Universität Stuttgart

# Degrees of Similarity between Spanish and Portuguese Varieties

DGfS 2022
CL-Poster Session

Shuxian Pan
Sabine Schulte im Walde

## Goals

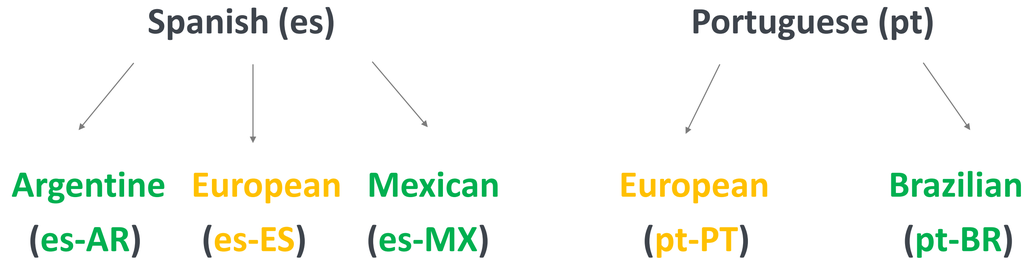
**Similarity of Spanish and Portuguese, from the (morpho)syntactic point of view: comparing the similarity between the varieties of Spanish and Portuguese;**

**(morpho)syntactic features distinguish one variety from the others most.**

## Meta hypothesis

## Origin and Development:

- Roman Empire (Celtic languages + Vulgar Latin)
- Establishment of Kingdoms and Re-coquest
- Colonization (varieties in America)




## Data:

**El corpus del español:** Web/Dialects (general/blog)
AR (≈ 177M), ES (≈ 415M), MX (≈ 247M)

**O corpus do português:** Web/Dalects (general/blog)
PT (≈ 320M), BR (≈ 591M)

**Hypothesis:** Differences in similarity across the varieties; different and shared (morpho)syntactic features in varieties

### verbs

a) alternative of conjugations
   cantas/vives (tú, es-ES or es-MX)
   versus cantás/vivés (vos, es-AR)
   cantas (tu, pt-PT)
   cantais (vós, pt-PT)
   versus canta (você, pt-BR)
   cantan (vocês, pt-BR)

b) tendency of tense and voice usage (es-AR)
   ha llegado < llegó
   [have arrived < arrived]
   hiciera < haga
   [would do < may do]

### clitics

a) proclisis versus enclisis (pt)
   A escola os treinou (...) (pt-BR)
   [The school trained them (...)]
   Ela contou-me a história (...) (pt-PT)
   [She told me the story (...)]

me: proclisis > enclisis  o: enclisis > proclisis (pt-BR)

b) mixture of nominative and accusative forms
   Eu vi ele (...) (pt-BR)  Eu vi-a (...) (pt-PT)
   [I saw him (...)]  [I saw her (...)]

### pronouns

a) alternative of pronouns
   tú (es-ES or es-MX) versus vos (es-AR)
   tu, vós (pt-PT) versus você, vocês (pt-BR)

b) possessive forms (pt-BR)
   seu versus de + ele/ela/eles/elas
   [his/her/their/your versus
   of him/of her/of them]

c) pro-drop versus. non-pro-drop
   Eu vi (...) (pt-BR)
   [I saw (...)]
   (...), (nós) podemos ver o cantor (...) (pt-PT)
   [(...), (we) can see the singer (...)]

## Methods

## Frequency of n-grams

## - Cosine Similarity

## - Chi square

## - Kullback-Leibler divergence (KLD):

corpora overall difference

## - Point-wise KLD/Z-score:

most significant features for the difference

| es-AR-es-ES | |
|---|---|
| ld n | 0.0138 |
| e ld | 0.0103 |
| o o | 0.0075 |
| n e | 0.0044 |
| vi e | 0.0033 |

| es-AR-es-MX | |
|---|---|
| ld n | 0.0078 |
| e ld | 0.0072 |
| n e | 0.0022 |
| vi e | 0.0019 |
| e o | 0.0017 |

| es-ES-es-AR | |
|---|---|
| v vp | 0.0070 |
| r r | 0.0027 |
| po v | 0.0025 |
| r j | 0.0019 |
| cs po | 0.0015 |

| es-ES-es-MX | |
|---|---|
| v vp | 0.0043 |
| r r | 0.0022 |
| li n | 0.0019 |
| r j | 0.0015 |
| po v | 0.0013 |

## Steps

## Current step:

| CosSim-me | pt-PT-B | pt-PT-G |
|---|---|---|
| pt-BR-B | 0.5160 | |
| pt-BR-G | | 0.7214 |

| Chi-verb-2p | pt-BR-G-1 |
|---|---|
| pt-BR-G-2 | 1.29 |
| pt-BR-G-8 | 11.18 |
| pt-PT-G-2 | 2.94 |
| pt-PT-G-5 | 21.77 |

## Further steps:

1) interpret the pointwise KLD list
2) apply other metrics
3) comparison between varieties of different languages
4) plot to visualize the similarity (vector or matrix) of varieties in 2-D space