

# THE PROCESS OF COMPILING A **TWITTER CORPUS** FOR **GALICIAN**

Jorge DIZ PICO  
jorge.diz@upf.edu  
@xurxodiz

# GALITUÍTER

- Tweets in Galician
- 1,500,000+ words
- Calendar year 2017
- Work in progress

# TWITTER: WHAT IS IT GOOD FOR?

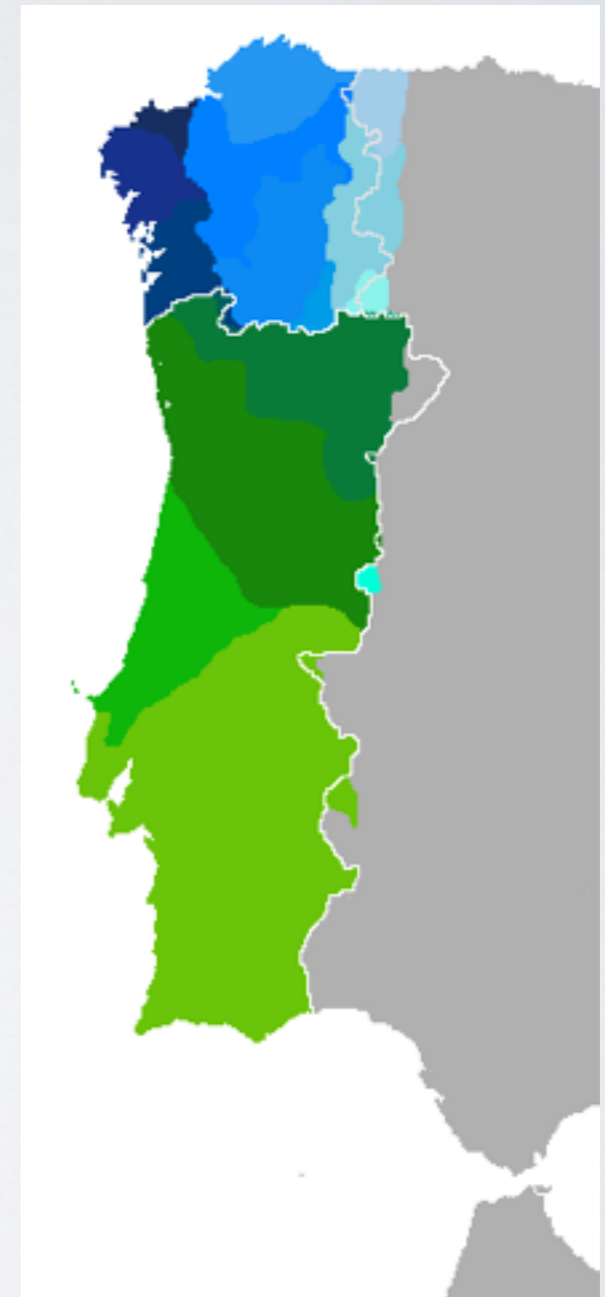
- Spontaneous text excerpts
  - Oral-like, informal productions
  - Not aware of observation

# TWITTER: WHAT IS IT GOOD FOR?

- Huge, constant volume
- Easily collectable (automatized)
- Boon for languages lacking in corpora
- Galician

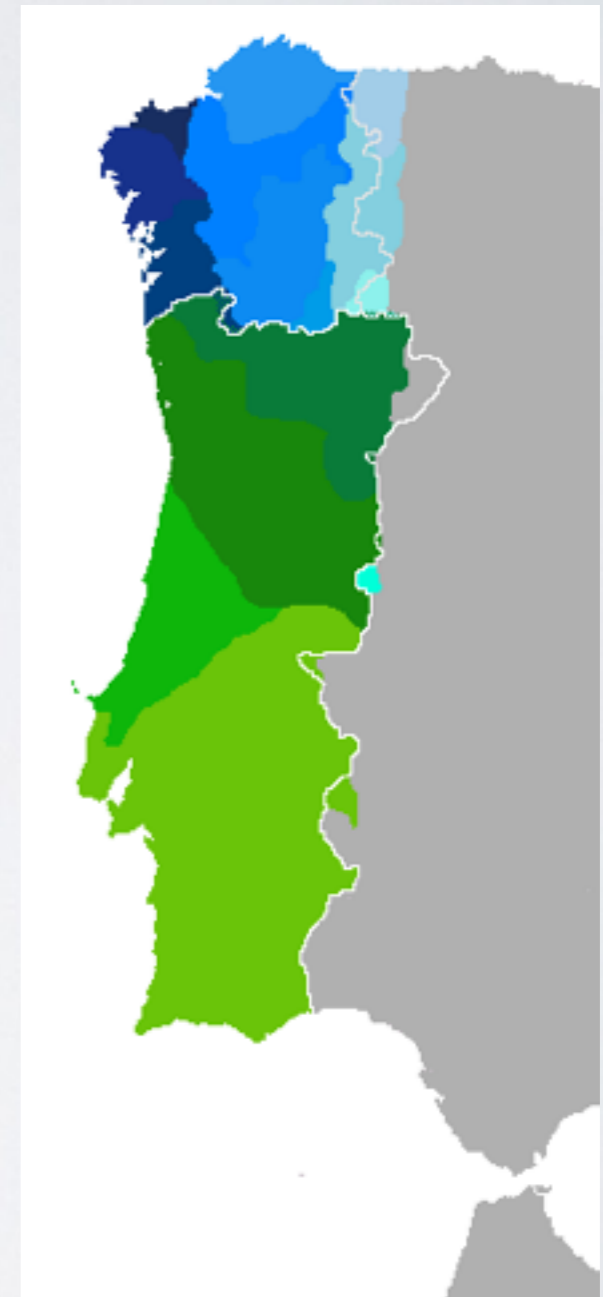
# GALICIAN HISTORY 101

- Romance language
- Northwest Iberian Peninsula
- Extended southwards: Portuguese
- Strong medieval literature



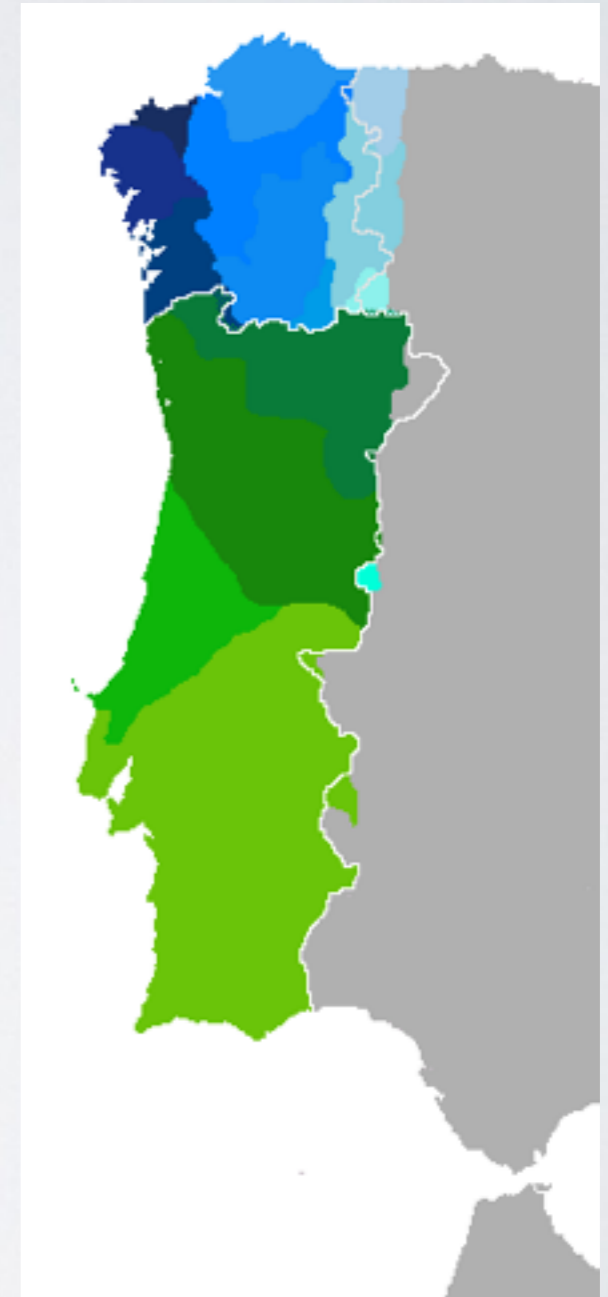
# GALICIAN HISTORY 102

- Portugal splits away, Castile annexes Kingdom
- Banned by Catholic Kings
- *Dark centuries*
- Spoken by 99%, but not written



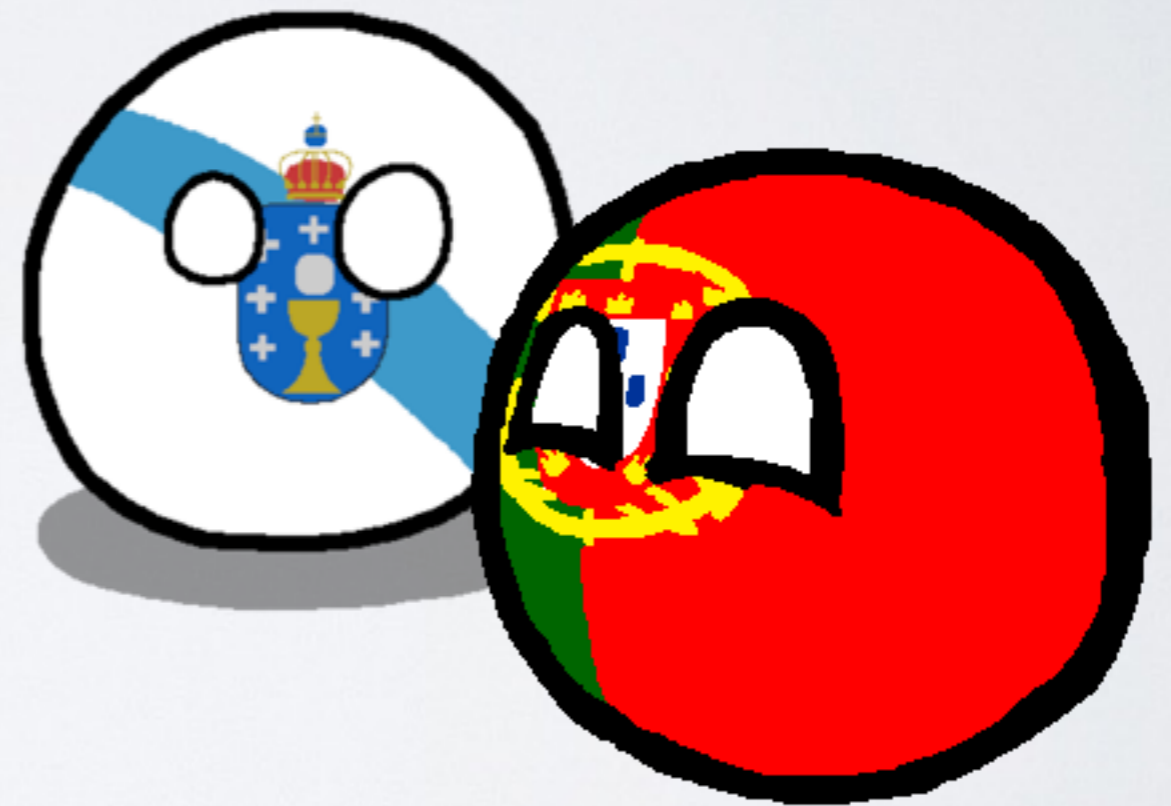
# GALICIAN HISTORY 103

- Literature again XIX c.
- Diversity of standards
- Settled 1982 by political manoeuvres
- ... or was it?



# SIBLINGS OR SPLIT PERSONALITY?

- Autonomistas  
(diverged languages)
- Reintegracionistas  
(dialects of same  
language)





# AUTONOMISTA ORTHOGRAPHY

- Guided by ‘Galicia as a language of its own’
- Based on mid-XIX Spanish-based uses
- Official, taught in schools, known by everyone

# REINTEGRACIONISTA ORTHOGRAPHY

- Guided by re-convergence with Portuguese
- Spiritual successor of medieval uses
- Minor but culturally and politically present

# EXAMPLE



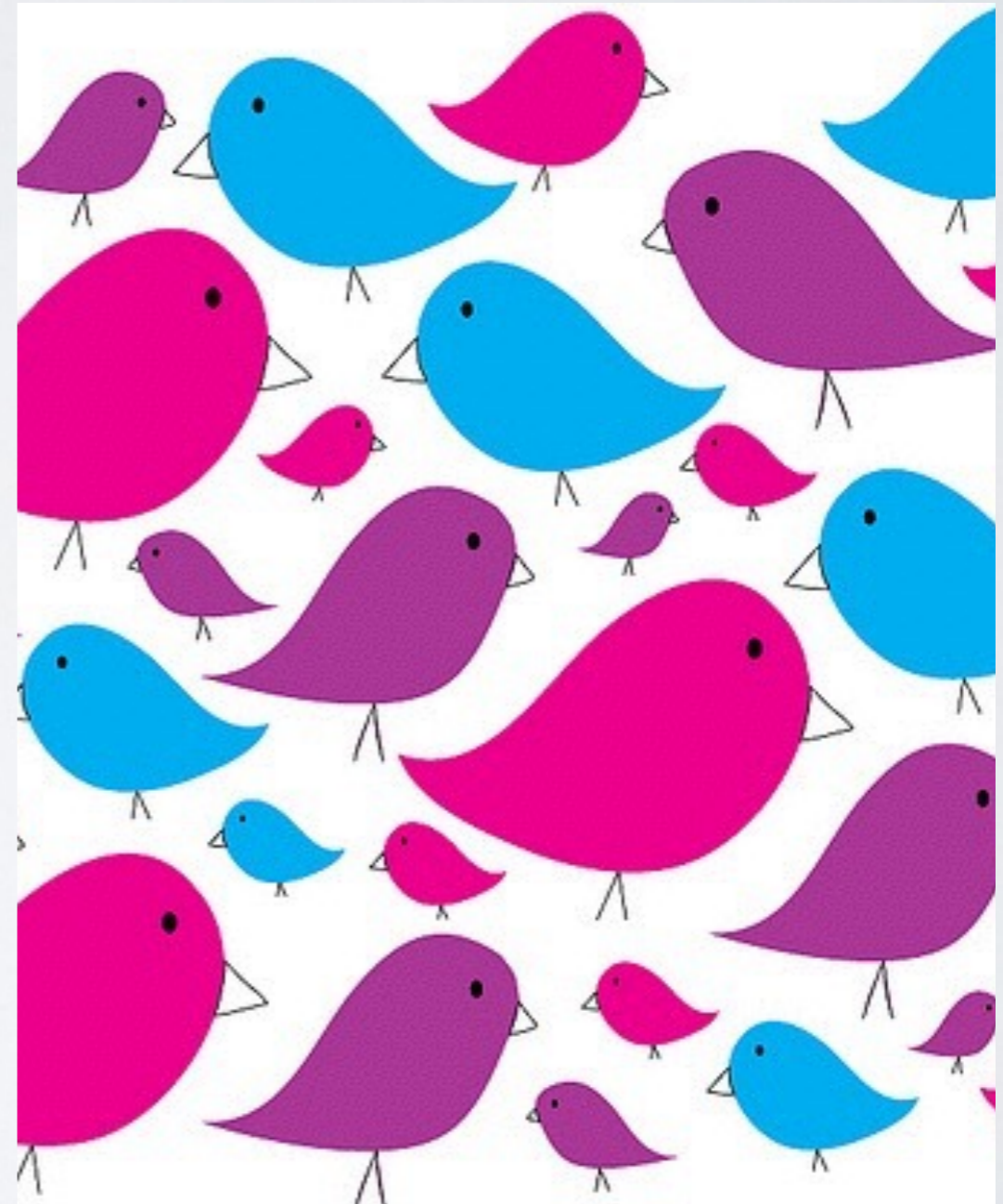
- (ES) Construcción de carteras de inversión. Tendencias.
- (GL) Construcción de carteiras de inversión. Tendencias.
- (GZ) Construção de carteiras de inversão. Tendências.
- (PT) Construção de carteiras de inversão. Tendências.

# CORPUS GOALS

- Full array of linguistic behaviours of Galicians
  - Autonomistas and reintegracionistas
- Specific sociolinguistic context of Galicia

# HOW TO GET TWEETS

1. By keywords
2. By language
3. By users
4. By coordinates



# BY KEYWORDS

- You need to know what you're looking for
- Well suited for specific events or topics
- Ill-suited for everything else

# BY LANGUAGE

- Twitter does not identify Galician tweets
  - it tags them as either Spanish or Portuguese
  - (...for reintegracionistas, that's kind of the point)
- But: it does allow for Galician to be set as interface language

# BY USERS

- How to find Galician-speaking users? By interface
- Setting stored in user metadata
- Compile a list and keep track of their tweets



# WHERE TO START?

- Handpick high profile local accounts
- Check followers and followers of those top 30
- Done by Álvaro Ordóñez (@chio\_en\_galego)



# LIST COMPILATION

- Filter + 12,000 users to top 5000
- Most followers = most active
- Every half hour, get last 100 tweets by them

# LANGUAGE DETECTION

- Galicians are heavily bilingual
- (Twitter) + LinguaKit + LangID + TextBlob
- Current heuristic: all three choose GL/PT

tw_lang	lk_lang	lg_lang	tb_lang
pt	pt	gl	pt
pt	pt	pt	pt
es	gl	gl	gl
pt	gl	pt	pt
es	gl	gl	gl

# METHOD PROBLEMS

- Reintegracionistas not that well represented
  - Many set PT as interface language
  - More likely to be removed from mainstream
- App changes language setting
  - Some might be EN, or ES (iPhone)


# METHOD PROBLEMS II

- Only have a tweet from 3049 of those 5000
  - Only 1789 with  $>2$  tweets and  $>2$  in Galician
- Many media, institutions in top tweeters
- Confined, endemic sample, not so spontaneous



## Galicia Confidencial

@Gconfidencial


 Seguir

Dende 2003 informando en galego. Facebook:  
<http://facebook.com/Gconfidencial> Tamén somos @GcTendencias  
e @GcDeportes Contacto: redessociais@galiciaconfidencial.com  
· <http://www.galiciaconfidencial.com>



## Galiciaé

@GaliciaeXornal


 Seguir

Xornal galego en liña <http://Galiciae.com>  
· <http://www.galiciae.com>



## TVG

@TVGalicia


 Seguir

Benvi@ ao twitter oficial da Televisión de Galicia. Tamén podes  
seguir a actualidade galega en @CRTVG e @radiogalega  
· <http://www.crtvg.gal/tvg>



## Praza Pública

@prazapublica


 Seguir

O xornal da Galicia que vén  
· <http://praza.gal>



## Galicia por Diante

@GxDRadioGalega

 Seguir

Saúde, moi bos días, Galicia! Na @radiogalega, de 6 a 12, de luns  
a venres. Presenta @kikonovoa. En Internet <http://radiogalega.gal>  
/ galiciapordiante@crtvg.gal

# BY COORDINATES

- Set a square around Galicia
- Every half hour, accumulate 50 tweets
- Save country code (to filter out PT)



# METHOD PROBLEMS

- Galician domain outside Galicia (Asturies, León)
  - it's OK (similar sociolinguistic context)
- Abysmal percentage of tweets in Galician: 3.8%
  - By-users ratio: 45.87%





**Fran**

@FranConGafas

Seguir



Just posted a photo @ RESURRECTION  
FEST [instagram.com/p/BWODK4DI8Nnj](https://www.instagram.com/p/BWODK4DI8Nnj) ...

23:04 - 6 xul, 2017 desde Viveiro, España



**Ramón Sas**

@RamonSas

Seguir



I'm at Meson O Pío in Narón, Galicia



**Ramon | Meson O Pío**

Get out. Explore. Download Swarm and live your life more  
checked in.

[swarmapp.com](http://swarmapp.com)

14:33 - 21 xuñ, 2017 desde Narón, España



**Vigo Weather**

@VigoES

Seguir



current weather in Vigo: clear sky, 21°C  
68% humidity, wind 4kmh, pressure 1016mb

07:01 - 20 xuñ, 2017 desde [Vigo, España](#)



**meteolouro**

@meteolouro1

Seguir



Wind 0,5 km/h NNW. Barometer 1002,17  
hPa, Rising. Temperature 26,9 °C. Rain today  
0,0 mm. Humidity 62%

22:00 - 19 xuñ, 2017 desde [O Porriño, España](#)



**Reloj de Vigo** @RelojVigo · 5 xul



Replying to @RelojVigo



Clan Clan



1



**Reloj de Vigo** @RelojVigo · 5 xul



Replying to @RelojVigo



Clan



1



**Reloj de Vigo** @RelojVigo · 5 xul



Replying to @RelojVigo



Clan Clan Clan Clan Clan Clan Clan Clan Clan Clan Clan Clan



1





**Reloj de A Coruña** @RelojMariaPita · 4 xul



Replying to @RelojMariaPita



Tintón Tintón Tintón Tintón Tintón Tintón Tintón Tintón Tintón Tintón Tintón Tintón



1



**Reloj de A Coruña** @RelojMariaPita · 4 xul



Replying to @RelojMariaPita



Tintón Tintón Tintón Tintón Tintón Tintón Tintón Tintón Tintón Tintón Tintón Tintón



1



**Reloj de A Coruña** @RelojMariaPita · 4 xul



Replying to @RelojMariaPita



Tintón Tintón Tintón Tintón Tintón Tintón Tintón Tintón Tintón Tintón Tintón



1



# STATUS QUO AND FUTURE PLANS

- Combined by-users and by-coordinates approach
  - Provides media-leaning group  
+ assorted random people
- Counterbalance coordinates abysmal ratio  
by increasing quantity fetched and filtering



OBRIGADO!

Questions and debate