Universität Konstanz

Changes in Null Subjects in Latin American Spanish: A Diachronic Corpus Study

Gemma McCarley

HiSoN 2022, 02.06.22

STARFISH

SOCIOLINGUISTIC TYPOLOGY AND RESPONSIVE FEATURES IN SYNTACTIC HISTORY

Background

• Spanish is a null subject language (NSL) which means it can have sentences like (1) that are perfectly grammatical

1. 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)

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

• In the literature, nullness has historically been linked with inversion, e.g. the NSP, because most consistent NSLs like Italian and Spanish also allow inversion (Rizzi 1982, 1986)

• This theoretical correlation tracks with findings that SV word order is also on the rise in varieties where overtness is too (Toribio 2000)

2. Papi, ¿qué ese letrero dice?
   (cf. Papi, ¿qué dice ese letrero?)
   'Daddy, what does that sign say?' (Toribio 2000: 322)

•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 (and SV word order).
- 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.

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

Research Questions

Main questions:

1. *do overtness and SV word order increase diachronically?*
2. *do they have higher rates from Spain > South America > Caribbean?*
3. *do certain genres have higher rates than others?*

Additional questions for pronoun realization:

1. *does switch-reference affect pronoun realization?*
2. *does person affect pronoun realization?*
3. *does clause type affect pronoun realization?*
4. *do any of these effects vary by country, century, or genre?*

Additional questions for word order:

1. *does clause type affect inversion?*
2. *does declarative vs. interrogative status affect inversion?*
3. *do either of these effects vary by country, century, or genre?*

Methodology: Corpus

- **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
- **The main sources for the texts are Cervantes Virtual, dLOC, and BDH**
- **Each texts 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*

Methodology: Tagset

Subject:

- dep(endency) type: "nsubj" (Nominal Subject)
- subpos (subject position): **SV/VS**
- POS: **NULL**

Subject pronouns:

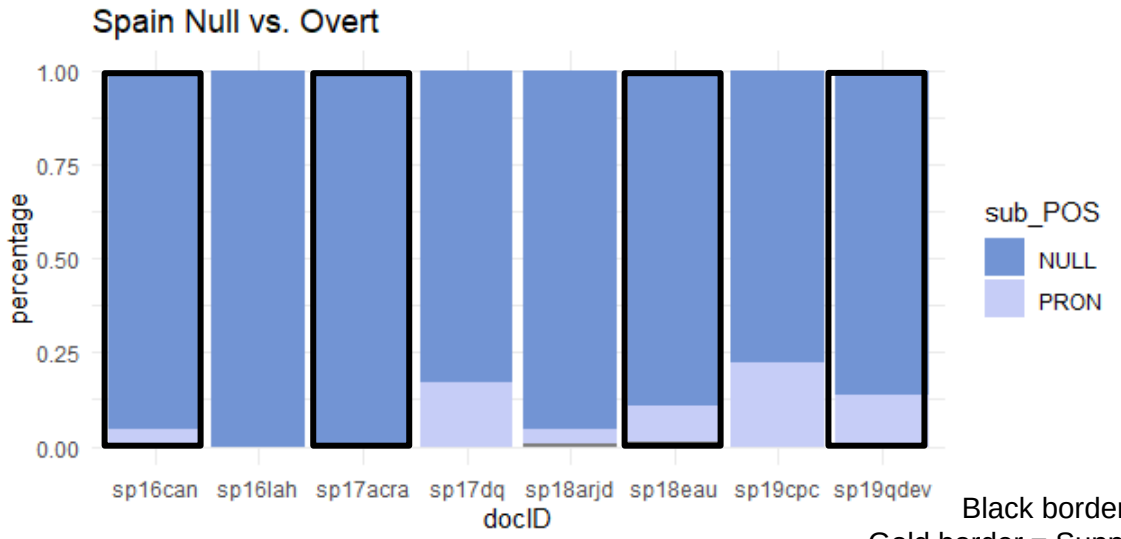
- morphology
  - person: 1/2/3/u (u is for 'usted/es')
  - number: s/p/v (v is for 'vos')
  - e.g. "nosotros" = 1p
- psub (previous subject): **same/diff**
  - this tags for the same referent as the immediately previous clause

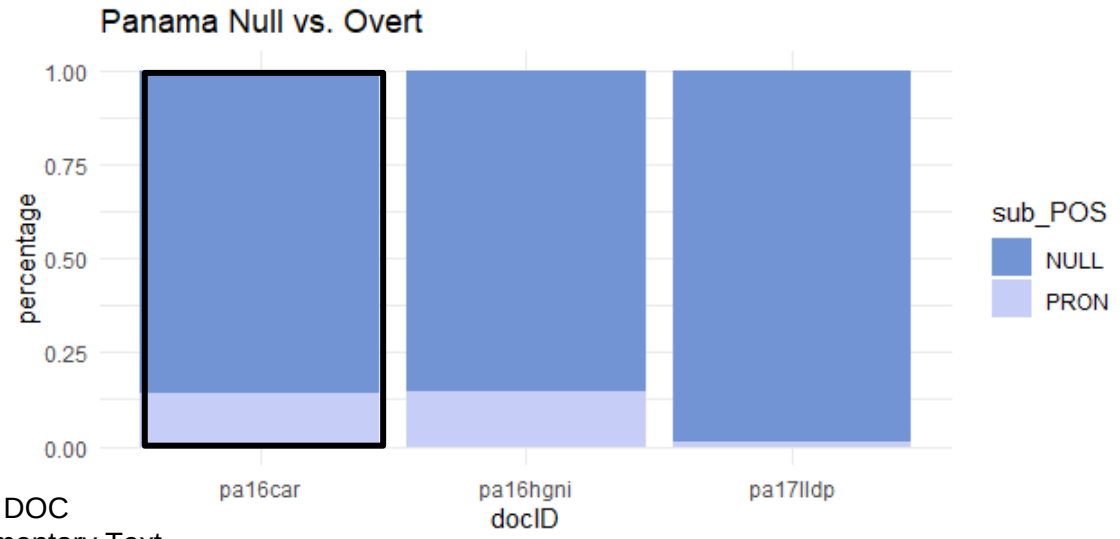
Finite Verbs:

- dep(endency) type: **root** (main clause) / **sub** (dependent clause) / **rel** (relative clause)
  - -INT for questions
- subid (subject ID): the ID of the corresponding subject's token
- morphology:
  - person: 1/2/3
  - number: s/p
  - tense
  - aspect
  - mood

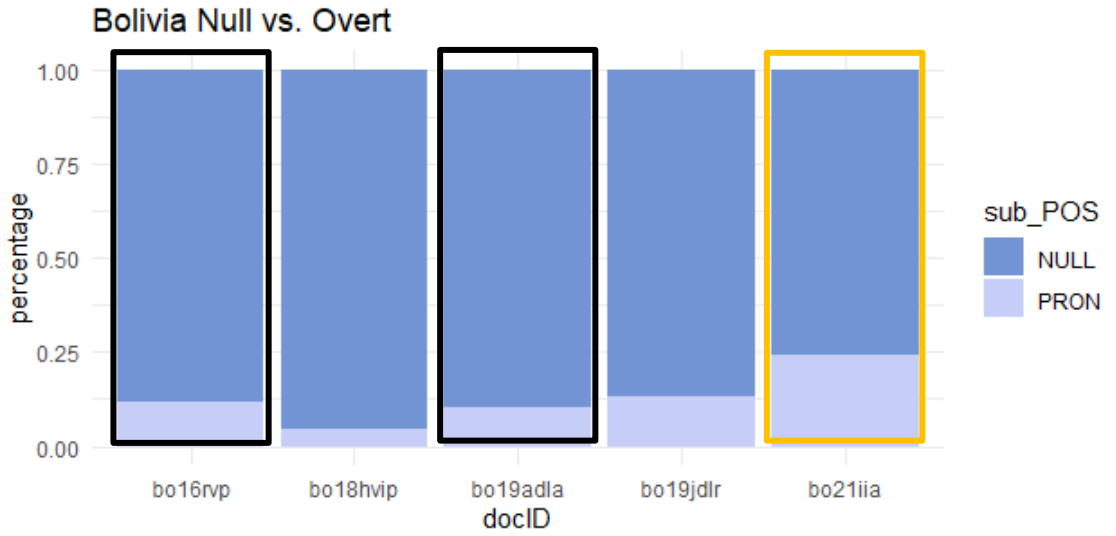
Pronoun Realization (Percent)

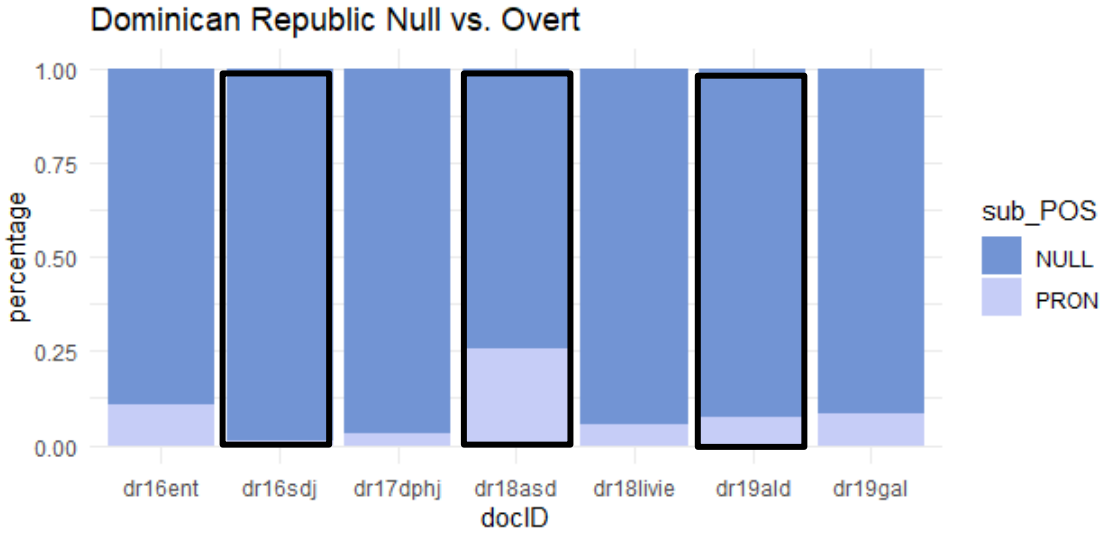
*Discourse switching rates were checked through the number of times a subject had the same or different referent as the previous subject (psub) and there was no correlation there.



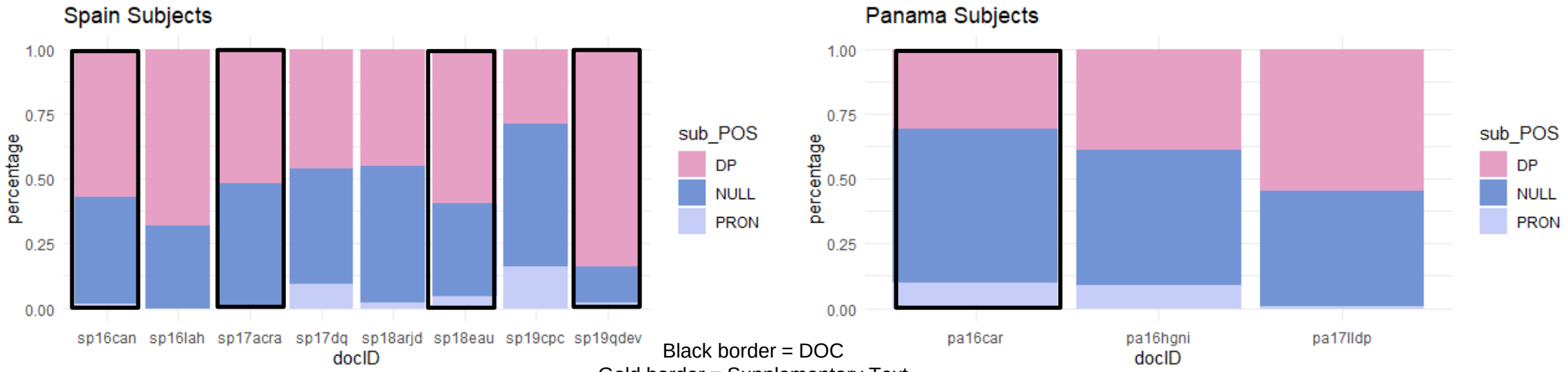


Black border = DOC
Gold border = Supplementary Text

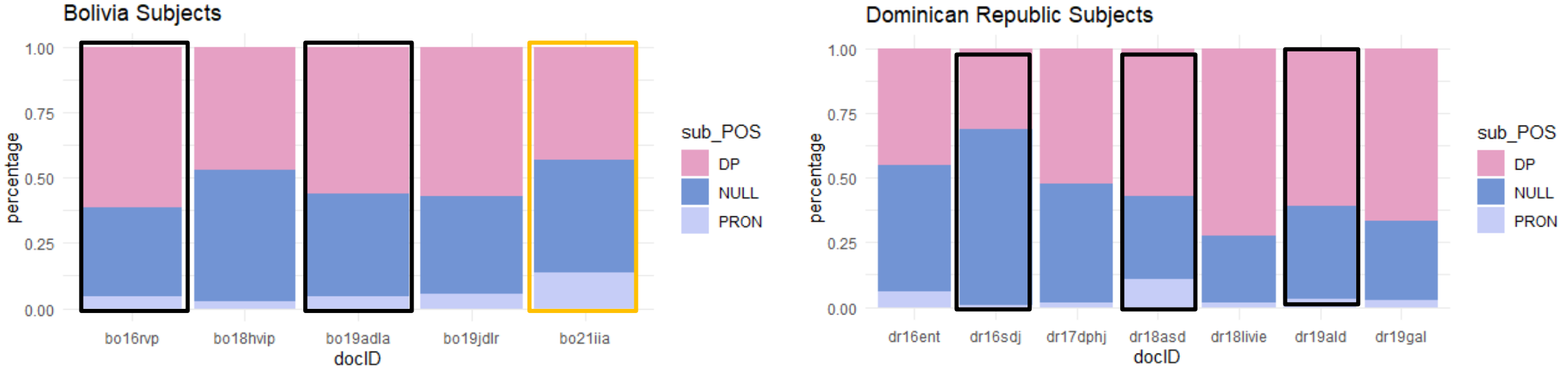




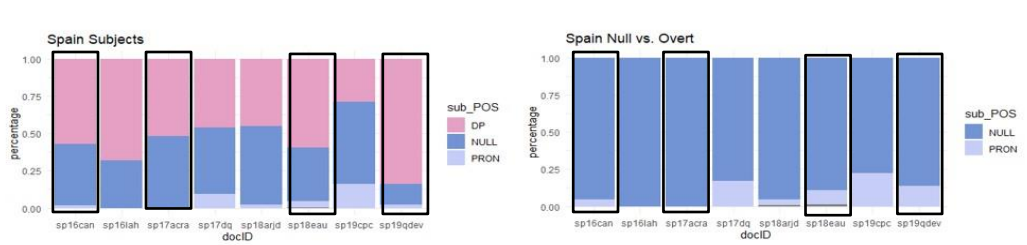
Subject Realization (Percent)

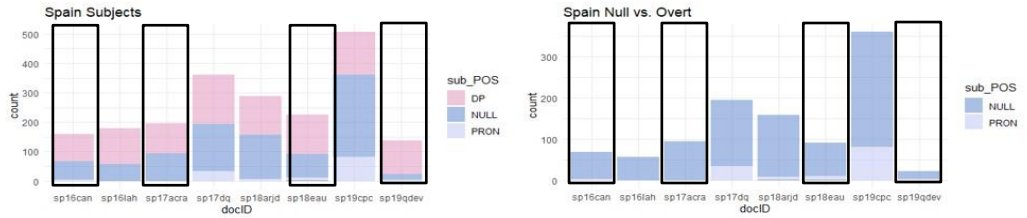


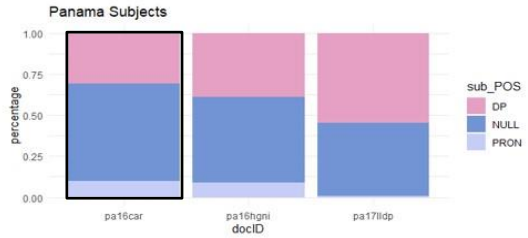
Black border = DOC
Gold border = Supplementary Text

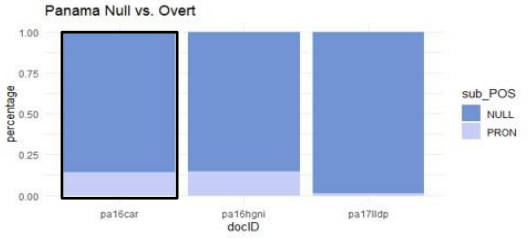


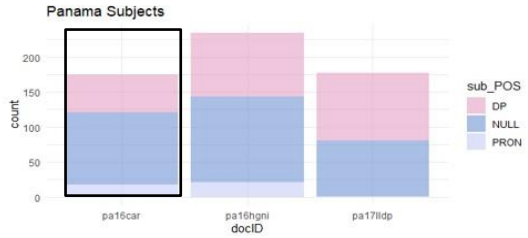
Subject Realization (Count)

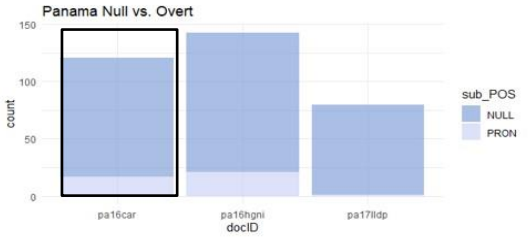




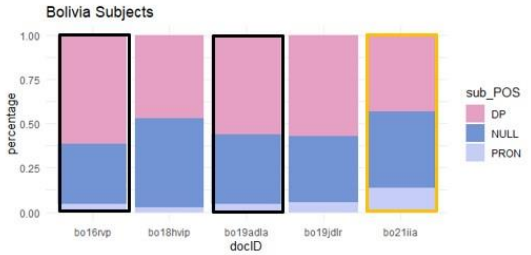


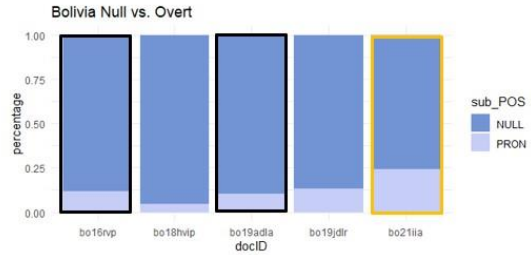


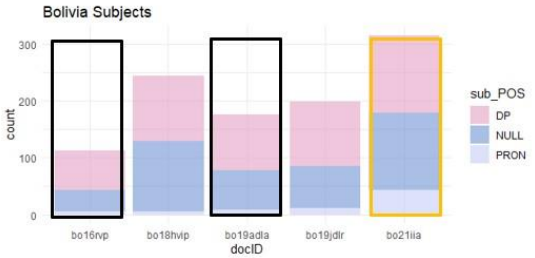


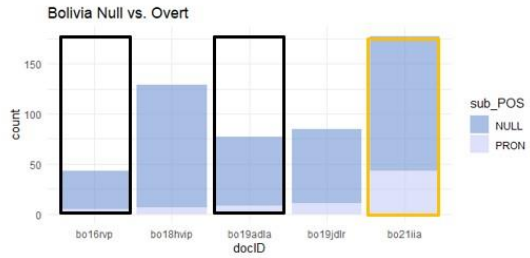


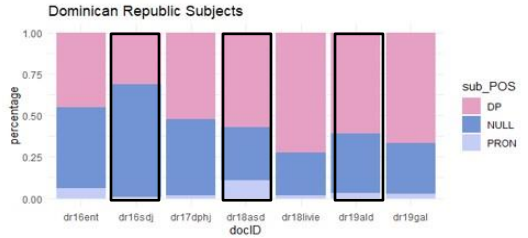
Black border = DOC
Gold border = Supplementary Text

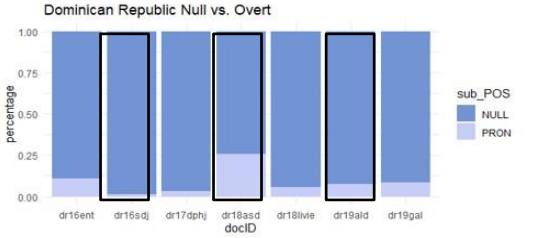


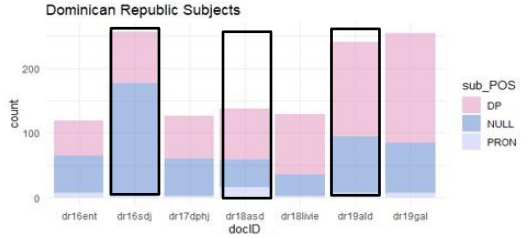


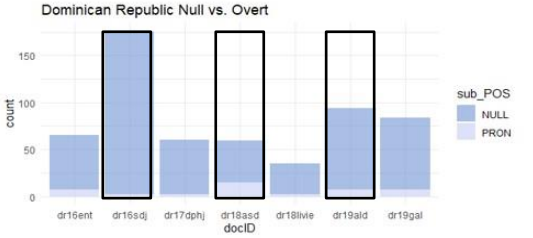




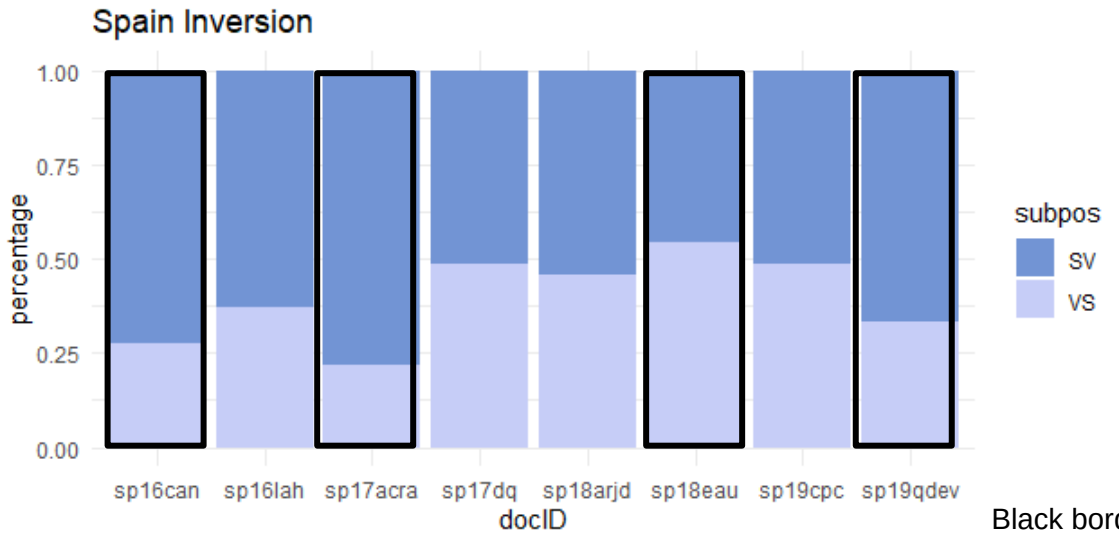


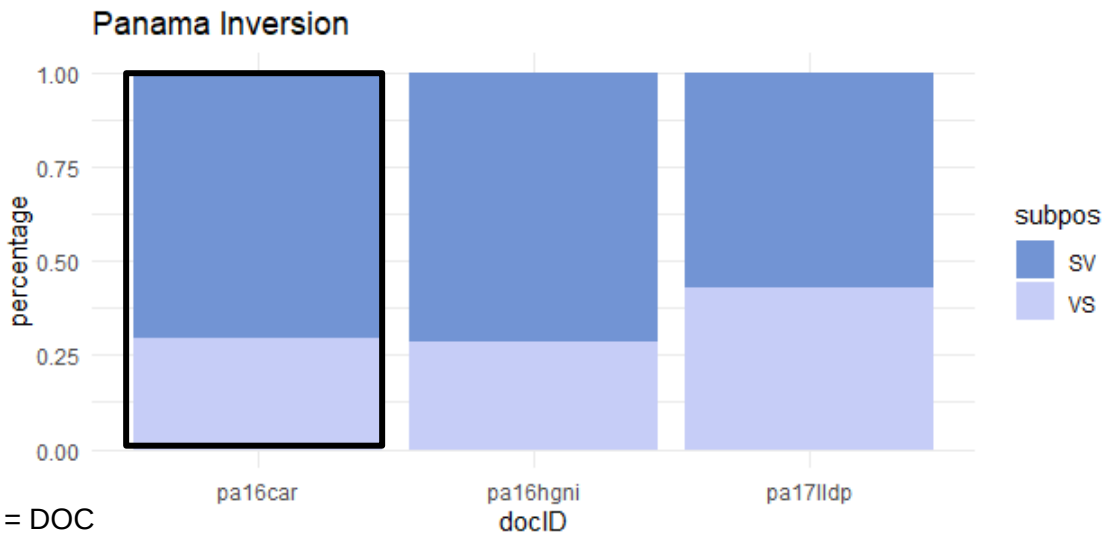




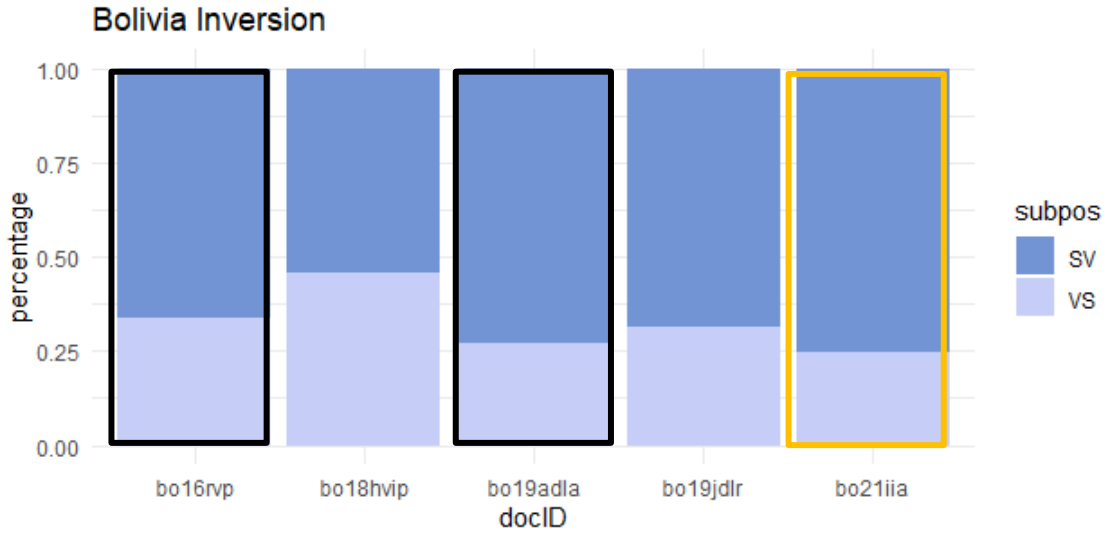


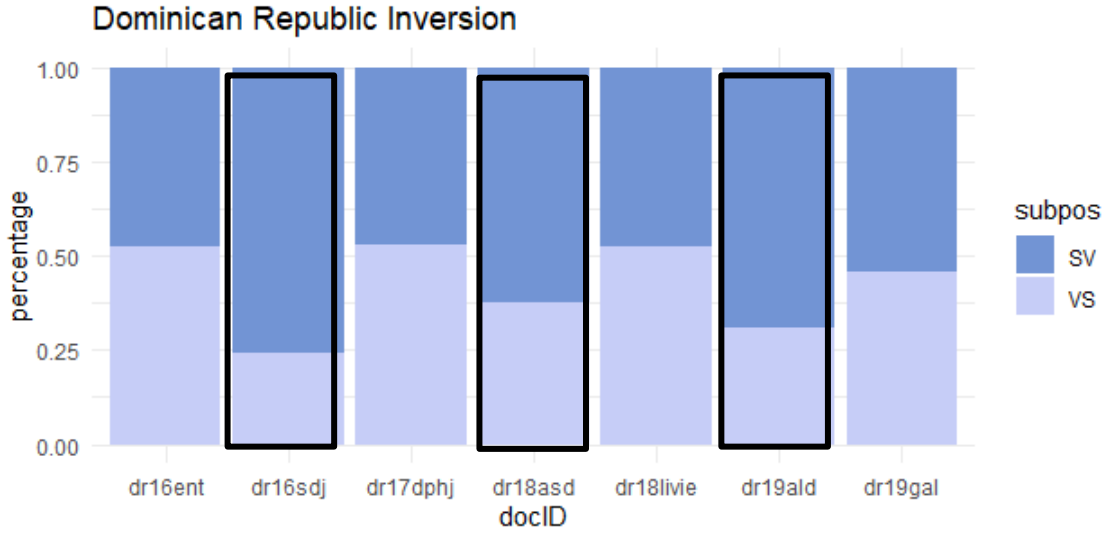
Word Order (Percent)



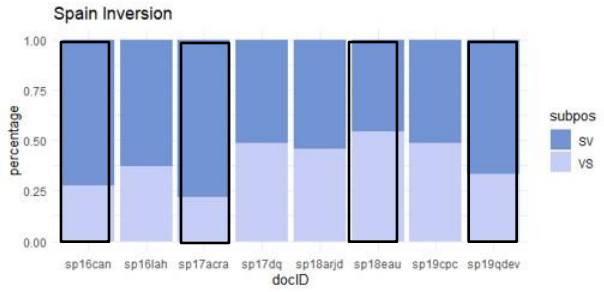


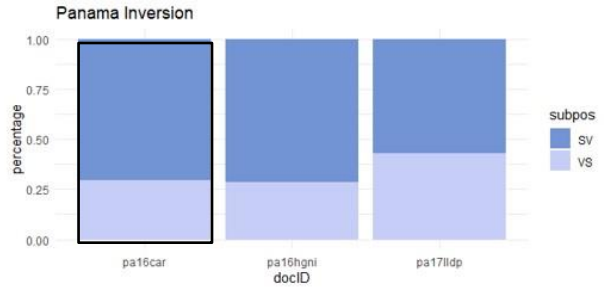
Black border = DOC
Gold border = Supplementary Text

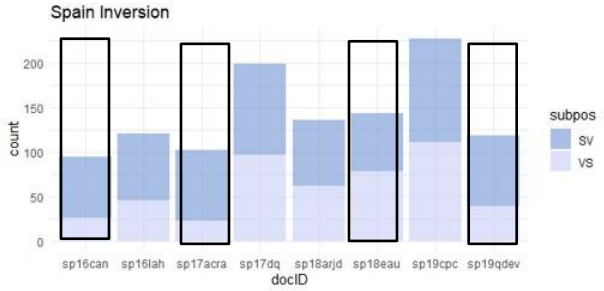


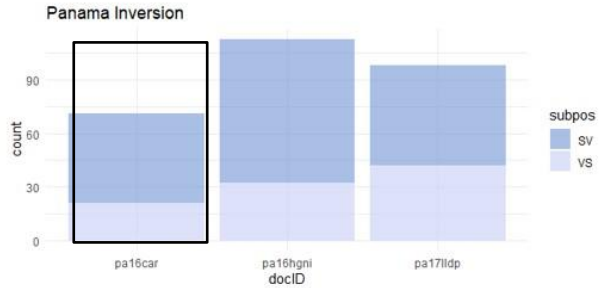


Word Order (count)

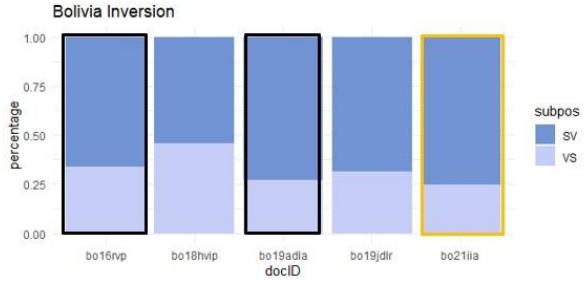


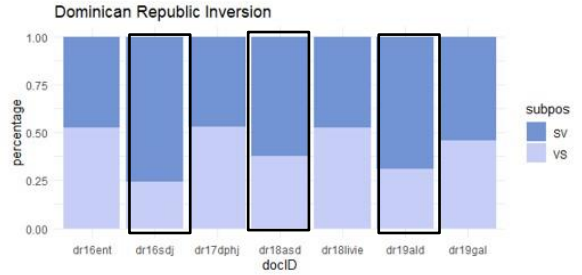


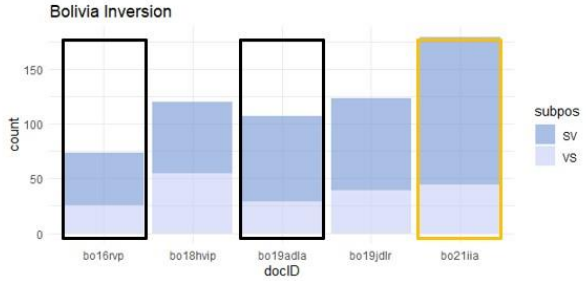


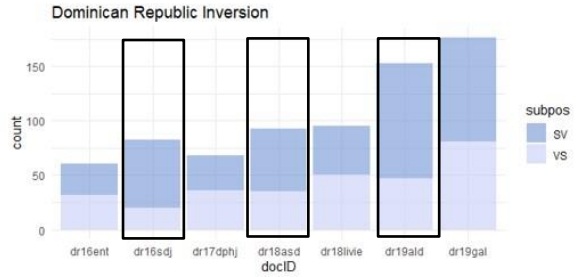


Black border = DOC
Gold border = Supplementary Text

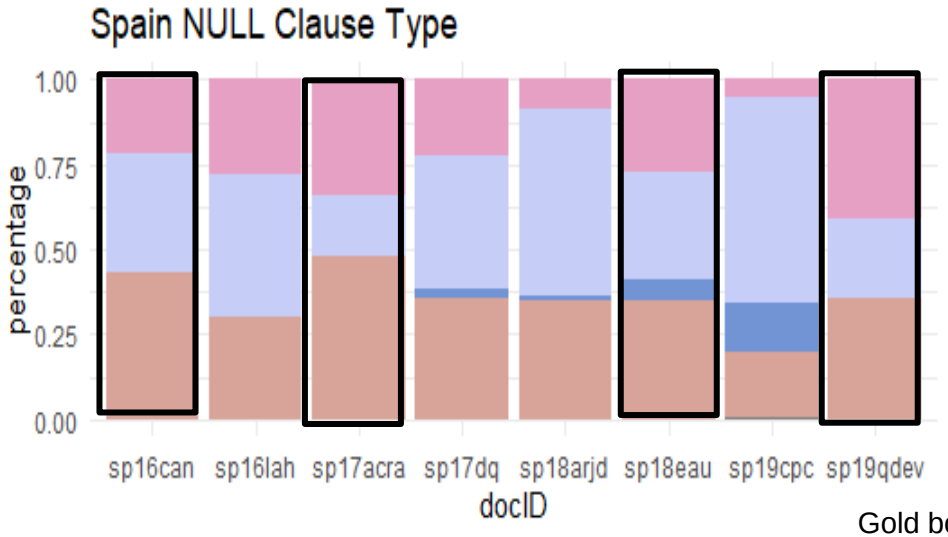


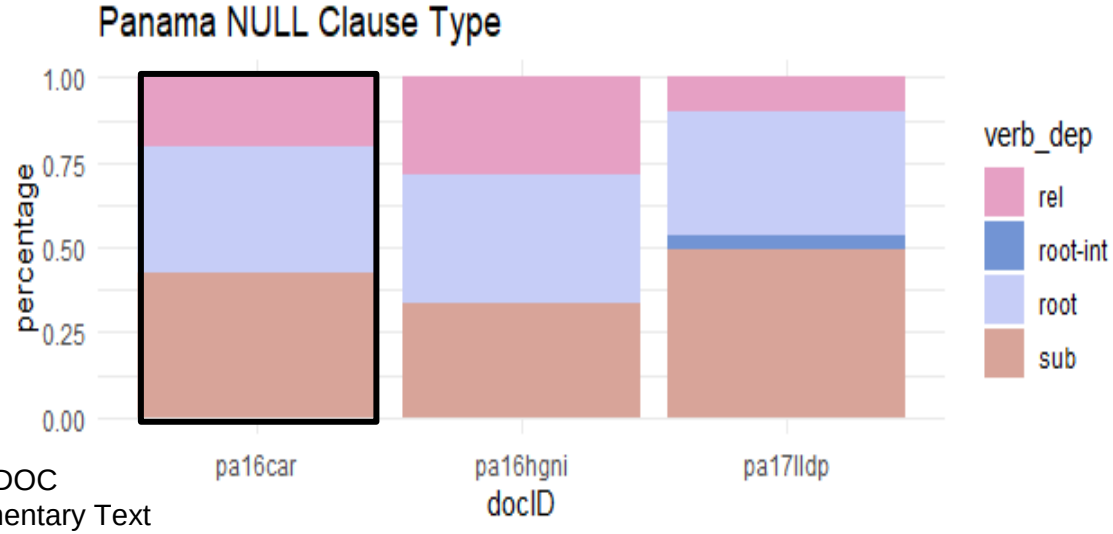




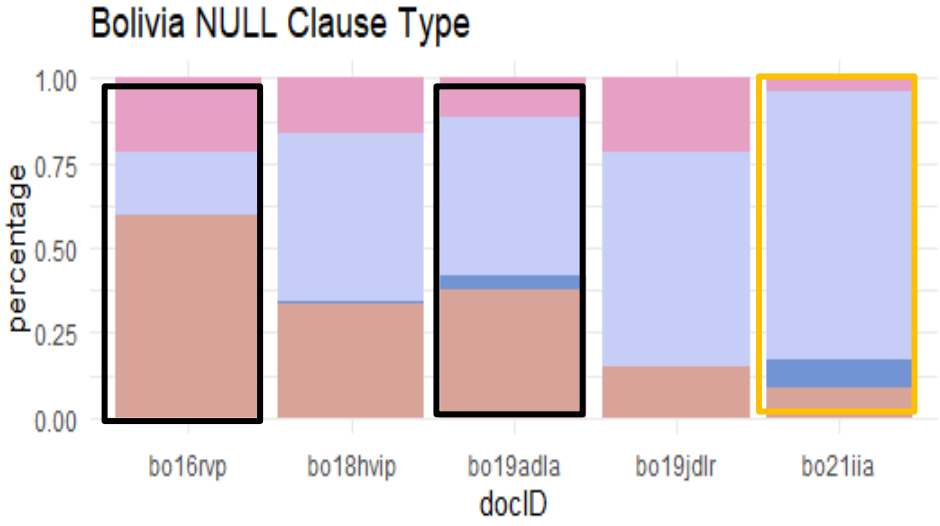


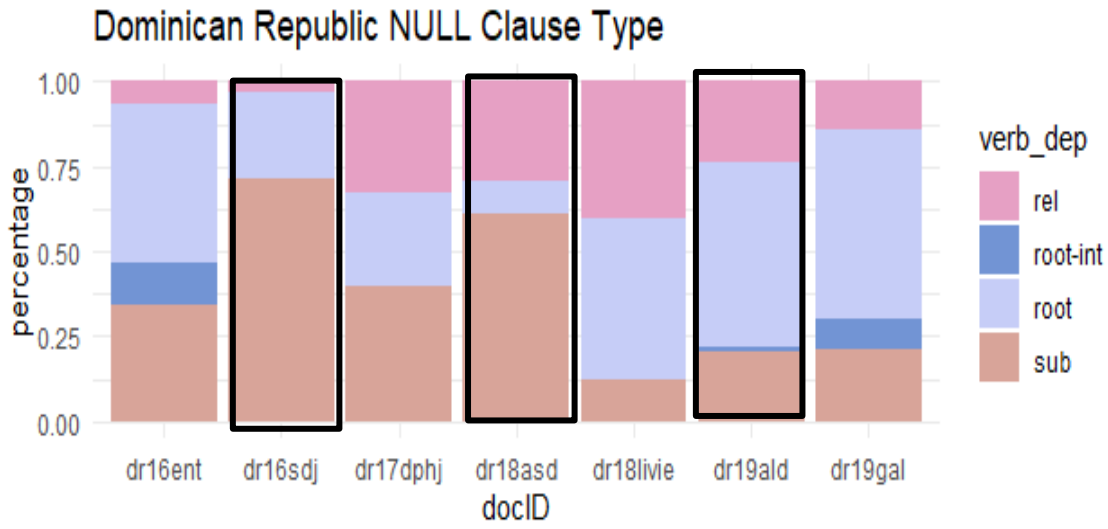
Clause Type (Null)





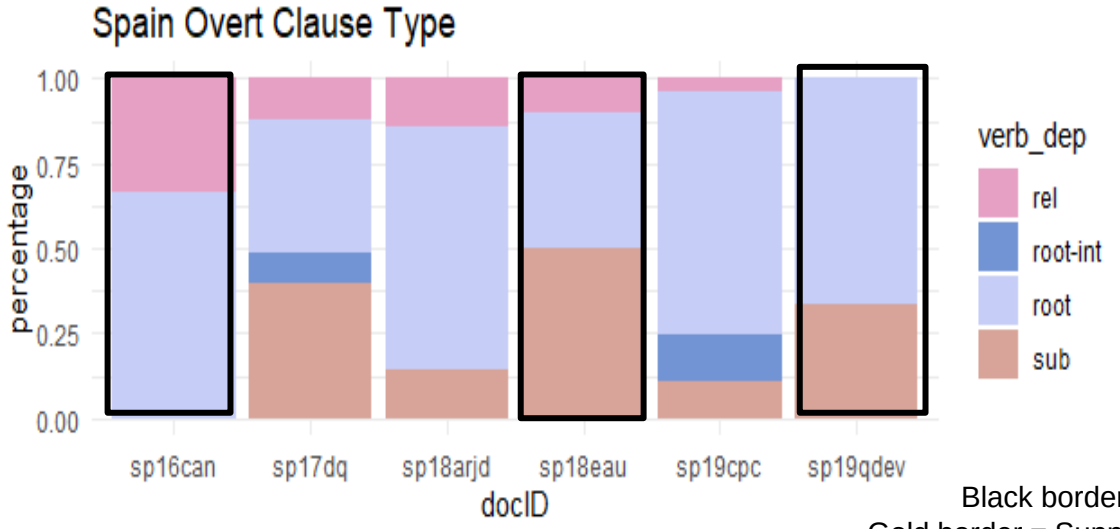
Black border = DOC
Gold border = Supplementary Text

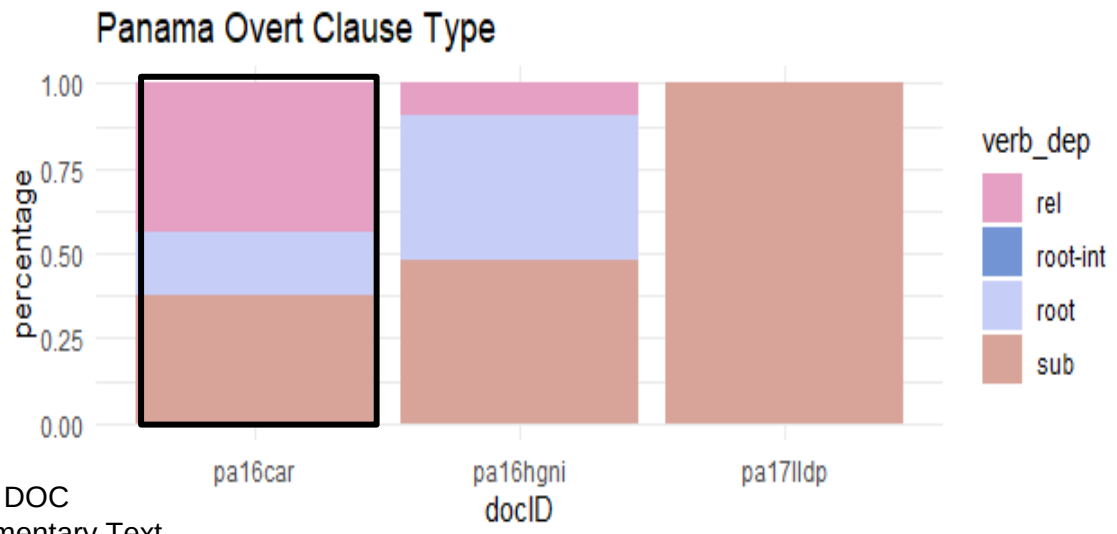




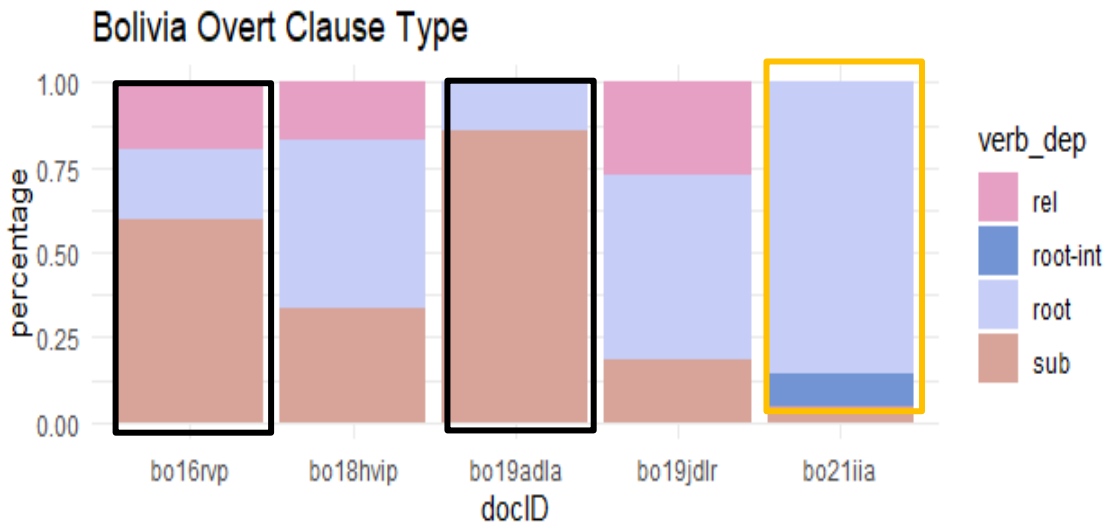
Clause Type (Overt)

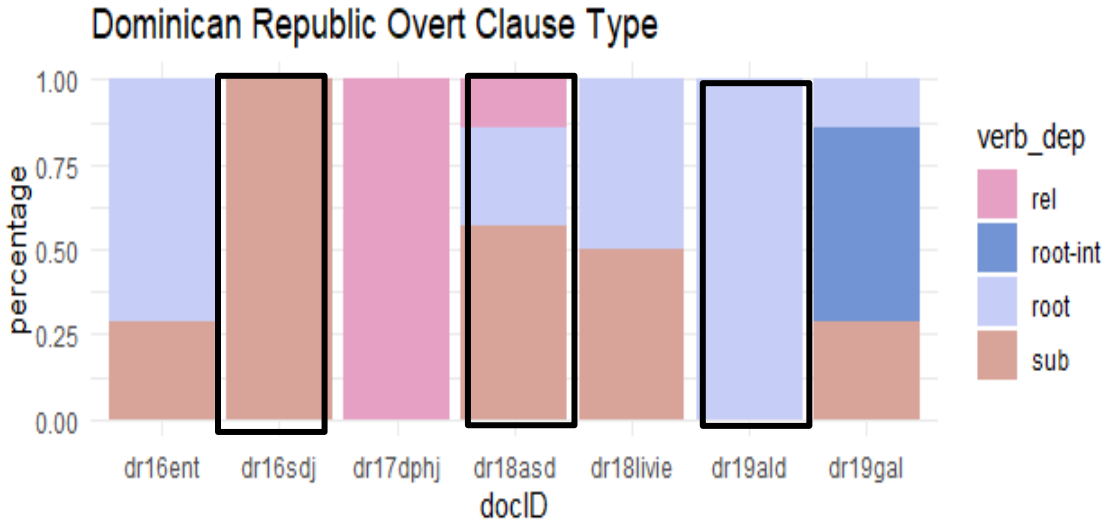
*Three texts from the Spain chart are missing because they don't have any overt subjects at all. Perhaps crucially, they are from the 16th and 17th centuries.



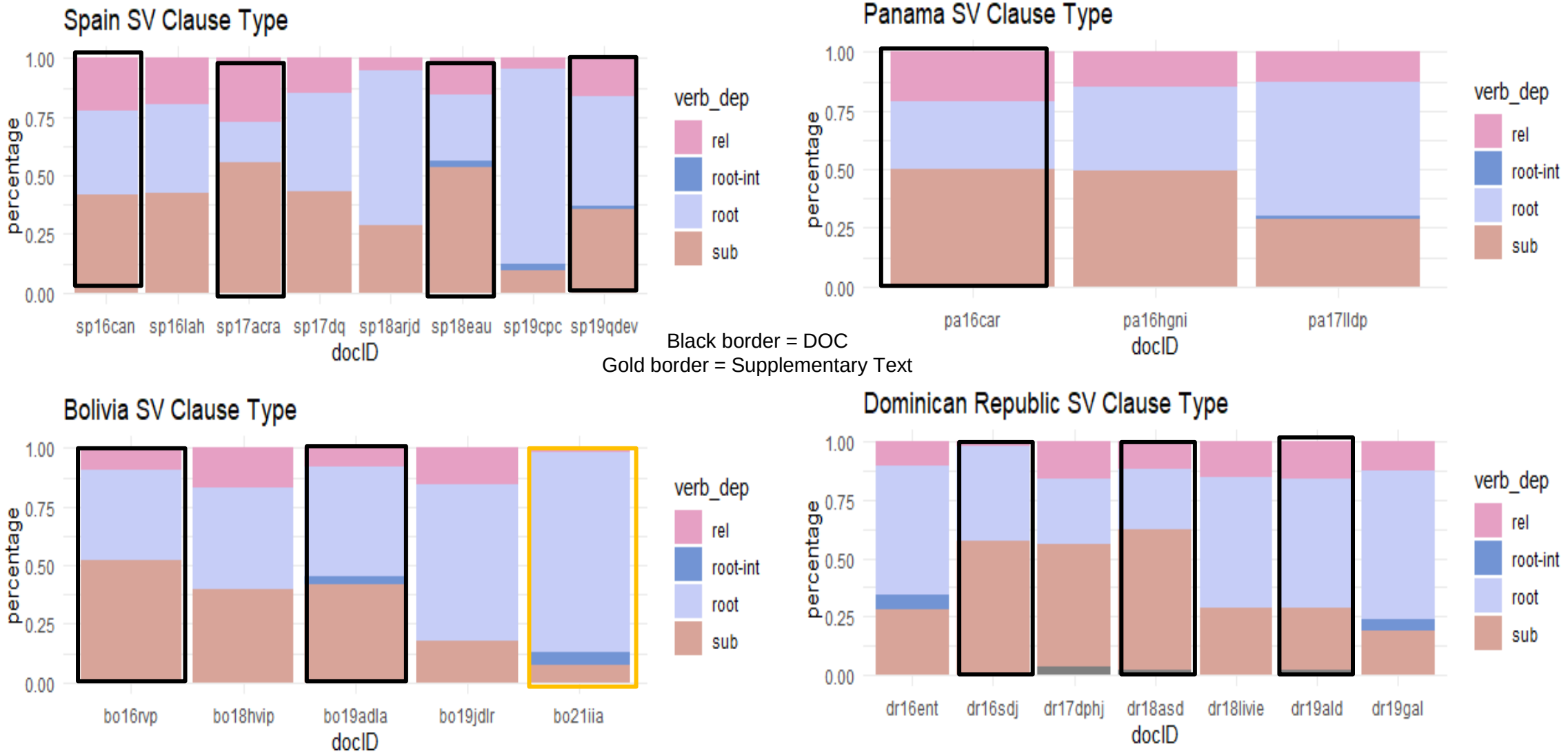


Black border = DOC
Gold border = Supplementary Text

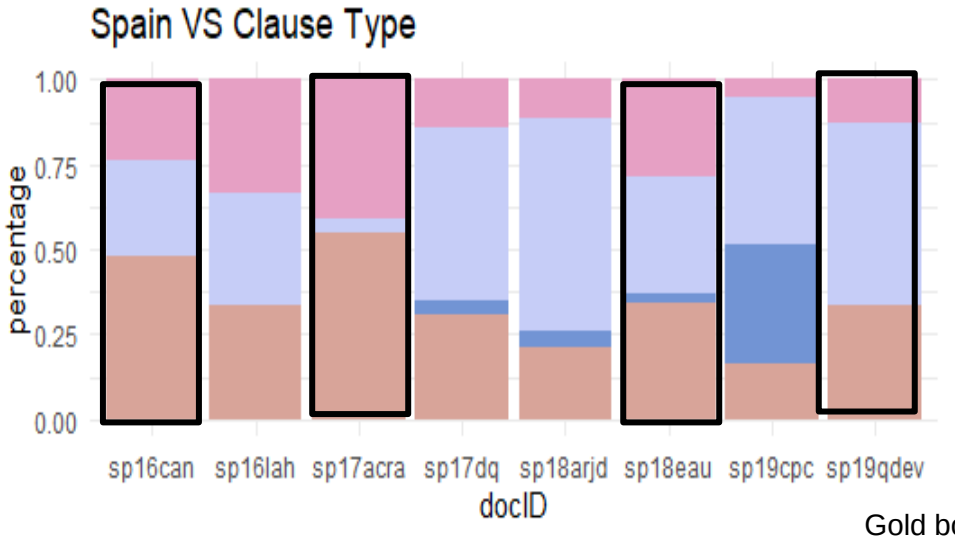


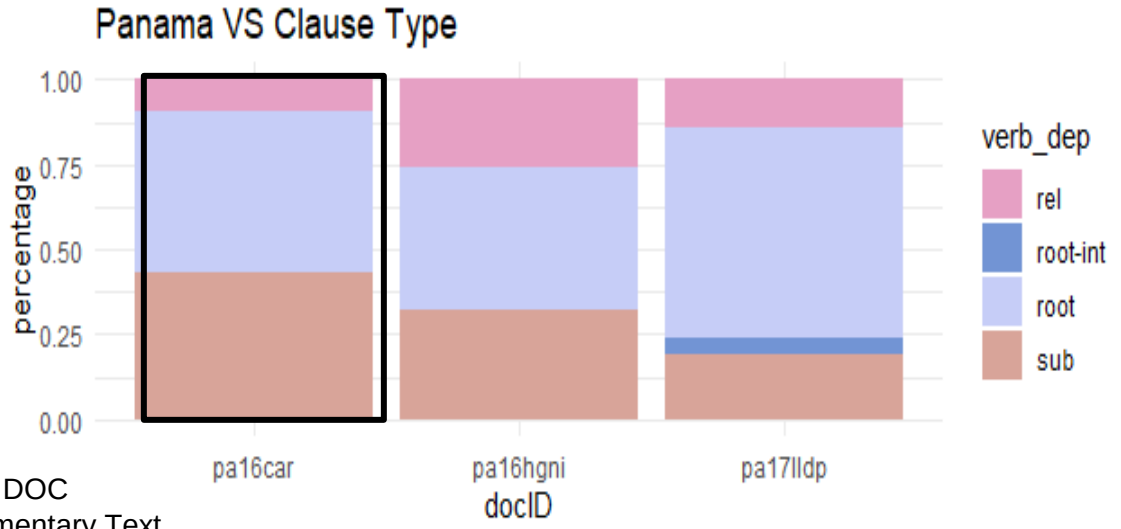


Clause Type (SV)

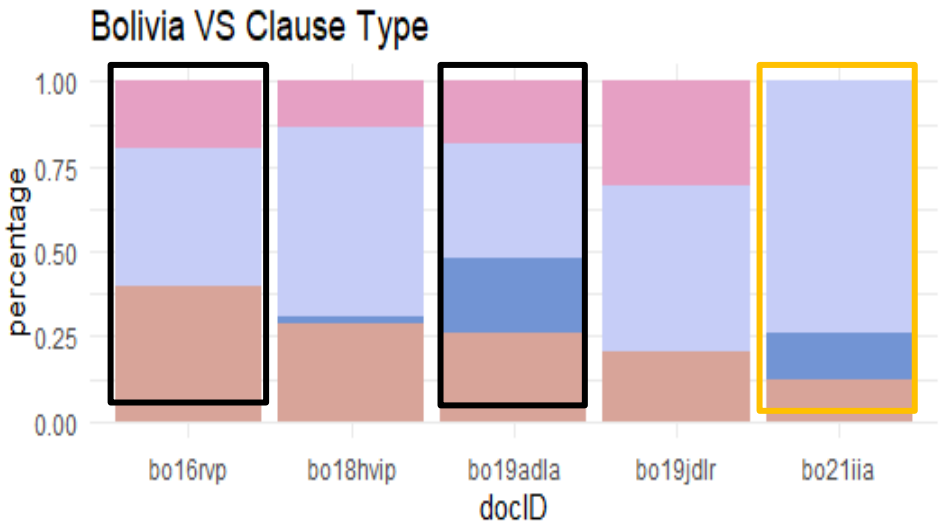


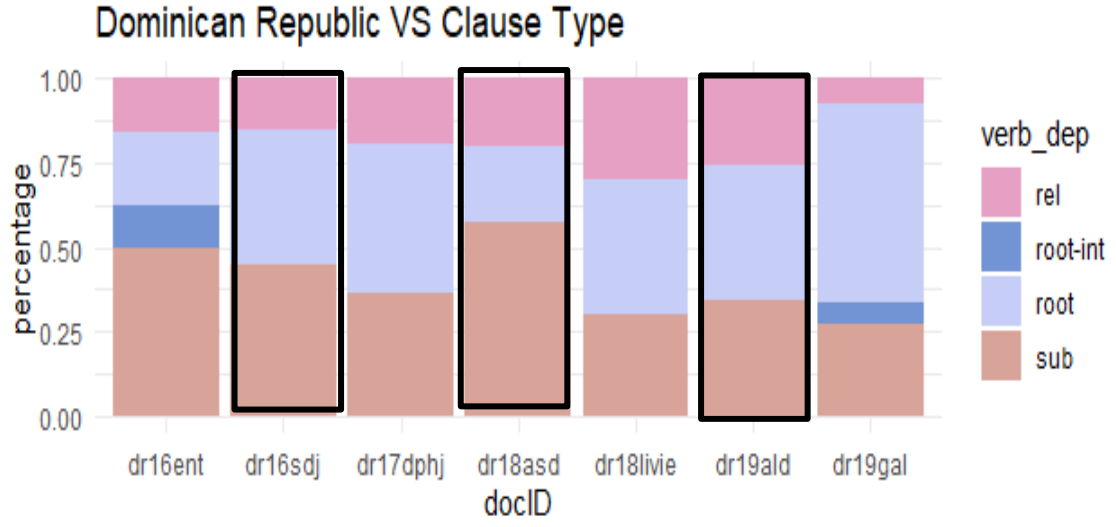
Clause Type (VS)





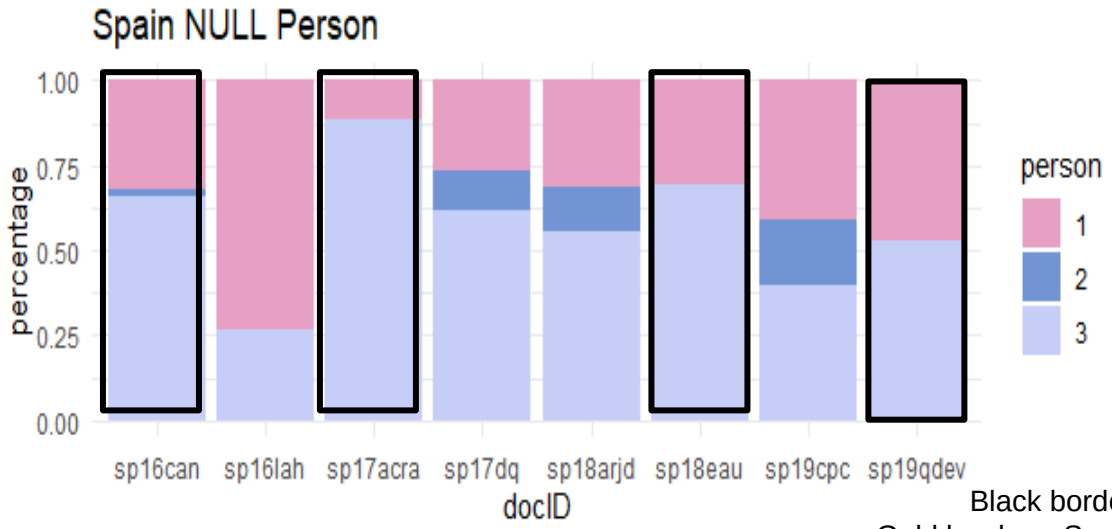
Black border = DOC
Gold border = Supplementary Text

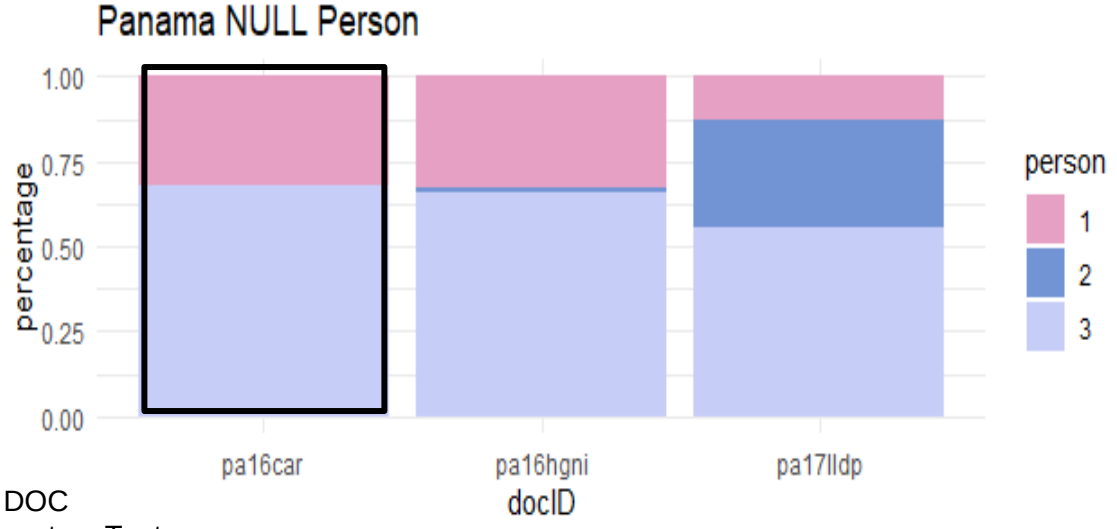




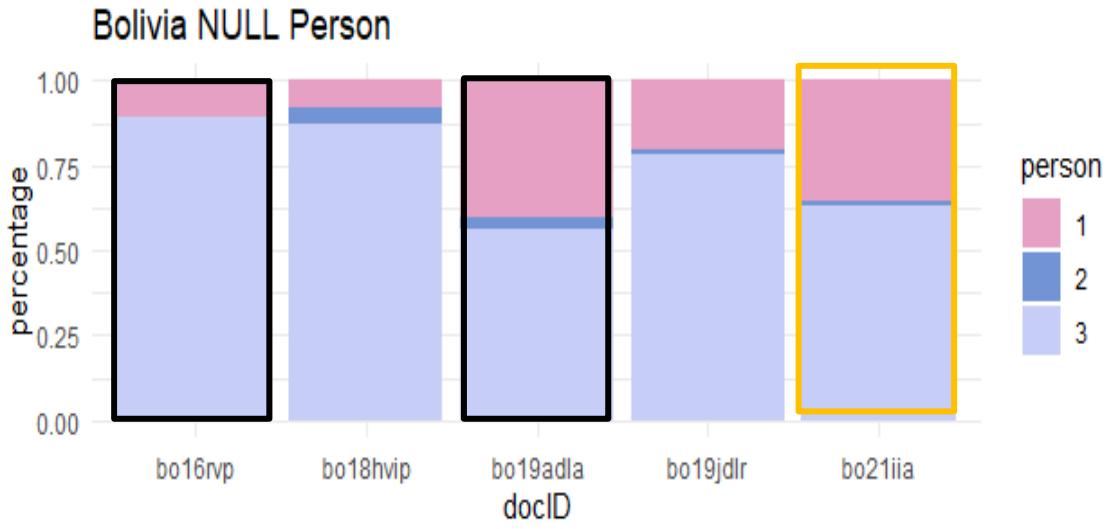
Person (Null)

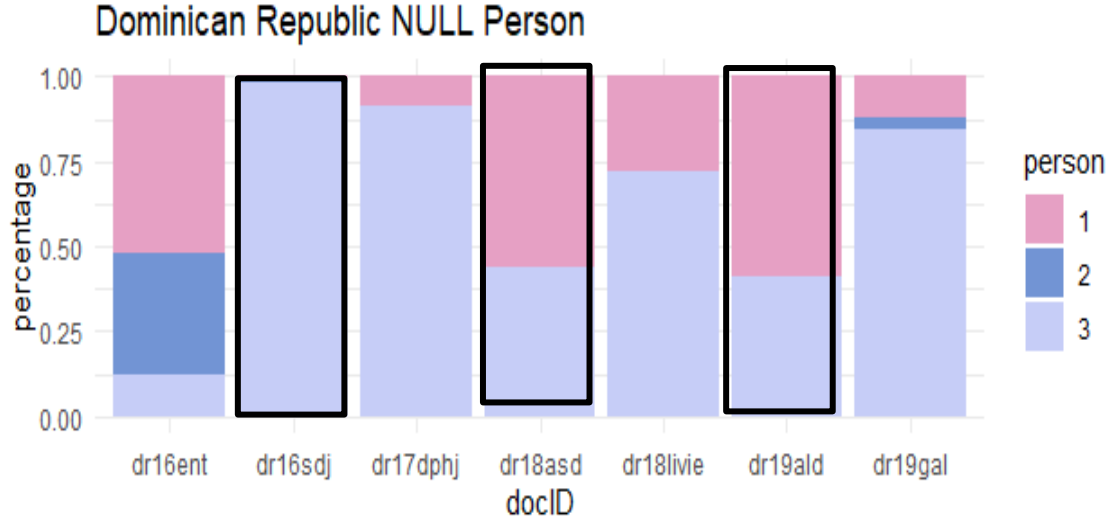
Caveat: 3rd person includes usted/ustedes here because it comes from the verbal morphology (since null subjects couldn't be tagged for their morphology)





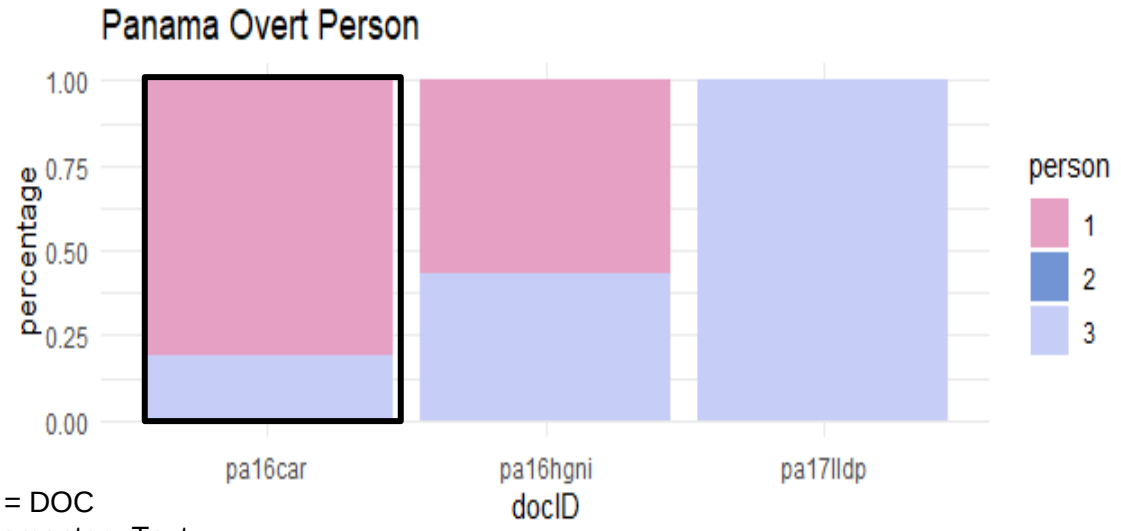
Black border = DOC
Gold border = Supplementary Text



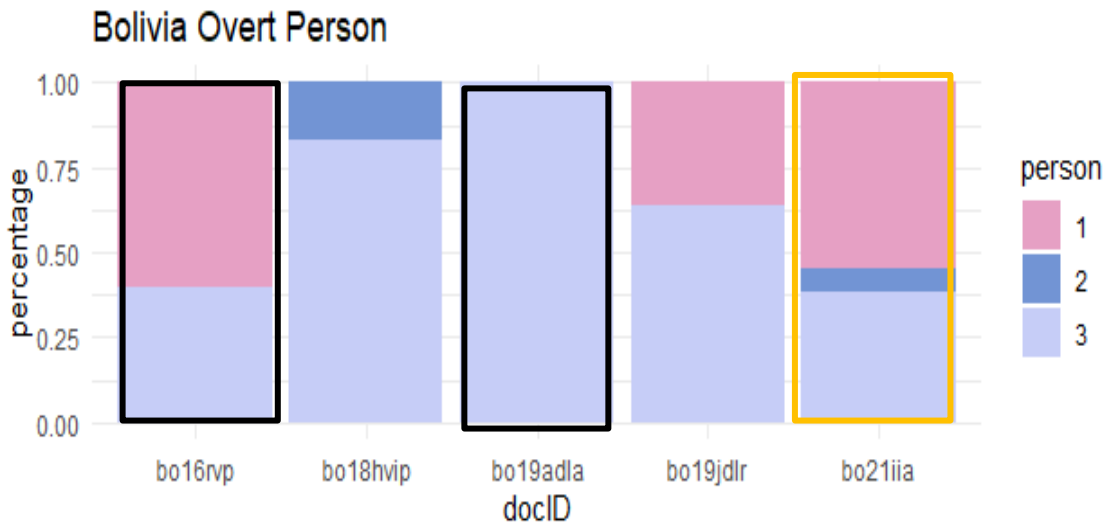


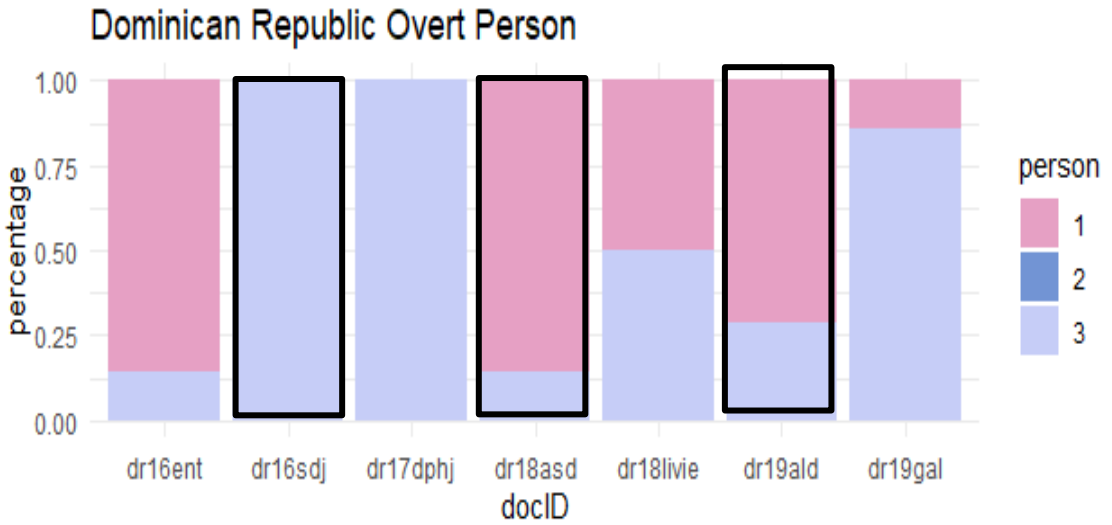
Person (Overt)

*Again, the same three Spain texts that don't have any overt subjects at all are missing.



Black border = DOC
Gold border = Supplementary Text





Mixed Models: Pronoun Realization

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: sub_POS ~ Country * Genre + Century + (1 | docID)
   Data: binary
Control: glmerControl(optimizer = "bobyqa")

     AIC      BIC   logLik deviance df.resid
  1494.0   1562.9   -735.0   1470.0     2284

Scaled residuals:
    Min      1Q  Median      3Q     Max
-0.5344 -0.4038 -0.2942 -0.1648  7.1139

Random effects:
 Groups Name        Variance Std.Dev.
 docID  (Intercept) 0.5594   0.7479
Number of obs: 2296, groups:  docID, 22

Fixed effects:
                           Estimate Std. Error z value Pr(>|z|)
(Intercept)                -2.66604    0.67974  -3.922 8.78e-05 ***
CountryDR                  -0.46910    0.80825  -0.580   0.5617
CountryPanam<e1>            0.83290    1.04422   0.798   0.4251
CountrySpain               -0.63319    0.81031  -0.781   0.4346
GenreLIT                   -0.64479    0.89361  -0.722   0.4706
Century17                  -0.44325    0.63780  -0.695   0.4871
Century18                   0.67977    0.58709   1.158   0.2469
Century19                   0.96158    0.54471   1.765   0.0775 .
CountryDR:GenreLIT          0.76798    1.14195   0.673   0.5013
CountryPanam<e1>:GenreLIT  -0.06033    1.34775  -0.045   0.9643
CountrySpain:GenreLIT       1.27829    1.10131   1.161   0.2458
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- **Models**
  - glmer from lme4 package in R
  - Looking at the fixed variables of Country, Genre, and Century and their interactions for pronoun realization and word order
  - Neither model would converge with Year as continuous variable (even when used as the only variable)
- **Pronoun Realization**
  - Country*Genre*Century : **no**
  - Country*Genre + Century : **yes (18th marginal)**
  - Country + Genre + Century : **yes (nothing close to significant)**
  - Country / Genre / Century: **yes (still nothing significant)**
  - So, the model doesn’t find anything.
  - We’ll see if that changes once the corpus is complete and there’s more data

Mixed Models: Word Order

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: subpos ~ Country * Genre + Century + (1 | docID)
   Data: inversion

     AIC      BIC   logLik deviance df.resid
  3408.9   3479.2  -1692.5   3384.9     2566

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.1457 -0.8051 -0.6490  1.0654  1.7194

Random effects:
 Groups Name        Variance Std.Dev.
 docID  (Intercept) 0.01842  0.1357
Number of obs: 2578, groups:  docID, 22

Fixed effects:
                            Estimate Std. Error z value Pr(>|z|)
(Intercept)                -0.94766    0.20828  -4.550 5.37e-06 ***
CountryDR                  -0.11821    0.24264  -0.487 0.626125
CountryPanam<e1>            0.07716    0.36015   0.214 0.830350
CountrySpain                0.06302    0.23327   0.270 0.787041
GenreLIT                    0.09230    0.26200   0.352 0.724624
Century17                   0.22535    0.16626   1.355 0.175295
Century18                   0.59706    0.16192   3.687 0.000227 ***
Century19                   0.17053    0.14917   1.143 0.252968
CountryDR:GenreLIT          0.72081    0.32627   2.209 0.027157 *
CountryPanam<e1>:GenreLIT   0.05566    0.43219   0.129 0.897528
CountrySpain:GenreLIT       0.37056    0.30880   1.200 0.230130
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Country*Genre*Century : **no**
- Country*Genre + Century : **yes**
  - **18th century**
  - **interaction between Genre and DR**

```
Generalized linear mixed model fit by maximum likelihood (Laplace
  Approximation) [glmerMod]
 Family: binomial  ( logit )
Formula: subpos ~ Country + Genre + Century + (1 | docID)
   Data: inversion

     AIC      BIC   logLik deviance df.resid
  3408.6   3461.3  -1695.3   3390.6     2569

Scaled residuals:
    Min      1Q  Median      3Q     Max
-1.0724 -0.8309 -0.6373  1.0651  1.6758

Random effects:
 Groups Name        Variance Std.Dev.
 docID  (Intercept) 0.03115  0.1765
Number of obs: 2578, groups:  docID, 22

Fixed effects:
                  Estimate Std. Error z value Pr(>|z|)
(Intercept)      -1.113364   0.189885  -5.863 4.54e-09 ***
CountryDR         0.288844   0.173493   1.665  0.09594 .
CountryPanam<e1>  0.007397   0.239295   0.031  0.97534
CountrySpain      0.256611   0.170935   1.501  0.13330
GenreLIT          0.485799   0.118692   4.093 4.26e-05 ***
Century17         0.188085   0.179426   1.048  0.29452
Century18         0.503938   0.170777   2.951  0.00317 **
Century19         0.114708   0.162789   0.705  0.48103
---
Signif. codes:  0 ‘***’ 0.001 ‘**’ 0.01 ‘*’ 0.05 ‘.’ 0.1 ‘ ’ 1
```

- Country + Genre + Century **: yes**
  - **18th still but less so**
  - **Genre in general**
  - Same results when each variable run individually
- Why the 18th century? I can't say other than that since year had to be adjusted to century, the model doesn't take into account that there's a diachronic relationship

Putting the Models into Perspective

	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

- **There will be more than double the data by the time the corpus is complete**
- **It is important to keep in mind that this is just preliminary data**
- **When the models have more to work with, they may yield some significant findings**

Conclusion

Main questions:

1. ***do overtness and SV word order increase diachronically?***
   - Not significantly in the data we have
   - Why not? Possibly these changes just were not captured in the written register
2. ***do they have higher rates from Spain > South America > Caribbean?***
   - No, there is a lot of inter- and intra-country variation
   - Why not? Again, possibly a register effect
3. ***do certain genres have higher rates than others?***
   - Yes, the “DOC” genre has a higher SV rate in each country, especially the DR
   - Why? Inversion is pragmatically determined in Spanish, used to mark emphasis and focus (Sánchez 2008)
     - Possibly non-literary texts mark emphasis and focus less than literary texts
     - Possibly literary texts prefer to introduce new information through subjects whereas documents favor using subjects as topics
   - Alternatively, post-verbal subjects seem to be preferred by subordinate clauses (Rivas 2013) which had higher rates in the document texts
     - Possibly, like Germanic languages, Spanish word order is determined in part by clause type
   - There is also the possibility of interference from verb class, e.g. unaccusatives prefer VS order
     - This will be investigated in a smaller random sample later on

Conclusion, cont.

Additional questions for pronoun realization:

1. ***does switch-reference affect pronoun realization?***
   - It doesn't seem to, switch-reference rate is pretty consistent diachronically and across countries
   - There is maybe an uptick in favor of "same" diachronically, but it doesn't vary by genre and doesn't correlate with the pronoun realization
2. ***does person affect pronoun realization?***
   - 1st person favors overt realization
   - But, there's an increase in overt 3rd person pronouns and a possible increase in null 1st person pronouns
   - Why? Previous studies have found that 1st and 2nd person have the highest overtness rates in Spanish (Cerrón-Palomino 2018)
     - Seems to be a preference for speaker/hearer historically. The increase in overt 3rd person could represent a levelling
3. ***does clause type affect pronoun realization?***
   - There seems to be a move toward more null subjects in main clauses diachronically
   - Overt pronouns seem to prefer sub clauses, but there's a lot more inter-text variation, especially in DR
   - Why? Possibly to further differentiate between the subjects of the main and sub clauses

Conclusion, cont.

Additional questions for word order:

1. ***does clause type affect inversion?***
   - There's the same trend of the number of main clauses increasing
   - As we already mentioned, VS is more common in "sub" and "rel" clauses which supports the idea that clause type plays a role in word order
2. ***does declarative vs. interrogative status affect inversion?***
   - As can be expected, interrogatives favor inversion, but there are still some instances of SV order cropping up

Summation:

- **The data doesn't show the diachronic rise in overt subjects and SV order in LAS that we expected (either from the bar charts or the models)**
- **However, there were interesting and unexpected trends in genre and person which suggest that clause type plays a larger role than anticipated**

References

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.

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.

Rivas, Javier. 2013. “Variable Subject Position in Main and Subordinate Clauses in Spanish: A Usage-Based Approach.” *Moenia* 19 (2013): 97-113.

Rizzi, Luigi. 1982*. Issues in Italian syntax*. Dordrecht: Foris.

Rizzi, Luigi. 1986. Null objects in Italian and the theory of pro. *Linguistic Inquiry* 17: 501–57.

Sánchez, M.E. 2008. “Tipos de cláusula, clases verbales y posición del sujeto en español.” *Lexis* XXXII/1, 83-105.

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!

STARFISH

SOCIOLINGUISTIC TYPOLOGY AND RESPONSIVE FEATURES IN SYNTACTIC HISTORY

gemma-hunter.mccarley@uni-konstanz.de

https://www.ling.uni-konstanz.de/en/walkden/starfish/

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

European Research Council

Established by the European Commission

Appendix: Full Tagset

Sentence:

- poem title/letter number (if applicable)
- speaker number/ character name (if applicable)

Subject:

- dep(endency) type: "nsubj" (Nominal Subject)
- subpos (subject position): **SV/VS**
- POS
  - 3p inanimate expletives: PRON-EXP or NULL-EXP
  - relative pronouns: PRON-REL (these get excluded)
  - passive 'se': XPOS-PASS, NULL-PASS*
  - passive 'se' expletives: NULL-EXP-PASS
  - impersonal 'se': NULL-IMP
  - impersonal expletives (e.g. *hay* 'there is/are'): change to NULL-EXP-IMP

Subject pronouns:

- morphology
  - person: 1/2/3/u (u is for 'usted/es')
  - number: s/p/v (v is for 'vos')
  - e.g. "nosotros" = 1p
- psub (previous subject): **same/diff** (different)/**imp** (impersonal)/ **amb** (ambiguous)
  - this tags for the same referent as the immediately previous clause
  - which means in a dialogue, the person morphology can change between speakers.
  - E.g. Maria: Qué haces? Juan: Tomo café. In this case, the psub is 'same' because the referent is Juan both times
- pp (previous pronoun): **overt/null**

Finite Verbs:

- dep(endency) type: **root** (main clause) / **sub** (dependent clause) / **rel** (relative clause)
  - -INT for questions
- subid (subject ID): the ID of the corresponding subject's token
- morphology: e.g. "me fuera": <morphology>1si-s</morphology>
  - person: 1/2/3
  - number: s/p
  - tense:
    - p=present
    - i=imperfect
    - r=preterite
    - f=future
  - aspect:
    - p=perfect
    - g=progressive
  - mood:
    - i=indicative
    - s=subjunctive
    - c=conditional
    - m=imperative