Investigating pronominal variation using Twitter data: advantages, challenges and lessons learned

AAAL 2021 VIRTUAL CONFERENCE

Camila Lívio
Bruna Sommer-Farias
Elisa Marchioro Stumpf
Larissa Goulart
Adriana Picoral

GIRL
Group of Inter-institutional Research in Languages

girlanguages.com

Introduction

The starting point:

- Digital and social media can be used as a rich source for the study of language patterns (Grieve et al., 2019; McCulloch, 2019)

Our goal:

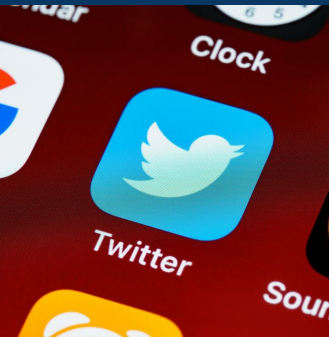
- Address challenges of working with internet language, text mining, and the analysis of linguistic variation using mixed-effect logistic regression modeling with sum contrasts in R
- Illustrate the use of second person subject pronouns in Brazilian Portuguese, including orthographic variations motivated by phonology, and other factors originated from computer-mediated communication

Data in Language Variation and Change (LVC) Studies

Sociolinguistic Interviews	Social Media
1. Historically used in Language Variation and Change studies (Labov, 2006) 2. Data from face-to-face interviews and naturally occurring speech 3. Time-consuming work to prepare data for analysis 4. A large number of interviews is required to extract a moderate amount of morphosyntactic tokens 5. Sample problem: friend of a friend (Millroy & Gordon, 2003) 6. Detailed demographic information is available	1. Not as explored (for exception see Grieve et al., 2019) 2. Data scraped in large amounts from social media (usually using APIs) 3. Data preparation is less time consuming 4. Possibility of extraction of large amounts of morphosyntactic tokens 5. Random sampling from public posts/tweets 6. Not all demographic information is retrievable

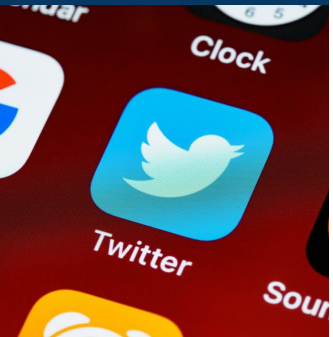
Both sources comprise naturalistic data (but observer's paradox might be a factor, see Tagliamonte, 2012)
Coding process (for the different factor groups) is time consuming
Focus on social dimensions in the analysis is essential

Why Tweets?



- Large volume of language data; and geographic metadata (Eisenstein, 2018; Grieve et al, 2018)
- Language variation can be studied more broadly, since “social media language is more colloquial and contains more linguistic variation” (Nguyen et al, 2016, p. 538)
- The study of informal writing has been neglected (McCulloch, 2019)
    - Twitter represents an informal written variety used when the speaker is not being monitored, akin to the type of data sought by sociolinguists

Why Tweets?



Social media data can be harder to automatically process due to the informal nature of texts (Nguyen et al, 2016)

bagulho

porque

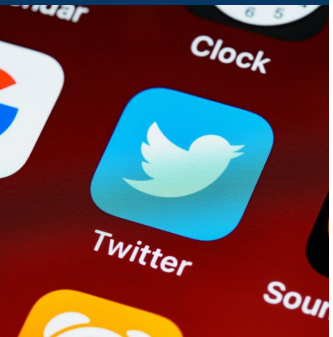
Se tu é rico(a) e usa essa porra de bgl caro é pq alguém em algum momento produziu, então quer dizer q tu se acha mlr q os outros pq tu DEPENDE (pra ser melhor) do trabalhador q com o suor produziu essa merda q tu ta consumindo filha da puta.

que

melhor

If you're rich and use this fucking expensive shit it's because someone, at some moment, made it, then it means that you think you're better than the others because you rely (to be better) on workers that, with their blood, sweat and tears, made this shit that you are using, asshole.

Why Tweets?



Tweets contain variation:

(1a) Claro mandei sim, **nós** estamos te mandando outro
*I mailed it, **we** are mailing you another one*

(1b) queria namorar uma cacheada p **nois** dividir os creme
*I wanted to date a curly so **we** can share hair products*

(2) o problema é que **a gente** conhece tudo eles pq **a gnt** faz pouca merda
*the problem is that **we** know them all because **we** have done some shit*

para

lack of number agreement

porque

Challenges to consider

- While English has been broadly researched, there are few studies on other languages, such as Portuguese
- “Representativeness is a major concern with social media data” (Eisenstein, 2018, p. 370)
    - language on Twitter is but one variety
    - there is little information about Brazilian Twitter users
- Most automatic annotation tools are not available for processing of social media language (other than English)

Challenges of using Twitter data

Orthography (or why automatic annotators can't do this)

(1) só que **nos** vivemos ataques das gringas por mais de um mês (included)
*it's that **we** under attack from abroad for over a month*

(2) vou ai qualquer hora pra **nos** toma uma gelada. (included)

*I'm dropping by sometime so that **we** can drink a beer*

(3) infelizmente esse direito **nos** foi tomado (not included)
*Unfortunately this right was taken from **us***

(4) E tu **nos** meus sonhos ♥ (not included)
*And you **in** my dreams*

(5) Mas **nós** últimos dias não consigo nem me ajudar (not included)
*But **in** the last fews days I can't even help myself*

Case Study

Our variables: pronominal variation between *tu* vs *você*

General 2nd person singular forms: ***tu*** and ***você***

Additional forms: ***cê, ocê -*** growing tendency on using variations of ***você*** (Othero, 2013)

Sociopragmatic (relationship among the speakers) and **geographic** variation

(Gonçalves, 2008; Ponzo Peres, 2006; Loregian-Penkal & Menon, 2012; Almeida Ferrari, 2013; Guimaraes, De Araújo & Pereira, 2018; Scherre, Andrade & Catão, 2020)

Other Romance languages:

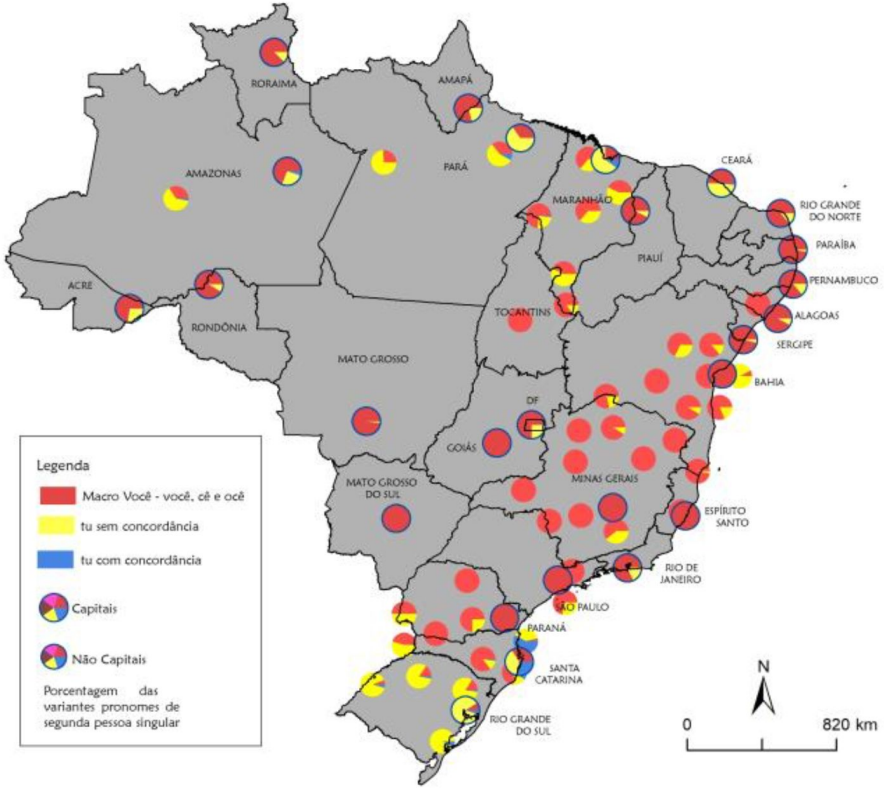
- Spanish (**both types of variation**, cf. Blas Arroyo, 2008; Moyna, Kluge & Simon, 2019);
- French (**more sociopragmatic**, cf. Brown & Gilman, 1960; Gardner-Chloros, 2007).

What we know about pronominal variation: *tu* vs *você*

Geographic variation

Scherre, Andrade & Catão (2020):

- Macro você - *você, ocê, cê:* generally used across Brazil
- *Tu* without agreement - *tu* ***vai*** (*3rd* p.sg.): mostly used in Rio Grande do Sul state, also present in some specific areas (e.g., Rio de Janeiro, Fortaleza)
- *Tu* with agreement - *tu* ***vais*** (*2nd* p.sg.): more common in Santa Catarina and Paraná area



Elaborado por Marta Scherre, Carolina Andrade e Rafael Catão
Fonte: Pesquisa Bibliográfica

Methods

Case Study Methods

- Data collected using the R package Rtweet (Kearney, 2019) between June 12 2020 and July 20 2020
- Geo-tagged to randomly collect tweets from major cities in Brazil in a 7-8 day window
    - Porto Alegre, Rio de Janeiro, Salvador, São Paulo
- The entire raw corpus consists of 21,445 tweets
- 5,002 tweets have been coded by hand to date
    - Approx. 40% of tweets retained for analysis (i.e., 60% of tweets were eliminated)

Why coded by hand?

- Tweets Excluded:
  - Retweets (automatically)
  - Manually:
    - Formulaic expressions with pronouns (é nóis)
    - Music lyrics
- Retained tweets were hand-coded:
  - Phonology (phonemic orthography)
    - Categories such as onomatopoeia, interjection, character extension, omission, etc.
  - Slang and abbreviations
    - generational and by community
  - Internet language use, Subject-Verb Order, type of reference, etc.

Quantitative Analysis

- Variationist Sociolinguistics
  - a pioneer in quantitative methods for linguistic analysis (Tagliamonte, 2016)
- Naturalistic data is often unevenly distributed
- ANOVA and linear regression in LVC (before 1970s)
- The switch to multivariate logistic regression analysis (1970s)

Logistic Regression: What is it and why to use it?

Multivariate logistic regression with sum contrasts outputs:

- the community probability of realizing a variable
- the distribution of the variable in a corpus and
- the effect of factors that favor (or not) that variable

"The goal of this method is to establish whether what is said (or not said) in what contexts and with what type and frequency of co-occurrence (in terms of phonological or syntactic environments)" (Picoral, 2020; Sankoff, 1982)

Results

Results Second Person

Tu vs. Você

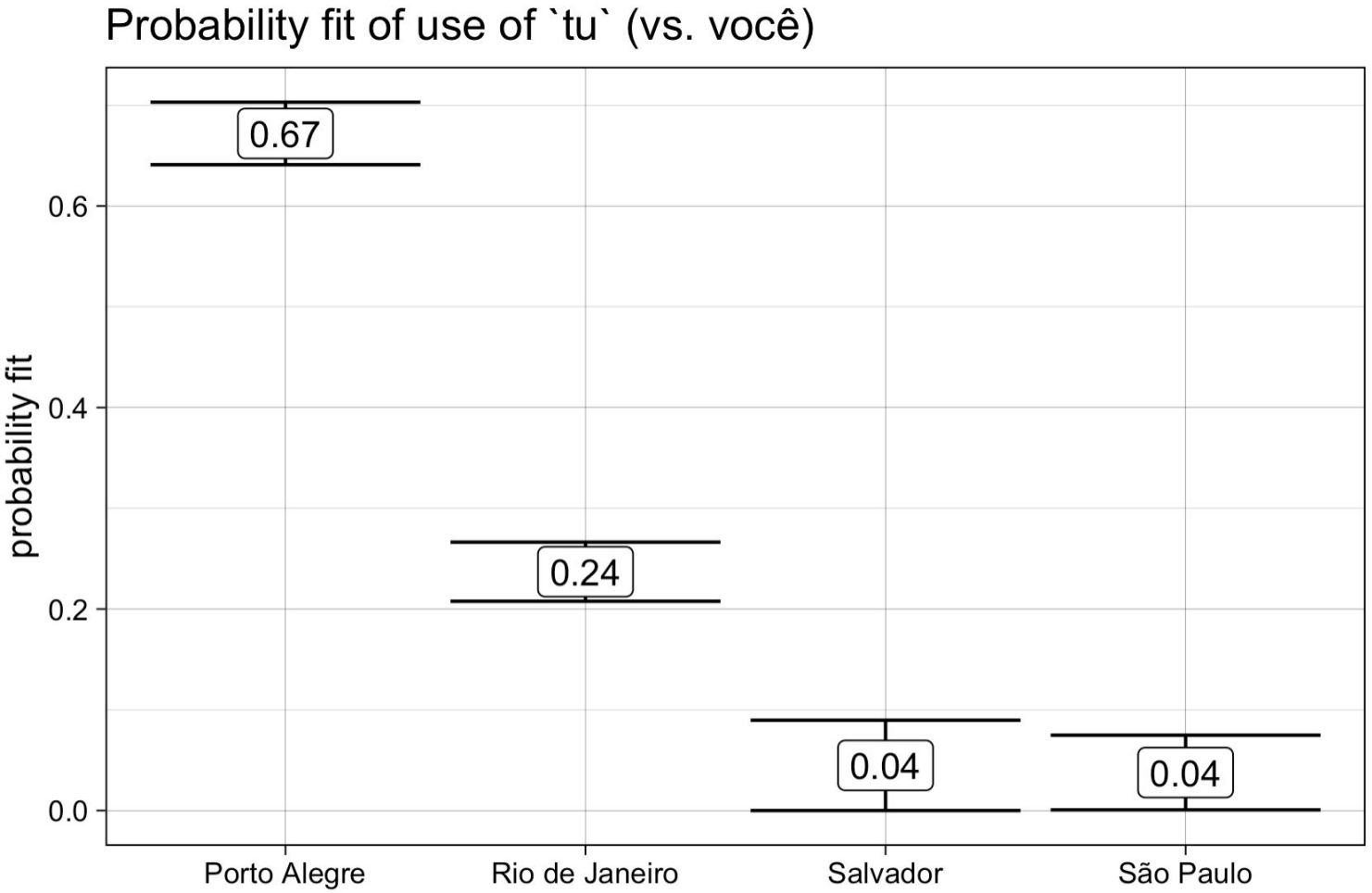
Table 1. Logistic regression of the factors conditioning “tu” for second person subject pronoun in Brazilian Portuguese tweets

Input: 0.11

Factor group	levels	n	proportion	logodds	weight	range
Location (p < .001)	Porto Alegre	621	66.02	2.34	0.91	
	Rio de Janeiro	673	22.29	0.37	0.59	
	Salvador	351	7.98	-1.19	0.23	
	São Paulo	410	4.88	-1.53	0.18	73
Subject Verb Order (p < .001)	subject-verb	1941	29.98	0.95	0.72	
	omitted verb	80	25.00	-0.04	0.49	
	verb-subject	34	17.65	-0.91	0.29	43
Internet Language Use (p < .05)	yes	977	21.70	0.17	0.54	
	no	1078	36.73	-0.17	0.46	08

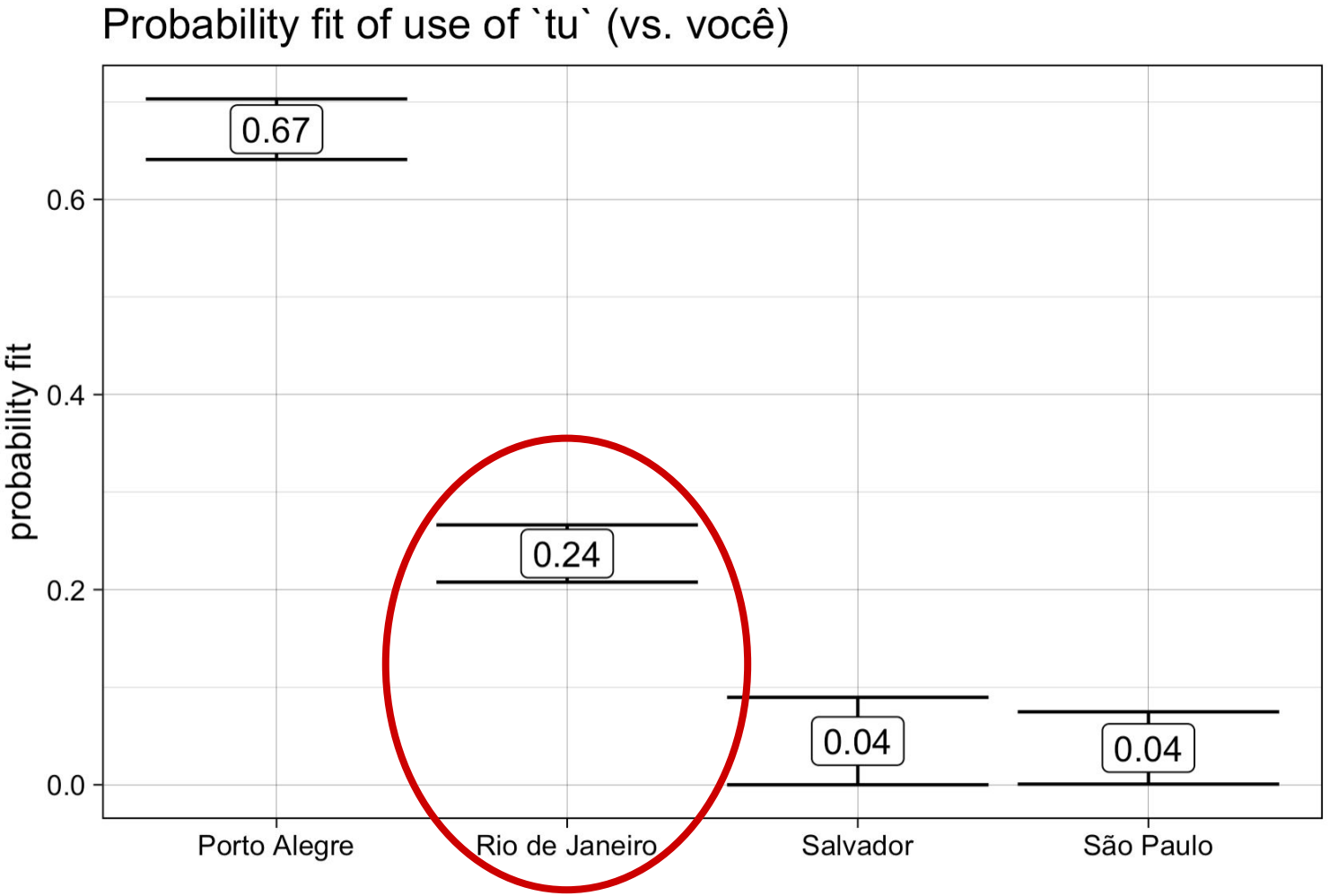
Results Second Person

Tu vs. Você

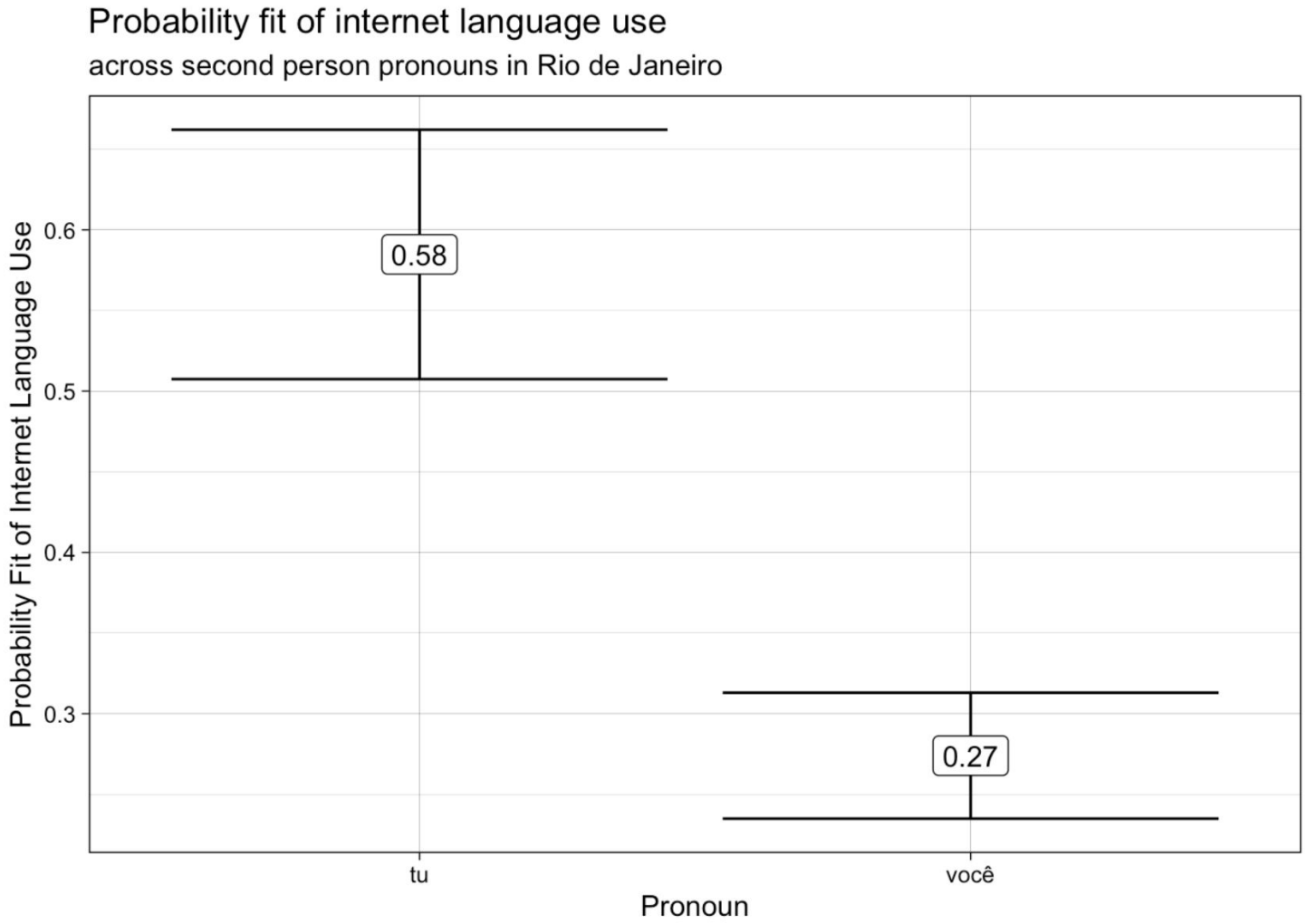


Results Second Person

Tu vs. Você



Internet Use in Second Person



Discussion

- Logistic regression shows the distribution of “tu” use across Twitter in Brazil, along with other factors that condition “tu”
  - Regional variation patterns on Twitter (see Centanin Bertho et al. 2021 for more details) corroborated previous (and more traditional) studies (see Scherre, Andrade & Catão, 2020)
- Use of internet-specific language:
  - favors the use of “tu”
  - indicates more informal context
  - is unique to this type of data (i.e., tweets)

Advantages, challenges & lessons learned

Advantages:

- Informal writing (natural language) that displays variation (see for example Grieve et al., 2019)
- Huge amounts of data publicly available
- Ability to "map" networks through linguistic practices/uses (McCulloch, 2019)

Challenges:

- Filtering the data (to exclude re-tweets and music lyrics) shrinks the corpus size, but filtering makes the corpus more representative
- Non-standard orthography requires hand coding

Lessons learned:

- Importance of agreeing on a coding scheme (this time consuming process does not go away)
- Coding by hand with multiple coders create a variety of code variations (solution: coding interface)

Implications for the field

- Normalize bigger research groups!
- Include more interdisciplinarity to account for interpreting variables and implementing research tools (and to account for the complexity of the data)
- Be part of open science (search our corpus at twitter-corpus.girlanguages.com)
- Reflect on the ethical considerations of social media data:
  - no need for IRB approval
  - use only publicly available posts
  - ensure user anonymity
  - report aggregate data, and when using examples, ensure no identifiable information is included

Our Research Group

girlanguages.com

Adriana Picoral
University of Arizona

Bruna Sommer-Farias
Michigan State University

Camila Lívio
University of Georgia

Elisa Marchioro Stumpf
Pelotas Federal University

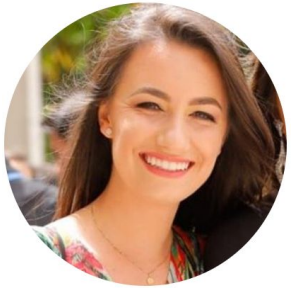
Isabella Calafate de Barros
University of Arizona

Larissa Goulart
Northern Arizona University

Laura Fontana Soares
University of Arizona

Mariana Centanin Bertho
University of Arizona

Marina Carcamo Garcia
University of Arizona



Marine Laísa Matte
Rio Grande do Sul Federal University

Investigating pronominal variation using Twitter data: advantages, challenges and lessons learned

AAAL 2021 VIRTUAL CONFERENCE

GIRL
Group of Inter-institutional Research in Languages

girlanguages.com

Camila Lívio camila.emidio25@uga.edu
Bruna Sommer-Farias fariasbr@msu.edu
Elisa Marchioro Stumpf elisa.stumpf@ufpel.edu.br
Larissa Goulart lg845@nau.edu
Adriana Picoral adrianaps@arizona.edu

Selected References

Eisenstein, J. E. (2018). Identifying Regional Dialects in On-Line Social Media. In C. Boberg, J. Nerbonne, & D. Watt (Eds.), *The Handbook of Dialectology* (pp. 368–383). Wiley.

Feagin, C. (2002). Entering the Community: Fieldwork. In J.K. Chamber, P. Trudgill, and N. Schilling-Estes (eds). The Handbook of Language Variation and Change. Oxford, Blackwell.

Grieve, J., Montgomery, C., Nini, A., Murakami, A., & Guo, D. (2019). Mapping lexical dialect variation in British English using Twitter. *Frontiers in Artificial Intelligence*, *2*, 11.

Grieve, J., Nini, A., & Guo, D. (2018). Mapping Lexical Innovation on American Social Media. *Journal of English Linguistics*, *46*(4), 293–319. https://doi.org/10.1177/0075424218793191

Kearney, M. W. (2019). rtweet: Collecting and analyzing Twitter data. Journal of Open Source Software, 4(42), 1829.

Kendall, T. (2008). On the History and Future of Sociolinguistic Data. *LANGUAGE AND LINGUISTICS COMPASS*, *2*, 332.

Labov, W. (2006). *The Social Stratification of English in New York City: Vol. 2nd ed*. Cambridge University Press.

McCulloch, G. (2020). *Because internet: Understanding the new rules of language*. Riverhead Books.

Nguyen, D., Doğruöz, A. S., Rosé, C. P., & de Jong, F. (2017). Computational sociolinguistics: A survey. *Computational linguistics*, *42*(3), 537-593.

Picoral, Adriana. (2020, December 20). Quantitative Language Data Analysis in R: Regression and Contrasts [Blog post]. Adriana Picoral. Retrieved from https://picoral.github.io/

Tagliamonte, S. (2016). *Making waves : the story of variationist sociolinguistics*. Wiley Blackwell.