Jojajovai
A Parallel Guarani-Spanish Corpus for MT Benchmarking

Luis Chiruzzo, Santiago Góngora, Aldo Alvarez,
Gustavo Giménez-Lugo, Marvin Agüero-Torales, Yliana Rodríguez

UNIVERSIDAD DE LA REPÚBLICA URUGUAY
PARAGUAY
UTFPR UNIVERSIDADE TECNOLOGICA FEDERAL DO PARANÁ
UNIVERSIDAD DE GRANADA

Introduction

We present *Jojajovai*, a parallel corpus for the Guarani-Spanish pair aligned at sentence level. It is structured as a collection of subsets from different sources, split into *training*, *dev*, and *test* sets. Even though Guarani, one of the official languages of Paraguay, is not a minority language in terms of its speakers, it is under-resourced and under-researched from a computational linguistics perspective.

The long history of contact between Guarani and Spanish in South America has resulted in many interesting language varieties. Besides focusing on standard **Guarani**, we consider the **Jehe'a** variant, where Spanish loanwords are incorporated with their morphology adapted to Guarani, and the **Jopara** variant, a deeper mixture that often involves code switching and Spanish loanwords that keep their original morphology.

Guarani is an indigenous South American language that belongs to the Tupi-Guarani family. It is spoken by around 10 million people, mainly in Paraguay, but also in other countries in the region, both by indigenous and non-indigenous people. It has been in contact with Spanish and Portuguese for around 500 years, resulting in many language varieties.

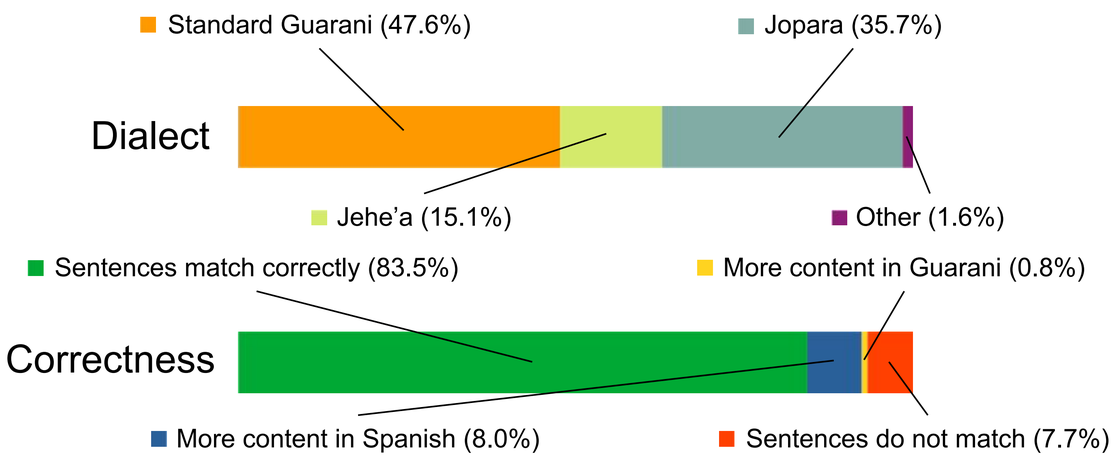
Composition and Analysis of the Corpus

The corpus contains about 30,000 pairs in eight subsets. A sample of each subset was annotated by native speakers to identify the dialects present and the correctness of the translation pairs.

Corpus available here:
https://github.com/pln-fing-udelar/jojajovai

Total Corpus Size

30,855 sentence pairs
20,207 train, 5,314 dev, 5,334 test
538,326 gn tokens
761,812 es tokens



abc
News text from *ABC Color* newspaper, automatically aligned and manually corrected.

16,492 sentence pairs
329,467 gn tokens
474,001 es tokens

dialect
correctness

anlp
Manual translation of a subset of XNLI corpus used in *AmericasNLP* workshop. Dev and test sets only.

2,000 sentence pairs
16,619 gn tokens
24,126 es tokens

dialect
correctness

blogs
Compilation of blog posts including articles, folktales and poems, from different sources such as the *lenguaguarani* blog.

2,444 sentence pairs
31,676 gn tokens
41,468 es tokens

dialect
correctness

hackaton
Manual translation of sentences from *Wikipedia* and *Tatoeba* created during a linguistic hackaton.

513 sentence pairs
2,370 gn tokens
3,607 es tokens

dialect
correctness

libro_gn
Books about terminology and translation guidelines for terms related to science and the internet.

1,423 sentence pairs
5,388 gn tokens
6,958 es tokens

dialect
correctness

libro_td
Issue of the journal *Territorio Digital*, discussing terms related to social networks.

1,016 sentence pairs
3,733 gn tokens
5,525 es tokens

dialect
correctness

seminario
Transcriptions of a seminar on low-resource languages translation, including workshop and papers.

2,179 sentence pairs
35,624 gn tokens
51,435 es tokens

dialect
correctness

spl
News text from the *Paraguayan Bureau of Linguistic Policies*, automatically aligned and manually corrected.

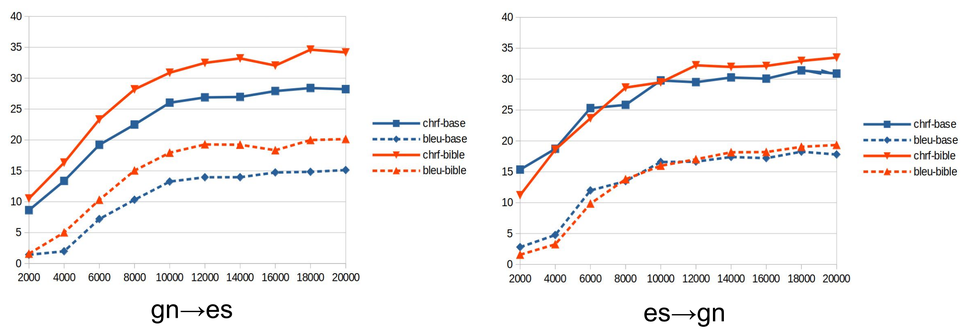
4,788 sentence pairs
113,440 gn tokens
154,692 es tokens

dialect
correctness

MT Baseline

We trained MT baselines with combinations of the *training* set and the Bible, using the OpenNMT tool with its default configuration. Models were trained for 20K steps, the best models checked against the *dev* set were then evaluated over the *test* set.

Dev Results



Test Results

Dir	Model	Metric	Global	abc	anlp	blogs	hackaton	libro_gn	libro_td	seminario	spl
gn→es	base	ChrF	31.84	40.25	14.77	24.71	19.35	17.15	24.02	23.15	41.68
		BLEU	19.06	20.84	1.55	11.89	6.45	5.40	10.25	6.37	25.93
	bible	ChrF	33.31	42.03	17.19	25.40	23.58	19.08	26.45	23.05	41.24
		BLEU	19.98	22.14	2.52	12.50	6.48	7.80	8.56	6.80	25.83
es→gn	base	ChrF	29.41	37.44	14.10	21.35	20.02	16.98	24.10	19.83	37.49
		BLEU	16.10	18.24	0.75	7.73	3.09	3.44	5.15	3.02	20.73
	bible	ChrF	35.28	46.14	18.67	25.45	23.39	19.15	28.25	22.32	39.63
		BLEU	20.77	24.48	1.76	11.26	3.06	7.46	3.38	5.15	23.51