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

Yliana Rodríguez, PhD
& Luis Chiruzzo, PhD

UNIVERSIDAD
DE LA REPÚBLICA
URUGUAY

SADiLaR

Sadilar Colloquium
October 12th 2022

science & innovation
Department:
Science and Innovation
REPUBLIC OF SOUTH AFRICA

Outline

Intoduction

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?

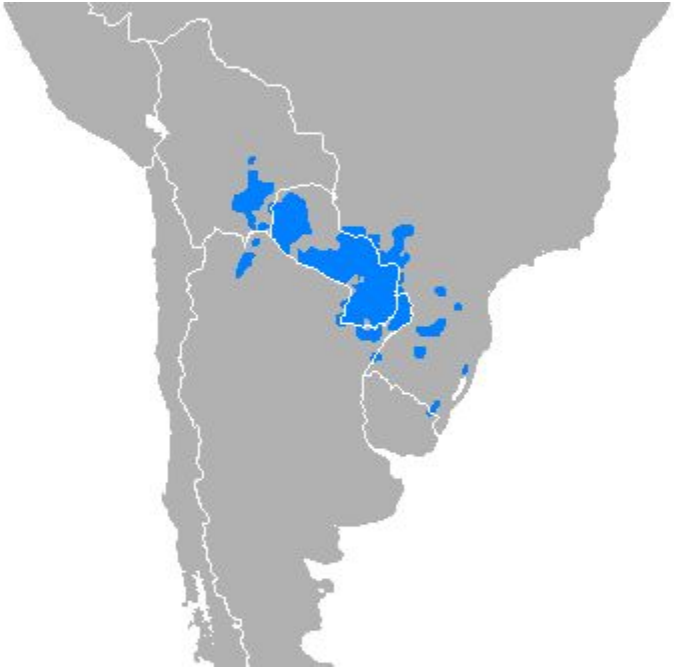
What are its implications?

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
- Guarni 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

Corpus

There is lack of data to build a corpus.

Minority languages tend to have little written output, and finding parallel texts is very difficult.

Language contact

Guarani and Spanish meet in Jopara, today's Guarani (but not for everyone).

Guarani has been in contact with Spanish for five centuries, that has consequences.

Spelling

The unbearable lightness of Guarani orthography.

There’s an official alphabet only since 2018. Though this story starts with the arrival of European.

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?**

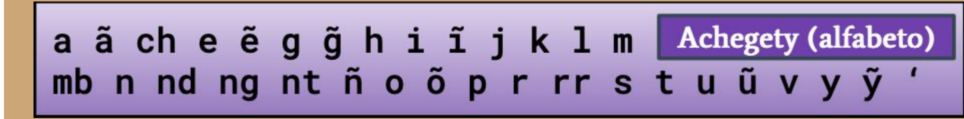
The spelling diversity & inconsistency

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:



Achegety (alfabeto)

a ã ch e ẽ g g̃ h i ĩ j k l m
mb n nd ng nt ñ o õ p r rr s t u ũ v y ỹ ʻ

What is Natural Language Processing?

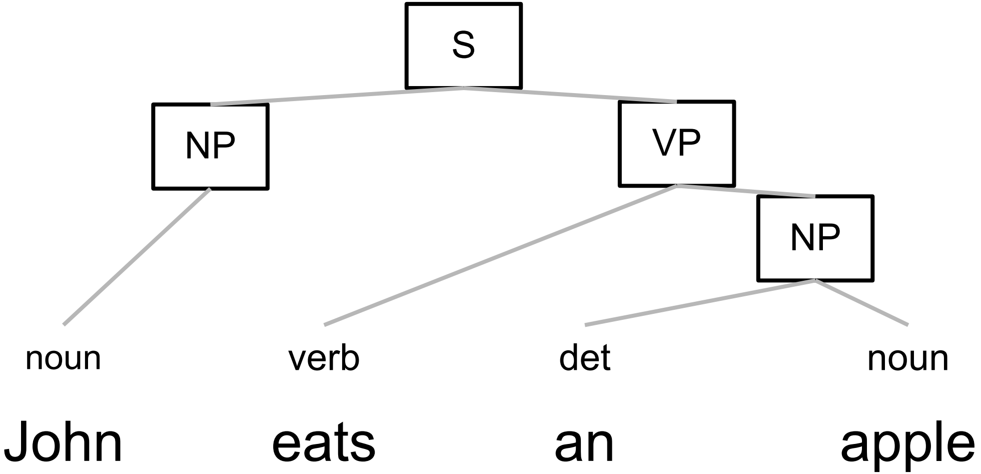
Methods and techniques so that computers can understand and generate human language

Examples:

- Web search
- Speech recognition
- Machine translation
- Question answering

What is Natural Language Processing?

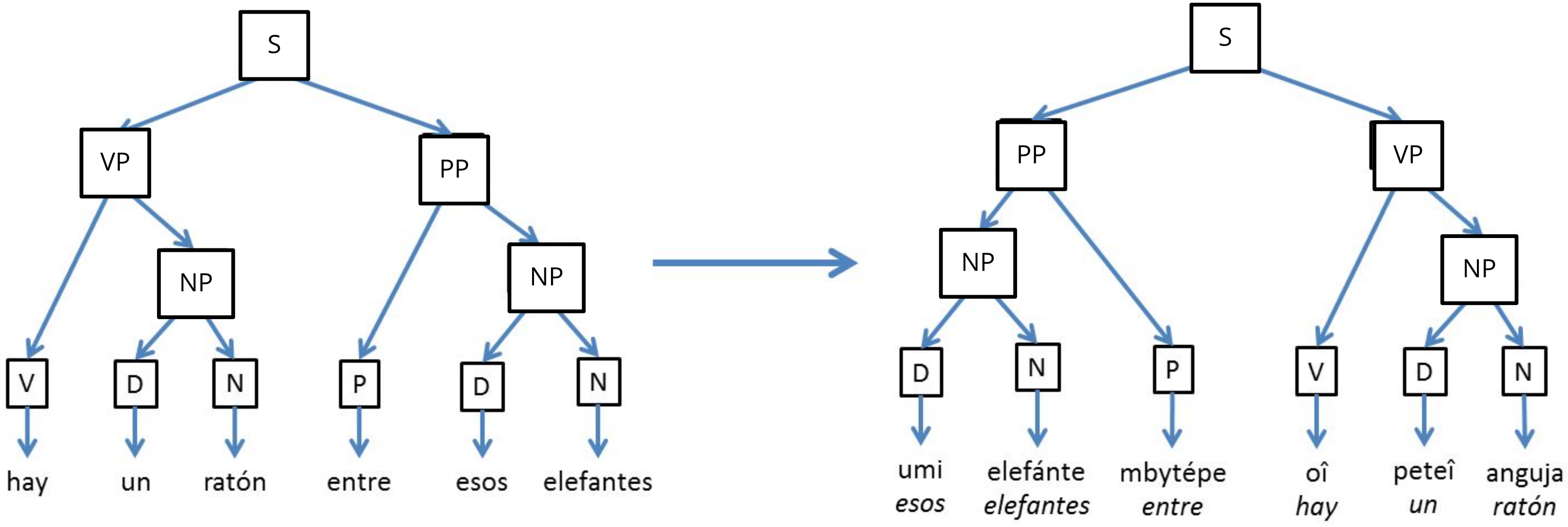
- Parsing
- POS Tagging



Machine Translation

- Rule based approaches
- Statistical models
- Neural models

Rule Based


S
VP
PP
NP
NP
V
D
N
P
D
N
hay
un
ratón
entre
esos
elefantes
S
PP
VP
NP
NP
D
N
P
V
D
N
umi
esos
elefánte
elefantes
mbytépe
entre
oî
hay
peteî
un
anguja
ratón

Modern Approaches

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

Large amounts of data (text)

Neural networks

Mathematical models

Performance metrics

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

From the text, the system *learns* to solve a task

We need a large volume of text!

Our Corpus

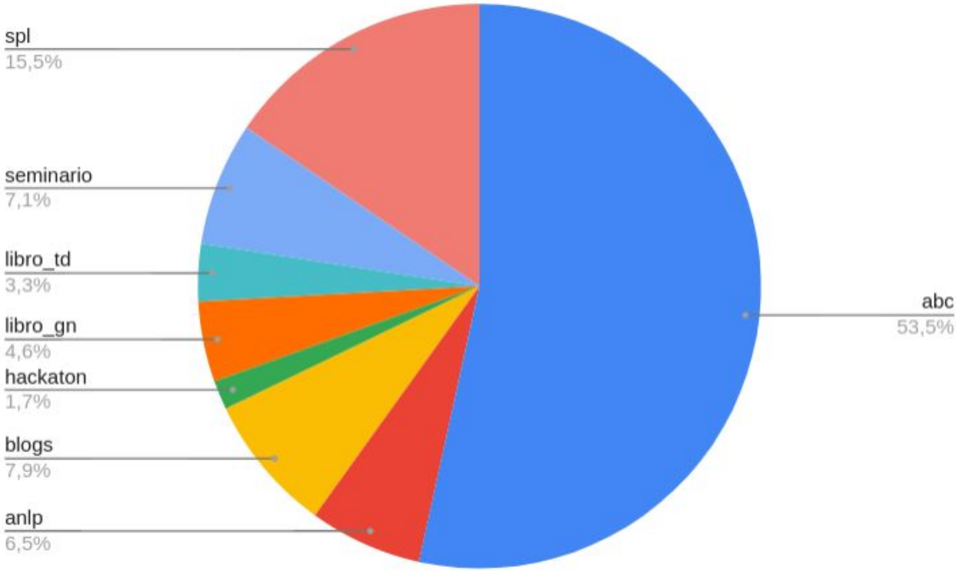
Crawling the web for content in Guarani and Spanish

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

Content that is translated manually

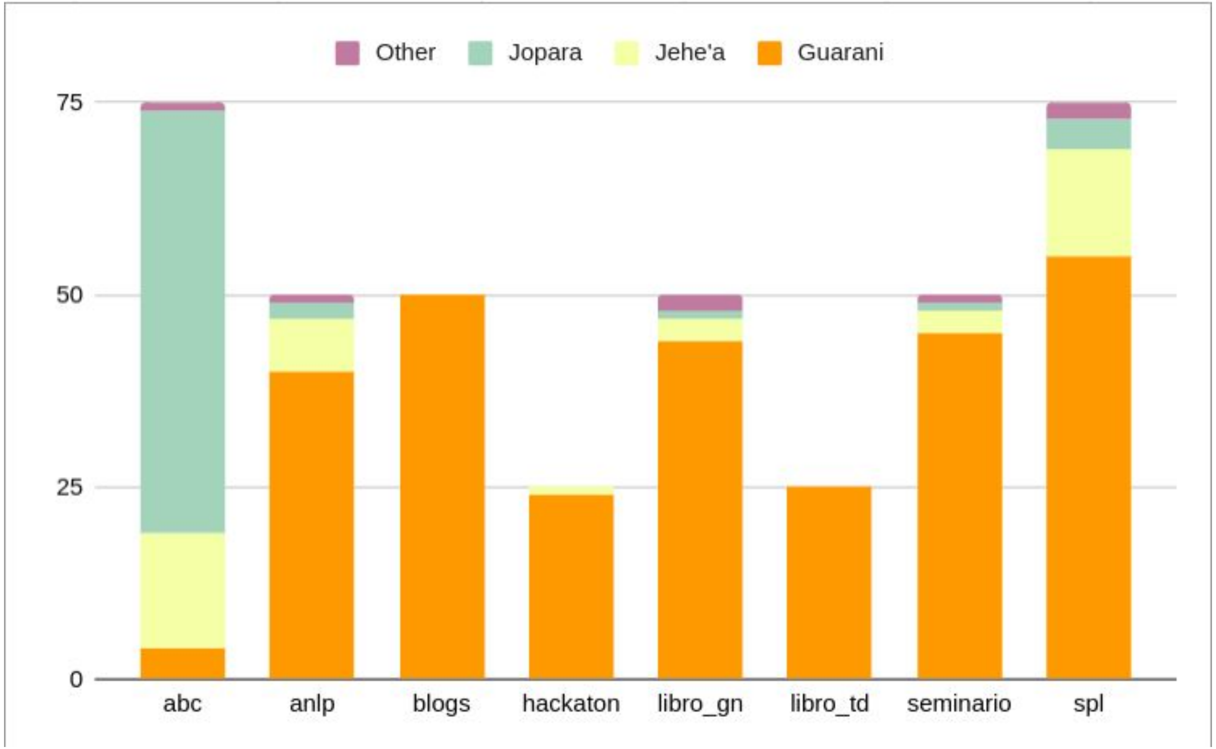
- Wikipedia articles
- Natural language understanding

Digitized books



Around 30K sentences total

Varieties


Other
Jopara
Jehe'a
Guarani
75
50
25
0
abc
anlp
blogs
hackaton
libro_gn
libro_td
seminario
spl

Varieties

- *Embohasamína ko marandu umi rehayhuvévape…*

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

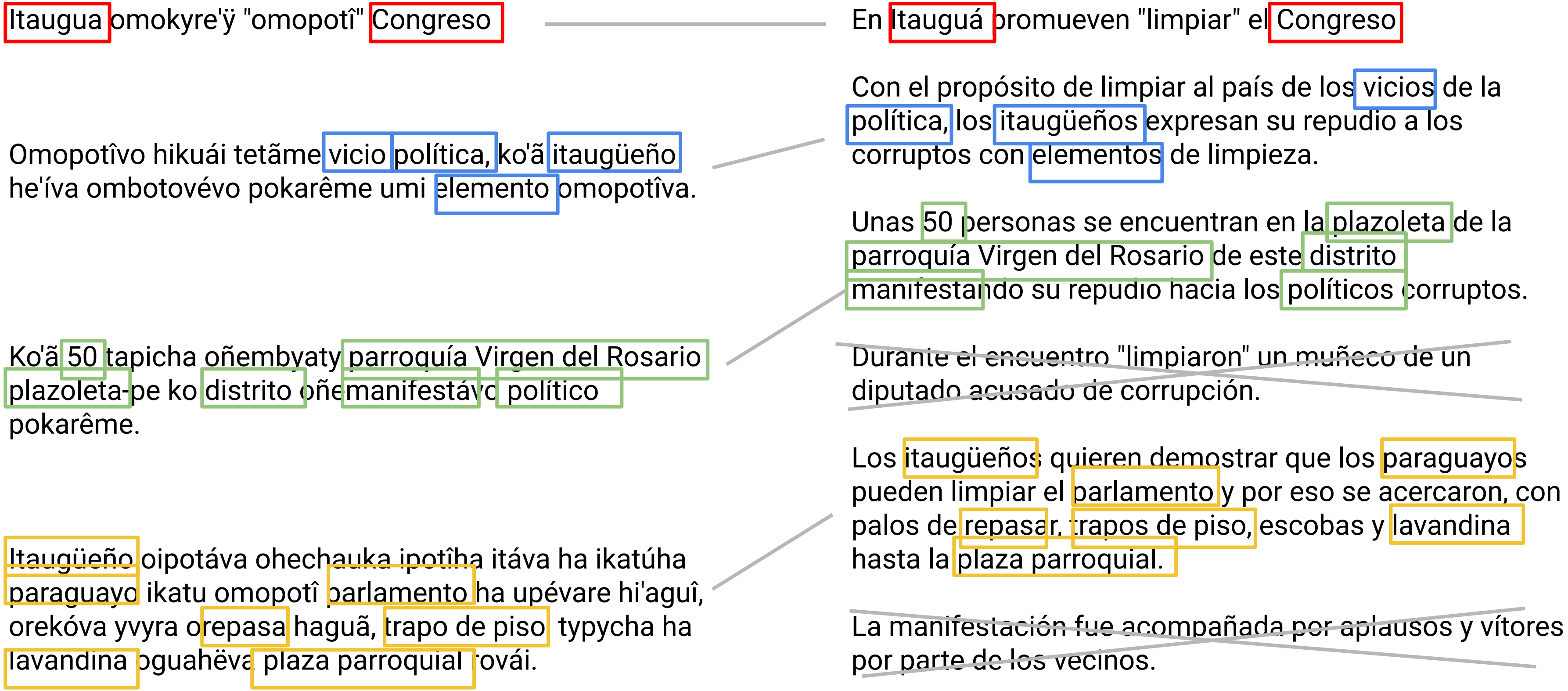
- *Afara orenunsiáta ko'êrõ*

  Afara renuncia mañana

- *Ojuhúma 52 allanamiento Argentina gotyo ha 21 detenido, 200.000 munición ha 2.500 fusil ojokóva.*

  En Argentina ya han realizado unos 52 allanamientos, 21 detenidos, 200.000 municiones con 2.500 fusiles secuestrados.

Alignment



Itaugua omokyre'ÿ "omopotî" Congreso

En Itauguá promueven "limpiar" el Congreso

Omopotîvo hikuái tetãme vicio política, ko'ã itaugüeño he'íva ombotovévo pokarême umi elemento omopotîva.

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.

Ko'ã 50 tapicha oñembyaty parroquía Virgen del Rosario plazoleta-pe ko distrito oñemanifestávo político pokarême.

Unas 50 personas se encuentran en la plazoleta de la parroquía 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.~~

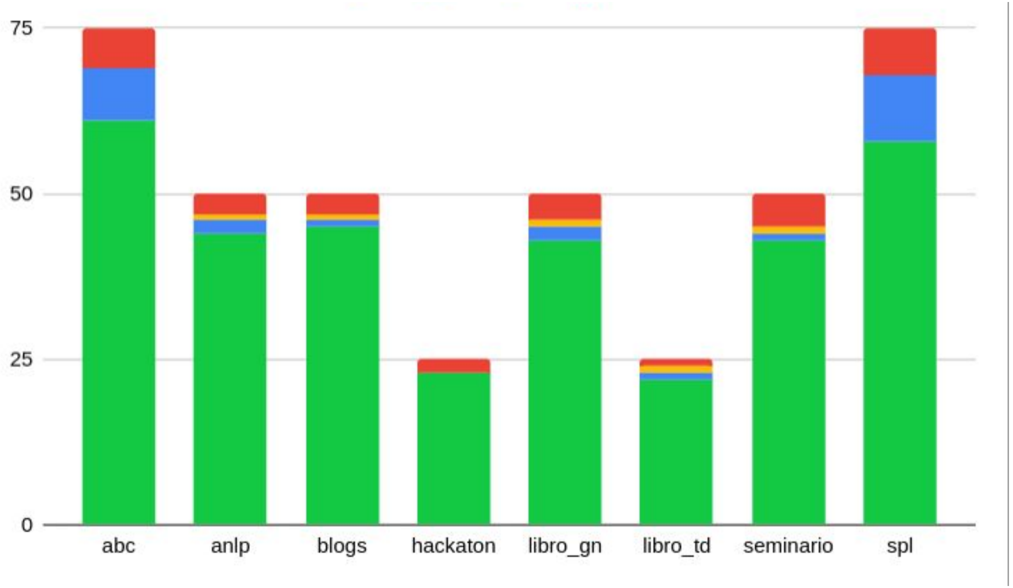
Itaugüeño oipotáva ohechauka ipotîha itáva ha ikatúha paraguayo ikatu omopotî parlamento ha upévare hi'aguî, orekóva yvyra orepasa haguã, trapo de piso typycha ha lavandina oguahëva plaza parroquial rovái.

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

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	03.09	3.44	5.15	03.02	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	03.06	7.46	3.38	5.15	23.51

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