Neural Machine Translation Models with Back-Translation for the Extremely Low-Resource Indigenous Language Bribri

Isaac Feldman
Dartmouth College
isaac.c.feldman.23@dartmouth.edu

Rolando Coto-Solano
Dartmouth College
rolando.a.coto.solano@dartmouth.edu

Abstract

This paper presents a neural machine translation model and dataset for the Chibchan language Bribri, with an average performance of BLEU 16.9±1.7. This was trained on an extremely small dataset (5923 Bribri-Spanish pairs), providing evidence for the applicability of NMT in extremely low-resource environments. We discuss the challenges entailed in managing training input from languages without standard orthographies, we provide evidence of successful learning of Bribri grammar, and also examine the translations of structures that are infrequent in major Indo-European languages, such as positional verbs, ergative markers, numerical classifiers and complex demonstrative systems. In addition to this, we perform an experiment of augmenting the dataset through iterative back-translation (Sennrich et al., 2016a; Hoang et al., 2018), by using Spanish sentences to create synthetic Bribri sentences. This improves the score by an average of 1.0 BLEU, but only when the new Spanish sentences belong to the same domain as the other Spanish examples. This contributes to the small but growing body of research on Chibchan NLP.

1 Introduction

State-of-the-art systems in neural machine translation require large amounts of rich and varied parallel data between the source and target language (Lui et al., 2019). While this is less of a problem for high-resource language pairs like English-German or English-Spanish, there are much fewer resources when one of the languages is an Indigenous language. While there are some researchers creating systems to optimize the training of NMT frameworks with lower resources, most work primarily from a "warm-start" where models are pre-trained on a small, but still significant amount of parallel data (He et al., 2016). For example even in papers geared towards developing techniques for low resource languages, training datasets can exceed millions of sentences (Edunov et al., 2018). Our motivation in adapting low-resource techniques to extremely low-resource situations is that they can help improve understanding of their underlying mechanisms and lower the barrier of entry for machine translation to be used for language revitalization purposes.

However, in the effective body of research, many papers use simulated low-resource scenarios where high resource parallel datasets are artificially truncated to test the efficacy of their techniques. While this does provide a level of standardization across different techniques, it is still artificial. This paper seeks to explore NMT in an authentic scenario, using an actual extremely low-resource language.

We leverage a variety of sources to produce a novel translation dataset in the indigenous Bribri language for demonstrating methods of neural machine translation (NMT) in extremely low-resource situations. To demonstrate usage of the dataset, we apply the technique of iterative back-translation with validation (Hoang et al., 2018). Finally, we present a translation analysis to show the unique challenges that this Bribri dataset presents.

Proceedings of the 28th International Conference on Computational Linguistics, pages 3965–3976
Barcelona, Spain (Online), December 8-13, 2020

2 Challenges for NMT in Extremely Low-Resource Languages

In extremely low resource situations, the drawbacks of NMT systems begin to show. NMT is a data-hungry approach to machine translation and much less efficient with respect to training data compared to other approaches (Koehn and Knowles, 2017), but with enough data, it has shown excellent results (Hassan et al., 2018). Still, these approaches assume that written data is continuously produced in the languages. However, there are 7000 languages in the world (Eberhard and Fennig, 2020); many of them are spoken in Indigenous communities, have small populations, and are scarcely used in social media or in other forms amenable to automatic scraping (Lillehaugen, 2019; Keegan et al., 2015; Ní Bhroin, 2015). In many of these communities, languages like English and Spanish have displaced the Indigenous languages in domains such as technology and chatting, and so the available data is curtailed. In addition to this, many Indigenous communities have chronic digital inequalities, which makes it difficult to generate crowd-sourcing campaigns for those languages. Finally, in many cases the data that is most valuable to speakers of the language is that collected from elders and knowledge keepers, but those elders might be the people who have the least access to technological means of communication.

NMT also struggles with out of domain translation (Chu and Wang, 2018) which is only exacerbated by a smaller dataset. Our primary solution to this is employing a popular dataset augmentation technique called iterative back-translation (Hoang et al., 2018). Additionally, NMT models are fairly opaque (Koehn and Knowles, 2017) in that their learning process is not deterministic and is difficult to interpret. Here, we try to provide a closer, linguistic analysis of the translations produced by the models.

2.1 Challenges for Bribri NLP and NMT

Bribri (Glottocode `brib1243`) is an Indigenous language spoken by the Bribri people in Costa Rica and Panama. It belongs to the Chibchan family and it has approximately 7000 speakers (INEC 2010). The language is vulnerable (Moseley, 2010; Sánchez Avendaño, 2013), which means that there are still children who speak it at home, but there are domains where Spanish is used instead.

There is very little material published in the Bribri language, so training sets will always remain small. There are two main groups of people who write Bribri (and indeed most Indigenous languages in the Americas): University-affiliated researchers, and community member school teachers and others related to didactic material creation. There is little to no usage of the language online, and practically all of the material exists only as printed books. The main sources of bitext are textbooks for Spanish speakers to learn the language (Constenla et al., 2004; Jara Murillo and García Segura, 2013), Spanish-Bribri dictionaries (Margery, 2005), grammar books (Jara Murillo, 2018a), collections of oral literature (Jara Murillo, 2018b; García Segura, 2016; Constenla, 2006; Constenla, 2006; Constenla, 1996), and schoolbooks for Bribri children (Sánchez Avendaño, 2020). There is one digital corpus that also contains traditional stories (Flores Solórzano, 2017). These sentences belong to general domains (e.g. *Íma be' kie?* 'How are you?'), but they also include specialized passages from traditional narrations.

There are numerous sources of internal variation in the data; these are summarized in table 1. First, Bribri has been studied over the last 50 years, but researchers have published materials using different writing systems. For example, a nasal vowel can be indicated by a line underneath the vowel (Constenla et al., 2004), by a tilde above the vowel (Jara Murillo and García Segura, 2013), or by a Polish hook (Margery, 2005). Moreover, this variation in orthography leads to many permutations of diacritic encodings. The word *ù̠* 'cooking pot' has a grave accent for high tone and a nasal vowel. If the vowel is expressed with the line below, the exact Unicode combining character varies amongst materials, and the tonal mark can be expressed as a single Unicode character with the 'u', or as separate characters.

While it is easy to standardize diacritics, it is more challenging to create an internally-standardized representation of lexemes that vary due to phonetic and phonological rules. For example, vowels in contact with nasal phonemes become phonetically nasalized (e.g. a̠mì̠ [ã˧ˈmĩ˥] 'mother'). This allophonic nasalization is sometimes represented in the orthography (e.g. a̠mì̠ (Constenla et al., 2004)) but some materials leave it out (e.g. *amì* (Jara Murillo, 2018a)). Another such rule is the deletion of unstressed vowels in word initial position (e.g. *mĩ̀* (Jara Murillo, 2018b)). It would be possible to standardize this to `amì` in an internal representation, but you would need to know all of the possible words that could

	Differences
Writing system	ù̲ ‘cooking pot’ (Constenla et al., 2004) ũ̀ (Jara Murillo, 2018a), ų̀ (Margery, 2005)
Diacritic encoding	ù̲ ‘cooking pot’: comb. grave (U+0300) comb. low line (U+0332) comb. grave (U+0300) comb. minus sign below (U+0320) latin small u with grave (U+00F9) comb. macron (U+0331)
Phonetics and phonology	Nasal assimilation: amì ∼ a̲mì̲ ‘mother’ Unstressed vowel deletion: mĩ̀ ∼ ãmì ‘mother’
Sociolinguistic and dialectal variation	ña̲là̲ (Amubri) ’road’ (Constenla et al., 2004) ñolõ̀ (Coroma) ’road’ (Jara Murillo, 2018a)
Orthographic variation	(a) ìë’pa rör këképa táìn ë. (MEP, 2017, 18) ie’pa dör akë́këpa ta̲î̲ë. (Equivalent in Constenla et al. (2004)) ‘They are important elders’. (b) E’kũẽ́k és ikíe dör (García Segura, 2016, 11) E’ ku̲é̲ki̲ e’s i kie dör. (Equivalent in Constenla et al. (2004)) ‘That’s why it is called like this’.

Table 1: Sources of variation in Bribri text input

undergo this variation in order to detect them.

Bribri has dialectal differences which are reflected in writing. For example, the nasal ‘a’ in the Amubri dialect often corresponds to the nasal ‘o’ in the Coroma dialect (*ña̲là̲* ∼ *ñolõ̀* ‘road’). These two dialects also have differences in tones (*sulû* [su˩'ɹu˥˩] ∼ *sulù* [su˧'ɹu˥] ‘bad’) and consonants (e.g. *Ye’ miátke* ∼ *Ye’ miátche* ‘Good bye’). The main issue for NLP is not in reducing these alternations to a single internal representation, but on deciding which representation should be given precedence. This is a political issue, and one of the thorniest in language revitalization and normalization, so much so that many communities have seen metaphorical orthography wars (Bermel, 2007; Hinton, 2014), where disagreements over writing systems and curricula carry on for years.

Perhaps the most challenging source of variation is that of orthographic variation amongst different writers. Bribri is not yet standardized, and this leads to different orthographic conventions when the language is actually written. Table 1 shows two examples of this. In example (a), the copula *dör* appears as *rör*, the word *ta̲î̲ë* ‘very much’ appears as two words *táìn ë*, and the nasalization is expressed with an n, not with a line. In example (b), the absolutive argument *i* ‘it’ appears in the same word as the verb *kíe* ‘to be called’, and the verb *kíe* itself has a tone marking that is not present in other standards. We want to stress that, even if this variation makes NLP work more difficult, this is a challenge that NLP has to deal with in order to do justice to these languages. When people write in Bribri and in any Indigenous language, they are expanding its domains of usage, and this contributes to their revitalization and normalization. Regardless of how much variation/“noise” there might be in the written text, the main priority is for the speakers of these languages to keep using them in their daily lives, regardless of how they write. There will be plenty of time later to debate standardization; right now all that matters is that the community can perpetuate the use of their language. While we have discussed some specifics for the Bribri case, all of these issues are present in many Indigenous languages around the world (Galla, 2016).

There is relatively little research on the NLP of Indigenous languages of the Americas (Mager et al., 2018), but Bribri has received some attention. Published research includes the design of virtual keyboards (Flores Solórzano, 2010) and finite-state machines for morphological analysis (Flores Solórzano, 2019; Flores Solórzano, 2017), and there is an online Spanish-Bribri dictionary (Krohn, 2020). Finally, there is work on machine learning: untrained forced alignment (Coto-Solano and Flores Solórzano, 2016; Coto-Solano and Flores Solórzano, 2017) was used to study the phonetics of the language.

3 NMT Training from Bribri Bitext

Using the resources cited above, we created a dataset with 5923 Bribri-Spanish sentence pairs. We performed diacritic and writing system standardization with a rule-based, deterministic system, with a result that smooths out some author-specific variation and that can be converted to either type of contemporary orthography (the Constenla et al. (2004) and the Jara Murillo (2018a) conventions). For example, the string `ùx` represents a nasal high-tone 'u', as the 'x' is not used as a letter in the language. (For simplicity, we will use the convention of indicating nasalization with a line underneath when showing Bribri text on this paper). We normalized forms to the Amubri orthography for phonological variations (e.g. ***ña̲là̲***∼*ñolõ̀* 'road') when there is a systematic way to convert them back to Coroma orthography, but preserved lexemes that are exclusive to each variant (Coroma: *alàralar* 'children'). Finally, we attempted to standardize common function words (e.g. copula *rör* to *dör*). After standardizing all sentence pairs, we encoded each sentence with separate byte pair encoding models for each language (Sennrich et al., 2016b) where each encoding had a vocabulary size of about 2200 tokens.

For each trial the dataset was split between 80% training, 10% validation, and 10% for testing. Aditionally, the data was randomly shuffled ten times, resplit and retrained for further validation. From these sets we trained Transformer encoder/decoder models which in previous research have shown good performance for translation tasks. We used near identical hyperparameters to the base model described in Vaswani et al. (2017) with the exceptions being the number of training steps: each model was trained 4000 steps with a batch size of 4096 tokens on a single GPU. The models were only trained 4000 steps to prevent overfitting on the small dataset. Each translation model was trained using the PyTorch port of the OpenNMT package (Klein et al., 2017) on Google Colaboratory instances with a range of different GPUs. The training took approximately 50 minutes per model. Each model was then evaluated against the testing data with the Bilingual Evaluation Understudy Score (BLEU) technique (Papineni et al., 2002) where the specific implementation was the multi-bleu script from Moses (Koehn et al., 2007).

The code for sentence standardization, training, and a sample of BRI-SPA sentence pairs can be found here: `http://github.com/rolandocoto/bribri-coling2020`.

3.1 BLEU scores

We built 3 different types of models: One based on all the data (5923 total pairs; henceforth the 6K model), one based on half of the data (2961 total pairs, henceforth the 3K model), and one with only a quarter (1480 total pairs, henceforth the 1.5K model). Table 2 summarizes the dataset splits, the repeated random validation average (10 trials) and the maximum BLEU values for all the validation tests.

Model	Training pairs	Validation pairs	Testing pairs	BRI→SPA		SPA→BRI	
				Avg	Max	Avg	Max
1.5K	1184	148	148	6.3 ± 2.0	9.9	8.7 ± 0.6	9.5
3K	2368	296	296	11.2 ± 1.9	14.9	11.1 ± 0.7	12.0
6K	4737	593	593	16.9 ± 1.7	19.8	14.2 ± 2.7	18.9

Table 2: Data splits and BLEU scores for Bribri and Spanish NMT with increasing amounts of data.

The Bribri→Spanish models had an average performance of BLEU 16.9. This was better than the Spanish→Bribri models, which had an average of BLEU 14.2. There was considerable variation in the training results, which is to be expected with such a small amount of data, as the specific shuffle of each trial will have a large effect on the results. For example, while the best performance for a SPA→BRI pair was 18.9, the average BLEU for the 6K models was 14.2 ± 2.7, and the worst performing model was BLEU 10.3. An obvious question would be: How did some of the models perform so well? The dataset might be helping us: The data itself is mainly from textbooks and grammar books, which are rich in examples that the models might be taking advantage of. This might be a positive development for the implementation of NMT for Indigenous languages, given that most of the available data is in this format.

3.2 Translation analysis: Bribri→Spanish

What is the model learning? Table 3 shows examples from the top performing Bribri→Spanish model. Many translations are correct, and when it makes mistakes it is because of mismatches in the grammar of the two languages. Example 3 has the pronoun *ie'* "he/she/singularThey", which is genderless in Bribri. In Spanish the pronoun's gender should match the name *Ana*, but because gender is underspecified in Bribri, the Spanish translation comes out with the wrong pronoun. There are also difficulties in matching the verbal morphologies of the two languages. Bribri has a middle voice, where the verb is performed without a specific actor. (Japanese also has these structures: *Rajio ga naotta* 'The radio got fixed'). The model could not translate Bribri middle voice correctly, and tried to match it with the Spanish active voice, giving it the random actor 'I' in example 4.

English	Bribri source	Spanish reference	Spanish traslation	Observations
1. You saw him.	Be'r ie' sú̱ .	Usted lo vió .	Usted lo vió .	Correct.
2. The bird was not sitting on the branch.	Dù kë̀ bák tkër kàlula ki̱ .	El pájaro no estuvo posado sobre la rama .	El pájaro no estuvo posado sobre la rama .	Correct
3. She is called Ana.	**Ie'** kie Ana .	**Ella** se llama Ana.	**Él** se llama Ana .	Wrong gender in Spanish
4. Rabbit meat has been eaten.	Sawë́ chka katárule .	Se ha comido carne de conejo .	he comido carne de conejo .	Verb is middle voice in BRI, active in SPA. (Translation meaning: 'I've eaten rabbit meat').

Table 3: Examples of Bribri→Spanish translations.

3.3 Translation analysis: Spanish→Bribri

Bribri is structurally very different from most Indo-European languages, and so we are interested in how the models are learning Bribri targets. Here we will focus on four grammatical phenomena: (i) positional verbs, (ii) ergativity, (iii) numerical classifiers and (iv) demonstratives.

Let's begin with the positional verbs. The sentence *Wë́pa* ***tso*** *ù a̱* 'The men are in the house' has the simple verb *tso* 'to be in'. Bribri can also use more specific verbs to provide specific details about the position of the subject. For example, in the sentence *Wë́pa* ***ië'te̱n*** *ù a̱* 'The men are standing in the house', the verb *ië'te̱n* 'to be standing in a place [plural]' tells you the position of the men relative to the house. Languages like German also have positionals (e.g. *Der Teddy* ***liegt*** *auf dem Boden.* lit: 'The teddy bear **lies** on the floor'). Bribri positional verbs include *tkër* 'to be sitting on', *tër* 'to be lying on', *a'r* 'to be hanging' and dur 'To be standing [singular]' amongst others. Table 4 shows the translations from the best performing Spanish→Bribri model, and compares them with the reference Bribri sentence. The system was successful in learning some positionals, but sometimes it overgeneralizes (replacing *a'r* with *tër* in example 3), and sometimes it fails to use the positional, as in example 4.

Bribri is a morphologically ergative language (Pacchiarotti, 2016). This means that the subjects of transitive sentences are marked with an affix. There are numerous other ergative languages in the world, such as Basque, Samoan and Warlpiri. In the case of Bribri, the morpheme *tö/dör/'r* marks the ergative word. Table 5 shows that, while the model has learned to place the ergative in some subjects of transitive sentences, it hasn't learned the general rule yet. Example 3 has the subject *be'* 'you', which is the doer of the action and should therefore be marked as ergative, but the model did not mark it as such.

Bribri has a type of word called *numerical classifiers*. These are number words, but each has a specific semantic class. For example, in Bribri, children are classified as persons, and so the word three would be *ma̱ñá̱l*. On the other hand, the word chicken is classified as a small thing, and so the word three would become *ma̱ñà̱t*. Another example are houses, which are classified as buildings. When counting buildings,

English	Bribri reference	Bribri translation	Observations
1. The bird is (sitting) on the branch.	Dù **tkër** kàlula ki̠ .	Dù **tkër** kàlula ki̠ .	Correct positional: *tkër*: to be sitting.
2. The dog is (lying down) by the edge of the river.	Chìchi **tër** di' jkö .	Chìchi **tër** ña̠là̠ jkö .	Correct positional: *tër*: to be lying down. Translation means: 'The dog is (lying down) by the edge of the road'.
3. The shirt is (hanging) over there.	Apàio **a'r** a̠wí̠e̠ ye' wa̠.	A̠@@wìe̠ apàio **tër**.	Wrong positional: *a'r*: hang; *tër*: lying down
4. He was (standing) in the house.	Ie' bák **dur** ù a̠ .	Ie' bák ù a̠ .	Missing positional: *dur*: to be standing. Translation means: 'He was in/by the house'

Table 4: Bribri positionals in Spanish→Bribri models (@ is an unknown token)

English	Bribri reference	Bribri translation	Observations
1. The tiger says: ”Alright”.	Na̠mù̠ **tö** i chè : “ ë̀kë̀kë ” .	Na̠mù̠ **tö** i che : “ ë̀kë̀kë ” .	Correct ergative marker *tö* Wrong tone for the verb.
2. You see her.	Be' **tö** i sa̠wè̠	Be' **tö** i sa̠wè̠	Correct ergative marker *tö*.
3. Do you see that house [up there far]?	Be**'r** ù aì sa̠wè̠ ?	Be' ù sa̠wè̠ ?	Missing ergative marker *dör/tö/'r*.

Table 5: Bribri ergatives in Spanish→Bribri models.

the word three is *ma̠ñá̠tkue*. There are numerous other numerical classes, such as flat objects, bunches and cylindrical objects. This phenomenon is not exclusive to Bribri, and is also present in languages like Mandarin Chinese and Japanese (e.g. *inu ippiki* 'one small animal of dog', *terebi ichidai* 'one machine of TV', *gakusei hitori* 'one person of student').

The model is relatively successful in learning the numerical classifiers of Bribri, as can be seen in the first two examples of table 6. Even when the model makes a mistake with the number, as in examples 4 and 6, it gets the classifier category correctly. The most common mistake seems to be confusing numbers with abstract quantities. In example 3, the reference sentence had the word “one [different] human”, but the target sentence has “a human”. In example 5, the word “five people” is replaced by “many”.

Finally, Bribri has a very complex system of demonstratives. Where English has the words *this* and *that*, Bribri demonstratives distinguish two spatial axes: closeness to the speaker (near, far) and relative position to the speaker (above, same level, underneath). This leads to structures like *dù aí* 'that bird up there, nearby', *dù aì* 'that bird up there, far away', *dù a̠wì̠* 'that bird as the same level as me, far away' and *dù dià* 'that bird down there, far away'.

The model seems to have difficulty coping with this complex system. Table 7 repeats previous examples for clarity and adds two more examples of issues with demonstratives. In examples 1 and 2, the system simplified the demonstrative to more general forms (*èt* 'one flat thing' instead of *aí* 'up there near'; *e'* 'that one' instead of *a̠wì̠* 'that one same level far'). In example 3 the demonstrative is simply not present in the target translation. Example 4 was very close to having the correct demonstrative, but the model got the tone wrong, changing the demonstrative from “near” to “far”.

One important detail to mention is that these errors can only partially be attributed to insufficient input from the Spanish source sentence. In sentence 3, the Spanish word *aquella* 'that one over there' might indeed be insufficient to determine the correct Bribri demonstrative. However, sentences 1, 2 and 5 have enough information in the Spanish source to perform an approximation of the Bribri demonstrative.

The issues discussed in this section serve as an illustration of some of the limits of the model. However, from the errors we can see that the model is using its very scarce data to learn some representations of Bribri morphological phenomena which have multiple mappings to the Spanish source sentences.

English	Bribri reference	Bribri translation	Observations
1. I have three chickens.	Ye’ wa̱ dakarò tso’ **ma̱ñà̱t** .	Ye’ wa̱ dakarò tso’ **ma̱ñà̱t** .	Correct classifier: *-t* ‘thing’
2. They have three children.	Ie’pa alà ta̱’ **ma̱ñá̱l** .	Ie’pa alà ta̱’ **ma̱ñá̱l** .	Correct classifier: *-l* ‘person’
3. Here comes another helper,	i’e̱ rö sini̱’ **èköl** dàtsi̱ ,	i’ rö i sini̱’ dàtsi̱ ,	Missing classifier: *èköl* ‘one person’ Locative error: *i’*: ‘this’, *i’e̱*: ‘here’.
4. Then there were three houses.	E’ ta̱ ù dé **ma̱ñá̱tkue** .	E’ ta̱ ù dé **bö̀tkue** .	Correct classifier: *-tkue* ‘building’. Wrong number: *ma̱ñá̱-* ‘three’, *bö̀-* ‘two’. Translation means: ‘Then there were two houses’.
5. There are five women in the house.	Alákölpa tso’ **ské̱l** ù a̱ .	Alákölpa tso’ **ta̱î̱** ù a̱ .	Wrong number: *ské̱l* ‘five’, *ta̱î̱* ‘many’. Translation means: ‘There are many women in the house’.
6. That bird (up, near)	dù **aí**	dù **èt**	Demonstrative *aí* ‘up there near’ replaced by correct classifier *èt* ‘one flat thing’.

Table 6: Bribri numerical classifiers in Spanish→Bribri models.

English	Spanish source	Bribri reference	Bribri target	Observations
1. That bird	aquel pájaro arriba cerca	dù aí	dù èt	Demonstrative *aí* ‘up, near’ replaced by correct classifier *èt* ‘one flat thing’
2. That house far away	aquella casa lejos	ù a̱wì̱	ù e’	Demonstrative *a̱wì̱* ‘same level far’ replaced with *e’* ‘that one’
3. Do you see that house?	Ves aquella casa?	Be’r ù aì sa̱wè̱?	Be’ ù sa̱wè̱?	Missing demonstrative: *aì* ‘up there, far away’
4. The shirt is hanging over there.	Allá tengo colgada la camisa.	Apàio a’r a̱wí̱e̱ ye’ wa̱.	A̱@@wì̱e̱ apàio tër .	Wrong demonstrative: falling tone *a̱wí̱*: ‘same level near’, high tone *a̱wì̱*: ‘same level far’

Table 7: Bribri spatial demonstratives in Spanish→Bribri models.

4 Iterative Back-Translation

Back-translation is the technique for leveraging large monolingual corpora and weak translation models to augment parallel datasets (Sennrich et al., 2016a). A translation model is built using real parallel data from the target language into the source language. Then, monolingual sentences in the target language are translated using that model to create synthetic bitext between the two languages. Finally, this synthetic bitext is concatenated with the real bitext and a final model is trained. This simple technique has been shown to increase scores in some models by almost 2 BLEU (Edunov et al., 2018). Iterative back-translation builds upon this technique by using the first synthetic bitext and real bitext to train a second target-to-source translation model to create new synthetic bitext that is, in theory, more accurate (Hoang et al., 2018).

4.1 Bribri Back-translation Results

In order to test the effect of back-translation, we used the structure in figure 1. First, we used a portion of our bitext corpus to train a Bribri→Spanish model. This provided a baseline for performance, and will henceforth be called the **Base** model. Second, we took the same bitext subset to train a Spanish→Bribri model. We used this to generate synthetic Bribri sentences out of real Spanish ones. We then combined these synthBribri+realSpanish pairs with the real bitext subset, and used this to train a Bribri→Spanish model that used both real and synthetic data. We will call this the **Synth1** model. Third, we used the concatenation of real bitext and synthBribri+realSpanish set to train a second Spanish→Bribri model. We then generate a second set of synthetic Bribri sentences, using the real Spanish as input again. We combined these new synthBribri+realSpanish pairs with the original bitext to train a second Bribri→Spanish model, which we will call **Synth2**. A separate testing set (from the original Bribri-Spanish bitext) remains the same throughout the cycle so that results from different models are comparable.

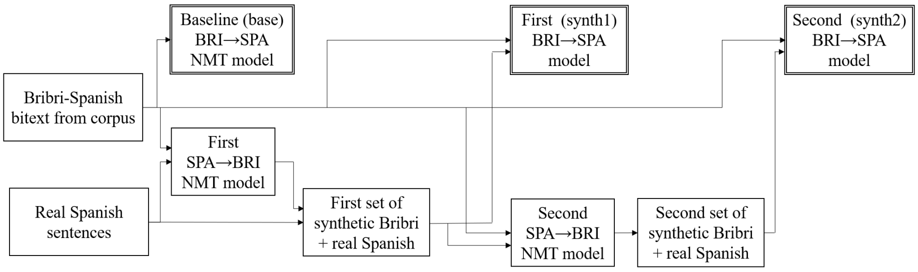

Figure 1: Structure of iterative back-translation.

We test this structure using three models:

1. 1.5K real bitext (1480 total pairs), plus 1.5K (1480) Spanish sentences from the actual Bribri-Spanish corpus. Here, we ignore the real Bribri sentences we have and generate new synthetic Bribri from the **in domain** Spanish data we collected in the corpus.

2. 3K real bitext (2961 total pairs), plus 1.5K (2961) Spanish sentences from the actual Bribri-Spanish corpus. Again, we ignore the real Bribri sentences and generate new synthetic Bribri from the **in domain** Spanish data from the corpus.

3. 6K real bitext (5923 total pairs), plus 6K (5923) Spanish sentences from an **out of domain** source. These sentences come from part of the News-Commentary dataset (Tiedemann, 2012), which contains headlines of daily news, and is therefore very different from the text in the corpus.

We tested each of these five times, using random reshuffling of the data. Table 4.1 shows the average and maximum gains in BLEU scores when we compared the Base and Synth1 models, and the Base and Synth2 models. Figure 2 shows the variation in BLEU gains for the different models.

Real pairs	Synth training	Domain of SPA	Base/Synth1		Base/Synth2	
(train, val, test)	pairs	for synth BRI	AvgΔ	MaxΔ	AvgΔ	MaxΔ
1.5K: 1184, 148, 148	1480	In domain	-2.5±0.8	-1.9	-2.4±1.5	-1.1
3K: 2368, 296, 296	2961	In domain	**1.0±0.9**	2.1	0.8±0.8	1.7
6K: 4737, 593, 593	5923	Out of domain	-1.9±0.5	-1.2	-1.7±0.8	0.1

Table 8: Changes in BLEU scores from adding backtranslated Bribri sentences.

The Synth1 model trained on 3K real pairs (2368 train, 296 val, 296 train) and 2961 synthetic pairs had the best gains relative to Base, with an average gain of BLEU 1 and a maximum gain of BLEU 2.1 (from 10.51 to 12.60). When compared to the 6K models on table 2 (avg: BLEU 16.9; max: BLEU 19.8), we can see that adding 3K synthetic sentences leads to a gain of 22%∼33% that of adding 3K real sentences. While this result is lower than the 67%∼83% reported by Edunov et al. (2018), it shows that synthetic data can produce gains even in very small datasets. The Synth2 models gained much less when compared to the Base, suggesting that there might be limits to the effectiveness of back-translation.

Both of the 1.5K models suffered losses in BLEU. The 1.5K models might simply have too little data to generate appropriate synthetic Bribri. This result is in line with (Przystupa and Abdul-Mageed, 2019), who found that there are limits to how much models can learn from back-translation.

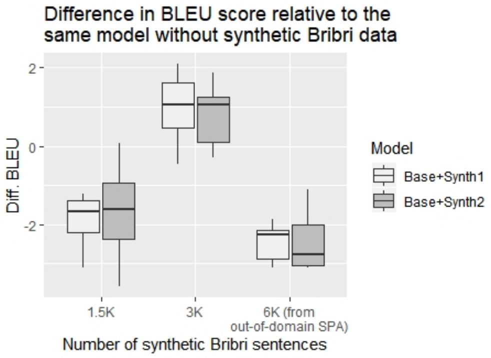

Figure 2: Variation in changes in BLEU scores from adding backtranslated Bribri sentences.

The 6K models also show BLEU losses, but for a different reason: The out-of-domain data used to generate synthetic Bribri might be too different from the real bitext, which is after all mostly composed of stories and classroom examples, not news. This poses a challenge for the effectiveness of back-translation: It might be the case that, in order to benefit from back-translation, Spanish input might need to come from other language learning books, or from traditional stories that are published in Spanish.

4.2 How did translations improve?

Table 9 shows changes in Bribri→Spanish translations between the Base and Synth1 models, built with 2961 real bitext pairs and 2961 pairs of synthetic Bribri and real in-domain Spanish (3K model). Example 1 shows that the system is improving on its understanding of basic sentence structure. Synth1 wrongly interprets the verb *ali'* 'to cook' as its homophone *ali'* 'cassava'. Even if it doesn't get the exact verb right, it generates a translation that has the correct argument structure. Example 2 shows an improved understanding of subject structure and ergativity. The word *bö'* 'you [ERG]' is correctly understood by Synth1, whereas the Base model thinks that the subject is 'we'. Finally, example 3 shows a sentence where the models went from a model without positional information, to one that successfully understood the positional *tulur* 'to be sitting [plural]'. These examples provide evidence that the back-translation is helping the model further its generalizations of Bribri grammar.

BRI source	SPA target	SPA from Base	SPA from Synth1
1. Ye'r i ali'.	Yo lo cociné 'I cooked it'.	Yo lo comió yuca . 'I it ate cassava'.	Yo lo hice. 'I did it'.
3. Ì ká̱wöta̱ ka̱nè̱wè̱ bö' î̱ñe̱?	Qué debes hacer hoy? 'What must you do today?'	Qué vamos a hacer en la tarde ? 'What are we doing this afternoon?'	Qué vas a hacer mañana ? 'What are you doing tomorrow?'
2. Pë' tulur kó̱pàkök.	La gente está sentada conversando. 'People are sitting down chatting'.	Ellos están conversando. 'They are chatting'	(Same as target)

Table 9: Improvements in translation from Base to Synth1 3K models.

5 Conclusions and Future Work

In this paper we described the challenges involved in building a 5923 bitext dataset for Bribri-Spanish, which combines textbook sentences with traditional stories, and standardizes some of the variation found in the data. We built a Transformer-based NMT model with a maximum BLEU of 19.8, and examined some of its errors in learning parts of Bribri grammar that are not found in Indo-European languages.

Finally, we trained a group of models augmented with iterative backtranslation. This technique produced a gain of 22%∼33% in BLEU compared to the gain made by augmenting with real data. More generally, the paper provides evidence that NMT techniques can have acceptable results with extremely low-resource languages, which could help in the process of language documentation.

In future work, this dataset could be expanded to test other NMT techniques such as transfer learning (Zoph et al., 2016) or some unsupervised techniques (Wu et al., 2019; Lample et al., 2017). Given the current amount of bilingual Bribri text, we estimate that the dataset could potentially double in size, but unfortunately there isn't yet enough Bribri text to augment it past that point. Another expansion could be done with Bribri monolingual data, of which there is little (probably less than 5K sentences), but which could be used for dual learning (He et al., 2016).

6 Acknowledgments

The authors wish to thank the Dartmouth College Neukom Institute for Computational Science for its support of this project. They also wish to thank Dr. Felix Stahlberg, Dr. Samantha Wray, Dr. Sofía Flores, Dr. Mario Coto, Dr. Beatriz Coto, Humberto Jiménez, Joan Saborío, Giancarlo Corrales, Viviana Rovira, Gosvami Zúñiga and the three anonymous reviewers for their invaluable help in improving this paper.

References

Neil Bermel. 2007. *Linguistic authority, language ideology, and metaphor: the Czech orthography wars*, volume 17. Walter de Gruyter.

Chenhui Chu and Rui Wang. 2018. A survey of domain adaptation for neural machine translation. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1304–1319, Santa Fe, New Mexico, USA, August. Association for Computational Linguistics.

Adolfo Constenla, Feliciano Elizondo, and Francisco Pereira. 2004. *Curso Básico de Bribri*. Editorial de la Universidad de Costa Rica.

Adolfo Constenla. 1996. *Poesía tradicional indígena costarricense*. Editorial Universidad de Costa Rica.

Adolfo Constenla. 2006. *Poesía bribri de lo cotidiano: 37 cantos de afecto, devoción, trabajo y entretenimiento*. Editorial Universidad de Costa Rica.

Rolando Coto-Solano and Sofía Flores Solórzano. 2016. Alineación forzada sin entrenamiento para la anotación automática de corpus orales de las lenguas indígenas de costa rica. *Kánina*, 40(4):175–199.

Rolando Coto-Solano and Sofía Flores Solórzano. 2017. Comparison of two forced alignment systems for aligning bribri speech. *CLEI Electron. J.*, 20(1):2–1.

Gary F. Simons Eberhard, David M. and Charles D. Fennig. 2020. Ethnologue: Languages of the world. twenty-third edition. sil international.

Sergey Edunov, Myle Ott, Michael Auli, and David Grainger. 2018. Understanding back-translation at scale. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium.

Sofia Margarita Flores Solórzano. 2017. *Un primer corpus pandialectal oral de la lengua bribri y su anotacion morfologica con base en el modelo de estados finitos*. Ph.D. thesis, Universidad Autónoma de Madrid.

Sofía Flores Solórzano. 2019. La modelización de la morfología verbal bribri - modeling the verbal morphology of bribri. *Revista de Procesamiento del Lenguaje Natural*, 62:85–92.

Sofía Flores Solórzano. 2010. Teclado chibcha: Un software lingüístico para los sistemas de escritura de las lenguas bribri y cabécar. *Revista de Filología y Lingüística de la Universidad de Costa Rica*, 36(2):155–161.

Sofía Flores Solórzano. 2017. Corpus oral pandialectal de la lengua bribri.

Candace Kaleimamoowahinekapu Galla. 2016. Indigenous language revitalization, promotion, and education: Function of digital technology. *Computer Assisted Language Learning*, 29(7):1137–1151.

Alí García Segura. 2016. *Ditsö Rukuö Identity of the Seeds: Learning from Nature*. IUCN.

Hany Hassan, Anthony Aue, Chang Chen, Vishal Chowdhary, Jonathan Clark, Christian Federmann, Xuedong Huang, Marcin Junczys-Dowmunt, William Lewis, Mu Li, Shujie Liu, Tie-Yan Liu, Renqian Luo, Arul Menezes, Tao Qin, Frank Seide, Xu Tan, Fei Tian, Lijun Wu, Shuangzhi Wu, Yingce Xia, Dongdong Zhang, Zhirui Zhang, and Ming Zhou. 2018. Achieving human parity on automatic chinese to english news translation. *CoRR*, abs/1803.05567.

Di He, Tao Xia, Yingceand Qin, Liwei Wang, Nenghai Yu, Tie-Yan Liu, and Wei-Ying Ma. 2016. Dual learning for machine translation. In *30th Conference on Neural Information Processing Systems (NIPS 2016)*, Barcelona, Spain.

Leanne Hinton. 2014. Orthography wars. In Michael Cahill and Keren Rice, editors, *Developing orthographies for unwritten languages*, pages 139–168. SIL International Publications.

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

Carla Victoria Jara Murillo and Alí García Segura. 2013. *Se' ttö' bribri ie Hablemos en bribri*. EDigital.

Carla Victoria Jara Murillo. 2018a. *Gramática de la Lengua Bribri*. EDigital.

Carla Victoria Jara Murillo. 2018b. *I Ttè Historias Bribris*. Editorial de la Universidad de Costa Rica, second edition.

Te Taka Keegan, Paora Mato, and Stacey Ruru. 2015. Using twitter in an indigenous language: An analysis of te reo māori tweets. *AlterNative: An International Journal of Indigenous Peoples*, 11(1):59–75.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proc. ACL*.

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, August. 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, June. Association for Computational Linguistics.

Haakon S. Krohn. 2020. Diccionario digital bilingüe bribri.

Guillaume Lample, Ludovic Denoyer, and Marc'Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. *CoRR*, abs/1711.00043.

Brook Danielle Lillehaugen. 2019. Tweeting in zapotec: social media as a tool for language activists. *Indigenous Interfaces: Spaces, Technology, and Social Networks in Mexico and Central America*, pages 201–226.

Ding Lui, Ning Ma, Fangtao Yang, and Xuebin Yang. 2019. A survey of low resource neural machine translation. In *4th International Conference on Mechanical, Control and Computer Engineering (ICMCCE)*, page 390, Hohhot, China.

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, August. Association for Computational Linguistics.

Enrique Margery. 2005. *Diccionario Fraseológico Bribri-Español Español-Bribri*. Editorial de la Universidad de Costa Rica, second edition.

MEP. 2017. *Los Bribri y Cabécares de Sulá, Tomo 1 - Minienciclopedia de los Territorios Indígenas de Costa Rica*. Dirección de Desarrollo Curricular, Educación Intercultural. Ministerio de Educación Pública.

Christopher Moseley. 2010. *Atlas of the World's Languages in Danger*. Unesco.

Niamh Ní Bhroin. 2015. Social media-innovation: The case of indigenous tweets. *The Journal of Media Innovations*, 2(1):89–106.

Sara Pacchiarotti. 2016. Verbal deponency in the chibchan family. In *49th Annual Meeting of the Societas Linguistica Europaea*.

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

Michael Przystupa and Muhammad Abdul-Mageed. 2019. Neural machine translation of low-resource and similar languages with backtranslation. In *Proceedings of the Fourth Conference on Machine Translation (Volume 3: Shared Task Papers, Day 2)*, pages 224–235, Florence, Italy, August. Association for Computational Linguistics.

Carlos Sánchez Avendaño. 2013. Lenguas en peligro en costa rica: vitalidad, documentación y descripción. *Revista Káñina*, 37(1):219–250.

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

Rico Sennrich, Barry Haddow, and Alexandra Birch. 2016b. 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, August. Association for Computational Linguistics.

Carlos Sánchez Avendaño. 2020. *Se' Dalì Diccionario y Enciclopedia de la Agricultura Tradicional Bribri*. DIPALICORI.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In Nicoletta Calzolari (Conference Chair), Khalid Choukri, Thierry Declerck, Mehmet Ugur Dogan, Bente Maegaard, Joseph Mariani, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey, may. European Language Resources Association (ELRA).

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Jiawei Wu, Xin Wang, and William Yang Wang. 2019. Extract and edit: An alternative to back-translation for unsupervised neural machine translation. *CoRR*, abs/1904.02331.

Barret Zoph, Deniz Yuret, Jonathan May, and Kevin Knight. 2016. Transfer learning for low-resource neural machine translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas, November. Association for Computational Linguistics.