DIMINUTIVE SUFFIX PRODUCTIVITY

Juan Berríos

LING 2340 – Data Science for Linguists

University of Pittsburgh

INTRODUCTION

GOALS OF THE PROJECT

- A cross-dialectal analysis of two competing Spanish diminutive suffixes (*-ito*, *-illo*) in terms of productivity; i.e., the extent to which a morphological pattern can be applied to new bases and form new words.
- This analysis also considers dialectal variation, given that varieties of Spanish might not necessarily display the same trends.
- Goals:
  1. Explore the cross-dialectal distribution of two competing diminutive suffixes in a representative, cross-dialectal corpus.
  2. Apply statistical measures of productivity to the data.
  3. Determine whether differences are reflected across varieties.

THE MORPHOLOGICAL PATTERN

- Diminutivization.
- Function: form a complex word denoting a smaller version of the base (Haspelmath & Sims, 2010).

a. *un hombre-cito*
a man-DIM.SG
"A little man."

b. *muy chiqu-ito*
very *small*-DIM
"Very small."

c. *com-iend-ito*
eat-PROG-DIM
"Eating."

d. *ahor-ita.*
now-DIM
"NOW."

By MarieM1FLERéunion – Own work, CC BY-SA 4.0, httpscommons.wikimedia.orgwindex.phpcurid=36689506

El principito

"The little prince"

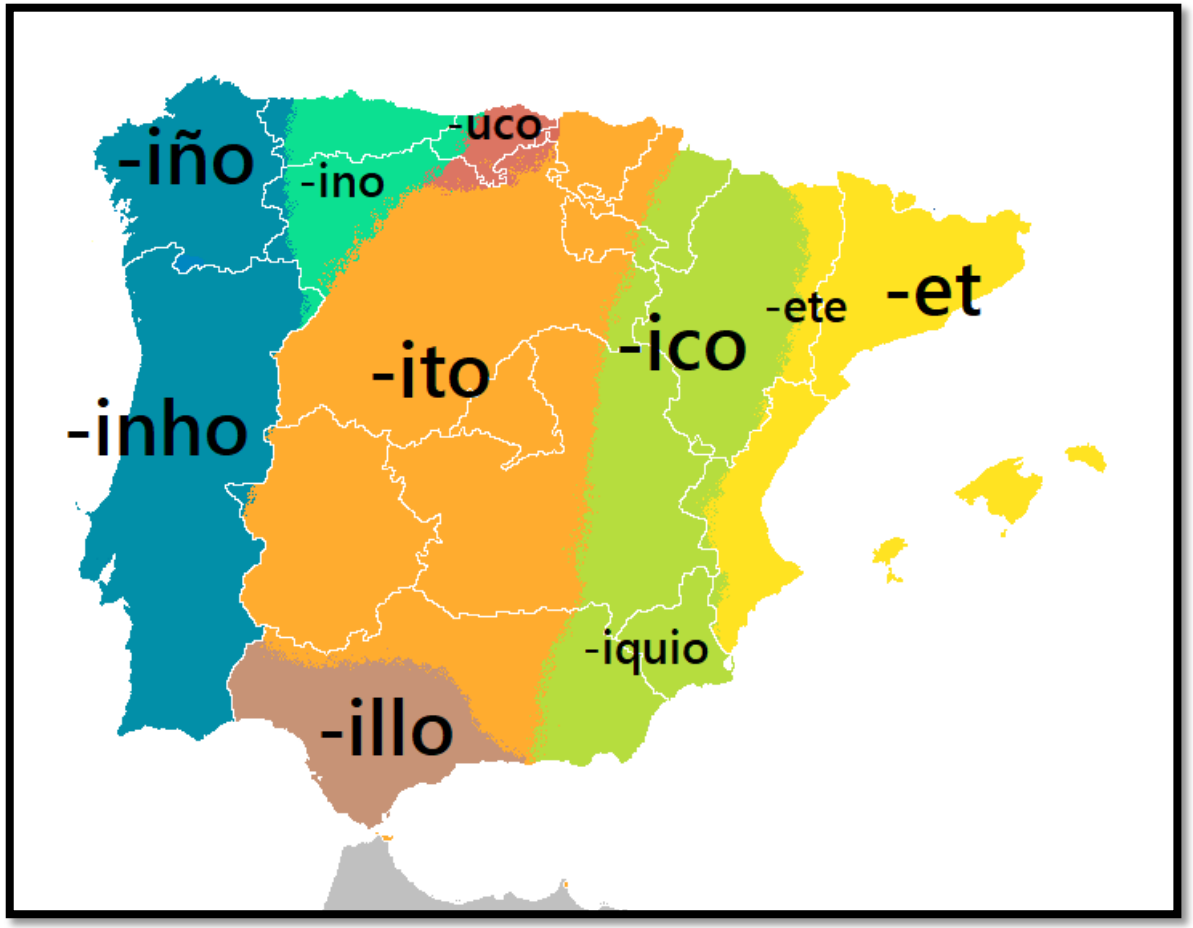
BACKGROUND

DIMINUTIVE FORMATION

- Several productive diminutive suffixes attested in Spanish; notably *–ito, -illo, -ico*, and *-ete* (Hualde, 2013).
- Attention in the literature from several perspectives, notably from competence-based approaches.
- Prior research on diminutive formation has focused on:
  - The distribution of the allomorphs of *-ito* (Bradley & Smith, 2011; Eddington, 2002, 2017).
  - Diminutive formation as explained by different theoretical frameworks, including lexical phonology (Castro, 1998), exemplar theory (Eddington, 2002, 2017), and optimality theory (Bradley & Smith, 2011; Colina, 2003; Elordieta & Carreira, 1996).

IN THIS PROJECT

- Focus on *–ito* and *–illo* because they are both attested across varieties and because a comparison between them can hence provide a more fine-grained analysis of productivity.
- Despite their similarity, *-ito* is widely agreed to be much more productive (Hualde, 2013; Lipski, 1994; Náñez Fernández, 2006).



https://corpuscuenta.wordpress.com/2016/09/10/los-diminutivos-variacion-formacion-y-usos/

University of Pittsburgh

MEASURING PRODUCTIVITY

- Statistical measures of productivity (Baayen, 2009):
- 1. Realized productivity: size of the morphological category.
  - Type count of the members of a morphological category in a corpus with N tokens.
- 2. Expanding productivity: the rate at which a category is attracting new members.
  - The number of words in a morphological category that occur only once in a corpus of N tokens; the hapax legomena.
- 3. Potential productivity: productivity as measured by the number of occasionalisms.
  - The number of hapax legomena in the corpus divided by the total number of tokens affected by the same category. Also known as the category-conditioned degree of productivity.

$$\mathcal{P} = V_{1,m} / N_m$$

$$\mathcal{P}^* = V_{1,m} / V_1$$

RESEARCH QUESTIONS

RESEARCH QUESTIONS

- What is the productivity of each suffix?
  - H1: *-ito* is claimed to be the more productive suffix by far. I expect this to be reflected in the data.
- Are the differences statistically significant?
  - H2: I also expect differences, particularly those of potential productivity, to be significant.
- Are differences reflected across varieties?
  - H3: one of the suffixes (*-illo*) is claimed to be more productive in Spain.

PROCEDURE

DATA

- Corpus del español
  - https://www.corpusdelespanol.org/
  - Searchable online
  - Full corpus available under license
- Web / Dialects
  - Created in 2016
- Size: 2 billion words
- 21 Spanish-speaking countries represented
- Fully lemmatized
- POS-tagged

CORPUS PROCESSING

- The data set is available in three formats: (i) Database (Structured Query Language), (ii) Word/lemma/PoS, and (iii) linear (raw) text. All are .txt files and the former two are tab-delimited.
- Challenges: size, structure of directories, and extraction of relevant rows.

textID	ID(seq)	word	lemma	PoS	
124	2511368388	@@124			
124	2511368389	Gran	gran	o	
124	2511368390	convocatoria	convocatoria	nfs	
124	2511368391	para	para	e	
124	2511368392	el	el	ld-ms	
124	2511368393	concurso	concurso	nms	
124	2511368394	docente	docente	jms	
124	2511368395	que	que	cs	
124	2511368396	se	se	po	
124	2511368397	realiza	realizar	vip-3s	
124	2511368398	en	en	e	
124	2511368399	la	la	ld-fs	
124	2511368400	Escuela	escuela	o	
124	2511368401	Normal	normal	o	
124	2511368402	Con	con	e	
124	2511368403	una	un	li-fs	
124	2511368404	inmensa	inmenso	jfs	
124	2511368405	convocatoria	convocatoria	nfs	
124	2511368406	de	de	e	
124	2511368407	docentes	docente	nmp	
124	2511368408	,	$,	y	

CORPUS PROCESSING

CORPUS PROCESSING

- Functions created:
  - `toDF`
  - `add_variety`
  - `remove_syms`
  - `remove_nondims`
- Master functions
  - `corpus_process`
  - `extract_hapax`

CORPUS PROCESSING

```
def corpus_process(fdir, country_df, variety):
country_df  =  pd.DataFrame(columns=['SourceID',  'TokenID',  'Word',
'Lemma', 'POS'])
    for fname in fdir:
        df = toDF(fname)
        df = remove_nondims(df)
        country_df = pd.concat([country_df, df], sort=True)
    add_variety(country_df, variety)
    return country_df

def extract_hapax(fdir, country_hapax):
country_hapax = set()
    for fname in fdir:
        df = toDF(fname)
        df = remove_syms(df)
        df = remove_redacted(df)
        hapax = set([w.lower() for w in df['Word']])
        for word in hapax:
            country_hapax.add(word)
    return country_hapax
```

CORPUS PROCESSING

Lemma	POS	SourceID	TokenID	Word	Variety
nikita	o	431270	2206403194	Nikita	ES
escrito	jms	431270	2206403206	escrito	ES
calladito	j	431290	2074527333	calladita	ES
sólito	jms	431290	2074527343	sólito	ES
necesitar	vip-3s	431310	2143630275	necesita	ES
...	...	...	...	...	
sencillo	jfs	1891249	1779969779	sencilla	ES
permitir	vsp-1/3s	1891249	1779969895	permita	ES
inscrito	j	1891249	1779970044	inscrito	ES
visita	nfp	1891249	1779970074	visitas	ES
visita	nfp	1891249	1779970084	visitas	ES

Not diminutives

CLEANING AND EXPLORATORY ANALYSIS

- Created a `master_DF` object.
- Removed categories to which the morphological pattern doesn't apply
- Refined the POS column
- Last step:
    - There were still many tokens that did not belong in the data frame because they are (i) lexicalized forms that have acquired a meaning of their own, or (ii) words that meet the word class and phonological requirement but simply do not happen to be diminutives.

CLEANING AND EXPLORATORY ANALYSIS

- Solution: extract a list of highly frequent forms from the corpus that end in the segments of interest to get a list that I later cross-compared with my data frame's `Lemma` column. For this purpose, I used a lexicon that is included with the corpus data, loaded it, derived a frequency-based list of lemmas I wanted to exclude from the data frame, and then actually excluded those from the `master_DF` object.

```
master_DF = master_DF[~master_DF['Lemma'].isin(lexicalized)]
```

CLEANING AND EXPLORATORY ANALYSIS

SourceID	TokenID	Lemma	Word	POS	Variety	POS_binary	Number	Gender
1020347	2416709509	monedita	monedita	n	PE	Noun	unknown	unknown
491500	459306545	papelito	papelito	n	ES	Noun	unknown	unknown
349116	2599105296	carretilla	carretilla	nfs	CU	Noun	singular	feminine
1573286	2665899879	morrillo	Morrillo	nms	CU	Noun	singular	masculine
523695	925421781	masilla	masilla	nfs	ES	Noun	singular	feminine

EXPLORATORY ANALYSIS

EXPLORATORY ANALYSIS

	Lemma	Word	Variety	POS_binary	Diminutive
count	1429012	1429012	1429012	1429012	1429012
unique	49526	62073	20	2	2
top	poquito	poquito	ES	Noun	-ito
freq	59520	54543	383969	1170533	1195810

Figure 1. Diminutive suffix tokens.

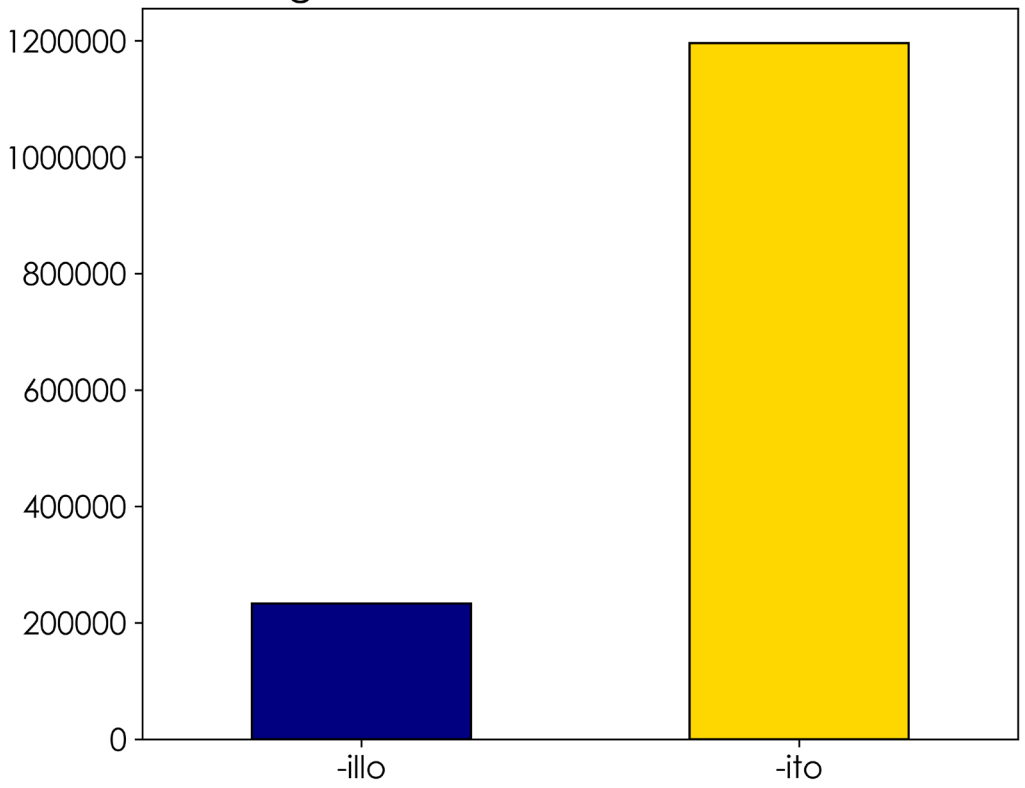


Figure 2. Diminutive suffix tokens by POS.

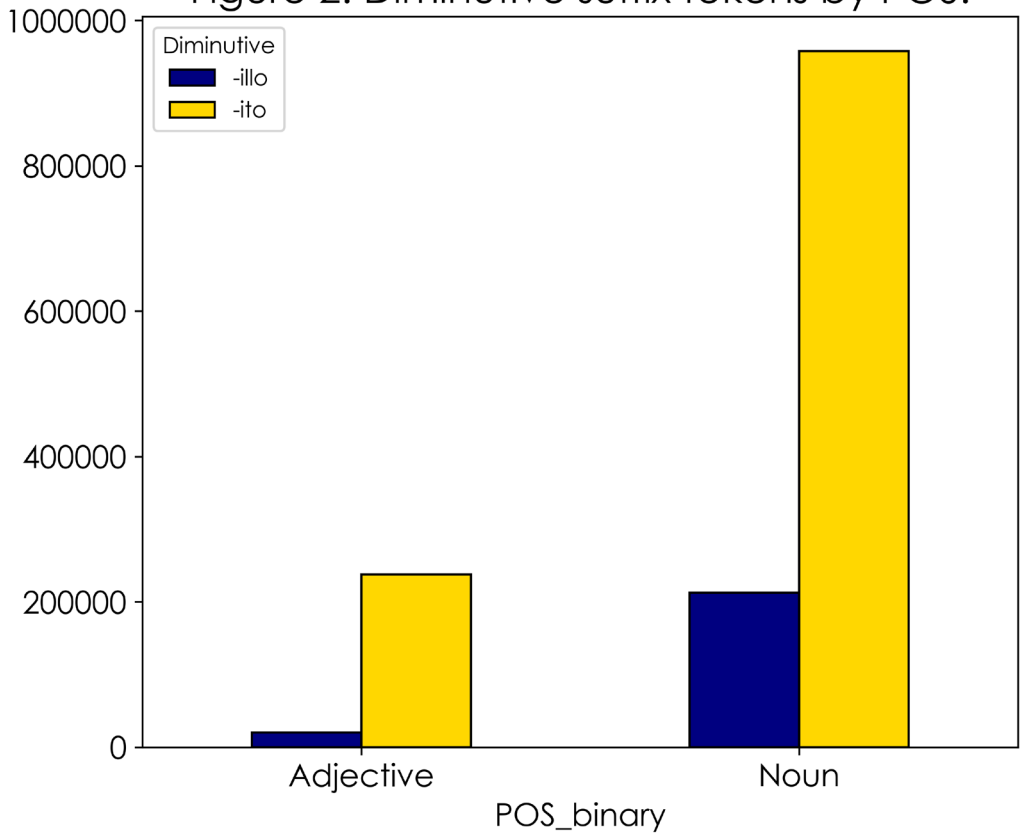


Figure 3. Diminutive suffix tokens by variety.

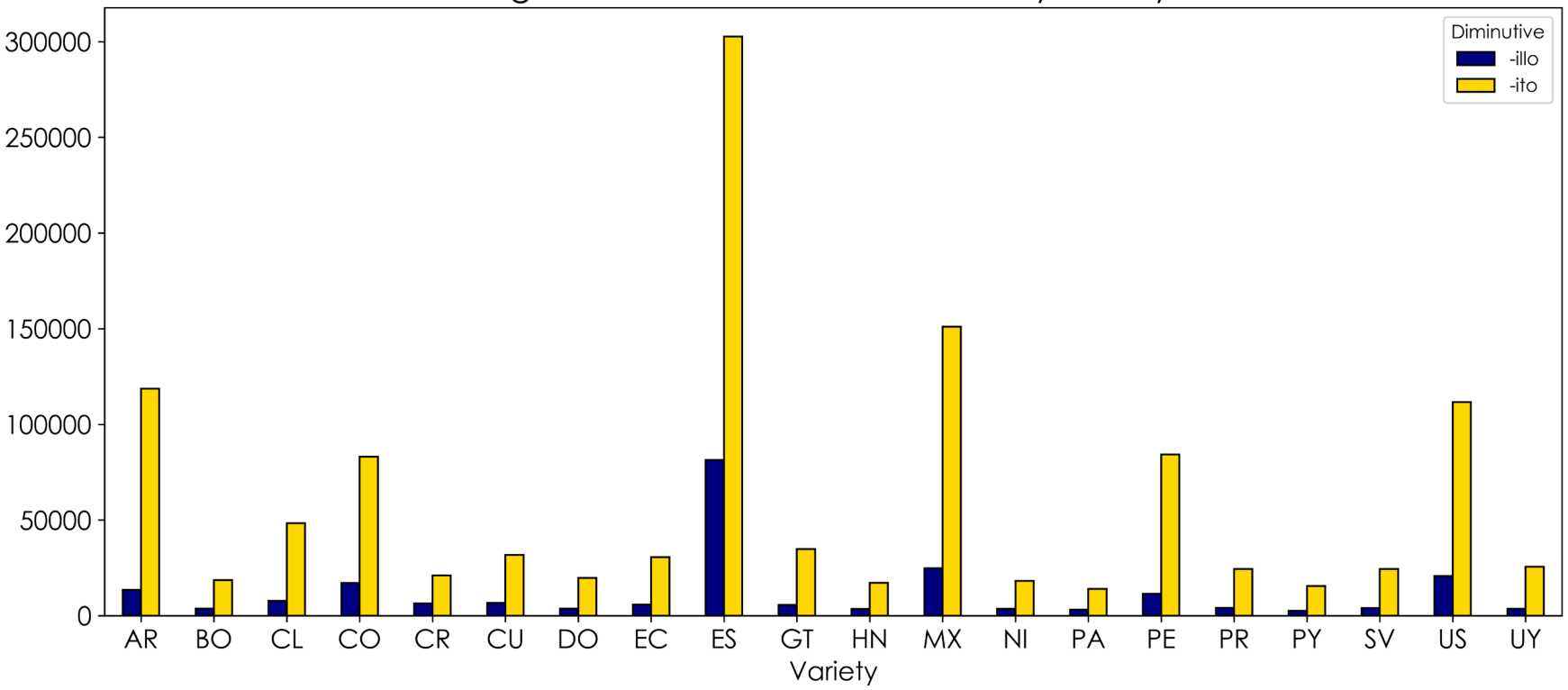
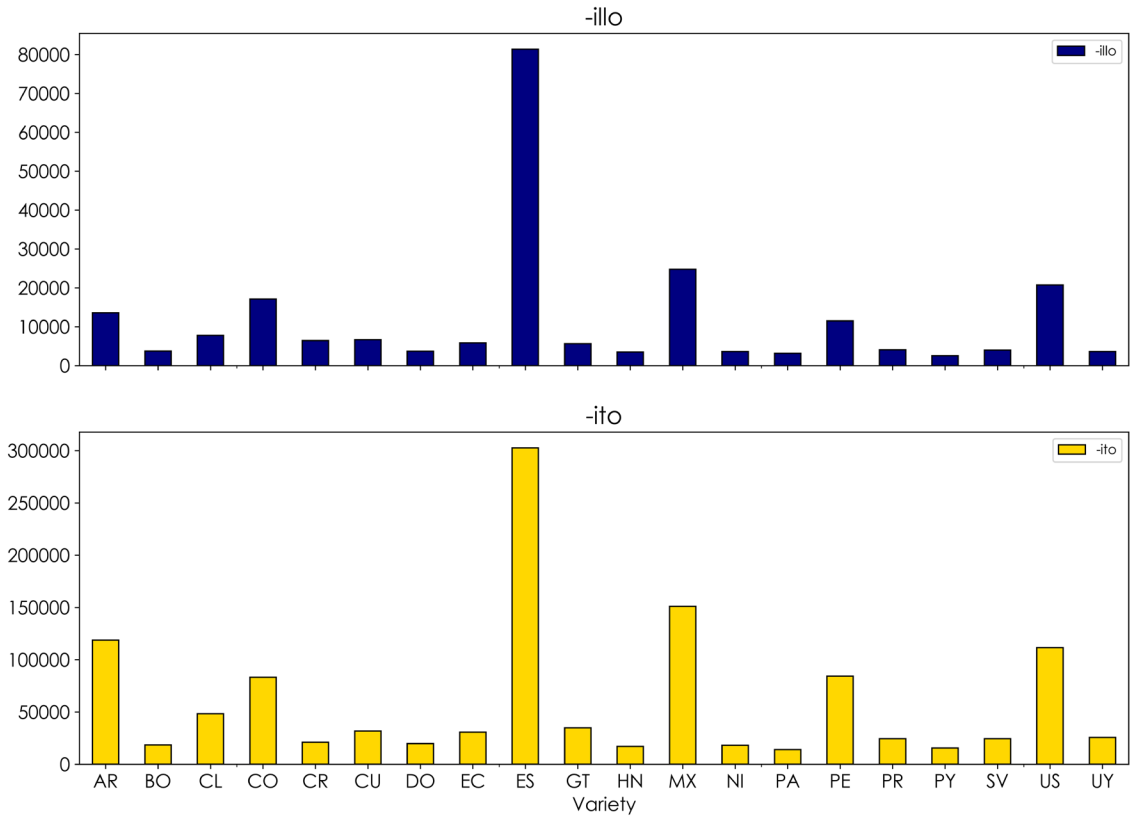


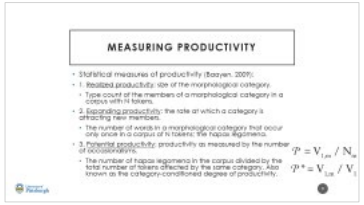
Figure 4. Diminutive suffix tokens by suffix and variety.



STATISTICAL MEASURES OF PRODUCTIVITY

CORPUS PROCESSING

- Extracted hapax legomena from the master data frame.
- Created new summary data frame objects including token counts, type counts, hapax legomena counts, *P*, and *P**.
- Plotted differences across and within varieties.



STATISTICAL MEASURES OF PRODUCTIVITY

Realized productivity

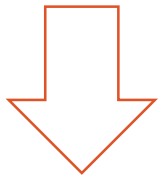
Diminutive	**Tokens**	**Types**	**Hapax**	***P***	***P****
-illo	233202	13157	6513	2.79286	0.121974
-ito	1195810	48930	26611	2.22535	0.498367

STATISTICAL MEASURES OF PRODUCTIVITY

Expanding productivity

Diminutive	**Tokens**	**Types**	**Hapax**	***P***	***P****
-illo	233202	13157	6513	2.79286	0.121974
-ito	1195810	48930	26611	2.22535	0.498367

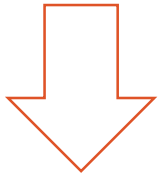
STATISTICAL MEASURES OF PRODUCTIVITY



Category-conditioned degree of productivity

Diminutive	**Tokens**	**Types**	**Hapax**	***P***	***P****
-illo	233202	13157	6513	2.79286	0.121974
-ito	1195810	48930	26611	2.22535	0.498367

STATISTICAL MEASURES OF PRODUCTIVITY



Hapax-conditioned degree of productivity

Diminutive	**Tokens**	**Types**	**Hapax**	***P***	***P****
-illo	233202	13157	6513	2.79286	0.121974
-ito	1195810	48930	26611	2.22535	0.498367

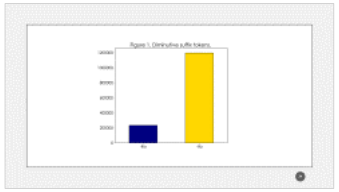


Figure 5. Diminutive suffix types.

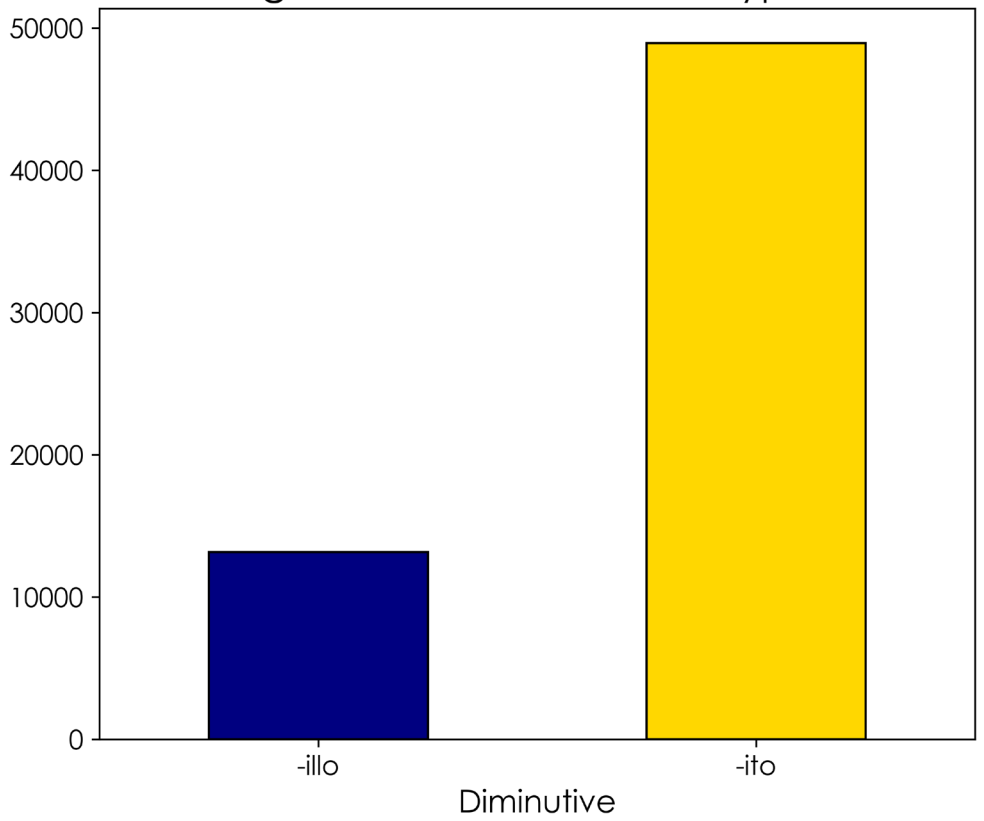


Figure 6. Diminutive suffix hapax legomena.

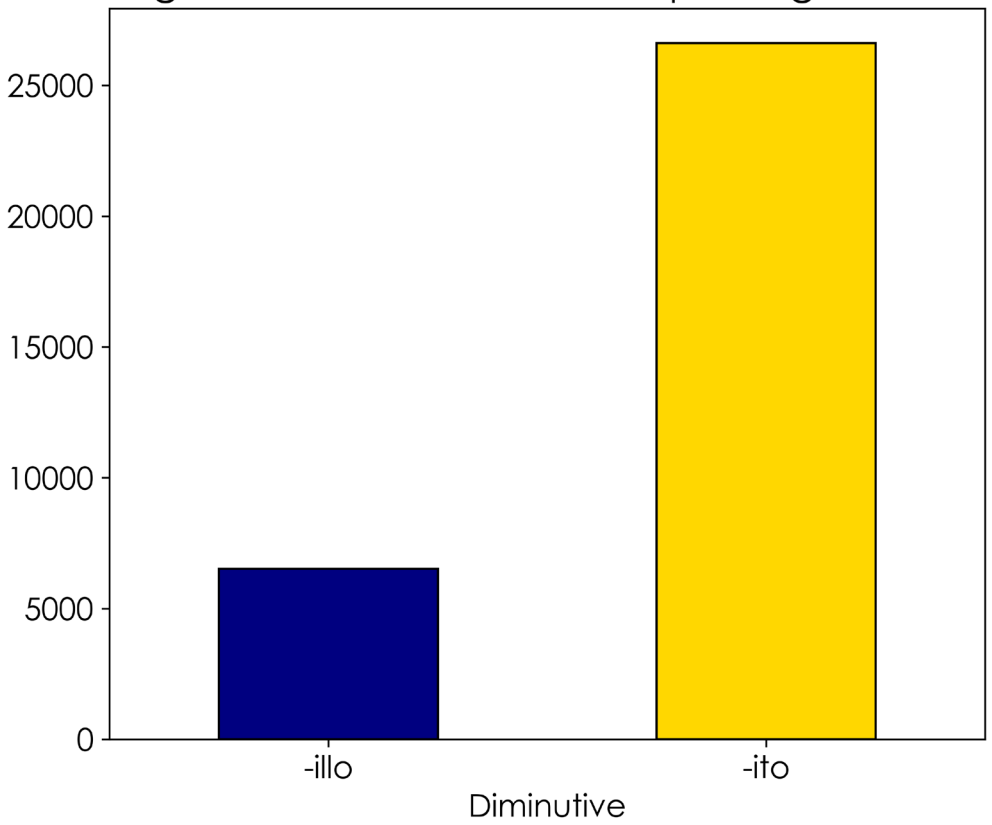


Figure 7. Category-conditioned productivity by variety.

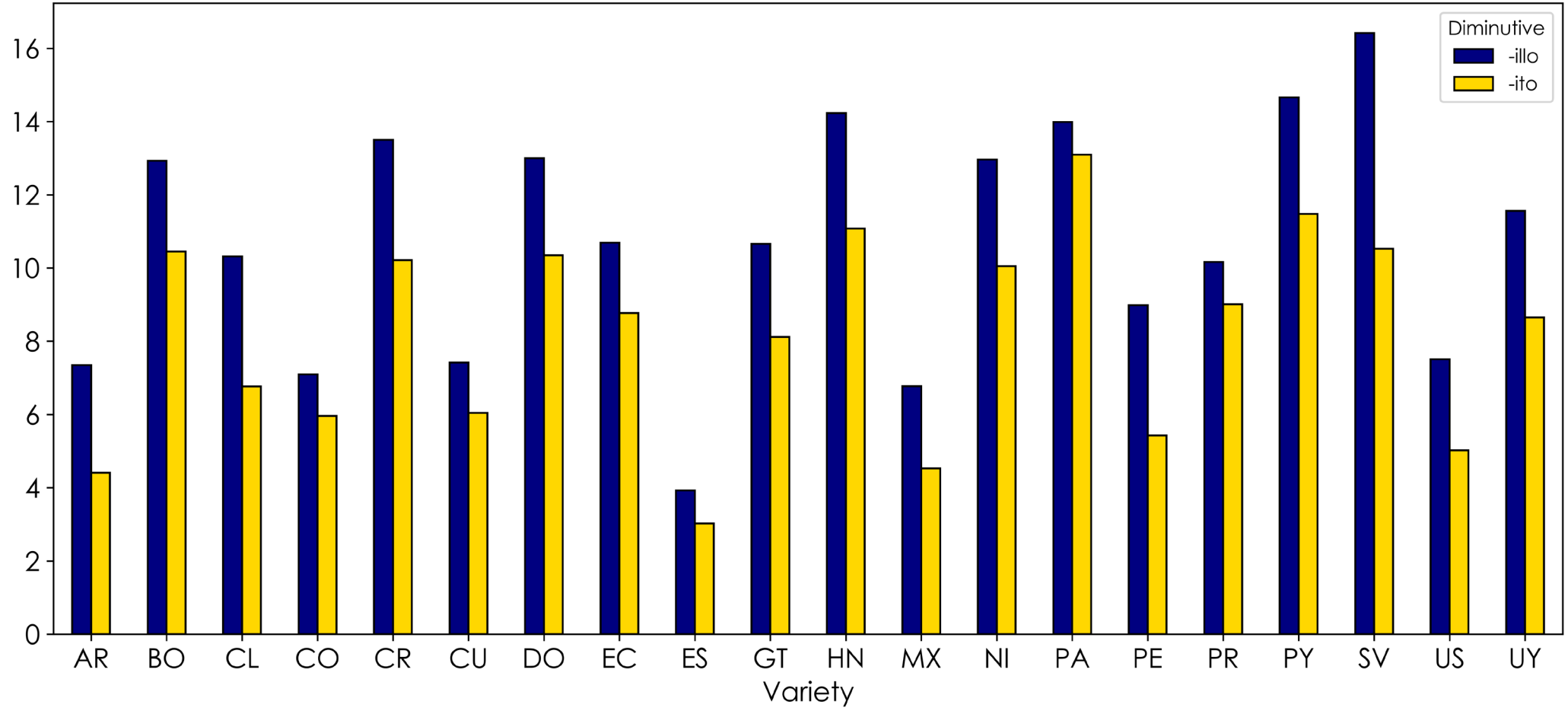
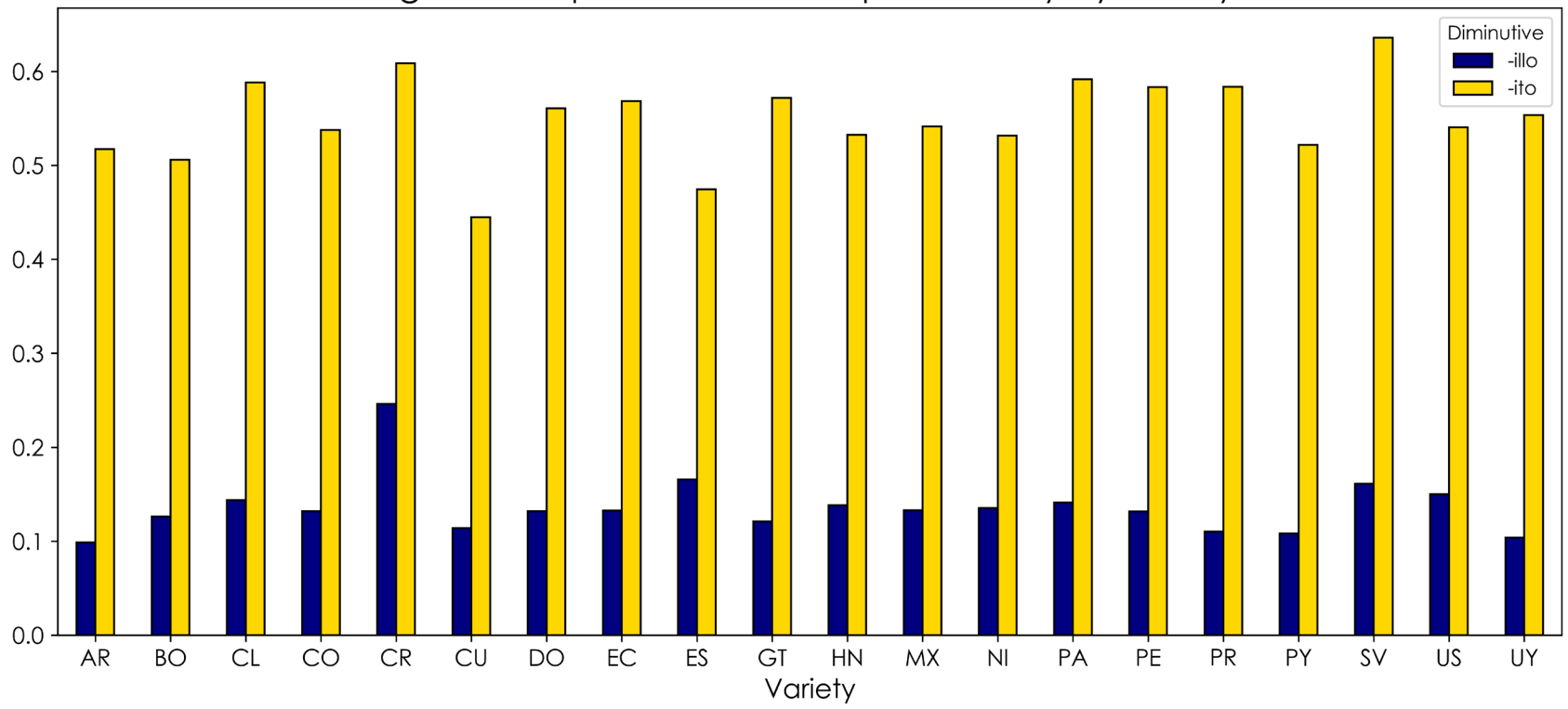


Figure 8. Hapax-conditioned productivity by variety.



CONCLUSIONS

CONCLUSIONS

- Across varieties: in terms of realized and expanding productivity, the numbers show that *-ito* is the bigger category and that it is overall attracting more new members.
- The category-conditioned degree of productivity doesn't appear to fully capture the difference between the two suffixes in terms of occasionalisms.
  - This might be due to the fact that *-illo* has notoriously fewer tokens. Alternatively, the hapax-conditioned degree of productivity shows a far more noticeable difference.

CONCLUSIONS

- Within varieties: the numbers by variety follow the same trend as those of the master data frame. A few differences are worth noting, however.
- Whereas *P* showed both suffixes in similar standing, the differences here are higher in favor *-illo*, particularly in countries such as Peru where it almost doubles *-ito*.
- For *P**, however, *-ito* remains the prevailing suffix in all varieties, although differences vary by country and might be worth looking at with inferential statistics.

ACKNOWLEDGEMENTS

- Dr. Na-Rae Han

- Dr. Jevon Heath

- Robert Henderson Language Media Center.
- Dr. Matthew Kanwit

- Ben Naismith

- LING 2340 Spring 2020 peers.

University of Pittsburgh

CONTACT

- Email: juanberrios@pitt.edu
- Website: pitt.edu/~jeb358
- GitHub: github.com/juanberrios

REFERENCES

Baayen, H. (2009). Corpus linguistics in morphology: morphological productivity. In A. Lüdeling & M. Kyto (Eds.), *Corpus Linguistics. An international handbook* (pp. 900-919). Berlin: Mouton De Gruyter.

Bradley, T. G., & Smith, J. (2011). The phonology-morphology interface in Judeo-Spanish diminutive formation: A lexical ordering and subcategorization approach. *Studies in Hispanic and Lusophone Linguistics, 4*(2), 247-300.

Castro, O. (1998). La formación del diminutivo en español y en gallego: Procesos morfológicos simples; implicaciones teóricas complejas. In A. Acosta Félix, Z. Estrada Fernández, M. Figueroa Estava, & G. López Cruz (Eds.), *IV Encuentro internacional de lingüística en el noroeste* (pp. 135–159). Sonora, México: Universidad de Sonora.

Colina, S. (2003). Diminutives in Spanish: A morpho-phonological account. *Southwest Journal of Linguistics, 22*(2), 45-88.

Eddington, D. (2002). Spanish diminutive formation without rules or constraints. *Linguistics, 40*(2), 395-419.

Eddington, D. (2017). Dialectal variation in Spanish diminutives: A performance model. *Studies in Hispanic and Lusophone Linguistics, 10*(1), 39.

Elordieta, G., & Carreira, M. M. (1996). An optimality theoretic analysis of Spanish diminutives. Papers from the Regional Meetings, Chicago Linguistic Society, *32*(1), 49-60.

REFERENCES

Haspelmath, M., & Sims, A. D. (2010). *Understanding morphology* (2nd ed.). London, UK: Hatchette.

Hualde, J. I. (2013). *Los sonidos del español*. Cambridge: Cambridge University Press.

Lipski, J. M. (1994). *Latin American Spanish*. London: Longmans.

Náñez Fernández, E. (2006). *El diminutivo. Historia y funciones en el español clásico y moderno*. Madrid: UAM Ediciones.