European Research Council
Established by the European Commission

Universität
Konstanz

Diachronic Null Subject Use across Latin American Spanish: Comparing corpora

Gemma McCarley

ICHL, 07.09.23

STARFISH
SOCIOLINGUISTIC TYPOLOGY
AND RESPONSIVE FEATURES
IN SYNTACTIC HISTORY

Outline

- Background
- My corpus: CorDELES
- Orality
- CORDIAM
- CDH
- Conclusion

Background

- **Spanish is a null subject language (NSL) which means it's perfectly grammatical to "drop" the subject pronoun:**

Spanish [consistent NSL]: (Nosotros) queremos ir a la playa

English [non-NSL (NNSL)]: *(We) want to go to the beach

- **It's been noticed that in Latin American Spanish (LAS) overt pronouns are being used at higher rates, e.g. Dominican Spanish (Toribio 2000):**

1. ***Ellos*** me dijeron que ***yo*** tenía anemia . . . Si ***ellos*** me dicen que ***yo*** estoy en peligro cuando ***ellos*** me entren la aguja por el ombligo, ***yo*** me voy a ver en una situación de estrés.

   'They told me that I had anemia . . . If they tell me that I am in danger when they put the needle in my belly-button, I am going to find myself in a stressful situation.' (Toribio, 2000:319, ex. 3e).

- **This could potentially represent an incipient process towards becoming a NNSL (Camacho 2013)**

- **Why might this be? One of the biggest characteristics of LAS is its history of significant language contact**

Background: Null Subject Acquisition & Simplification

- **When we talk about language contact, we are really talking about language acquisition.**
- **It has been well-noted in the acquisition literature that null subjects are harder to acquire, particularly for L2 speakers (Bini 1993, Pérez-Leroux & Glass 1999, Margaza & Bel 2006, Sorace 2011, Tsimpli & Lavidas 2019)**
- In that case, increasing the use of overt pronouns seems to be an act of *simplification*
- **Language contact, then, is often an impetus for simplification when the simplifying feature is difficult to acquire.**
- Especially when that contact takes the form of short-term, loose-knit, adult language learning (Trudgill 2011, Walkden & Breitbarth 2019)
- **That is exactly the context for African learners of Spanish in colonial Latin America**

Background: AHLAs

- **Specifically, during the colonial period enslaved Africans were brought over to Latin America.**
- **These adult learners of L2+ Spanish might have struggled acquiring the L2-difficult null subject system, preferring overt pronouns.**
- **Their children would then have nativized this system.**
  - This is exactly the scenario Sandro Sessarego (2013) proposes for Latin American Spanish where AHLAs are these nativized varieties.
- **So, the next step would be to look into the diachronic trajectory of pronoun realization and word order in Latin American Spanish.**
- **I'm in the process of creating a corpus of 60+ texts to do just that.**

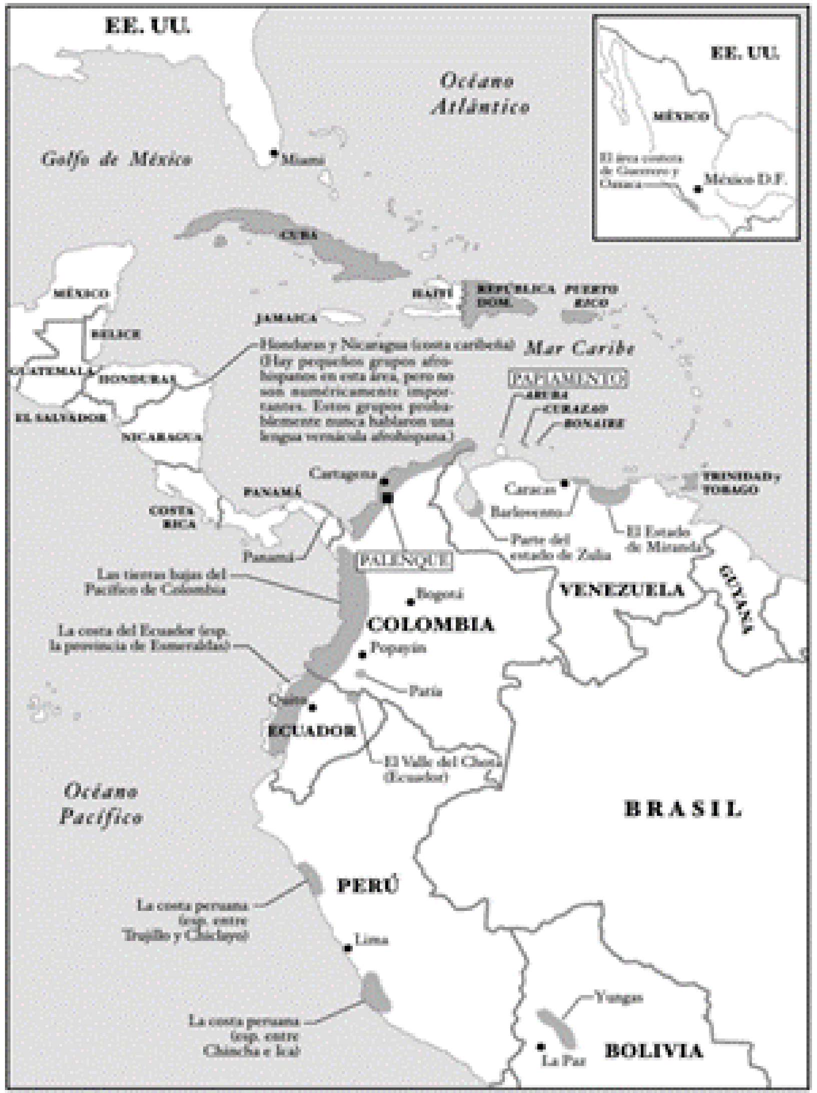

EE. UU.
Océano Atlántico
Golfo de México
Miami
EE. UU.
MÉXICO
El área costera de Guerrero y Oaxaca
México D.F.
CUBA
MÉXICO
HAITÍ
REPÚBLICA DOM.
PUERTO RICO
JAMAICA
BELICE
Honduras y Nicaragua (costa caribeña) (Hay pequeños grupos afro-hispanos en esta área, pero no son numéricamente importantes. Estos grupos probablemente nunca hablaron una lengua vernácula afrohispana.)
Mar Caribe
GUATEMALA
HONDURAS
PAPIAMENTO
ARUBA
CURAZAO
BONAIRE
EL SALVADOR
NICARAGUA
Cartagena
PANAMÁ
COSTA RICA
Caracas
Barlovento
TRINIDAD y TOBAGO
Parte del estado de Zulia
El Estado de Miranda
Panamá
PALENQUE
Las tierras bajas del Pacífico de Colombia
Bogotá
VENEZUELA
GUYANA
COLOMBIA
La costa del Ecuador (esp. la provincia de Esmeraldas)
Popayán
Patía
Quito
ECUADOR
El Valle del Chota (Ecuador)
Océano Pacífico
BRASIL
PERÚ
La costa peruana (esp. entre Trujillo y Chiclayo)
Lima
Yungas
La costa peruana (esp. entre Chincha e Icá)
La Paz
BOLIVIA

Figure 1: Afro-Hispanic areas of Latin America (Klee & Lynch 2009:6)

Research Questions

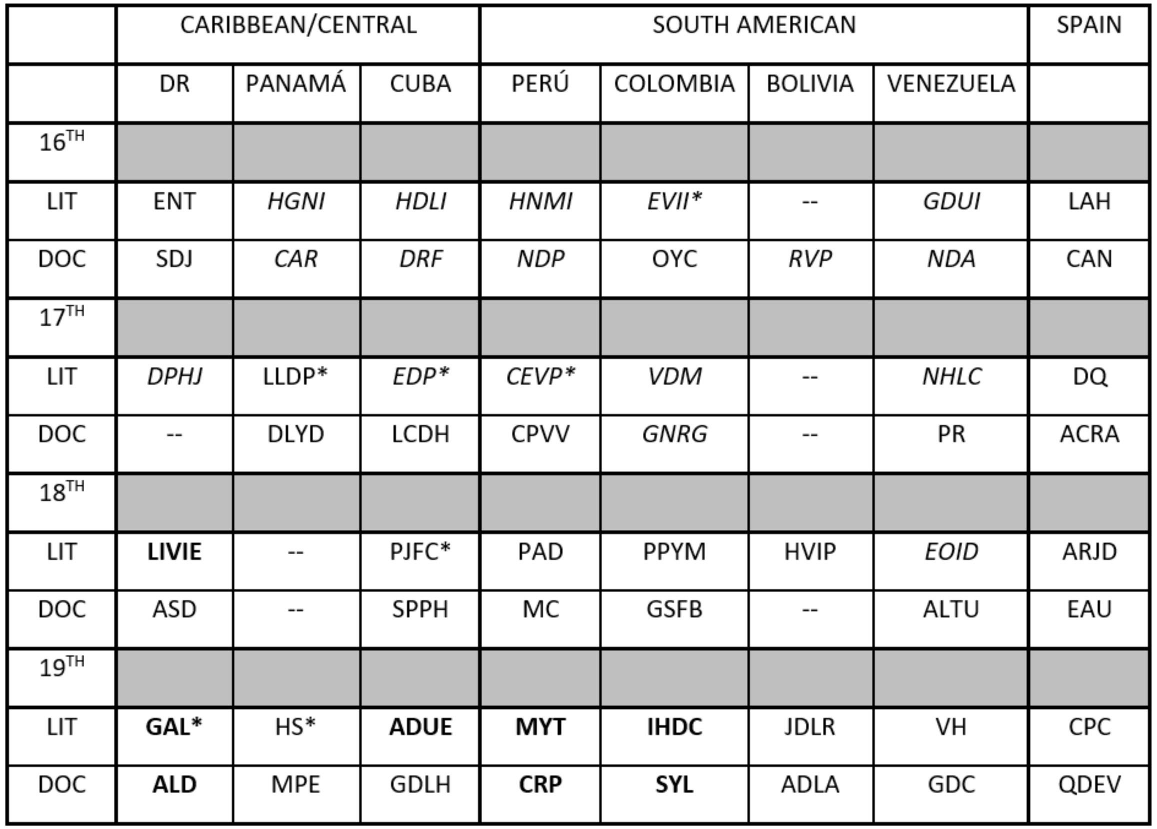
1. *does overtness increase diachronically?*

2. *does it have higher rates from Spain > South America > Caribbean?*

CorDELES

Methodology: CorDELES

- **This is the main historical corpus covering 57 texts (~2-3k words each) from 8 countries during the 16th-19th centuries**
  - I selected 7 countries from the Caribbean and Central and South America (plus Spain as a control)
  - They were selected for their high Afro-Hispanic populations
- **For each century + country combination, there are ideally 2 texts, one from each genre:**
  - Literature (e.g. novels, plays, poetry)
  - Documents (e.g. newspapers, legal documents, letters)
- **In addition to this corpus, I have also set aside:**
  - A transcript of an interview in Afro-Bolivian from 2010
  - 4 texts from Candelario Obeso, a 19th century writer and speaker of Afro-Colombian
- **The main sources for the texts are Cervantes Virtual, dLOC, and BDH**
- **Each text has been transcribed by myself or my research assistant, parsed by the Stanford Parser, and then annotated by hand**



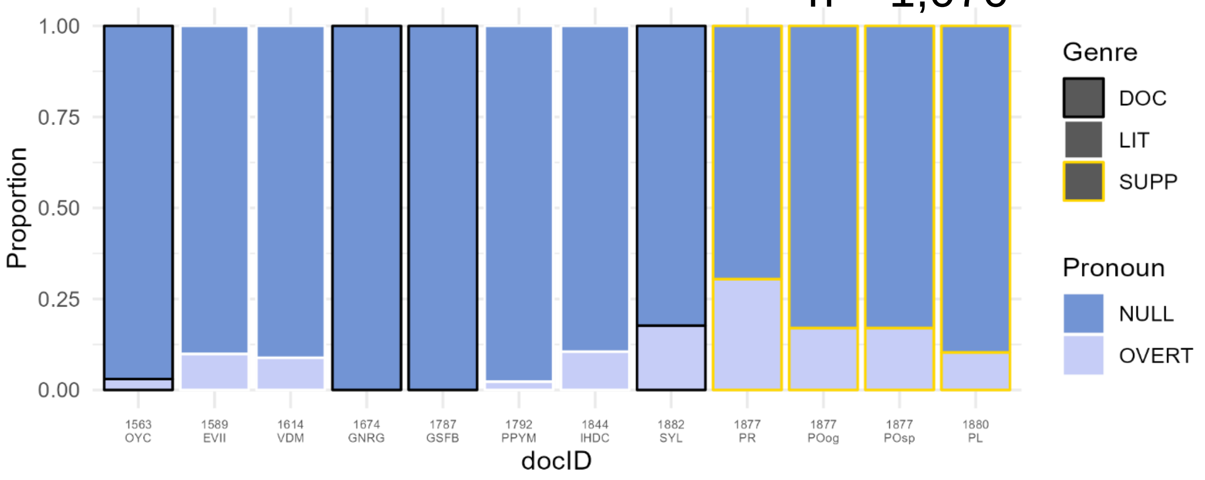
	CARIBBEAN/CENTRAL			SOUTH AMERICAN				SPAIN
	DR	PANAMÁ	CUBA	PERÚ	COLOMBIA	BOLIVIA	VENEZUELA	
16TH								
LIT	ENT	_HGNI_	_HDLI_	_HNMI_	_EVII*_	--	_GDUI_	LAH
DOC	SDJ	_CAR_	_DRF_	_NDP_	OYC	_RVP_	_NDA_	CAN
17TH								
LIT	_DPHJ_	LLDP*	_EDP*_	_CEVP*_	_VDM_	--	_NHLC_	DQ
DOC	--	DLYD	LCDH	CPVV	_GNRG_	--	PR	ACRA
18TH								
LIT	**LIVIE**	--	PJFC*	PAD	PPYM	HVIP	_EOID_	ARJD
DOC	ASD	--	SPPH	MC	GSFB	--	ALTU	EAU
19TH								
LIT	__GAL*__	HS*	**ADUE**	**MYT**	**IHDC**	JDLR	VH	CPC
DOC	**ALD**	MPE	GDLH	**CRP**	**SYL**	ADLA	GDC	QDEV

Table 1: Corpus Composition | **AH** | *Born in Spain* | Verse*

CorDELES: Pronoun Realization

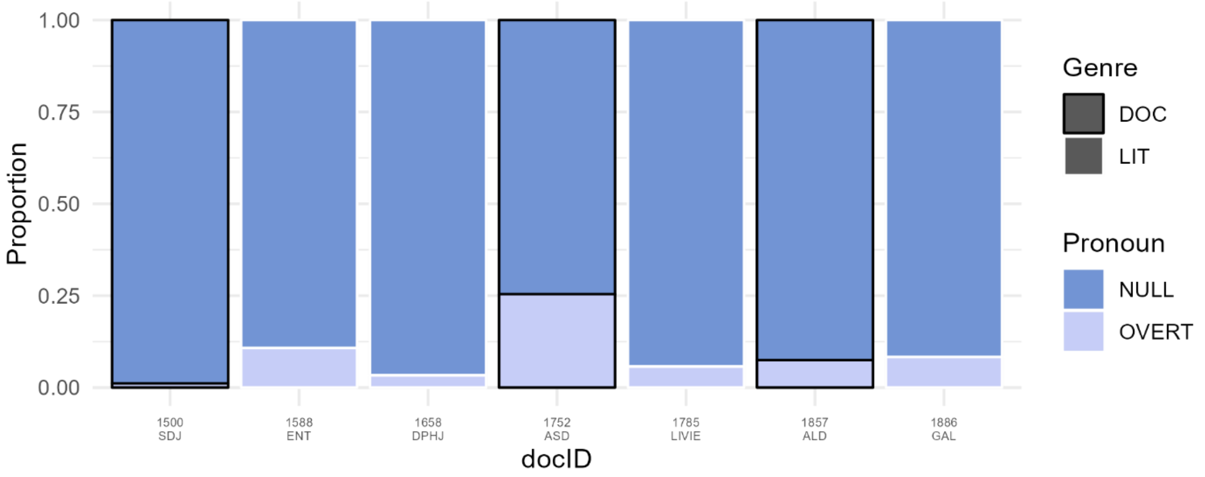
Pronoun Realization: Colombia

n = 1,070



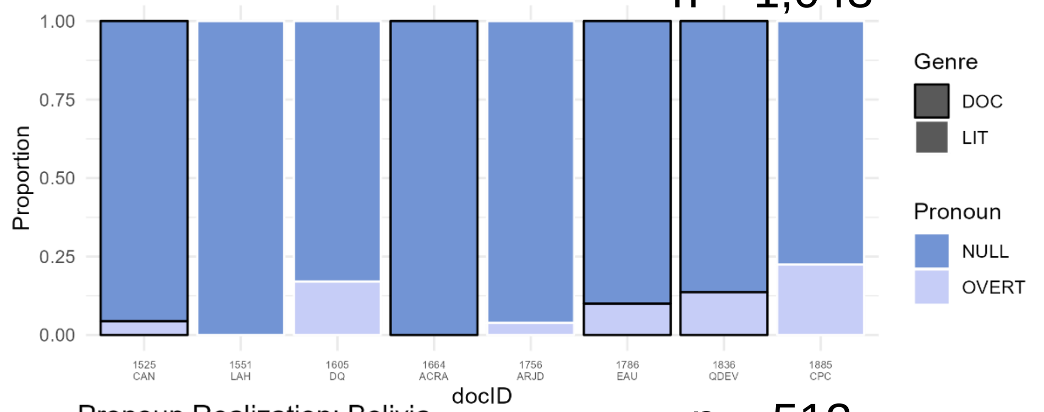
Pronoun Realization: Dom. Rep.

n = 573



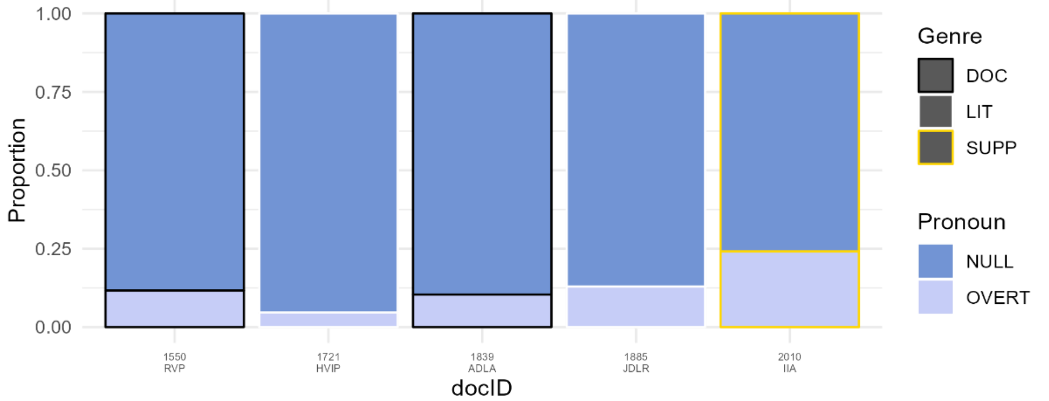
Pronoun Realization: Spain

n = 1,043



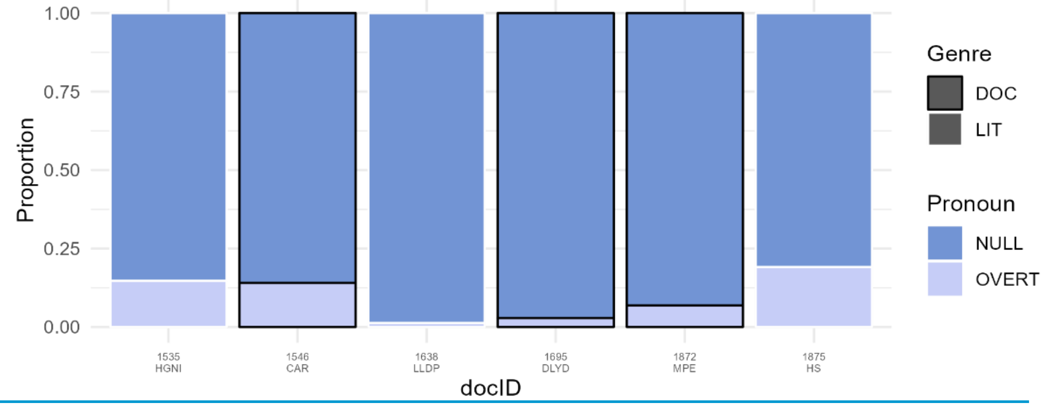
Pronoun Realization: Bolivia

n = 512



Pronoun Realization: Panamá

n = 575



Measuring Orality

•Rosemeyer (2019) measured orality levels in a diachronic corpus of Brazilian Portuguese plays:

- The plays followed a shift toward reflecting spoken speech over the centuries

•Rosemeyer (2019) variables:

•Present progressive

•Demonstrative neuter pronouns

•Time and place adverbs

•Discourse markers

•Private verbs

•My variables:

•Progressive

•Demonstrative neuter pronouns

- *esto/eso/aquello*

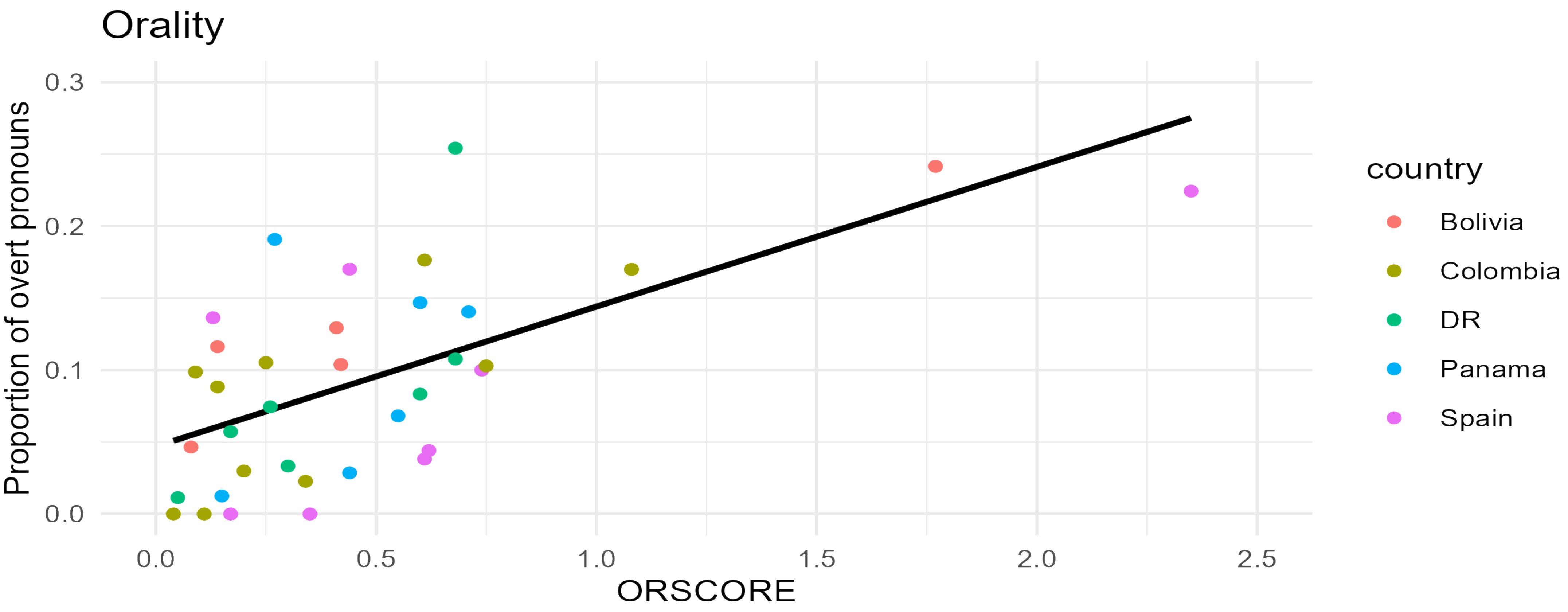
•Time and place adverbs

- *aqui/ahora*

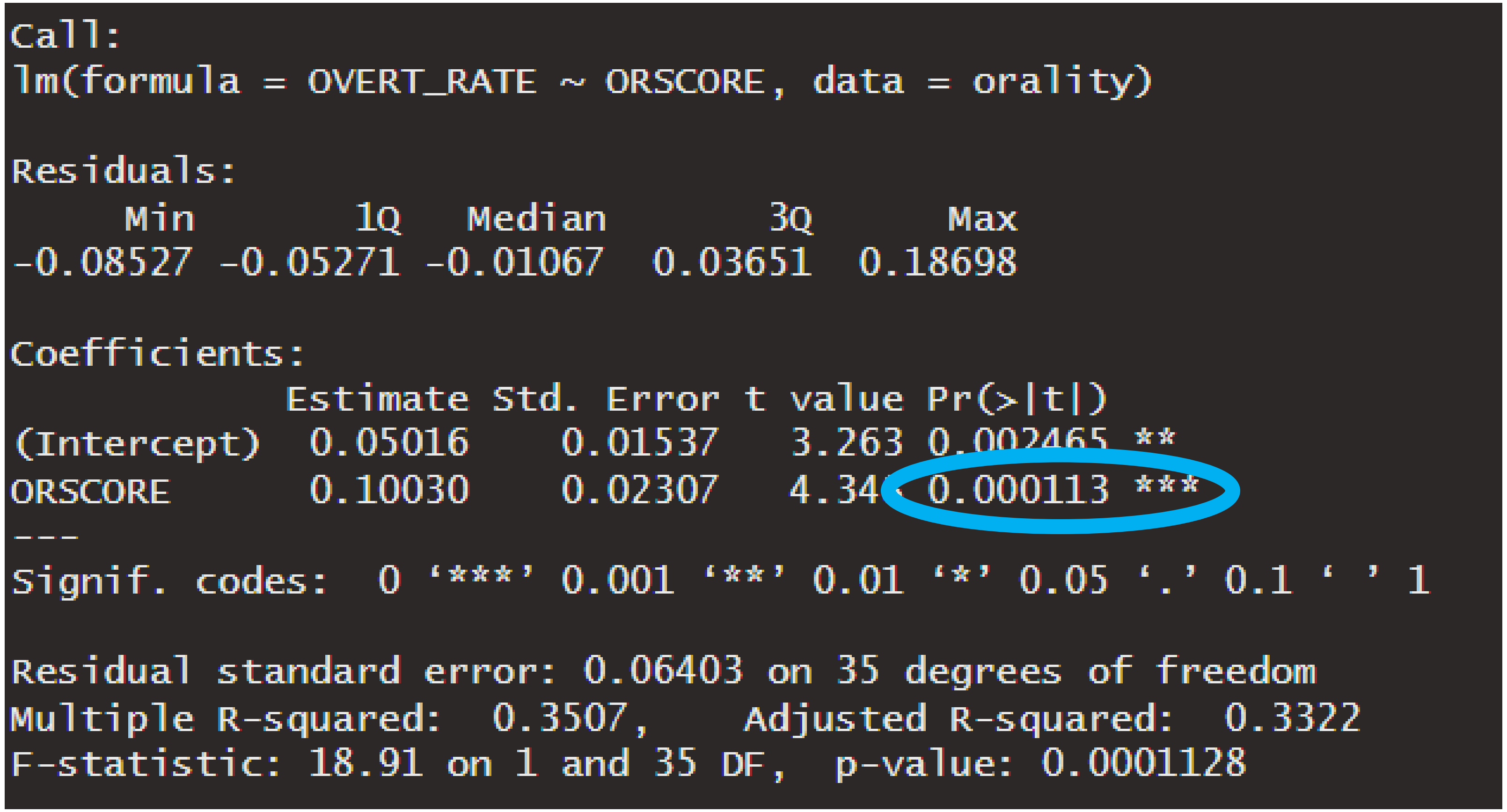
•Private verbs

- *pensar* ‘to think’ / *creer* ‘to believe'

Plotting Orality Against Overtness Rates



Modelling Orality



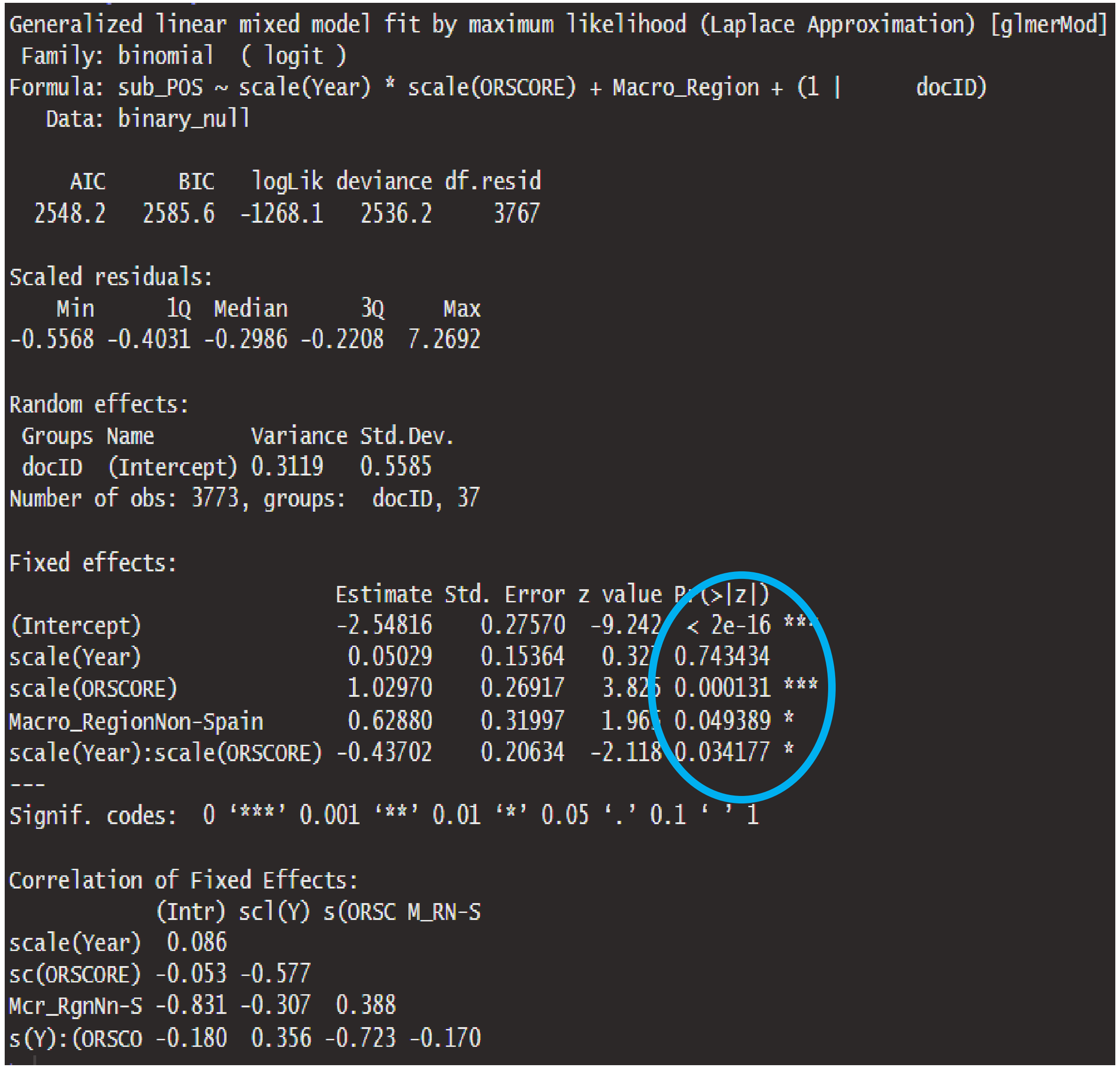
```
Call:
lm(formula = OVERT_RATE ~ ORSCORE, data = orality)

Residuals:
     Min       1Q   Median       3Q      Max 
-0.08527 -0.05271 -0.01067  0.03651  0.18698 

Coefficients:
            Estimate Std. Error t value Pr(>|t|)    
(Intercept)  0.05016    0.01537   3.263 0.002465 ** 
ORSCORE      0.10030    0.02307   4.348 0.000113 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Residual standard error: 0.06403 on 35 degrees of freedom
Multiple R-squared:  0.3507,    Adjusted R-squared:  0.3322 
F-statistic: 18.91 on 1 and 35 DF,  p-value: 0.0001128
```

CorDELES: Model



```
Generalized linear mixed model fit by maximum likelihood (Laplace Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: sub_POS ~ scale(Year) * scale(ORSCORE) + Macro_Region + (1 |      docID)
   Data: binary_null

     AIC      BIC   logLik deviance df.resid
  2548.2   2585.6  -1268.1   2536.2     3767

Scaled residuals:
    Min      1Q  Median      3Q     Max
-0.5568 -0.4031 -0.2986 -0.2208  7.2692

Random effects:
 Groups Name        Variance Std.Dev.
 docID  (Intercept) 0.3119   0.5585
Number of obs: 3773, groups:  docID, 37

Fixed effects:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                -2.54816    0.27570  -9.242  < 2e-16 ***
scale(Year)                 0.05029    0.15364   0.327 0.743434
scale(ORSCORE)              1.02970    0.26917   3.826 0.000131 ***
Macro_RegionNon-Spain       0.62880    0.31997   1.965 0.049389 *
scale(Year):scale(ORSCORE) -0.43702    0.20634  -2.118 0.034177 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

Correlation of Fixed Effects:
            (Intr) scl(Y) s(ORSC M_RN-S
scale(Year)  0.086
sc(ORSCORE) -0.053 -0.577
Mcr_RgnNn-S -0.831 -0.307  0.388
s(Y):(ORSCO -0.180  0.356 -0.723 -0.170
```

Model

- glmer from lme4 package in R
- Fixed effects: Year (z-scored), ORSCORE (z-scored), Macro-Region; interaction between Year and ORSCORE
- docID as a random effect

- ➢ Year is no longer significant on its own
- ➢ But the interaction between Year and ORSCORE is!
  - $p < 0.034$
- ➢ Still no effect of Country or Region; but Macro-Region (Spain vs. Non-Spain) comes out just significant
  - $p < 0.049$

Research Questions

1. ***does overtness increase diachronically?***

2. ***does it have higher rates from Spain > South America > Caribbean?***

3. ***are these trends the same across corpora?***

CORDIAM

CORDIAM: Data Collection

Searched for 8 forms of *creer* ‘to think/believe’

- *creo* (1sg.pres.) – 1791 = 76%
- *creemos* (1pl.pres.) – 389 = 17%
- *crees* (2sg.pres.) – 27 = 1.2%
- *creíamos* (1pl.imperf.) – 35 = 1.5%
- *creías* (2sg.imperf.) – 2 = <1.0%
- *creí* (1sg.pret.) – 63 = 2.7%
- *creímos* (1pl.pret.) – 37 = 1.6%
- *creíste* (2sg.pret.) – 1 = <.05%

2,345 tokens

Reasons for excluding 3rd person:

- Ambiguous with subjunctive forms of *crear* ‘to build’
- Can have relative pronouns or full noun phrases as subjects
- Impersonals
- Semantic person split between *usted* ‘you.f’ and *él*/*ella* ‘he/she’

Reasons for excluding subjunctive and 1sg.imperf.:

- Too many forms ambiguous with 3rd person

CORDIAM: *creer* counts

	16th		17th		18th		19th	
	NULL	OVERT	NULL	OVERT	NULL	OVERT	NULL	OVERT
ARGENTINA	2	0	1	0	6	0	22	5
CHILE	3	0	8	7	39	5	12	1
COLOMBIA	34	2	1	0	17	4	27	4
MEXICO	214	24	53	4	69	6	23	1
PERU	78	35	33	4	65	10	247	15
VENEZUELA	13	0	3	0	3	1	103	22

Religious vs. discourse *creer*

creer ‘to believe/think’ has two senses

1. *creo que la prueba es el lunes* ‘I think the quiz is on Monday’
2. *creo en Dios* ‘I believe in God’

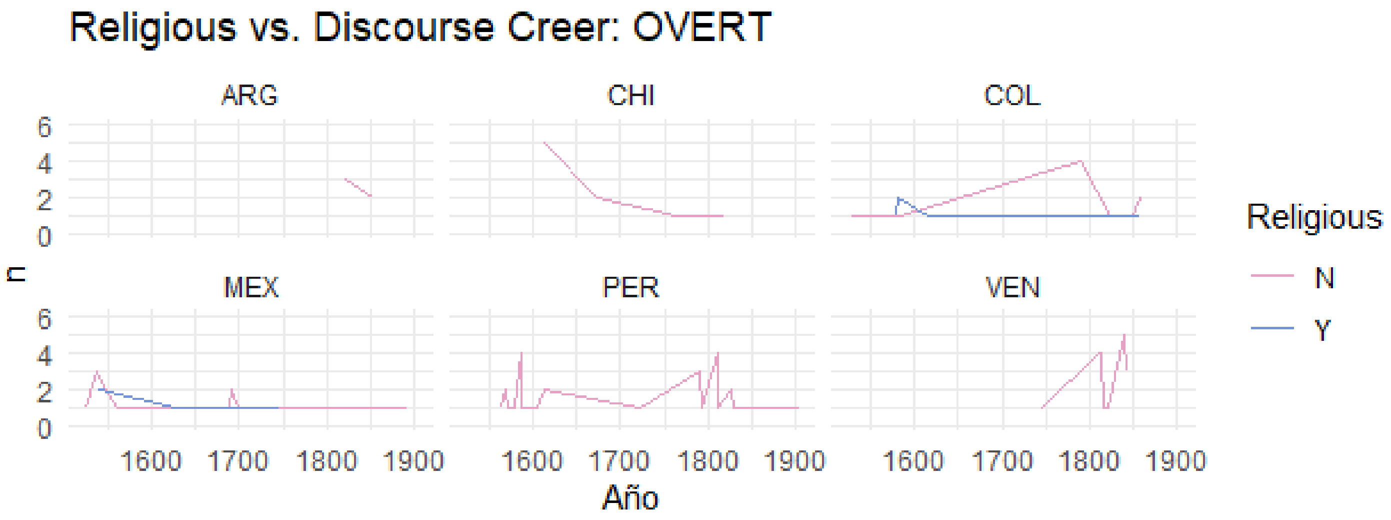
Why exclude the second?

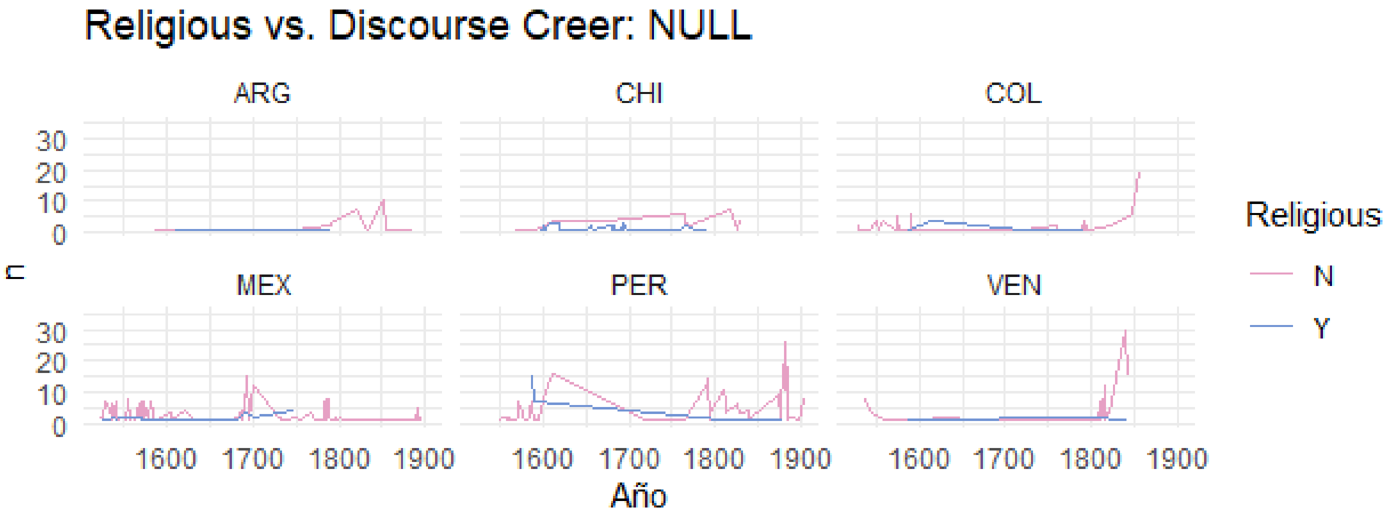
- Religious *creer* can be very formulaic and repetitive:
  - *Creyendo como Creo en el Misterio de la Santísima Trinidad Padre, Hijo y Espíritu santo* (Año 1690, Argentina, Documentos administrativos, CORDIAM) → 14 repetitions
    - ‘Believing as I Believe in the Mystery of the Holy Trinity of the Father, Son, and the Holy Ghost’
- Doesn’t mark orality as well as the first sense

The data should then reflect this by showing a lower overtness rate for religious *creer* than discourse *creer*

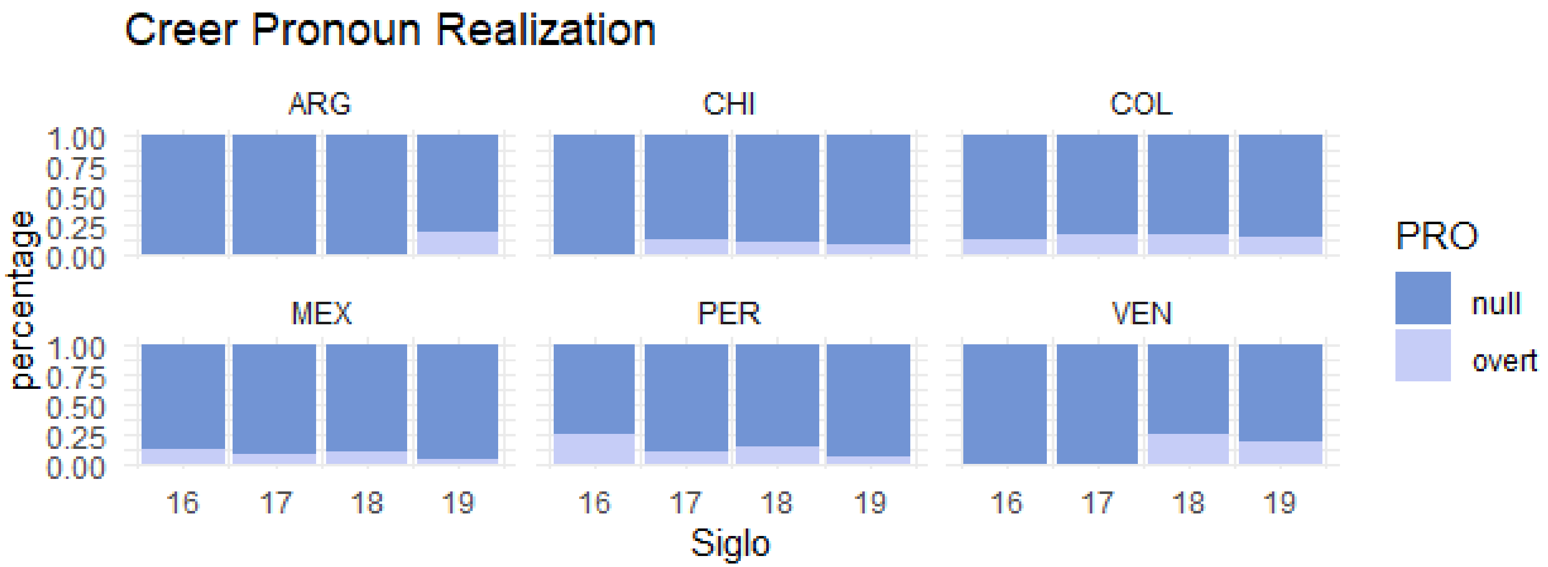
➢ **Added tag for religious vs. discourse *creer***

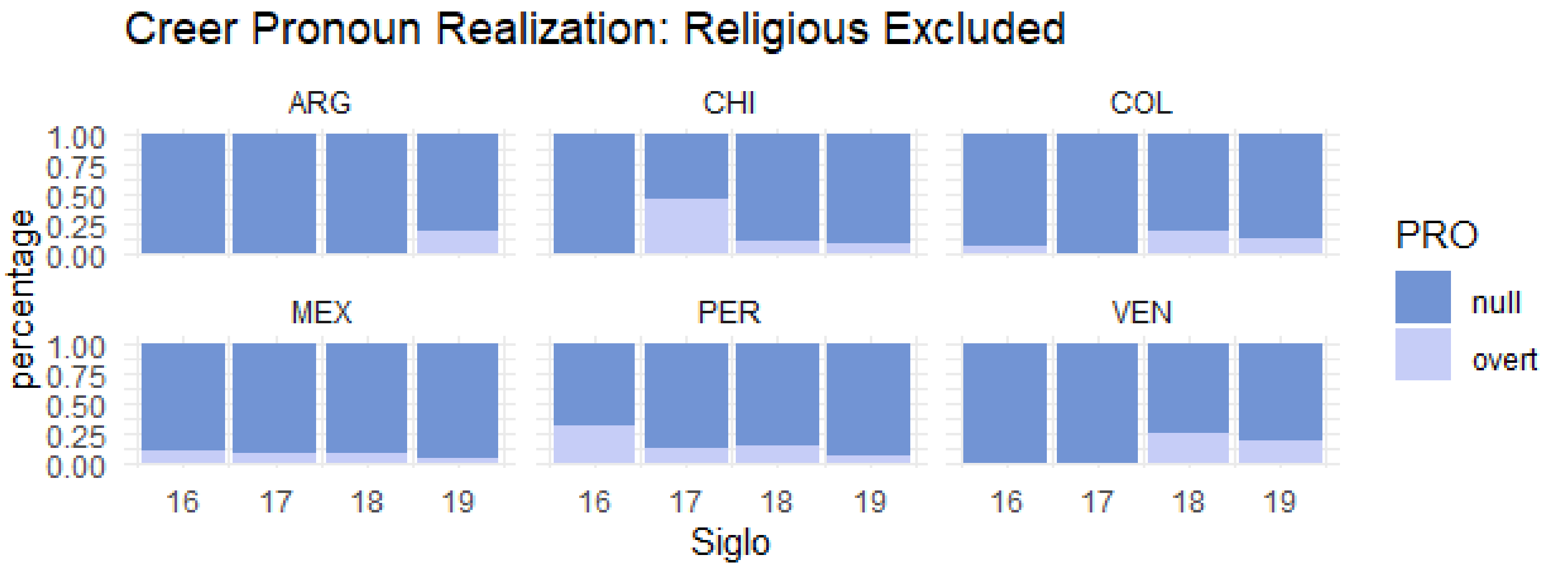
Religious vs. discourse *creer* plotted



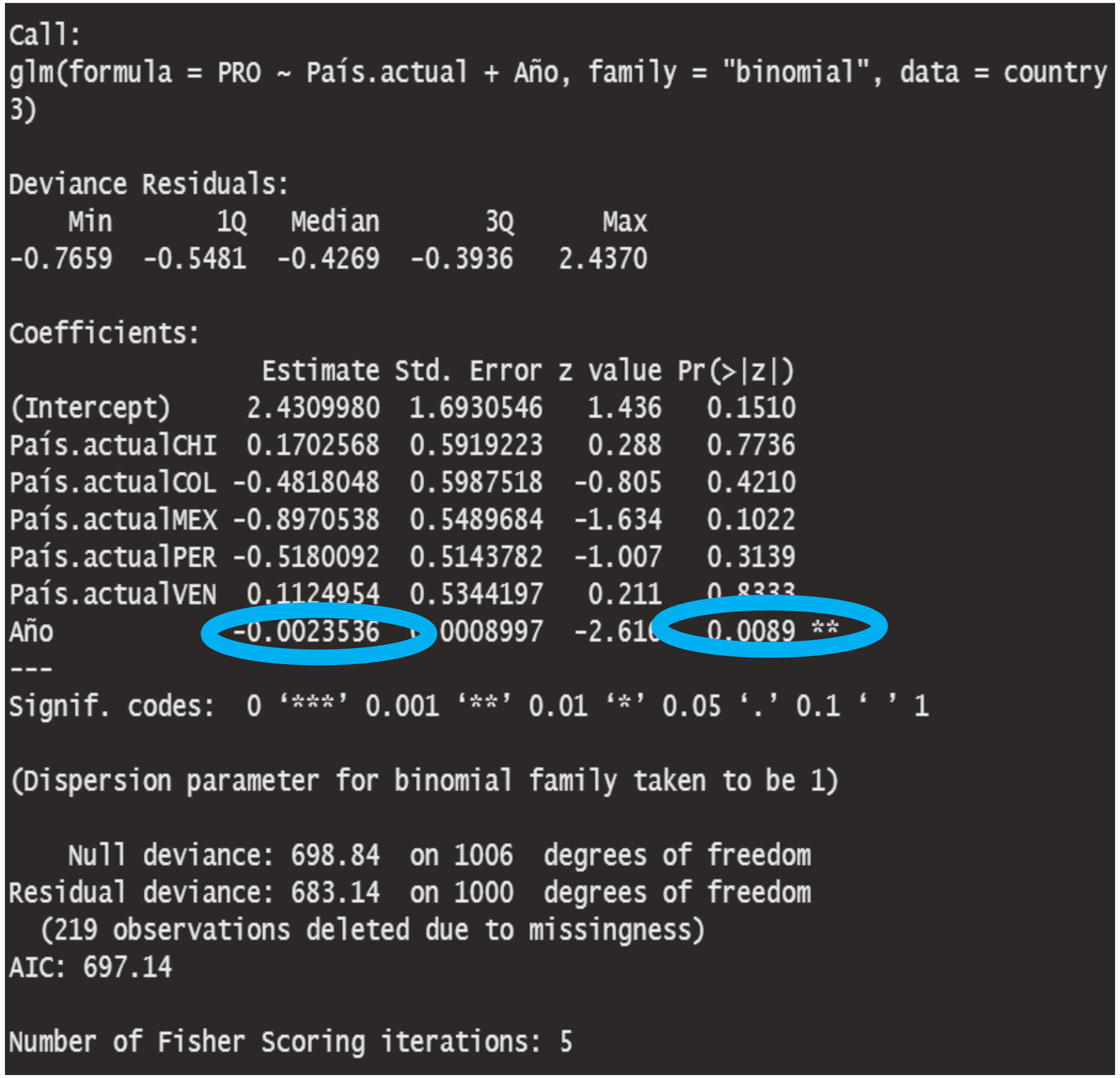


CORDIAM: *creer* bar charts





CORDIAM: Model



```
Call:
glm(formula = PRO ~ País.actual + Año, family = "binomial", data = country
3)

Deviance Residuals:
    Min       1Q   Median       3Q      Max
-0.7659  -0.5481  -0.4269  -0.3936   2.4370

Coefficients:
                 Estimate Std. Error z value Pr(>|z|)
(Intercept)     2.4309980  1.6930546   1.436   0.1510
País.actualCHI  0.1702568  0.5919223   0.288   0.7736
País.actualCOL -0.4818048  0.5987518  -0.805   0.4210
País.actualMEX -0.8970538  0.5489684  -1.634   0.1022
País.actualPER -0.5180092  0.5143782  -1.007   0.3139
País.actualVEN  0.1124954  0.5344197   0.211   0.8333
Año            -0.0023536  0.0008997  -2.616   0.0089 **
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 698.84  on 1006  degrees of freedom
Residual deviance: 683.14  on 1000  degrees of freedom
  (219 observations deleted due to missingness)
AIC: 697.14

Number of Fisher Scoring iterations: 5
```

Model

- glm from lme4 package in R
- Fixed effects: Year, Country
- ➢ Country isn’t significant
- ➢ But Year is
  - p < 0.0089
  - Negative co-efficient: -0.002

CDH

CDH: Data Collection

CDH

- 19 countries: Argentina, Bolivia, Chile, Colombia, Costa Rica, Cuba, Dominican Republic, Ecuador, Filipinas, Guatemala, Honduras, Mexico, Nicaragua, Paraguay, Peru, Puerto Rico, Spain, Uruguay, Venezuela
- 6 centuries: 16th-21st

Mass Queries

- *creo* = 61,136 tokens
- *creo* + *yo* = 7,826 tokens
- *creo* + *que* = 40,304 tokens
- *creo* + *yo* + *que* = 7,409 tokens

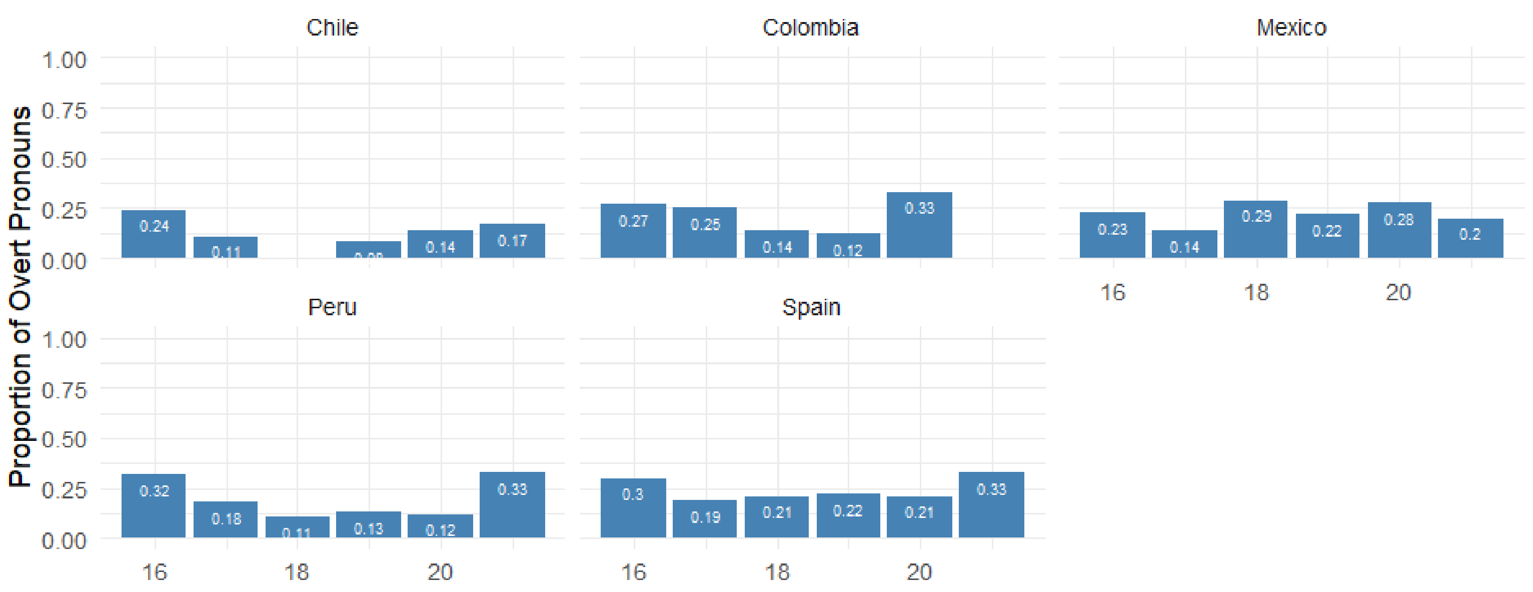
Cons of not going through by hand:

- False positives
- Undercount for overt realization
- No accounting for religious *creer*
  - *creo que* workaround
  - There are potential exceptions though: *creo que Jesús es el hijo de Dios* ‘I believe that Jesus is the son of God’

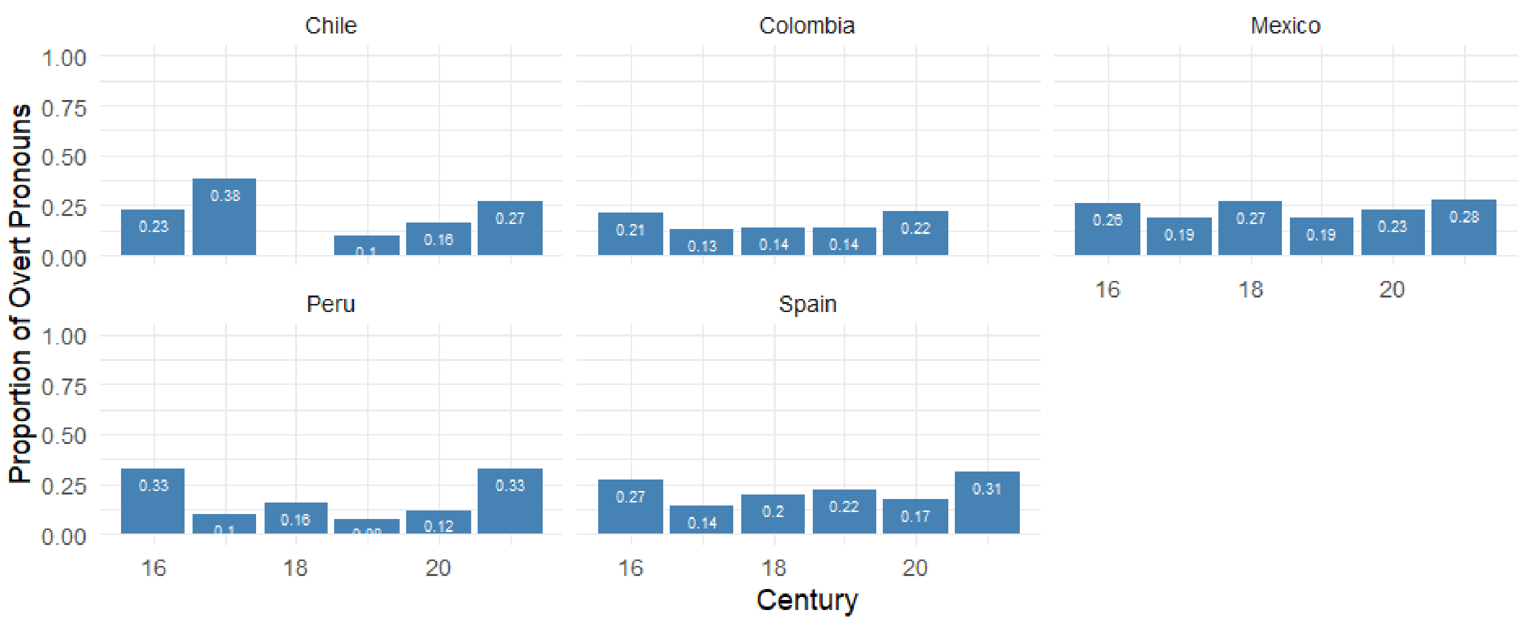
CDH: *creo* bar charts

CDH	***creo + que***
Spain	27,272
Chile	1,710
Colombia	1,042
Mexico	2,689
Peru	1,096

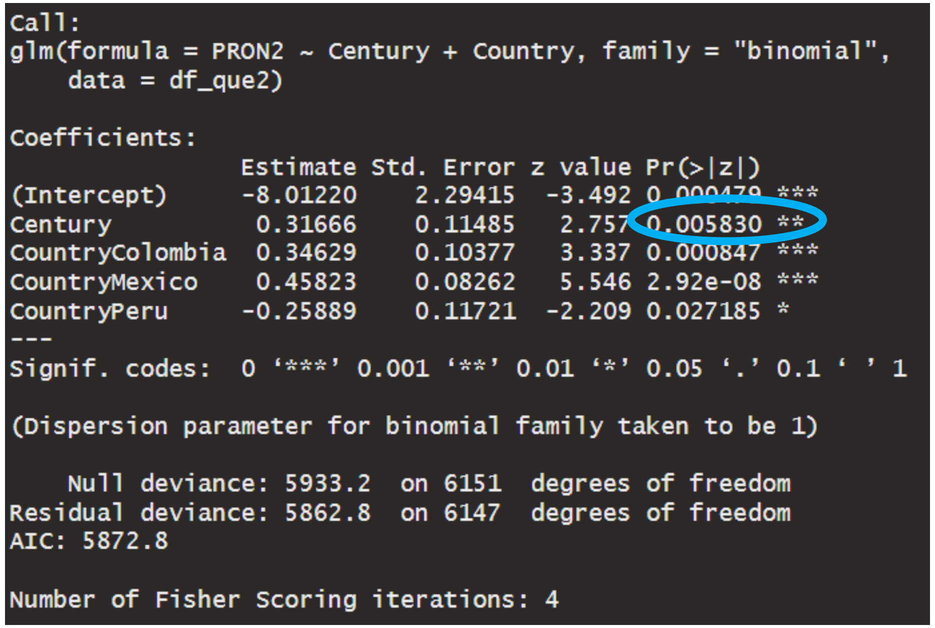
Pronoun Realization: CDH (creo)



Pronoun Realization: CDH (creo que)



CDH: Models (19th-21st Centuries)



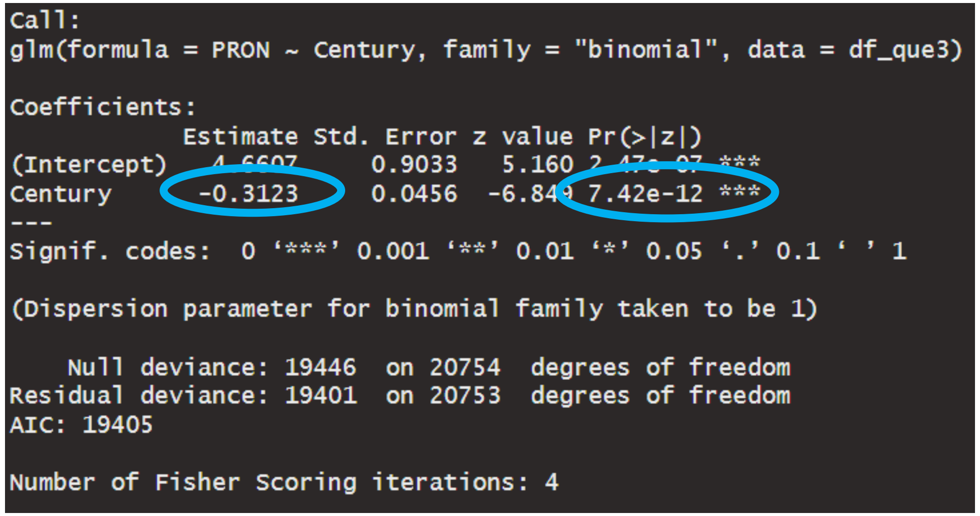
```
Call:
glm(formula = PRON2 ~ Century + Country, family = "binomial",
    data = df_que2)

Coefficients:
                Estimate Std. Error z value Pr(>|z|)
(Intercept)     -8.01220    2.29415  -3.492 0.000479 ***
Century          0.31666    0.11485   2.757 0.005830 **
CountryColombia  0.34629    0.10377   3.337 0.000847 ***
CountryMexico    0.45823    0.08262   5.546 2.92e-08 ***
CountryPeru     -0.25889    0.11721  -2.209 0.027185 *
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 5933.2  on 6151  degrees of freedom
Residual deviance: 5862.8  on 6147  degrees of freedom
AIC: 5872.8

Number of Fisher Scoring iterations: 4
```



```
Call:
glm(formula = PRON ~ Century, family = "binomial", data = df_que3)

Coefficients:
            Estimate Std. Error z value Pr(>|z|)
(Intercept)   4.6607     0.9033   5.160 2.47e-07 ***
Century      -0.3123     0.0456  -6.849 7.42e-12 ***
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

    Null deviance: 19446  on 20754  degrees of freedom
Residual deviance: 19401  on 20753  degrees of freedom
AIC: 19405

Number of Fisher Scoring iterations: 4
```

Model (Chile, Colombia, Mexico, Peru)

- glm from lme4 package in R
- Fixed effects: Century, Country
- ➢ Every country is significant!
- ➢ Century is significant!
  - p < 0.0058

Model (Spain)

- glm from lme4 package in R
- Fixed effects: Century
  - Spain was separated as it has 10x the tokens, affecting the model results when included
- ➢ Century is significant!
  - p < 7.42e-12
  - Negative co-efficient: --0.3123

Conclusion

I accounted for orality and compared pronoun realization across 3 corpora:

1. **CorDELES (ORSCORE accounted for, by hand)**
   - Diachrony: increase in overt subjects
   - Region: Spain vs. non-Spain
2. **CORDIAM (discourse *creer* only, by hand)**
   - Diachrony: decrease in overt subjects
   - Region: none
3. **CDH (*creo que* only, mass counts)**
   - Diachrony: increase in overt subjects from 19th-21st centuries
   - Region: Spain vs. non-Spain

Why 19th century?

- ➢ New access to publication after independence and abolition?

References

Bates, Douglas, Mächler, Martin, Bolker, Ben & Walker, Steve. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software* 67(1): 1–48.

Bini, M. (1993). "La adquisición del italiano: Más allá de las propiedades sintácticas delparámetro pro-drop." In J. M. Liceras (Ed.), *La lingüística y el análisis de los sistemasno nativos:* 126–139. Dovehouse Editions Canada.

Camacho, José. 2013. *Null subjects*. Cambridge: Cambridge University Press.

Cerrón-Palomino, Álvaro. 2018. "Variable subject pronoun expression in Andean Spanish: a drift from the acrolect". *Onomázein* 1 (42): 53-73.

[*CORDIAM*] Academia Mexicana de la Lengua, *Corpus Diacrónico y Diatópico del Español de América*, <www.cordiam.org>

Klee, C.A. & Lynch, A. 2009. *El español en contacto con otras lenguas*. Washington DC: Georgetown University Press.

Margaza, P., & Bel, A. (2006). "Null subjects at the syntax–pragmatics interface: Evidence from Spanish interlanguage of Greek speakers." In M. Grantham O'Brien, C. Shea, &J. Archibald (Eds.), *Proceedings of the 8th Generative Approaches to Second Language Acquisition Conference (GASLA 2006):* 88–97. Cascadilla Proceedings Project.

Pérez-Leroux, A. T., & Glass, W. R. (1999). "Null anaphora in Spanish second language acquisition: Probabilistic versus generative approaches." *Second Language Research*, 15 (2): 220–249.

R Core Team. (2021). R: *A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing.

Real Academia Española. (2013). *Corpus del Diccionario histórico de la lengua española (CDH)* [en línea]. https://apps.rae.es/CNDHE

Rosemeyer, Malte. 2019. "Actual and apparent change in Brazilian Portuguese wh-interrogatives." *Language Variation and Change* 31(2): 165–191. CUP.

Sessarego, Sandro. 2013. "Afro-Hispanic Contact Varieties as Conventionalized Advanced Second Languages". *IBERIA* 5 (1): 99-125.

Sorace, Antonella. 2011. "Pinning down the concept of "interface" in bilingualism". *Linguistic Approaches to Bilingualism* 1(1): 1-33.

Toribio, Almeida J. 2000. "Setting parametric limits on dialectal variation in Spanish". *Lingua: International Review of General Linguistics* 110 (5): 315–341.

Trudgill, Peter. 2011. Sociolinguistic typology: *Social determinants of linguistic complexity.* Oxford: OUP.

Tsimpli, Iantha Maria and Lavidas, Nikolaos. 2019. "Object Omission in Contact: Object Clitics and Definite Articles in the West Thracian Greek (Evros) Dialect". *Journal of Language Contact* 12: 141-190.

Walkden, George and Breitbarth, Anne. 2019. "Interpreting (un)interpretability" *Theoretical Linguistics* 45 (3-4): 309-317.

Universität
Konstanz

Thank you
for listening!

gemma-hunter.mccarley@uni-konstanz.de
https://gemmamccarley.com/
projectstarfish.eu

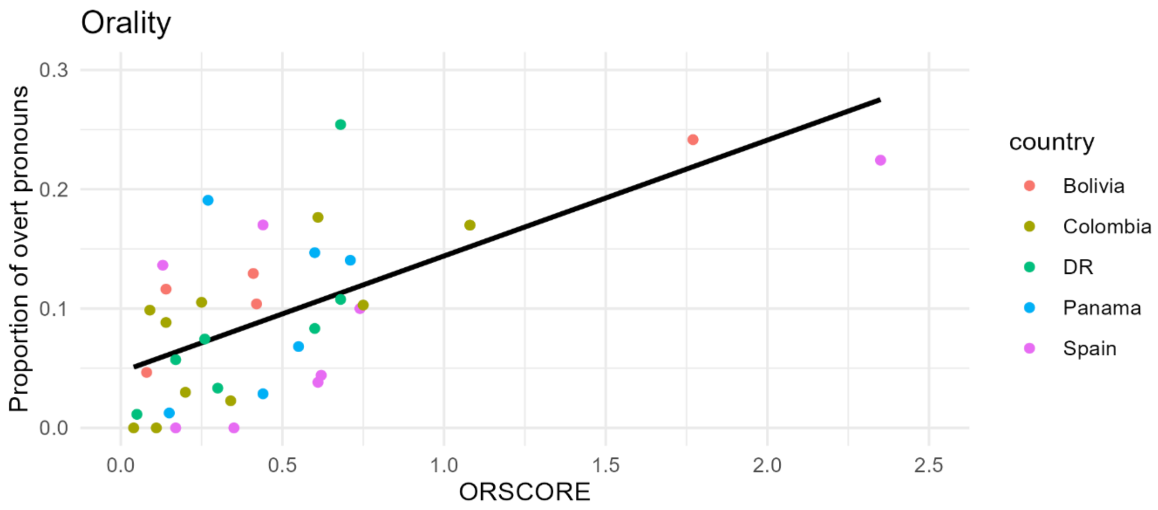
STARFISH

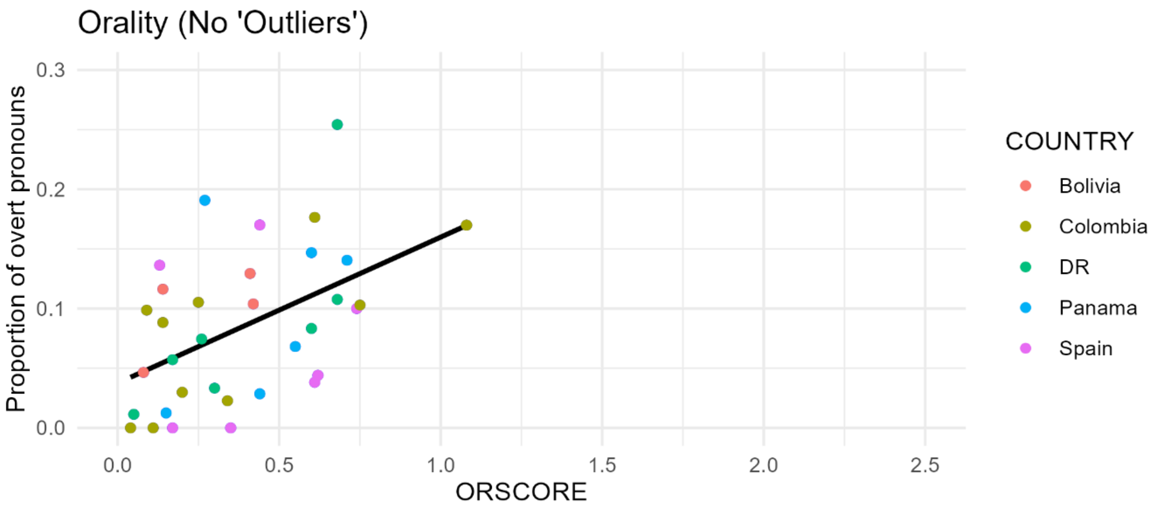
SOCIOLINGUISTIC TYPOLOGY
AND RESPONSIVE FEATURES
IN SYNTACTIC HISTORY

This project has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement No. 851423

erc
European Research Council
Established by the European Commission

Plotting Orality Against Overtness Rates: Regression





CDH: *creo* + *que* counts

CDH	16	17	18	19	20	21
Spain	4,360	1,224	933	3,465	17,261	29
Chile	13	13	0	92	1,577	15
Colombia	52	8	7	57	879	39
Mexico	105	54	26	211	2,254	39
Peru	49	20	38	48	890	51