Considering language varieties and language contact in Natural Language Processing and Machine Translation: the case of Guarani

Yliana Rodríguez, PhD & Luis Chiruzzo, PhD

Sadilar Colloquium
October 12th 2022
# Outline

## Introduction

Introduction to language contact, Guarani and NLP.

## The challenges

1. There is lack of data to build a corpus.
2. Guarani has been in contact with Spanish for five centuries, that has consequences.
3. Spelling is inconsistent

## What we are doing

We have been attempting to tackle these challenges in different manners.

Our corpus and some experiments.
Language contact

What is it?
Is it an exception?
What are its implications?
What about languages and their varieties?
How can we set the limits of dialects and languages?
How can we tell one from another one?
How does that affect NLP and MT?
Where is Guarani spoken?
This map, as any linguistic map, is just an approximation to reality.

Where is Paraguay?
Paraguay is in the Southern Cone of South America. Its neighbours are Bolivia, Brazil and Argentina.
First, some facts about Guarani

- Grammars and dictionaries have been published
- Guarani is one of the most widely spoken and studied SA language
- It has received little attention from the technological perspective (Mager et al. 2018)
- Even though it is not a minority language in terms of its speakers, it is under-resourced (Krauwer 2003) and under-researched from a computational linguistics perspective.
First, some facts about Guarani

- Together with Spanish, Guarani is an official language of Paraguay, and it is also widely spoken by its non-indigenous population (Estigarribia 2015).
- Its co-existence with Spanish resulted in the emergence of new varieties and language mixing, which can be traced back to colonial times in the Jesuits' notes, e.g. (Dobrizhoffer 1783).
What we are dealing with

<table>
<thead>
<tr>
<th>Corpus</th>
<th>Language contact</th>
<th>Spelling</th>
</tr>
</thead>
<tbody>
<tr>
<td>There is lack of data to</td>
<td>Guarani and Spanish meet in Jopara, today's Guarani (but not for everyone).</td>
<td>The unbearable lightness of Guarani orthography.</td>
</tr>
<tr>
<td>build a corpus.</td>
<td>Guarani has been in contact with Spanish for five centuries, that has consequences.</td>
<td>There’s an official alphabet only since 2018. Though this story starts with the arrival of European.</td>
</tr>
<tr>
<td>Minority languages tend to</td>
<td></td>
<td></td>
</tr>
<tr>
<td>have little written output,</td>
<td></td>
<td></td>
</tr>
<tr>
<td>and finding parallel texts</td>
<td></td>
<td></td>
</tr>
<tr>
<td>is very difficult.</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Lack of data to build a corpus

Minority languages tend to be written less than other languages, plus, finding parallel texts is an extra difficulty.

- Even when only searching inside the country’s domain (”.py”), the rule-based Guarani detected text is scarce (Góngora et al. 2021).
Spanish–Guarani contact

There are several varieties of Guarani.

- Like all historical languages, Guarani is embodied in a number of varieties, and Jopará is the main dialect used in Paraguay.

- How does that interfere in MT?
The spelling diversity & inconsistency

Until the arrival of Europeans, Guarani did not have a written code. Today, that code is still unstable.

- How did the alphabet develop?
- How does that affect NLP and MT?

The alphabet for Guarani presented by Academia de la Lengua Guaraní in 2018:

```
a ä c h e ĕ g ţ h i ĭ j k l m
m b n nd ng n t ţ ŕ o ŕ r r s t u ū v y ŭ y'
```
What is Natural Language Processing?

Methods and techniques so that computers can understand and generate human language

Examples:
- Web search
- Speech recognition
- Machine translation
- Question answering
What is Natural Language Processing?

- Parsing
- POS Tagging

John eats an apple

```
S
  NP noun
  VP verb
  NP det noun
```

- Parsing
- POS Tagging
Machine Translation

- Rule based approaches
- Statistical models
- Neural models
Rule Based

```
S
  VP
    NP
      V
        hay
    D
    N
      un
      ratón
  PP
    NP
      P
        entre
      D
      N
      esos
      elefantes
  PP
    NP
      D
      N
      umi
      esos
      elefante
      elefantes
      mbytẹpe
      entre
      oї
      hay
      petei
      un
      anguja
      ratón
```
Modern Approaches

Large amounts of data (text)

Mathematical models

Performance metrics

From the text, the system learns to solve a task

Guarani-Spanish parallel text (it has to be aligned)

Neural networks

BLEU and ChrF: how much does the candidate sentence resemble the expected sentence?

We need a large volume of text!
Our Corpus

Crawling the web for content in Guarani and Spanish

- Pages linked to the other language
- Articles that combine both languages
- News sites, blogs

Content that is translated manually

- Wikipedia articles
- Natural language understanding

Digitized books

Around 30K sentences total
Varieties
● Embohasamína ko marandu umi rehayhuvévape...

Por favor, pasa este mensaje a las personas que estimas...

● Afara orenunsiáta ko'êrō

Afara renuncia mañana

● Ojuhúma 52 allanamiento Argentina gotyo ha 21 detenido, 200.000 munición ha 2.500 fusilo jokóva.

En Argentina ya han realizado unos 52 allanamientos, 21 detenidos, 200.000 municiones con 2.500 fusiles secuestrados.
En Itauguá promueven "limpiar" el Congreso

Con el propósito de limpiar al país de los vicios de la política, los itaugüeños expresan su repudio a los corruptos con elementos de limpieza.

Unas 50 personas se encuentran en la plaza de la parroquia Virgen del Rosario de este distrito manifestando su repudio hacia los políticos corruptos.

Durante el encuentro "limpiaron" un muñeco de un diputado acusado de corrupción.

Los itaugüeños quieren demostrar que los paraguayos pueden limpiar el parlamento y por eso se acercaron, con palos de repasar, trapos de piso, escobas y lavandina hasta la plaza parroquial.

La manifestación fue acompañada por aplausos y vítores por parte de los vecinos.
Alignment

- A: Perfect match
- B: Match but Spanish has more info
- C: Match but Guarani has more info
- D: Mismatch
## Machine Translation experiments

<table>
<thead>
<tr>
<th>Dir</th>
<th>Model</th>
<th>Metric</th>
<th>Global</th>
<th>abc</th>
<th>anlp</th>
<th>blogs</th>
<th>hackaton</th>
<th>libro_gn</th>
<th>libro_td</th>
<th>seminario</th>
<th>spl</th>
</tr>
</thead>
<tbody>
<tr>
<td>gn→es</td>
<td>base</td>
<td>ChrF</td>
<td>31.84</td>
<td>40.25</td>
<td>14.77</td>
<td>24.71</td>
<td>19.35</td>
<td>17.15</td>
<td>24.02</td>
<td>23.15</td>
<td>41.68</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BLEU</td>
<td>19.06</td>
<td>20.84</td>
<td>1.55</td>
<td>11.89</td>
<td>6.45</td>
<td>5.40</td>
<td>10.25</td>
<td>6.37</td>
<td>25.93</td>
</tr>
<tr>
<td></td>
<td>bible</td>
<td>ChrF</td>
<td>33.31</td>
<td>42.03</td>
<td>17.19</td>
<td>25.40</td>
<td>23.58</td>
<td>19.08</td>
<td>26.45</td>
<td>23.05</td>
<td>41.24</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BLEU</td>
<td>19.98</td>
<td>22.14</td>
<td>2.52</td>
<td>12.50</td>
<td>6.48</td>
<td>7.80</td>
<td>8.56</td>
<td>6.80</td>
<td>25.83</td>
</tr>
<tr>
<td>es→gn</td>
<td>base</td>
<td>ChrF</td>
<td>29.41</td>
<td>37.44</td>
<td>14.10</td>
<td>21.35</td>
<td>20.02</td>
<td>16.98</td>
<td>24.10</td>
<td>19.83</td>
<td>37.49</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BLEU</td>
<td>16.10</td>
<td>18.24</td>
<td>0.75</td>
<td>7.73</td>
<td>0.309</td>
<td>3.44</td>
<td>5.15</td>
<td>0.302</td>
<td>20.73</td>
</tr>
<tr>
<td></td>
<td>bible</td>
<td>ChrF</td>
<td>35.28</td>
<td>46.14</td>
<td>18.67</td>
<td>25.45</td>
<td>23.39</td>
<td>19.15</td>
<td>28.25</td>
<td>22.32</td>
<td>39.63</td>
</tr>
<tr>
<td></td>
<td></td>
<td>BLEU</td>
<td>20.77</td>
<td>24.48</td>
<td>1.76</td>
<td>11.26</td>
<td>0.306</td>
<td>7.46</td>
<td>3.38</td>
<td>5.15</td>
<td>23.51</td>
</tr>
</tbody>
</table>
Thanks for your attention

Luis Chiruzzo
luischir@fing.edu.uy

Santiago Góngora
sgongora@fing.edu.uy

Yliana Rodríguez
ylianarodriguez@gmail.com
Twitter: @ylirodriguez
IG: @ylianavirginiarodriguez