BPE vs. Morphological Segmentation: A Case Study on Machine Translation of Four Polysynthetic Languages

Manuel Mager[♢] **Arturo Oncevay**[♡] **Elisabeth Mager**[♯]
Katharina Kann[♠] **Ngoc Thang Vu**[♢]
[♢]University of Stuttgart [♯]Universidad Nacional Autónoma de México
[♠]University of Colorado Boulder [♡]University of Edinburgh

Abstract

Morphologically-rich polysynthetic languages present a challenge for NLP systems due to data sparsity, and a common strategy to handle this issue is to apply subword segmentation. We investigate a wide variety of supervised and unsupervised morphological segmentation methods for four polysynthetic languages: Nahuatl, Raramuri, Shipibo-Konibo, and Wixarika. Then, we compare the morphologically inspired segmentation methods against Byte-Pair Encodings (BPEs) as inputs for machine translation (MT) when translating to and from Spanish. We show that for all language pairs except for Nahuatl, an unsupervised morphological segmentation algorithm outperforms BPEs consistently and that, although supervised methods achieve better segmentation scores, they under-perform in MT challenges. Finally, we contribute two new morphological segmentation datasets for Raramuri and Shipibo-Konibo, and a parallel corpus for Raramuri–Spanish.

1 Introduction

Polysynthetic languages are known because of their rich morphology, that encodes most parts of the semantics into verbs, leading to a high morpheme-per-word rate. The resulting combinations of morphemes and roots result in extreme type sparsity. Thus, polysynthetic languages represent a challenging environment for NLP methods (Klavans, 2018). Subword segmentation has been a common method to reduce sparsity (Vania and Lopez, 2017). Moreover, as these languages are mostly extremely low-resource (ELR), the challenge is even harder. Some of the reasons behind this is that most of them are endangered and spoken by minority groups (Mager et al., 2018; Littell et al., 2018).

But what impact does morphological segmentation have on downstream tasks like machine translation (MT), when translating from or into fusional languages? Linguistically inspired segmentation was considered to be the best option to handle rich morphology (Koehn et al., 2005; Virpioja et al., 2007) until the appearance of Byte-Pair Encodings (BPEs; Sennrich et al., 2016) and has been adopted as the default segmentation technique. BPEs earned this status for its good results, unsupervised training and language independence. Saleva and Lignos (2021) show that there is no significant gain when using an unsupervised morphological segmentation for the input over BPEs when evaluating those methods in moderate LR scenarios for Nepali–English and Kazakh–English, contradicting initial findings of Ataman and Federico (2018). However, how would BPEs perform for polysynthetic languages in ELR scenarios? Schwartz et al. (2020) compare BPE, with Morfessor (Smit et al., 2014) and Rule-Based morphological analyzers for medium resourced Inuktitut–English, and for the ELR Yupik–English and Guarani–Spanish. Their results show that BPEs outperform Morfessor and the morphological analyzer in all MT cases (but with better Language Modeling capabilities of morphological models over BPEs). However, most of these studies only rely on the usage of a limited set of segmentation methods and do not consider the quality of the used morphological segmentation methods.

This study aims to answer the following research questions: i) is morphological segmentation beneficial for MT where one language is polysynthetic and ELR?; and ii) is higher morphological segmentation quality correlated with higher MT scores?

To answer these questions, we perform segmentation experiments on four polysynthetic languages:[1] Nahuatl (`nah`), Raramuri (`tar`), Shipibo-Konibo (`shp`) and Wixarika (`hch`) and apply those segmentations to MT paired with Spanish (`spa`). First, we revisit a wide set of supervised and unsupervised methods and apply them to the input of MT transformer models. This study is the first

[1]We choose the languages for this study based on the availability of a morphological segmentation dataset.

Findings of the Association for Computational Linguistics: ACL 2022, pages 961 - 971

	train		dev		test	
	tar	spa	tar	spa	tar	spa
S	13,102		587		1,030	
N_{spa}/N_{tar}	1.692		1.794		1.689	
N	73,022	93,410	3,183	4,133	5,847	7,547
V	19,044	16,220	1,713	1,771	2,793	2,803
V1	12,894	10,021	1,402	1,365	2,221	2,120
V/N	0.261	0.174	0.538	0.429	0.478	0.371
V1/N	0.177	0.107	0.440	0.330	0.380	0.281
OOV			573	434	1,037	779
%OOV			0.334	0.245	0.371	0.277

Table 1: Parallel corpus' description: S = number of sentences; N_{spa}/N_{tar} = ratio of tokens between Spanish and Rarámuri; N = number of tokens; V = vocabulary size; V1 = number of tokens occurring once (hapax); V/N = vocabulary growth rate; V1/N = hapax growth rate; OOV = out-of-vocabulary words w.r.t. train set.

	shp			tar		
	train	dev	test	train	dev	test
Words	604	163	329	504	136	274
SegWords	437	114	228	323	87	178
Morphs	1215	321	642	1028	273	563
UniMorphs	476	181	319	474	181	287
Seg/W	0.72	0.69	0.69	0.64	0.64	0.65
Morphs/W	2.01	1.97	1.95	2.04	2.01	2.06
MaxMorphs	5	5	5	5	5	5
OOV-M		93	179		93	163

Table 2: Number of words, segmentable words (SegWords), total morphemes (Morphs), and unique morphemes (UniMorphs) in our new datasets. Seg/W: proportion of words consisting of more than one morpheme; Morphs/W: morphemes per word; MaxWords: maximum number of morphemes found in one word; OOV-M: morphemes in evaluation not seen in training.

to show that strong unsupervised morphological approaches outperform BPEs consistently on ELR polysynthetic languages, except for nah. These results are related to Ortega et al. (2020), that found that a morphologically guided BPE can improve the MT performance for Guarani–Spanish. On the other hand, even when supervised morphological segmentation methods achieve better results for the segmentation task, when it comes to MT systems they under-perform all other approaches. We hypothesize that this might be due to overfitting the clean and out-of-domain morphological training set. To make all these experiments possible we introduce additionally two new morphologically annotated datasets for tar and shp; and one parallel dataset for spa–tar[2].

Polysynthetic languages. A polysynthetic language is defined by the following linguistic features: the verb in a polysynthetic language must have an agreement with the subject, objects and indirect objects (Baker, 1996); nouns can be incorporated into the complex verb morphology (Mithun, 1986); and, therefore, polysynthetic languages have agreement morphemes, pronominal affixes and incorporated roots in the verb (Baker, 1996), and also encode their relations and characterizations into that verb.

[2]The datasets are available under http://turing.iimas.unam.mx/wix/mexseg

2 Descriptions of Novel Datasets

2.1 Raramuri–Spanish Parallel Dataset

For the dataset, we manually extract phrases that had a translation into Spanish from the Brambila (1976) dictionary. Additionaly, given that the orthography in this book is out of use, we normalized it to a modern version used in (Caballero, 2008). The book does not specify the dialect of the sentences. Table 1 shows the characteristics of the dataset, and the dataset splits.

2.2 Morphological Segmentation Datasets

We also introduce two new morphologically annotated datasets. For Raramuri we manually extracted segmented morphemes from a specialized linguistics paper (Caballero, 2010) and thesis (Caballero, 2008) that contain segmented and non-segmented words. Both sources annotate the Raramuri variant of the village of Choguita.

For Shipibo-Konibo, we adapted annotated sentences for lemmatization and part-of-speech tagging (Pereira-Noriega et al., 2017), and from a treebank (Vasquez et al., 2018), which was segmented in morphemes due to a particular phenomenon for clitics in the dependencies annotation.

3 Experimental Setup

3.1 Resources

For the machine translation experiment we use the following parallel datasets: the hch–spa translation of the fairy tales of Hans Christian Andersen (Mager et al., 2017); the Shipibo-Konibo–Spanish translations from a bilingual dictionary and educational material (Galarreta et al., 2017); and for nah–spa, the Axolotl dataset (Gutierrez-Vasques et al., 2016). This dataset contains several variants of Nahuatl. On top of that we also use our collected tar–spa Parallel corpora (§2.1). The details of the data splitting are described in Table 5 in the appendix. For morphological segmentation we use the nah and hch annotated datasets from Kann et al. (2018b) and additionally we use the shp and tar datasets introduced in section 2.2. We use the same splits as reported by the original sources.

3.2 Metrics

For machine translation we use the standard BLEU (Papineni et al., 2002) and chrF (Popović, 2015) metrics from the SacreBLEU implementation (Post, 2018). To evaluate morphology, we compare all outputs against the gold annotated test sets calculating accuracy and the EMMA F1 metric (Spiegler and Monson, 2010).

3.3 Subword Segmentation

BPEs (BPEs; Sennrich et al., 2016) is our reference system we use the sentence piece implementation (Kudo and Richardson, 2018) of BPEs. We tune the vocabulary size on a vanilla transformer small for each language, and take the best model evaluated on the development set.

Morfessor (Morfessor; Smit et al., 2014) As an unsupervised method we use Morfessor 2.0, that is a statistical model for the discovery of morphemes using minimum description length optimization.

FlatCat (FC; Grönroos et al., 2014), is a variant of Morfessor. It consists of a category-base hidden Markov model and a flat lexicon structure for segmentation.

LMVR (Ataman et al., 2017) modify the FC implementation by adding a lexicon size restriction and increase the tendency of the model to increase segmentation of commonly seen words.

CRFs (CRFs) As our first supervised model we use the conditional random fields (CRFs; Lafferty et al., 2001) segmentation model of Ruokolainen et al. (2014). We also investigate the capabilities of semiCRFs (Sarawagi and Cohen, 2005) for this particular task. For this, we use the Chipmunk implementation (Cotterell et al., 2015).

Seq2seq We also use a vanilla RNN sequence-to-sequence model with attention. The first variant (s2s) employs a supervised neural model. Additionally, we use the most promising extension proposed by Kann et al. (2018b) adding random generated strings in an auto-encoding fashion (s2s+multi).

Pointer–Generator Networks (PtrSeg; See et al., 2017) are commonly used in task where copying part of the input to the output is part of the task. This model has been used successfully for canonical segmentation (Mager et al., 2020).

3.4 NMT System

As our translation models, we use an encoder-decoder transformer model (Vaswani et al., 2017) with the hyperparameters proposed by Guzmán et al. (2019) as a baseline for low-resource languages. We use the vanilla version of this transformer without any further back-translation or other enhancements, so that we can remove any additional variables from the experiment, and focus only on the input segmentation. We use a 5k[3] vocabulary size for all sides using BPE. We use fairseq (Ott et al., 2019) for all translation experiments. The polysynthetic languages are segmented with the different investigated segmentation methods and Spanish always uses BPE in both translation directions.

4 Results

Morphology Table 3 shows that BPEs, a model that is not intended for morphological segmenta-

[3] We searched for the best vocabulary size using 2k, 4k, 5k, 6k and 8k.

system	hch	nah	tar	shp
BPEs	53.17	53.38	62.54	71.41
Morfessor	61.51	60.48	59.05	59.45
FC	<u>62.28</u>	58.94	64.65	<u>67.95</u>
LMVR	61.27	<u>60.55</u>	<u>65.46</u>	67.58
semiCRFs	68.10	81.92	81.22	-
CRFs	82.43	**87.83**	89.79	-
s2s	82.42	84.62	88.47	82.25
s2s+multi	**83.75**	84.90	88.37	**85.99**
PtrSeg	65.60	83.85	**90.13**	78.22

Table 3: Test results of surface segmentation for hah, nah and tar, and canonical segmentation for shp. Values are F1 scores, bold numbers are the best systems overall, underscored are the best unsupervised systems.

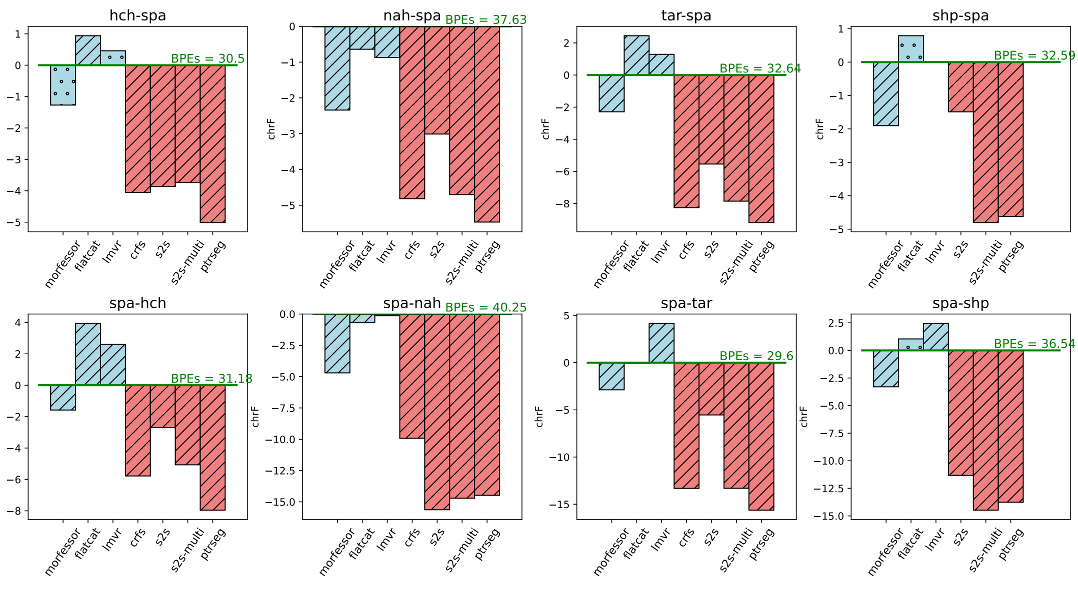

Figure 1: chrF score difference for all morphological segmentation when compared to BPEs on the test sets for both translation directions. We run a paired approximation test with 10000 trials using the BPEs system output as the baseline. Diagonals indicates a p-value ≤ 0.05, while stars indicates a p-value > 0.05. Blue systems are unsupervised, while Red ones are supervised.

tion, perform worst on all languages as expected, with exception of tar. The unsupervised morphological segmentation models (Morfessor, FC and LMVR) are consistently the worst performing models among the morphologically inspired models. The best performing systems are supervised, with s2s+multi showing best results for hch (83.7 F_1) and shp (85.99 F_1). CRFs achieved the best result for nah with 87.8 F_1 and PtrSeg achieved the best scores for tar with 90.13 F_1.

4.1 Discussion

MT Figure 1 shows the chrF score difference against the BPEs baseline in all directions[4]. We first observe that the supervised segmentation approaches under-perform in contrast with the unsupervised ones in all the settings.

Moreover, with the polysynthetic languages in the source side, FC has a significantly higher score for hch-spa and tar-spa, and a statistical tie in shp-spa; whereas LMVR obtains similar results to BPEs in hch-spa and shp-spa. In the other direction, with the polysynthetic languages as targets, LMVR is the method that significantly surpasses the baseline for more language pairs: spa-hch, spa-tar and spa-shp; whereas FC obtains the maximum score in spa-hch and statistical ties in spa-tar and spa-shp. We conclude that both methods are robust alternatives for translating from and to a polysynthetic language.

Despite the good results of s2s, s2s+multi or PtrSeg in morphological segmentation, for MT they have the worst performance. We argue that these kind of methods innovate new subwords in their output, which can aid for morphological segmentation, but for MT only adds noise in the input for the model.

Overall, we notice that in contrast to other languages (Saleva and Lignos, 2021), segmentation methods matter for polysynthetic ones. Poor suited methods can strongly decrease the performance of down-stream tasks like MT. However, the question on which segmentation method is better for MT is still open.

4.2 Analysis

To better understand the current results, we explore the outputs of different systems. For simplicity, we choose the best performing segmentation system for each of the segmentation paradigms. For unsupervised morphological inspired segmentation, we use LMVR, s2s+multi for supervised morphological segmentation, and BPEs for frequency-based segmentation.

First, we explore the impact of morphological richness on each of the systems. We use

[4] See Table 6 for the specific scores, BLEU ones included.

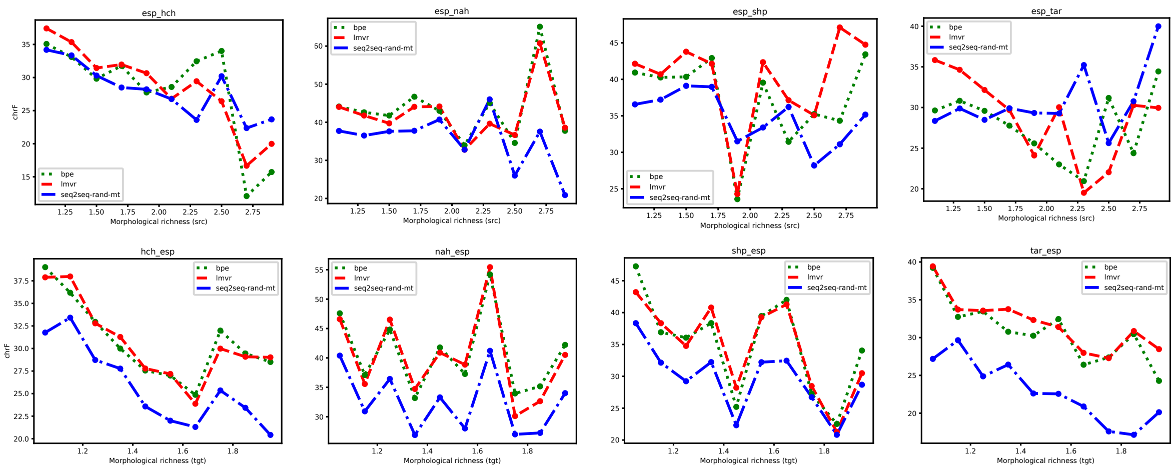

Figure 2: Relation between morphological richness of each polysyntetic language with relation to its chrF score, in each translation direction. The scores are analysed for `BPEs`, `LMVR` and `s2s+multi`.

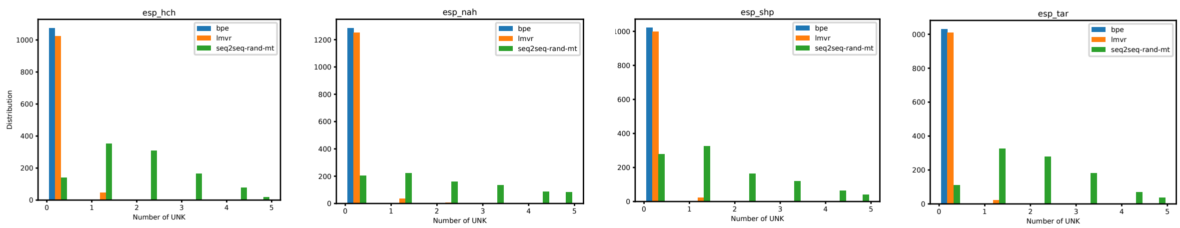

Figure 3: Number of out-of-vocabulary tokens (UNK) found for each polysynthetic language classified by system. The scores are analysed for `BPEs`, `LMVR` and `s2s+multi`.

`Morfessor` to infer the segmentation for each polysynthetic language data point and divide the number of found morphemes by the total number of tokens. Figure 2 shows that there is no clear correlation between morphological richness and systems' performance for `nah` and for `shp`. However, for `hch` we observe that a richer morphology implies a loss in translation quality. The same correlation can be seen for the `tar-esp` direction. This correlation is stronger when the polysynthetic language is in the source and weaker when it is in the target. Overall, a similar behavior can be observed between `LMVR` and `BPEs`.

Second, we explore the impact of out-of-vocabulary (UNK) tokens that each segmentation model introduces because having a high number of UNK tokens can negatively influence the MT results. In figure 3, we show the number of UNK tokens that each segmentation has when used with the dictionary of an MT system. The supervised `s2s+multi` has the highest amount of UNK symbols. We suggest that the reasons behind this phenomena could be the strong generative power of such systems and well-known artifacts that such models introduce (i.e., string repetitions). However, `LMVR` has a slightly higher number of UNK tokens, leaving `BPEs` the best vocabulary coverage. This can explain the surprisingly low performance of supervised models.

5 Conclusion

In this paper, we compared a wide set of morphological segmentation models with BPEs when applied to the input of Neural Machine Translation systems for extreme low-resource polysynthetic languages. We found that unsupervised morphological segmentation outperformed BPEs significantly on 5 out of 8 language pairs, setting a consistent overall performance. Surprisingly SOTA supervised morphological segmentation achieved the lowest performance of all systems. In future, we will explore Adaptor-Grammars (Johnson et al., 2006; Narasimhan et al., 2015; Eskander et al., 2020) for segmentation, and also the way to make unsupervised segmentation more robust and suitable for MT including the reduction of produced UNK symbols.

Ethical Considerations

The datasets introduced in this paper for machine readable training and evaluations are extracted from previous specialized linguistic work. We stick to the ethical standards giving credit to the original author in the spirit of *fair scientific usage*. We further strongly encourage future work that use these resources to cite also the original sources of the data. Additionally we found another ethical risks of this work: for the down-stream task of MT, a translation system should not be deployed with low quality translations, as it can mislead the user, and have implicit biases. Finally, want to state that the authors of this paper have a long record of working with the studied indigenous languages. Some have conducted field studies with the communities in the past, and Manuel Mager is part of the Wixarika community. This allows the authors to have a better understanding of the concerns of the communities that speak the discussed languages.

Acknowledgments

We want to thank all the anonymous reviewers as well as Pavel Denisov for their helpful comments and suggestions. This project has benefited from financial support to Manuel Mager by a DAAD Doctoral Research Grant.

References

Duygu Ataman and Marcello Federico. 2018. An evaluation of two vocabulary reduction methods for neural machine translation. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 97–110, Boston, MA. Association for Machine Translation in the Americas.

Duygu Ataman, Matteo Negri, Marco Turchi, and Marcello Federico. 2017. Linguistically motivated vocabulary reduction for neural machine translation from turkish to english. *arXiv preprint arXiv:1707.09879*.

Mark C Baker. 1996. *The polysynthesis parameter*. Oxford University Press.

Rachel Bawden, Alexandra Birch, Radina Dobreva, Arturo Oncevay, Antonio Valerio Miceli-Barone, and Philip Williams. 2020. The university of edinburgh's english-tamil and english-inuktitut submissions to the wmt20 news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 92–99.

David Brambila. 1976. *Diccionario rarámuri-castellano (tarahumar)*. Obra Nacional de la buena Prensa.

Gabriela Caballero. 2008. Choguita rarámuri (tarahumara) phonology and morphology.

Gabriela Caballero. 2010. Scope, phonology and morphology in an agglutinating language: Choguita rarámuri (tarahumara) variable suffix ordering. *Morphology*, 20(1):165–204.

Ryan Cotterell, Thomas Müller, Alexander Fraser, and Hinrich Schütze. 2015. Labeled morphological segmentation with semi-Markov models. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 164–174, Beijing, China. Association for Computational Linguistics.

Mathias Creutz and Krista Lagus. 2002. Unsupervised discovery of morphemes. In *CoNLL–SIGMORPHON*.

Mathias Creutz and Krista Lagus. 2007. Unsupervised models for morpheme segmentation and morphology learning. *ACM TSLP*, 4(1):3.

Robert MW Dixon and Alexandra Y Aikhenvald. 1999. *The amazonian languages*, volume 20. Cambridge University Press Cambridge.

Ramy Eskander, Francesca Callejas, Elizabeth Nichols, Judith Klavans, and Smaranda Muresan. 2020. MorphAGram, evaluation and framework for unsupervised morphological segmentation. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 7112–7122, Marseille, France. European Language Resources Association.

Ramy Eskander, Judith Klavans, and Smaranda Muresan. 2019. Unsupervised morphological segmentation for low-resource polysynthetic languages. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 189–195, Florence, Italy. Association for 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.

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 *COLING*, pages 1177–1185.

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).

Francisco Guzmán, Peng-Jen Chen, Myle Ott, Juan Pino, Guillaume Lample, Philipp Koehn, Vishrav Chaudhary, and Marc'Aurelio Ranzato. 2019. The FLORES evaluation datasets for low-resource machine translation: Nepali–English and Sinhala–English. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 6098–6111, Hong Kong, China. Association for Computational Linguistics.

Zellig Sabbettai Harris. 1951. *Methods in structural linguistics*. Chicago University Press.

INEGI. 2020. Censo de población y vivienda.

Mark Johnson, Thomas Griffiths, and Sharon Goldwater. 2006. Adaptor grammars: A framework for specifying compositional nonparametric bayesian models. *Advances in neural information processing systems*, 19.

Katharina Kann, Jesus Manuel Mager Hois, Ivan Vladimir Meza Ruiz, and Hinrich Schütze. 2018a. Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. In *NAACL-HLT*, volume 1, pages 47–57.

Katharina Kann, Jesus Manuel Mager Hois, Ivan Vladimir Meza-Ruiz, and Hinrich Schütze. 2018b. Fortification of neural morphological segmentation models for polysynthetic minimal-resource languages. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 47–57, New Orleans, Louisiana. Association for Computational Linguistics.

Judith L. Klavans. 2018. Computational challenges for polysynthetic languages. In *Proceedings of the Workshop on Computational Modeling of Polysynthetic Languages*, pages 1–11, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Rebecca Knowles, Darlene Stewart, Samuel Larkin, and Patrick Littell. 2020. NRC systems for the 2020 Inuktitut-English news translation task. In *Proceedings of the Fifth Conference on Machine Translation*, pages 156–170, Online. Association for Computational Linguistics.

Tom Kocmi. 2020. Cuni submission for the inuktitut language in wmt news 2020. In *Proceedings of the Fifth Conference on Machine Translation*, pages 171–174.

Philipp Koehn et al. 2005. Europarl: A parallel corpus for statistical machine translation. In *MT summit*, volume 5, pages 79–86. Citeseer.

Oskar Kohonen, Sami Virpioja, and Krista Lagus. 2010. Semi-supervised learning of concatenative morphology. *ACL*, page 78.

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.

John Lafferty, Andrew McCallum, and Fernando CN Pereira. 2001. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *ICML*.

Patrick Littell, Anna Kazantseva, Roland Kuhn, Aidan Pine, Antti Arppe, Christopher Cox, and Marie-Odile Junker. 2018. Indigenous language technologies in Canada: Assessment, challenges, and successes. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2620–2632, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Manuel Mager, Özlem Çetinoğlu, and Katharina Kann. 2020. Tackling the low-resource challenge for canonical segmentation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5237–5250, Online. Association for Computational Linguistics.

Manuel Mager, Dionico Gonzalez, and Ivan Meza. 2017. Probabilistic finite-state morphological segmenter for wixarika (huichol).

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

Marianne Mithun. 1986. On the nature of noun incorporation. *Language*, 62(1):32–37.

Karthik Narasimhan, Regina Barzilay, and Tommi Jaakkola. 2015. An unsupervised method for uncovering morphological chains. *Transactions of the Association for Computational Linguistics*, 3:157–167.

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

Myle Ott, Sergey Edunov, Alexei Baevski, Angela Fan, Sam Gross, Nathan Ng, David Grangier, and Michael Auli. 2019. fairseq: A fast, extensible toolkit for sequence modeling. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 48–53, Minneapolis, Minnesota. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

José Pereira-Noriega, Rodolfo Mercado-Gonzales, Andrés Melgar, Marco Sobrevilla-Cabezudo, and Arturo Oncevay-Marcos. 2017. Ship-lemmatagger: Building an nlp toolkit for a peruvian native language. In *Text, Speech, and Dialogue*, pages 473–481, Cham. Springer International Publishing.

Hoifung Poon, Colin Cherry, and Kristina Toutanova. 2009. Unsupervised morphological segmentation with log-linear models. In *NAACL-HLT*, pages 209–217. 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.

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.

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

Teemu Ruokolainen, Oskar Kohonen, Sami Virpioja, and Mikko Kurimo. 2013. Supervised morphological segmentation in a low-resource learning setting using conditional random fields. In *CoNLL*, pages 29–37.

Teemu Ruokolainen, Oskar Kohonen, Sami Virpioja, and Mikko Kurimo. 2014. Painless semi-supervised morphological segmentation using conditional random fields. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics, volume 2: Short Papers*, pages 84–89, Gothenburg, Sweden. Association for Computational Linguistics.

Jonne Saleva and Constantine Lignos. 2021. The effectiveness of morphology-aware segmentation in low-resource neural machine translation. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop*, pages 164–174, Online. Association for Computational Linguistics.

Sunita Sarawagi and William W Cohen. 2005. Semi-markov conditional random fields for information extraction. In *Advances in neural information processing systems*, pages 1185–1192.

Lane Schwartz, Francis Tyers, Lori Levin, Christo Kirov, Patrick Littell, Chi-kiu Lo, Emily Prud'hommeaux, Kenneth Steimel, Rebecca Knowles, Jeffrey Micher, et al. 2020. Neural polysynthetic language modelling. *arXiv preprint arXiv:2005.05477*.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

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.

Peter Smit, Sami Virpioja, Stig-Arne Grönroos, and Mikko Kurimo. 2014. Morfessor 2.0: Toolkit for statistical morphological segmentation. In *Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics*, pages 21–24, Gothenburg, Sweden. Association for Computational Linguistics.

Alexey Sorokin. 2019. Convolutional neural networks for low-resource morpheme segmentation: baseline or state-of-the-art? In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 154–159, Florence, Italy. Association for Computational Linguistics.

Sebastian Spiegler and Christian Monson. 2010. EMMA: A novel evaluation metric for morphological analysis. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1029–1037, Beijing, China. Coling 2010 Organizing Committee.

Clara Vania and Adam Lopez. 2017. From characters to words to in between: Do we capture morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2016–2027.

Alonso Vasquez, Renzo Ego Aguirre, Candy Angulo, John Miller, Claudia Villanueva, Željko Agić, Roberto Zariquiey, and Arturo Oncevay. 2018. Toward Universal Dependencies for Shipibo-konibo. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 151–161, Brussels, Belgium. Association for Computational Linguistics.

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.

Sami Virpioja, Jaako J Väyrynen, Mathias Creutz, and Markus Sadeniemi. 2007. Morphology-aware statistical machine translation based on morphs induced in an unsupervised manner. In *Proceedings of Machine Translation Summit XI: Papers*.

Linlin Wang, Zhu Cao, Yu Xia, and Gerard de Melo. 2016. Morphological segmentation with window LSTM neural networks. In *AAAI*.

Yaofei Yang, Shupin Li, Yangsen Zhang, and Hua-Ping Zhang. 2019. Point the point: Uyghur morphological segmentation using pointernetwork with gru. In *China National Conference on Chinese Computational Linguistics*, pages 371–381. Springer.

A Appendix

A.1 Data set splitting

	Train	Dev.	Test
hch–spa	665	167	553
nah–spa	540	134	449
tar–spa	604	163	329
shp–spa	504	136	274

Table 4: Data splitting (in number of instances) used for out the Morphological Segmentation experiments for all languages.

	Train	Dev.	Test
hch–spa	7442	447	1075
nah–spa	14208	644	1291
tar–spa	12987	582	1021
shp–spa	13102	587	1030

Table 5: Data splitting (in number of phrases) used for out Machine Translation experiments, from and to Spanish.

A.2 The Languages of new collected datasets

Raramuri (also known as Tarahumana) is a Yuto-Aztecan language, spoken in the northern part of the Mexican Sierra Madre Occidental by 89,503 speakers (INEGI, 2020). Raramuri is a polysynthetic and agglutinative language and has a Subject-Object-Verb (SOV) word order with morphonological fusion indicated by verbal suffixes (Caballero, 2008).

Shipibo-Konibo is a Panoan language spoken by around 26,000 people in the Amazonian region of Perú. This language is polysynthetic, with a strong tendency to agglutination, but also with certain degree of fusion. Its word order is mainly SOV (Dixon and Aikhenvald, 1999).

A.3 Additional related work

Morphological segmentation was firs introduced by Harris (1951). Unsupervised methods are popular with the Morfessor (Creutz and Lagus, 2002, 2007; Poon et al., 2009) family of segmentors. They also have a semi-supervised version (Kohonen et al., 2010; Grönroos et al., 2014). Recently Adaptor Grammars have been applied with great success to the task (Eskander et al., 2019, 2020). Supervised methods have achieved the best results with methods like CRFs (Ruokolainen et al., 2013), LSTM taggers (Wang et al., 2016), seq2seq RNNs (Kann et al., 2018a), CNNs (Sorokin, 2019), pointer networks (Yang et al., 2019), and pointer generator networks (Mager et al., 2020).

For the MT down-stream task, few research has been done (Schwartz et al., 2020; Roest et al., 2020). New research has been done in context of the WMT 2020 shared task on Inuktitut-English Bawden et al. (2020); Kocmi (2020); Knowles et al. (2020); Roest et al. (2020).

A.4 Machine translation results

Table shows the translation results using BLEU[5] and chrF[6].

[5]BLEU + case.mixed + numrefs.1 + smooth.exp + tok.13a + v.1.5.0

[6]chrF2 + numchars.6 + space.false + v.1.5.0

system	hch-spa		nah-spa		tar-spa		shp-spa	
	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF
bpe	15.04	30.50	**15.37**	**37.63**	11.44	32.64	11.85	32.59
morfessor	15.12	29.23	13.84*	35.29*	12.05	30.35*	9.65*	30.69*
flatcat	15.89*	**31.44***	14.89	36.99*	**15.55***	**35.09***	**12.29**	**33.38**
lmvr	**16.61***	30.96	14.78*	36.76*	12.97*	33.93*	11.14	32.60
crfs	10.66*	26.45*	12.48*	32.81*	8.42*	24.38*	-	-
seq2seq	9.23*	26.64*	12.13*	34.62*	7.69*	27.10*	10.27*	31.10*
seq2seq-rand-mt	11.46*	26.77*	12.22*	32.93*	8.31*	24.79*	9.51*	27.79*
pointernet	10.33*	25.49*	11.78*	32.16*	7.85*	23.46*	8.91*	27.97*
system	spa-hch		spa-nah		spa-tar		spa-shp	
	BLEU	chrF	BLEU	chrF	BLEU	chrF	BLEU	chrF
bpe	16.98	31.18	**13.29**	**40.25**	10.70	29.60	10.84	36.54
morfessor	12.26*	29.60*	8.52*	35.55*	5.95*	26.72*	5.00*	33.24*
flatcat	**18.70***	**35.12***	12.42*	39.59*	8.66*	29.52	11.68	37.58
lmvr	17.44	33.79*	12.26*	40.11	**12.88***	**33.76***	**12.84**	**38.99***
crfs	9.37*	25.40*	6.41*	30.33*	2.27*	16.28*	-	-
seq2seq	9.64*	28.48*	1.29*	24.62*	2.96*	24.06*	0.77*	25.21*
seq2seq-rand-mt	7.76*	26.11*	3.79*	25.54*	1.16*	16.29*	0.13*	22.06*
pointernet	4.22*	23.22*	2.76*	25.77*	0.76*	13.97*	0.06*	22.78*

Table 6: Translation results on test for both directions. Maximum scores are in bold. We run a paired approximation test with 10000 trials using the `BPEs` system output as the baseline, and "*" indicates a p-value < 0.05.