Grammar-based Data Augmentation for Low-Resource Languages: The Case of Guarani-Spanish Neural Machine Translation

Agustín Lucas †
agustin.lucas@fing.edu.uy

Alexis Baladón †
alexis.baladon@fing.edu.uy

Victoria Pardiñas †
victoria.pardinas@fing.edu.uy

Marvin Agüero-Torales *§
maguero@correo.ugr.es

Santiago Góngora †
sgongora@fing.edu.uy

Luis Chiruzzo †
luischir@fing.edu.uy

† Instituto de Computación, Facultad de Ingeniería, Universidad de la República, Uruguay
* Global CoE of Data Intelligence, Fujitsu, Spain
§ Universidad de Granada, Spain

Abstract

One of the main problems low-resource languages face in NLP can be pictured as a vicious circle: data is needed to build and test tools, but the available text is scarce and there are not powerful tools to collect it. In order to break this circle for Guarani, we explore if text automatically generated from a grammar can work as a Data Augmentation technique to boost the performance of Guarani-Spanish Machine Translation (MT) systems. After building a grammar-based system that generates Spanish text and syntactically transfers it to Guarani, we perform several experiments by pretraining models using this synthetic text. We find that the MT systems that are pretrained with synthetic text perform better, even outperforming previous baselines.

1 Introduction

From an early age, millions of South Americans grow up and build their reality speaking languages that were used in their region for hundreds of years before the first contact with Europeans. Many of them also have formal schooling in their indigenous languages, as a way to encourage them to forge their identity according to that world vision, carrying with them the rich culture of their communities and passing it on to the next generations. Although some of these indigenous languages are massively spoken, their speakers might not benefit from the recent advances in language technologies (e.g. high-performance machine translators, large language models) due to the lack of data to train models and other difficulties (Mager et al., 2023; Ebrahimi et al., 2023). Most of these languages are considered low-resource languages in NLP research (Joshi et al., 2020).

Low-resource languages suffer from a vicious circle that is hard to break: there is not enough data to build tools, but also there are no tools to massively collect text from the web to train them. That is the case of Guarani (*Avañe'ẽ*), an indigenous language spoken by nearly 10 million people mainly in Paraguay, but also in some regions of Argentina, Bolivia and Brazil. Back in the 17th century, the Tupi-Guarani language family was so widespread in the region that it worked as a *lingua franca* (Boidin, 2020). That legacy remains to this day, as evidenced by loanwords used in regions where Guarani is not directly spoken, like Uruguay (Rodríguez, 2015). Many years of contact with Spanish resulted in a considerable number of varieties, Jopara being the most widely spoken nowadays, used daily by millions of speakers in Paraguay (Estigarribia, 2015). Jopara is a variety that exhibits frequent code-switching between Guarani and Spanish (the two official languages of Paraguay), as well as many adapted and unadapted loanwords and phrases. The resources we use have a strong prevalence of this variety.

In this work, we present an approach that aims to break the aforementioned low-resource vicious circle: through generating synthetic parallel data based on grammar knowledge, we try to augment the available data to enhance Guarani-Spanish Machine Translation (MT). We then train Guarani-Spanish MT systems using different combinations of the datasets in a pretraining and fine-tuning strategy, to understand which datasets could contribute the most to improve the translation and in what scenarios. All the resources used and developed for this work are available on GitHub[1].

2 Related Work

Although Guarani is still considered a low-resource language in the NLP community, it is currently an active area of research and some work has focused on this language in the past few years. In

[1] https://github.com/pln-fing-udelar/guarani-grammar-NAACL2024

Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 6385–6397
June 16-21, 2024 ©2024 Association for Computational Linguistics

particular, considerable effort has been made to build Guarani-Spanish parallel corpora. Chiruzzo et al. (2020) created a parallel corpus based on Paraguayan news, which was later expanded and enriched using other parallel data sources (Alvarez, 2019; Góngora et al., 2021; Mager et al., 2021), resulting in the Jojajovai Guarani-Spanish benchmarking corpus (Chiruzzo et al., 2022a). Since that corpus was made up of several subsets, the text represents different Guarani varieties that show varying degrees of Guarani-Spanish code-switching, such as Jopara or Jehe'a. Furthermore, Guarani is included in Meta's initiative *No Language Left Behind* as an available language in the dataset (Costa-jussà et al., 2022). There is also a small BERT model trained on the Guarani Wikipedia (Agüero-Torales et al., 2023).

Previous efforts have been made to specifically work on Guarani-Spanish MT. Vázquez et al. (2021) won the AmericasNLP 2021 shared task (Mager et al., 2021), which included the Guarani-Spanish pair. Borges et al. (2021) tried to take advantage of the rich morphology of Guarani, incorporating morphological information to the models. Góngora et al. (2022) performed translation experiments to evaluate if pretrained static word embeddings made a difference in MT performance. Google Translate includes Guarani in the available languages since June 2022 (Bapna et al., 2022).

In low-resource scenarios, it is usual to explore strategies to increase the available data by creating synthetic text or reusing the existing text in clever ways. This process is known as Data Augmentation (DA). One of the most used techniques for DA is back-translation, consisting in translating monolingual data to the target language using an already available MT system for the considered language pair (Sennrich et al., 2016). Some works explore more complex strategies to improve the effect of back-translation (Burchell et al., 2022; Ebrahimi et al., 2022). Other works explore completely different approaches for DA, such as obtaining extra sentence pairs by changing common words for rare words in some original sentence pairs from the training set (Fadaee et al., 2017), transforming sentence pairs by cautiously changing the order of words or phrases (Sánchez-Cartagena et al., 2021), generating new data using different statistic-based algorithms (He et al., 2023), or creating multilingual lexicons and following different strategies to use them (Jones et al., 2023).

In this work we follow a different approach: we try to create a large parallel corpus of synthetic data from grammar knowledge, and use it to pretrain a neural translation model. As far as we know, this is the first time an ad hoc grammar is used as DA process for a South American indigenous language.

3 Synthetic Text Corpora

In preliminary experiments, we observed that MT systems for the Guarani-Spanish pair showed low performance when dealing with short sentences. A similar observation can be inferred from Koehn and Knowles (2017), where the BLUE score for the short-length buckets is lower than those for the mid and long-length buckets. In our case, this could be caused by the nature of the used parallel data (Chiruzzo et al., 2020), consisting mainly of long, complex and somewhat formal sentences often found in journalistic writing.

Obtaining pairs of short sentences is not always easy, since they are generally found in social media (usually very noisy), subtitles (usually not available for low-resource languages) or in simple texts for children, which in general have no translation available, and are often hard to find (as digital text) for a low-resource language like Guarani. To tackle this problem, we decided to build a synthetic collection of short texts with their translations, starting from a formal grammar based on the official Guarani grammar (Academia de la Lengua Guaraní, 2018).

<table>
<tr><th>Spanish</th><th>Guarani</th></tr>
<tr><td colspan="2">Part-of-speech: V (verb)</td></tr>
<tr><td>Type: M (main), A (auxiliary), S (semi-auxiliary)</td><td></td></tr>
<tr><td colspan="2">Mood: I (indicative), S (subjunctive), M (imperative), P (participle), G (gerund), N (infinitive)</td></tr>
<tr><td colspan="2">Tense: P (present), I (past imperfective), F (future), S (past perfect), C (conditional)</td></tr>
<tr><td colspan="2">Person: 1 (first), 2 (second), 3 (third)</td></tr>
<tr><td colspan="2">Number: S (singular), P (plural)</td></tr>
<tr><td rowspan="2">Gender: M (male), F (female), C (common)</td><td>Inclusiveness (only for first-person plural): I (inclusive), E (exclusive)</td></tr>
<tr><td>Pronoun position (only for third-person plural): B (before the verb), A (after the verb), 0 (not relevant)</td></tr>
<tr><td colspan="2">Transitivity: I (intransitive), T (transitive), D (ditransitive)</td></tr>
</table>

<table>
<tr><th>Spanish</th><th>Guarani</th></tr>
<tr><td colspan="2">Part-of-speech: N (noun)</td></tr>
<tr><td colspan="2">Type: C (common), P (proper)</td></tr>
<tr><td>Gender: M (male), F (female), C (common)</td><td>Gender: 0 (there is no gender for nouns)</td></tr>
<tr><td colspan="2">Number: S (singular), P (plural), N (invariable)</td></tr>
<tr><td></td><td>Nasalization: N (nasal), O (oral)</td></tr>
</table>

Table 1: Tags used for verbs (top) and nouns (bottom), comparing the ones used for Guarani and Spanish.

Our technique employs two simple grammars, for Guarani and Spanish, that can roughly model the same sentences, and a set of syntactic transfer rules for going from Spanish to Guarani trees. We first use the Spanish grammar to generate short sentences together with their parse trees, and then apply the transfer and morphological generation rules to create the Guarani translation.

3.1 Guarani-Spanish Lexicon

In order to build the grammars, we needed as many Guarani-Spanish word pairs as possible. To obtain them we used a mix of different resources, like Guarani-Spanish bilingual dictionaries (Chiruzzo et al., 2023a), and Spanish words with their morphosyntactic information from the Freeling library (Padró and Stanilovsky, 2012), enriched with annotations from the AnCora-Verbs dataset (Aparicio et al., 2008). Then, we automatically generated the annotations for the Guarani words, according to the morphological rules explained in the Guarani grammar (Academia de la Lengua Guaraní, 2018). Table 1 shows the tags used to annotate verbs and nouns, comparing those needed for Spanish, Guarani or both. The final result of this process is a bilingual lexicon, where both the Guarani and the Spanish words in the pair are annotated with the appropriate tags.

3.2 Generation of the Synthetic Parallel Corpus

We built the grammars using NLTK's feature grammars (FG) (Bird and Loper, 2004), inspired in HPSG grammars (Pollard and Sag, 1994) but with many simplifications[2]. The use of FGs allows to indicate the necessary morphological features for each word in both languages, and also to establish agreement constraints, which are not always the same on each side. For example, both languages have subject-verb agreement on person and number, but Guarani also includes nasal/oral agreement (a feature related to the pronunciation of some vowels and consonants), while Spanish includes determiner-noun gender and number agreement. We model short sentences with appropriate combinations of verbs, nouns, adjectives, pronouns, determiners[3], adpositional phrases and negation.

The types of sentences modeled can have a noun phrase or a pronoun as subject (null subjects are also possible), a verb phrase, and optionally an adpositional phrase. This leaves a total of six basic sentence rules for Spanish, with each rule having one or more transfer rules to Guarani, adding up to twelve. The transfer rules can include more changes, for example some pronouns in Guarani (although not all of them) should be written after the verb, changing them from the natural order in Spanish. One important difference between Guarani and Spanish that the transfer rules take into account is that the latter uses prepositional phrases, while the former uses postpositional phrases. Furthermore, negation is handled differently in both languages, in Spanish there is a separate adverb "no", while in Guarani it is denoted with a circumfix around the verb.

Since NLTK does not include a generation algorithm for FGs, we transform it into a Context-free Grammar (CFG), assigning terminal weights to the words in the vocabulary according to their frequency in the Jojajovai corpus, and generate candidate instances using this CFG[4]. Then, we use the FG to discard the sentences generated with the CFG that are syntactically incorrect. As a result, we obtain correctly generated sentences, enriched with their syntax tree.

Finally, we get the Guarani translation by a transfer approach: we apply rules that transform the Spanish tree into a Guarani tree, following a recursive bottom-up approach, and also inserting intermediate symbols that are solved later at the morphological generation stage (e.g. affix concatenation). Table 2 shows some examples of this process, while appendix A presents a detailed example of a transformation from a Spanish noun phrase to its Guarani counterpart using our process.

We used this process to generate 277,842 sentence pairs, comprising roughly 1M Guarani tokens, which make up what we call the **Synthetic Grammar corpus**.

3.3 Evaluation of the Synthetic Grammar

A sample of 70 sentence pairs generated by our system was evaluated by a native Guarani speaker, fluent both in Spanish and Guarani. For each pair, the evaluator was asked to answer "Yes" or "No"

[2] For instance, we only percolate the morphosyntactic head features we need, and also leave out semantic features.

[3] Determiners in Guarani are modeled using other POS, but for simplicity we make them correspond to the Spanish determiners in the bilingual lexicon.

[4] We use Eli Bendersky's algorithm to generate random strings from a probabilistic CFG – https://eli.thegreenplace.net/2010/01/28/generating-random-sentences-from-a-context-free-grammar

Spanish	**Guarani - Raw transfer**	**Guarani - Final**	**English**
yo no volvía	che ndajevyimi#	che ndajevyimi	I wasn’t going back
tu adolescente creció	ne mitãrusu okakuaakuri #	ne mitãrusu okakuaakuri	your teenager grew up
nuestra ironía no encoge de esta pieza	ore ñembohory nomocha’ĩri # ko kotypy _gua	ore ñembohory nomocha’ĩri ko kotypygua	our irony doesn’t shrink from this room
ellas pasarán	#hikuái ojehuta #	ojehuta hikuái	they will pass

Table 2: Examples for the Guarani-Spanish translation by syntactic transfer. The *Raw transfer* column shows the immediate transformation, including the “move pronoun if needed” (#) and the “concatenate” (_) symbols. A postprocessing step solves those pending symbols, thus generating the sentences shown in the *Guarani - Final* column.

to each of the following statements:

- **Structural Correctness:** The Guarani sentence is structurally correct.
- **Known words:** Every word in the Guarani sentence is a known word.
- **Same information:** The Spanish sentence and the Guarani sentence convey the same information.
- **Word choice:** A Guarani native speaker would express this information choosing those same words.

In the **Structural correctness** category, 87.1% of the sentence pairs looked syntactically correct. Given the morphosyntactic complexity of Guarani, we consider this one of the best aspects of the synthetic text. We have a lower score of 51.4% for the **Known words** category, which is also related to morphological complexity. It seems that some grammar rules were wrongly applied, or in a different order than expected, so some word affixes or contractions did not sound natural. The **Same information** category got an average score of 45.7%, mainly because our process has no way to select a word in context when considering synonyms, so on many occasions a wrong sense was selected. Finally, the fluency metric **Word choice** was the lowest one with 31.4%. This was expected to be very low, as the syntactic transfer approaches generally lack the subtleties of more fluency-oriented methods like those using language models.

Additionally, the evaluator provided qualitative comments to address specific linguistic nuances, with valuable insights into specific areas of improvement. Notable comments included suggestions for refining verb endings, selecting more appropriate vocabulary, and addressing minor syntactic issues.

The overall evaluation results show a positive reception of the synthetic Guarani text, especially considering we are not using this tool to translate sentences directly but as a previous step to train other MT systems, so translations at this stage need not be perfect.

3.4 Translation of the AnCora corpus

The methodology previously described in section 3.2 does not take into account any semantic property of the generated text: it just yields random combinations of words that preserve the grammar but do not necessarily have any meaning. As a strategy to make Guarani text that is more grounded on real examples, we used our grammar to analyze and translate the Spanish text available in the AnCora corpus[5] (Taulé et al., 2008) to Guarani, similarly to the approach presented in (Chiruzzo et al., 2022b). If during the transfer approach no translation is found for a specific word or phrase, then the Spanish words are preserved. As a result we obtained a syntactically correct silver-standard corpus, named **Synthetic AnCora corpus**, to use as extra text to train Guarani-Spanish MT systems. This corpus consists of 14,120 sentence pairs.

Table 3 shows the statistics of both the Synthetic Grammar and the Synthetic AnCora corpora. It is interesting to observe that using AnCora we produce very long sentences in comparison to those generated by the grammar as a standalone system. While our grammar was designed to generate only short sentences, the AnCora sentences are considerably longer, since they are taken from real

[5] Note we did not use the syntactic annotations present in AnCora, but parsed the sentences with our own simple grammar.

	Synthetic Grammar	Synthetic AnCora
Sentence pairs	277,842	14,120
Guarani tokens	999,398	334,976
Guarani tokens/sent	3.60	23.72
Guarani vocabulary	41,202	44,869
Spanish tokens	1,215,305	392,921
Spanish tokens/sent	4.37	27.83
Spanish vocabulary	32,758	34,701

Table 3: Statistics of the synthetic corpora.

Dataset	Sentence Pairs	Guarani Tokens	Spanish Tokens
Jojajovai Train	20,213	309,920	459,629
Jojajovai Dev	5,315	73,414	108,604
Jojajovai Test	5,335	73,111	109,806
Synthetic Grammar	277,842	999,398	1,215,305
Synthetic AnCora	14,120	334,976	392,921
Bible	21,979	372,166	497,815
All (G+A+B)	312,690	1,699,886	2,098,350

Table 4: Size of the different corpora used in the neural translation experiments. *All* refers to the union of the *Synthetic Grammar*, *Synthetic AnCora* and *Bible* datasets.

sources (i.e. news text, similar to the text comprising most of the Jojajovai corpus). Also, the mechanism of keeping the Spanish word if no translation was found generates mixed Guarani-Spanish text. Intuitively this can be thought of as an artificial code-switching, although it is not the same as the real phenomenon observed in previous works for the Jopara or Jehe'a varieties (Estigarribia, 2015; Chiruzzo et al., 2023b).

We did an evaluation of this Synthetic AnCora set in a way similar to the Synthetic Grammar set, having a native speaker evaluate 70 samples of the corpus. The evaluation categories are the same, with the difference being that the **Known words** category now means that all the words *translated* by the process are recognizable in Guarani, keeping in mind that there are many words that were kept in Spanish. The results of this evaluation are largely similar to the previous one, for example 91.4% of the sentences looked **Structurally correct**, even considering some parts were untranslated from Spanish. The **Same information** category was slightly better at 48.6%, and the **Word choice** was slightly worse at 28.6%. However, the category that got a significant improvement was **Known words**, where 84.3% of the translated words were deemed as correct Guarani (up from 52.4%). This is interesting, and might indicate that the words used in these actual examples could be easier to translate than words generated randomly. It is possible that less ambiguous words and more common verb tenses with less morphological complexity are used, so the transfer process does a better job at finding the Guarani correspondence.

However, there were some interesting mistakes in the process that were spotted during this evaluation. For example, in Spanish the prepositions *como* and *para* can also be forms of the verbs *comer* (*to eat*) and *parar* (*to stop*) respectively. The process systematically considers these prepositions as if they were verbs, translating them into Guarani as *ajepy'ajoko* (*I eat* or *I take a bite*) and *ojoko* (*he/she stops*). This could be easily solved by using POS information on the Spanish side, but our current process does not include this.

4 Neural Translation Experiments

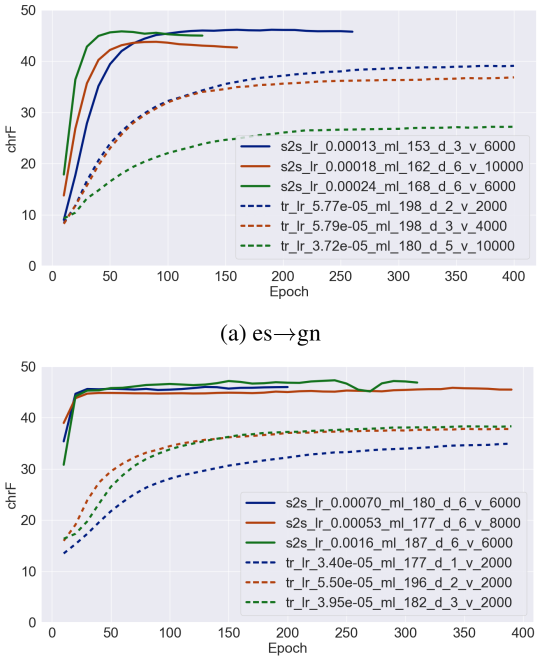
Our neural translation experiments were done in three phases: first we trained simple models with default parameters using the Jojajovai training data to establish basic baselines; secondly we performed a hyperparameter tuning phase where we tried to find the best possible configurations for both the transformer and the seq2seq architectures; finally we experimented with scenarios using separate pretraining and fine-tuning stages, varying the pretraining data and the number of pretraining steps, and then fine-tuning with the Jojajovai training data. All our experiments were done using the MarianNMT (Junczys-Dowmunt et al., 2018) framework, which allows to train neural translation models based on seq2seq (LSTM or GRU with attention) and transformer based models. Table 4 shows the size of the corpora we used in these experiments. All the experiments were run in the ClusterUY (Nesmachnow and Iturriaga, 2019) cluster environment using P100 GPUs, using up to four parallel tasks with 60GB RAM, and an estimated total of 3500 computing hours dedicated to experiments.

4.1 Default and Tuned models

We trained seq2seq and transformer based models using the MarianNMT framework. We first trained baseline models in both directions using MarianNMT's default configuration, and then performed a hyperparameter tuning phase to find the best possible configuration. During this second phase, we

Direction	Model	Epoch	lr	ml	depth	vocab	BLEU	ChrF
es→gn	Default s2s	240	1.00e-4	50	(1,1)	16000	5.51	25.70
	Tuned s2s	240	1.60e-3	187	(6,6)	6000	25.37	47.32
	Default tr.	920	1.00e-4	50	(6,6)	16000	3.01	18.70
	Tuned tr.	920	5.77e-5	198	(2,2)	2000	15.73	40.35
gn→es	Default s2s	190	1.00e-4	50	(1,1)	16000	5.83	30.31
	Tuned s2s	190	1.30e-4	153	(3,3)	6000	23.32	46.20
	Default tr.	800	1.00e-4	50	(6,6)	16000	5.58	26.24
	Tuned tr.	800	3.95e-5	182	(3,3)	2000	15.88	39.72

Table 5: Results of the baseline models trained with the default configuration (Default), and the best models found during the hyperparameter tuning phase (Tuned), evaluated over the Jojajovai dev split.



(a) es→gn

(b) gn→es

Figure 1: Performance of some models over the dev split during training. The transformer models generally took much longer to train, so we cropped the last epochs to fit the graph. They only showed marginal improvements in the remaining epochs.

did a random search varying the following hyperparameters: learning rate (`lr`), max sentence length (`ml`), encoder and decoder `depth`, and vocabulary size (`vocab`). More details about the baseline and hyperparameter tuning phase can be seen in appendix B.

Table 5 shows the results over the Jojajovai dev split, according to the ChrF (Popović, 2015) and BLEU (Papineni et al., 2002) (calculated with the sacreBLEU library[6] (Post, 2018)) metrics, of the default and tuned models, showing the hyperparameter values for each configuration. As can be seen in the table, the transformer based models underperformed in general compared to the seq2seq models. This can also be seen in Fig. 1, which shows the evolution of ChrF performance over the dev set during training for some experimental configurations of the random search. One possible explanation for this could be the lower learning rates we needed to use in order to keep the transformers training stable (see appendix B).

4.2 Pretraining and fine-tuning

Once we finished the previous phase, we obtained sets of hyperparameters that got promising results for the development set, on top of the default parameters that were used for the baseline experiments. This means we have two architectures (seq2seq and transformer), with two hyperparameter configurations each (default or tuned), for each of the two translation directions: a total of eight experimental configurations.

For each of these configurations, we tried a pretraining and fine-tuning strategy: first pretrain using a combination of datasets (all of them larger than Jojajovai but of different nature), and then fine-tune using the Jojajovai training data. The datasets we considered for pretraining are the Synthetic Grammar and Synthetic AnCora sets described in section 3, and the Bible dataset, widely used in low-resource MT and also used in the Jojajovai benchmark experiments. We also performed pretraining experiments using the union of all sets, which as shown in Table 4 has around 1.7M Guarani tokens and 2.1M Spanish tokens.

Besides the selection of pretraining data to use in each experiment, another important variable is how many pretraining epochs are used. In our experiments, we tried a grid combination of the

[6] https://github.com/mjpost/sacrebleu

Hyperp.	Model	Epochs	Synthetic Grammar				Synthetic AnCora				Bible				All			
es→gn		0	1	4	16	64	1	4	16	64	1	4	16	64	1	4	16	64
Default	s2s	18.49	29.45	31.62	28.08	20.79	18.17	18.53	29.35	35.01	18.44	18.95	18.77	17.46	30.38	33.88	34.78	30.24
Default	tr.	22.22	22.40	21.43	19.64	17.72	21.77	22.37	22.42	21.71	20.96	20.87	18.71	18.38	22.44	23.75	21.00	18.61
Tuned	s2s	46.20	43.64	45.79	45.64	44.98	43.83	44.52	46.44	47.77	42.75	43.28	44.96	45.56	43.93	46.09	48.16	**49.07**
Tuned	tr.	40.35	34.01	34.99	38.41	39.02	34.34	34.88	37.69	41.69	34.55	34.30	35.32	35.73	34.32	35.61	40.14	43.80
gn→es		0	1	4	16	64	1	4	16	64	1	4	16	64	1	4	16	64
Default	s2s	22.77	32.77	30.26	28.16	24.66	21.00	21.15	23.78	36.70	22.80	20.86	21.38	20.09	28.42	32.48	33.15	27.05
Default	tr.	25.40	25.95	24.91	22.59	20.18	26.30	27.14	26.66	27.37	25.58	25.39	23.29	22.19	26.96	27.66	25.31	22.91
Tuned	s2s	47.32	47.38	46.63	43.16	41.39	45.92	47.03	47.78	48.14	45.17	46.38	47.82	46.25	47.14	**48.33**	47.28	44.94
Tuned	tr.	39.72	35.54	36.90	38.47	38.43	35.39	35.38	36.58	40.88	35.61	36.05	36.37	36.99	36.10	37.27	39.80	43.49

Table 6: ChrF performance of the pretraining and fine-tuning experiments over the dev split. We show the results for each model (seq2seq or transformers) with or wihtout tuned hyperparameters. The results for 0 epochs mean there is no pretraining, then we show the results when pretraining during 1, 4, 16 or 64 epochs with the four combinations of pretraining corpora.

	es→gn		gn→es	
Model	BLEU	ChrF	BLEU	ChrF
s2s - No pretraining	24.40	47.96	25.53	47.24
s2s - Synthetic Grammar	23.48	46.71	25.63	47.44
s2s - Synthetic AnCora	*26.17*	*49.65*	*26.83*	*49.11*
s2s - Bible	24.12	47.40	25.97	47.87
s2s - All	**26.64**	**50.34**	26.20	49.09
tr. - No pretraining	15.96	40.65	4.08	18.50
tr. - Synthetic Grammar	15.00	40.07	14.78	39.36
tr. - Synthetic AnCora	17.70	43.28	17.24	42.44
tr. - Bible	12.08	36.83	14.11	37.96
tr. - All	19.81	45.26	20.37	45.38
Google Translate	19.31	48.92	**26.96**	**50.95**
Jojajovai Base	16.10	29.41	19.06	31.84
Jojajovai Bible	20.77	35.28	19.98	33.31

Table 7: Results of our models and external baselines over the Jojajovai test split. In each column, bold values indicate the best result, and italic values indicate the second best result.

pretraining datasets and number of epochs between 0, 1, 4, 16, and 64. When pretraining 0 epochs, it means using no pretraining at all, which would be analogous to the experiments in the previous phase.

Table 6 shows the results of these experiments. To simplify, we are showing only ChrF values in this table, but the BLEU measures were largely correlated. We can see that in all scenarios, the models with tuned hyperparameters beat the default models, which was expected. Also, the seq2seq models achieved better performance than the transformer based models in all scenarios with tuned parameters, but not always when using default parameters.

As for the use of pretraining data, we can see that when using default parameters, any combination of pretraining data was enough to improve the performance. This was not the case, however, when using tuned parameters, in these cases only pretraining with the Synthetic AnCora corpus and the union of all corpora gave consistent improvements. The best models in both directions were seq2seq models that used all corpora as pretraining, during 64 epochs for the es→gn direction, and 4 epochs for the gn→es direction. It is interesting to see that the union of all corpora gave better results, which could be explained in part by the size of the corpus, but perhaps also due to the combination of contents: large noisy but syntax-preserving data, artificially code-switched data, and archaic but correctly translated data.

5 Results

Section 4 presented the results of our experiments against the dev split of Jojajovai. In this section, we evaluate the models over the test split and compare it to the results found on the original Jojajovai benchmark experiments (Chiruzzo et al., 2022a). We also ran the test data on the Google Translate API on October 15, 2023. As of this date, Google Translate is the only available translation platform that includes the Guarani-Spanish pair. Table 7 shows the test results.

First of all, our best models (pretrained over all the corpora) beat the previous Jojajovai benchmark baselines, and also beat the Google Translate baseline for the es→gn direction. For the opposite direction, our models are very close to the Google baseline, but note that when translating into Spanish, Google is very likely to have a much larger language model that would yield much better results (Bapna et al., 2022), i.e. our models were probably trained on a fraction of that data, and still have almost as good results.

We also notice that the models pretrained only using the Synthetic AnCora set, come in a very close second position for all metrics. This is very interesting, as this dataset is not very large but seemingly still manages to obtain great results, perhaps because it could leverage the linguistically in-

Dir	Metric	Model	abc	anlp	blogs	hackathon	libro_gn	libro_td	seminario	spl
es→gn	ChrF	s2s - All	**58.76**	24.58	32.30	34.69	**30.16**	**39.38**	28.88	48.50
		s2s - AnCora	58.34	23.59	31.55	31.65	28.93	37.00	29.71	46.99
		Google Translate	56.61	**37.05**	**39.38**	**41.71**	28.82	28.15	**35.94**	**49.49**
		Jojajovai Base	37.44	14.10	21.35	20.02	16.98	24.10	19.83	37.49
		Jojajovai Bible	46.14	18.67	25.45	23.39	19.15	28.25	22.32	39.63
	BLEU	s2s - All	**31.45**	3.01	16.10	5.47	7.72	**10.49**	7.78	29.58
		s2s - AnCora	31.16	2.66	15.34	3.67	**10.86**	8.63	8.76	28.38
		Google Translate	23.56	**6.01**	**16.27**	**5.75**	8.30	3.09	**9.00**	**30.01**
		Jojajovai Base	18.24	0.75	7.73	3.09	3.44	5.15	3.02	20.73
		Jojajovai Bible	24.48	1.76	11.26	3.06	7.46	3.38	5.15	23.51
gn→es	ChrF	s2s - All	56.17	21.48	34.54	31.09	28.56	36.02	30.15	**48.61**
		s2s - AnCora	56.31	21.17	33.37	30.34	30.36	**37.69**	30.64	48.58
		Google Translate	**56.73**	**42.04**	**45.25**	**46.32**	**31.88**	29.62	**36.73**	44.49
		Jojajovai Base	40.25	14.77	24.71	19.35	17.15	24.02	23.15	41.68
		Jojajovai Bible	42.03	17.19	25.40	23.58	19.08	26.45	23.05	41.24
	BLEU	s2s - All	30.06	4.33	18.44	14.69	9.70	15.69	9.95	30.59
		s2s - AnCora	**30.83**	4.04	18.13	10.86	10.50	**18.41**	10.10	**31.21**
		Google Translate	30.81	**19.80**	**24.45**	**18.44**	**11.29**	9.02	**13.16**	23.58
		Jojajovai Base	20.84	1.55	11.89	6.45	5.40	10.25	6.37	25.93
		Jojajovai Bible	22.14	2.52	12.50	6.48	7.80	8.56	6.80	25.83

Table 8: Test results broken down by subset, showing our two best models compared to the external baselines. For each subset and translation direction, the best results are shown in bold, both for BLEU and ChrF.

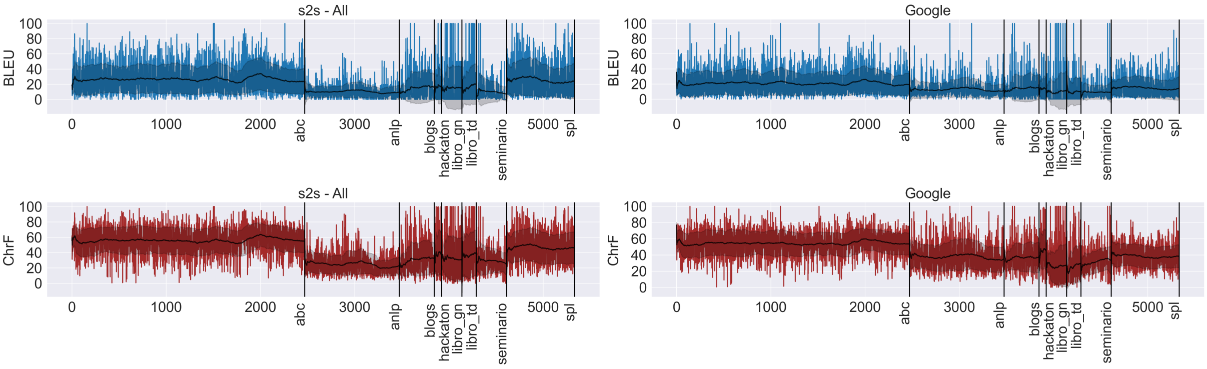


Figure 2: Comparison of BLEU (blue, top) and ChrF (red, bottom) metrics for all samples (X axis) in the test split for our best model (left) and Google Translate (right) in the es→gn direction.

spired syntactic-transfer information with the real-world AnCora text, creating a kind of artificial code-switched data that was very suitable for this dataset. In order to understand better if this is the case, we performed an analysis of performance broken down by subsets in Jojajovai, as shown in Table 8.

Our models beat the Google Translate baseline on the abc, libro_gn and libro_td subsets, and have almost the same performance on the spl subset. The abc and spl subsets are by far the largest of the corpus, and also abc is composed of news text that has frequent code-switching between Guarani and Spanish. We think that our dataset, especially the Synthetic AnCora, would be able to emulate data that is similar to this code-switched data (e.g. sentences with Guarani structure but where named entities and other non-frequent words are kept in Spanish) which could explain the performance gain in this type of text. On the other hand, the dataset for which we have the lowest performance is anlp, from the AmericasNLP (Mager et al., 2021) shared task, which is very different in nature and was built with the intention of having as little interference from Spanish words as possible. Fig. 2 shows a graphical comparison of the Spanish→Guarani translations over the test split for our model and Google's, where we can clearly see the performance differences across the subsets.

6 Conclusions

In this work we presented two main contributions. First, we built grammars for Guarani and Spanish that model roughly the same sets of simple sentences, and transfer rules between the languages, with a strong inspiration in the HPSG formalism. The grammar can be transformed into a probabilistic CFG, with subsequent FG filtering and syntactic transfer, that allows to generate Guarani-Spanish sentence pairs. Additionally, this grammar-based system was applied to the Spanish AnCora corpus to obtain a silver standard Guarani-Spanish parallel corpus.

Second, we evaluate if the generated text has enough quality to boost the performance of Guarani-Spanish MT systems when used as pretraining data, i.e. if grammar-based text generation is a suitable alternative as a DA methodology. We find that the seq2seq model pretrained on all the synthetic text plus the contents of the Bible, and later fine-tuned on the Jojajovai corpus, outperformed the previous Jojajovai baselines in both directions. Our results also outperformed the Google Translate translations for the es→gn setting, while showing competitive performance for the gn→es case.

Even if the sentences in the synthetic corpus sometimes did not sound completely natural to a native speaker, there is no doubt that their content had a positive impact on the models. Therefore, we think this approach can benefit other languages with a well-documented grammar, and also that it can be combined with other strategies, such as back-translation. Unlike back-translation, this synthetic text generation strategy does not need previously trained models or even digitized text to train them. We hope this work can inspire researchers working in low-resource settings, showing that research on grammars for MT could still be very relevant, especially as a way to alleviate data scarcity.

7 Limitations

As evidenced by the evaluation of the synthetic text, the linguistic quality of the synthetic grammar is far from perfect, even having some words that are not recognised by a native speaker. Although we suppose that improving the grammars would improve the text and hence provide a boost during pretraining, we did not test that hypothesis (i.e. we did our best effort to build a single version of the grammar, without testing other versions). Additionally, we did not try alternatives to the decision of keeping the Spanish word if no Guarani translation was found (during the translation of the AnCora sentences using our syntactic transfer approach), so we cannot make observations about that. Therefore, further work would be needed to check if greater effort on a larger grammar is worth it or if it just does not make a difference during pretraining.

All our neural machine translation experiments were done using MarianNMT. However, the Jojajovai benchmark baselines (that we used to guide the development of our models) were trained using OpenNMT (Klein et al., 2017). If we had additionally tried that framework, we could have compared the results as in an ablation process, to check the actual impact of using synthetic text during pretraining.

References

Academia de la Lengua Guaraní. 2018. *Gramática guaraní*. ".Editorial Servilibro".

Marvin M Agüero-Torales, Antonio G López-Herrera, and David Vilares. 2023. Multidimensional affective analysis for low-resource languages: A use case with guarani-spanish code-switching language. *Cognitive Computation*, 15(4):1391–1406.

Aldo Alvarez. 2019. Linguistic hackathon: Accelerating bilingual data generation through collaboration for guarani-spanish language pair. In *Presentation at 8th Podlasie Conference on Mathematics (8th PCM)*, Białystok, Poland.

Juan Aparicio, Mariona Taulé, and M. Antònia Martí. 2008. AnCora-verb: A lexical resource for the semantic annotation of corpora. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Ankur Bapna, Isaac Caswell, Julia Kreutzer, Orhan Firat, Daan van Esch, Aditya Siddhant, Mengmeng Niu, Pallavi Nikhil Baljekar, Xavier Garcia, Wolfgang Macherey, Theresa Breiner, Vera Saldinger Axelrod, Jason Riesa, Yuan Cao, Mia Chen, Klaus Macherey, Maxim Krikun, Pidong Wang, Alexander Gutkin, Apu Shah, Yanping Huang, Zhifeng Chen, Yonghui Wu, and Macduff Richard Hughes. 2022. Building machine translation systems for the next thousand languages. Technical report, Google Research.

James Bergstra and Yoshua Bengio. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(10):281–305.

Steven Bird and Edward Loper. 2004. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pages 214–217, Barcelona, Spain. Association for Computational Linguistics.

Capucine Boidin. 2020. Beyond linguistic description: territorialisation. Guarani language in the missions of Paraguay (17th-19th centuries). In Linda A. Newson, editor, *Cultural Worlds of the Jesuits in Colonial Latin America*, pages 127–145. London University Press. Institute of Latin American Studies.

Yanina Borges, Florencia Mercant, and Luis Chiruzzo. 2021. Using guarani verbal morphology on guarani-spanish machine translation experiments.

Laurie Burchell, Alexandra Birch, and Kenneth Heafield. 2022. Exploring diversity in back translation for low-resource machine translation. In *Proceedings of the Third Workshop on Deep Learning for Low-Resource Natural Language Processing*, pages 67–79, Hybrid. Association for Computational Linguistics.

Luis Chiruzzo, Marvin Agüero-Torales, Aldo Alvarez, and Yliana Rodríguez. 2023a. Initial experiments for building a Guarani WordNet. In *Proceedings of the 12th Global Wordnet Conference*, pages 197–204, University of the Basque Country, Donostia - San Sebastian, Basque Country. Global Wordnet Association.

Luis Chiruzzo, Marvin Agüero-Torales, Gustavo Giménez-Lugo, Aldo Alvarez, Yliana Rodríguez, Santiago Góngora, and Thamar Solorio. 2023b. Overview of gua-spa at iberlef 2023: Guarani-spanish code switching analysis. *Procesamiento del Lenguaje Natural*, 71(0):321–328.

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

Luis Chiruzzo, Santiago Góngora, Aldo Alvarez, Gustavo Giménez-Lugo, Marvin Agüero-Torales, and Yliana Rodríguez. 2022a. Jojajovai: A parallel Guarani-Spanish corpus for MT benchmarking. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2098–2107, Marseille, France. European Language Resources Association.

Luis Chiruzzo, Euan McGill, Santiago Egea-Gómez, and Horacio Saggion. 2022b. Translating Spanish into Spanish Sign Language: Combining rules and data-driven approaches. In *Proceedings of the Fifth Workshop on Technologies for Machine Translation of Low-Resource Languages (LoResMT 2022)*, pages 75–83, Gyeongju, Republic of Korea. Association for Computational Linguistics.

Marta R. Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, Anna Sun, Skyler Wang, Guillaume Wenzek, Al Youngblood, Bapi Akula, Loic Barrault, Gabriel Mejia Gonzalez, Prangthip Hansanti, John Hoffman, Semarley Jarrett, Kaushik Ram Sadagopan, Dirk Rowe, Shannon Spruit, Chau Tran, Pierre Andrews, Necip Fazil Ayan, Shruti Bhosale, Sergey Edunov, Angela Fan, Cynthia Gao, Vedanuj Goswami, Francisco Guzmán, Philipp Koehn, Alexandre Mourachko, Christophe Ropers, Safiyyah Saleem, Holger Schwenk, and Jeff Wang. 2022. No language left behind: Scaling human-centered machine translation.

Abteen Ebrahimi, Manuel Mager, Arturo Oncevay, Vishrav Chaudhary, Luis Chiruzzo, Angela Fan, John Ortega, Ricardo Ramos, Annette Rios, Ivan Vladimir Meza Ruiz, Gustavo Giménez-Lugo, Elisabeth Mager, Graham Neubig, Alexis Palmer, Rolando Coto-Solano, Thang Vu, and Katharina Kann. 2022. AmericasNLI: Evaluating zero-shot natural language understanding of pretrained multilingual models in truly low-resource languages. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6279–6299, Dublin, Ireland. Association for Computational Linguistics.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219, Toronto, Canada. Association for Computational Linguistics.

Bruno Estigarribia. 2015. Guarani-spanish jopara mixing in a paraguayan novel: Does it reflect a third language, a language variety, or true codeswitching? *Journal of Language Contact*, 8(2):183–222.

Marzieh Fadaee, Arianna Bisazza, and Christof Monz. 2017. Data augmentation for low-resource neural machine translation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 567–573, Vancouver, Canada. Association for Computational Linguistics.

Santiago Góngora, Nicolás Giossa, and Luis Chiruzzo. 2021. Experiments on a Guarani corpus of news and social media. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 153–158, Online. Association for Computational Linguistics.

Santiago Góngora, Nicolás Giossa, and Luis Chiruzzo. 2022. Can we use word embeddings for enhancing Guarani-Spanish machine translation? In *Proceedings of the Fifth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 127–132, Dublin, Ireland. Association for Computational Linguistics.

Zexue He, Graeme Blackwood, Rameswar Panda, Julian McAuley, and Rogerio Feris. 2023. Synthetic pre-training tasks for neural machine translation. In *Findings of the Association for Computational Linguis-*

tics: ACL 2023, pages 8080–8098, Toronto, Canada. Association for Computational Linguistics.

Alexander Jones, Isaac Caswell, Orhan Firat, and Ishank Saxena. 2023. GATITOS: Using a new multilingual lexicon for low-resource machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 371–405, Singapore. Association for Computational Linguistics.

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.

Guillaume Klein, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander Rush. 2017. OpenNMT: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, pages 67–72, Vancouver, Canada. 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.

Manuel Mager, Rajat Bhatnagar, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2023. Neural machine translation for the indigenous languages of the Americas: An introduction. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 109–133, Toronto, Canada. Association for Computational Linguistics.

Manuel Mager, Arturo Oncevay, Abteen Ebrahimi, John Ortega, Annette Rios, Angela Fan, Ximena Gutierrez-Vasques, Luis Chiruzzo, Gustavo Giménez-Lugo, Ricardo Ramos, Ivan Vladimir Meza Ruiz, Rolando Coto-Solano, Alexis Palmer, Elisabeth Mager-Hois, Vishrav Chaudhary, Graham Neubig, Ngoc Thang Vu, and Katharina Kann. 2021. Findings of the AmericasNLP 2021 shared task on open machine translation for indigenous languages of the Americas. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 202–217, Online. Association for Computational Linguistics.

Sergio Nesmachnow and Santiago Iturriaga. 2019. Cluster-uy: Collaborative scientific high performance computing in uruguay. In *Supercomputing*, pages 188–202, Cham. Springer International Publishing.

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 of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.

Carl Pollard and Ivan A Sag. 1994. *Head-driven phrase structure grammar*. University of Chicago Press.

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.

Yliana Rodríguez. 2015. Vestiges of an amerindian-european language contact: Guarani loanwords in uruguayan spanish. In *18e Rencontres Jeunes Chercheurs en Sciences du Langage*.

Víctor M. Sánchez-Cartagena, Miquel Esplà-Gomis, Juan Antonio Pérez-Ortiz, and Felipe Sánchez-Martínez. 2021. Rethinking data augmentation for low-resource neural machine translation: A multi-task learning approach. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8502–8516, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

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

Mariona Taulé, M. Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel annotated corpora for Catalan and Spanish. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Raúl Vázquez, Yves Scherrer, Sami Virpioja, and Jörg Tiedemann. 2021. The Helsinki submission to the AmericasNLP shared task. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 255–264, Online. Association for Computational Linguistics.

A Grammar Transfer Example

As seen in section 3.2, the synthetic corpus generation first creates randomized Spanish sentences and uses the grammar and the transfer rules to create their Guarani counterparts. For example, suppose the generator created the Spanish noun phrase *"nuestra amistad"* ("our friendship") and we want to generate the Guarani version to include the pair in the corpus. Our method would first parse the Spanish phrase using the appropriate grammar rule:

```
NP[NUM=?n,GEN=?g] →
D[NUM=?n,GEN=?g] N[NUM=?n,GEN=?g]
```

This includes the agreement features, which in the Spanish case would include the gender and number features. Then the process will look for a Guarani rule associated to this grammar rule. Notice in this case the gender agreement is no longer required but there is a new nasal/oral agreement:

```
NP[NUM=?n,NASAL=?na] →
D[NUM=?n,NASAL=?na] N[NUM=?n,NASAL=?na]
```

Table 9 shows how the method proceeds to find the translation for words and apply the rules to form the noun phrases. In this table, the left side is Spanish while the right side is Guarani. First of all, the bilingual lexicon is used to obtain the word translations. In the first row the Spanish determiner *nuestra* is transformed into its corresponding Guarani words *ore*, *ñande*, and *ñane*, which are possible translations that differ in terms of the nasality and the inclusiveness features. In the second row, the Spanish noun *amistad* is transformed into the Guarani alternatives *joayhu* (oral) and *ñoirũ* (nasal)[7]. These words are combined to form noun phrases, so on the Spanish side we form *"nuestra amistad"* and on the Guarani side we form, respecting nasality constraints, *"ore joayhy"*, *"ore ñoirũ"*, *"ñande joayhy"*, and *"ñane ñoirũ"*. The process preserves all valid translations, while discarding the ones that do not respect agreement rules. In the table we only show the valid combinations.

B Hyperparameter Tuning

As baseline experiments, we trained seq2seq and transformer based models using the MarianNMT framework in its default configuration, in both translation directions. The rows marked as 'Default' in Table 5 show the result of these baselines over the Jojajovai dev split. The first thing we noticed is that these baseline results are much lower than the results obtained in the Jojajovai benchmark experiments (Chiruzzo et al., 2022a), especially for the BLEU metric, where they originally report 19.06 in the gn→es and 16.10 in the es→gn, although we must take into account that their evaluation is over the test split, and in this phase our evaluation is over the dev split. One difference to point out is that work uses the OpenNMT (Klein et al., 2017) framework instead of MarianNMT.

In order to improve these results, we carried on a hyperparameter tuning phase. First we did a series of experiments to gauge which hyperparameters would be more useful and what value ranges we should use for them. In this case, we varied only one hyperparameter at a time, keeping the rest in their default values. We tried varying the encoder and decoder depths, the type of cell for the seq2seq models, the vocabulary size, and the maximum sentence length before cropping. After this stage, we decided to keep only the GRU units for seq2seq models, as the difference with using LSTM units was very small.

Once this was over, we performed a random search for both translation directions and both architectures, training several instances in order to find the hyperparameter combinations that would yield the best performance in each scenario. In these experiments, we varied the following parameters, based on recommendations by (Bergstra and Bengio, 2012) and taking into consideration earlier experiments:

- learning rate (`lr`): between 10^{-4} and 10^{0} for seq2seq, and between 10^{-5} and 10^{-4} for transformers.

- max sentence length (`ml`): normal distribution with mean 200, which was the best value found in the earlier experiments.

- encoder and decoder `depth`: up to 6.

- vocabulary size (`vocab`): between 2000 and 12000, using SentencePiece unigram tokenization.

[7] There is a third valid alternative in Guarani, *tekoayhu*, which also translates as friendship, but it was not present on our bilingual lexicons so it is not covered by the method.

$$
\begin{array}{c|cccc}
\text{es} & \multicolumn{4}{c}{\text{gn}} \\
\hline
\begin{bmatrix}\text{det} & \\ \text{NUM} & \text{s} \\ \text{GEN} & \text{f} \\ \text{POSSPER} & 1 \\ \text{POSSNUM} & \text{p}\end{bmatrix} & \multicolumn{2}{c}{\begin{bmatrix}\text{det} & \\ \text{NUM} & \text{s} \\ \text{NASAL} & \text{-} \\ \text{POSSPER} & 1 \\ \text{POSSNUM} & \text{p} \\ \text{INC} & \text{e}\end{bmatrix}} & \begin{bmatrix}\text{det} & \\ \text{NUM} & \text{s} \\ \text{NASAL} & \text{o} \\ \text{POSSPER} & 1 \\ \text{POSSNUM} & \text{p} \\ \text{INC} & \text{i}\end{bmatrix} & \begin{bmatrix}\text{det} & \\ \text{NUM} & \text{s} \\ \text{NASAL} & \text{n} \\ \text{POSSPER} & 1 \\ \text{POSSNUM} & \text{p} \\ \text{INC} & \text{i}\end{bmatrix} \\
\textit{nuestra} & \multicolumn{2}{c}{\textit{ore}} & \textit{ñande} & \textit{ñane} \\
+ & + & + & + & + \\
\begin{bmatrix}\text{noun} & \\ \text{NUM} & \text{s} \\ \text{GEN} & \text{f} \\ \text{PER} & 3\end{bmatrix} & \begin{bmatrix}\text{noun} & \\ \text{NUM} & \text{s} \\ \text{NASAL} & \text{o} \\ \text{PER} & 3\end{bmatrix} & \begin{bmatrix}\text{noun} & \\ \text{NUM} & \text{s} \\ \text{NASAL} & \text{n} \\ \text{PER} & 3\end{bmatrix} & \begin{bmatrix}\text{noun} & \\ \text{NUM} & \text{s} \\ \text{NASAL} & \text{o} \\ \text{PER} & 3\end{bmatrix} & \begin{bmatrix}\text{noun} & \\ \text{NUM} & \text{s} \\ \text{NASAL} & \text{n} \\ \text{PER} & 3\end{bmatrix} \\
\textit{amistad} & \textit{joayhu} & \textit{ñoirũ} & \textit{joayhu} & \textit{ñoirũ} \\
\downarrow & \downarrow & \downarrow & \downarrow & \downarrow \\
\begin{bmatrix}\text{np} & \\ \text{NUM} & \text{s} \\ \text{GEN} & \text{f} \\ \text{PER} & 3\end{bmatrix} & \begin{bmatrix}\text{np} & \\ \text{NUM} & \text{s} \\ \text{NASAL} & \text{o} \\ \text{PER} & 3\end{bmatrix} & \begin{bmatrix}\text{np} & \\ \text{NUM} & \text{s} \\ \text{NASAL} & \text{n} \\ \text{PER} & 3\end{bmatrix} & \begin{bmatrix}\text{np} & \\ \text{NUM} & \text{s} \\ \text{NASAL} & \text{o} \\ \text{PER} & 3\end{bmatrix} & \begin{bmatrix}\text{np} & \\ \text{NUM} & \text{s} \\ \text{NASAL} & \text{n} \\ \text{PER} & 3\end{bmatrix} \\
\textit{nuestra amistad} & \textit{ore joayhu} & \textit{ore ñoirũ} & \textit{ñande joayhu} & \textit{ñane ñoirũ}
\end{array}
$$

Table 9: Example of translation of a Spanish noun phrase into Guarani using the grammar method.

We also used other techniques to prevent overfit and the vanishing and exploding gradient problems, such as dropout, gradient clipping, and sharing embedding weights at the input and output layers. Notice we used much lower learning rates for transformers, as higher values often resulted in unstable training performance. This resulted in slower training for transformer models, and also possibly obtaining suboptimal local minima.

The random search comprised 20 iterations for each architecture and each translation direction, with a total of 80 experiments. The rows marked as 'Tuned' in Table 5 show the results for the best models found after this phase, and the hyperparameter values used for those configurations. In this case, the best performing seq2seq models already seem to beat the Jojajovai benchmark baselines both in terms of BLEU and ChrF and, as seen in section 4.2, after the fine-tuning phase these resulted in the best performing models.