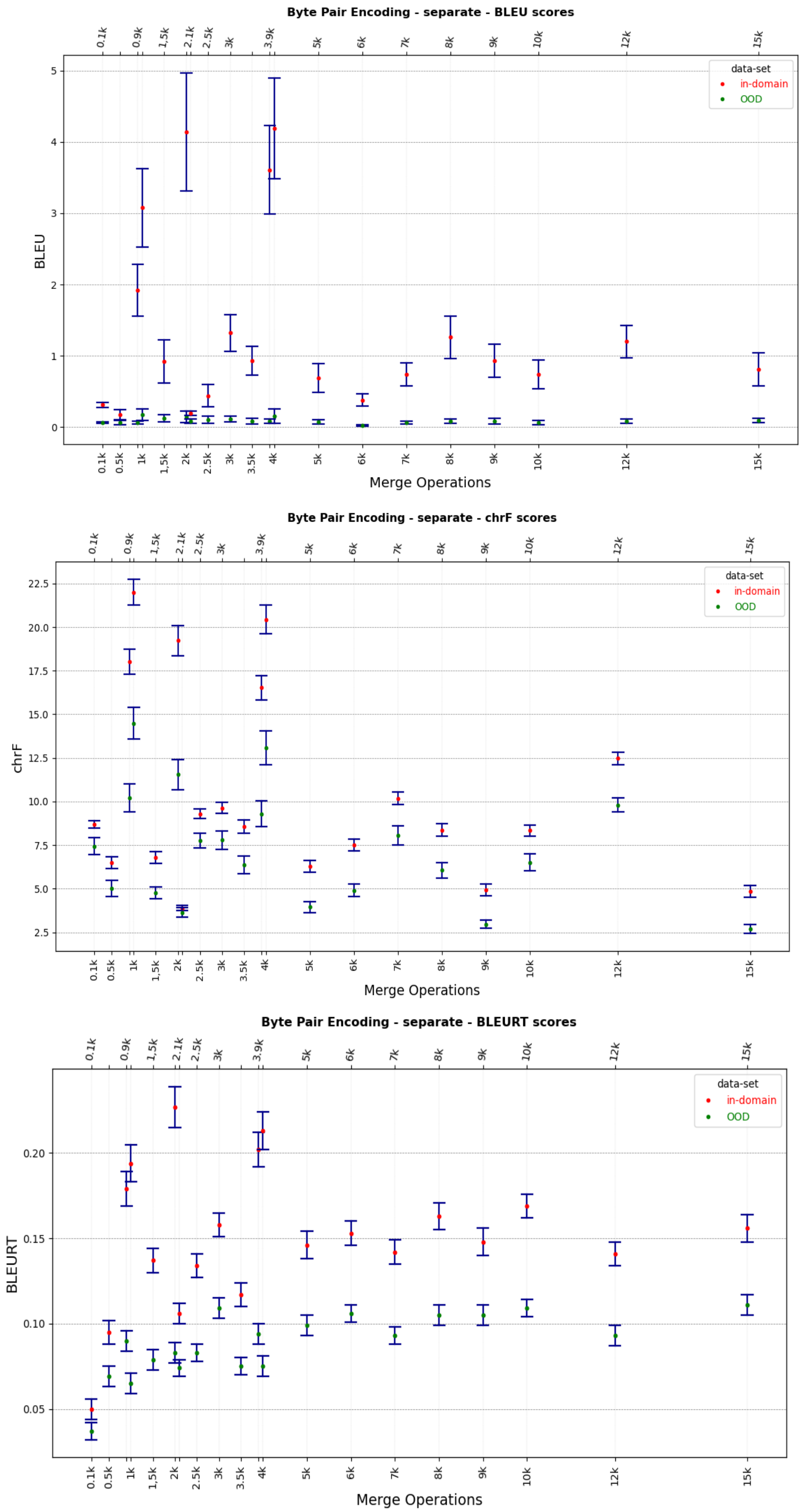

Figure 3: Automatic evaluation scores of the translations with the religious, in-domain and below the OOD-test set of the Transformer **SUBW-bpe** system trained with different BPE merge operations in a seperate vocabulary setting. The confidence intervals were obtained via bootstrap resampling

C.2 The Use of Automatic Metrics

Results in Section 6 show very low scores for the automatic metrics. Notice, that even if improvements with respect to the baselines are statistically significant, different metrics point to different rankings of the systems. This problem appears generally with low scores and with small differences between systems, both issues we encounter in Wayúunaiki–Spanish translation. As result, metrics do not correlate well with each other. The Pearson correlation among pairs of metrics (BLEU, chrF, BLEURT) is r < 0.6, being far from linearity. We show in Table 4 the scores of all our systems projected into the 2D spaces for BLEU-chrF (black crosses, r = 0.534, ρ = 0.451), BLEU-BLEURT (red stars, r = 0.571, ρ = 0.720) and chrF-BLEURT (green dots, r = 0.498, ρ = 0.377).

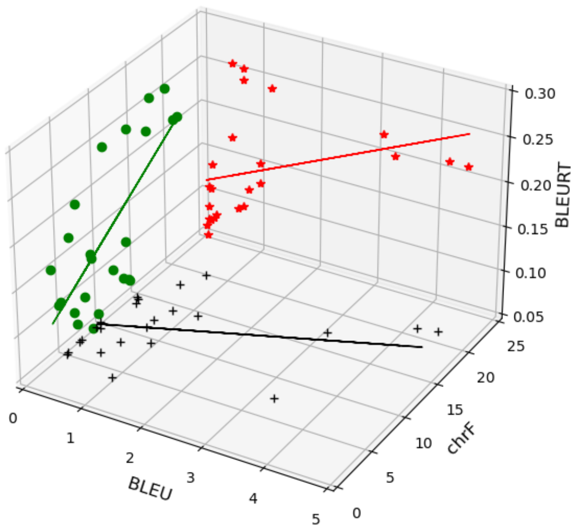

Figure 4: Correlation between the metrics used in the automatic evaluation. We include all of the model scores reported in Tables 4, 5 and 6.