Parallels between spoken and CMC language: Do tweets reflect spoken language choices?

Adriana Picoral — University of Arizona

Elisa Marchioro Stumpf — Universidade Federal do Rio Grande do Sul

Larissa Goulart — Northern Arizona University

Isabella Calafate de Barros — University of Arizona

Bruna Sommer-Farias — Michigan State University

Marine LaísaMatte — Universidade Federal do Rio Grande do Sul

Marina Cárcamo García — University of Arizona

Mariana Centanin Bertho — University of Arizona

Introduction

- Language Variation and Change (LVC)
- Collection of large-scale, random sampling of spoken language production from different communities
- CMC data (see for example De Decker, Vandekerckhove, and Sandra, 2016, for a study on Flemish and Dutch; and Willis, 2020, on Welsh)
- Pronominal variation in Brazilian Portuguese has been extensively studied in spoken language

Pronominal variation in Brazilian Portuguese

- Innovative pronouns have been introduced to Brazilian Portuguese through **grammaticalization of nouns and other pronouns**, with the function of **individualizing or disambiguating the referent** (Vianna, & dos Santos Lopes, 2012)
- Choice of **Subject pronouns** in modern Brazilian Portuguese is determined by **linguistic** (e.g., verb tense) and **extra-linguistic factors** (e.g., location)
- **Two variables:**
  - First person (we): *nós* vs. *a gente*
  - Second person singular (you): *tu* vs. *você*

Our Corpus

- Tweets randomly sampled by location (multistage sampling) between June 20 and July 10, 2020
- Geolocation codes for seven major metropolitan areas in Brazil
- We assume that our corpus represents a younger (mid 20s on average) and whiter (50%) population compared to the general demographics in Brazil (Huang et al., 2020)

Our Corpus

	1st person plural	2nd person singular
Florianópolis	355	144
João Pessoa	419	247
Porto Alegre	215	606
Recife	397	236
Rio de Janeiro	554	634
Salvador	118	409
São Paulo	420	477
TOTAL	2,478	2,753

Table 1: Token count for first person plural and second person singular tokens across regions in Brazil.

N=1,679 for first person

N=1,999 for second person

Envelope of Variation

- Only contexts that (Tagliamonte, 2012)
  - allow for variation
  - are not categorically encoded with one of the variants
- Expressed subject pronouns only (and spelling variations)
- Greedy regex that retrieved other word types (e.g., prepositions and object pronouns), with hand coding of each tweet

Envelope of Variation

Tokens **included** in the analysis:

a. só que **nos** vivemos ataques das gringas por mais de um mês
only that **we** lived attacks from foreigners for more than a month
'It's just that we've lived under foreign attack for more than a month'

b. vou ai qualquer hora pra **nos** toma uma gelada
(I) go there any time for **us** drink a cold
'I'll be there anytime for us to drink a beer'

Tokens **eliminated** from the analysis:

a. infelizmente esse direito **nos** foi tomado
unfortunately this right **us** was taken
'unfortunately this right was taken from us'

b. E tu **nos** meus sonhos
And you **in+the** my dreams
'And you in my dreams'

c. Mas **nós** últimos dias não consigo nem me ajudar
But **in+the** last days not can not even me help
'But lately I can't help even myself'

Analysis

- Logistic regression (innovative pronouns are the reference for both variables)
- Reported here are probabilities (the estimates transformed from log odds to probabilities) by region
- Not reported here: factor weights (sum contrasts, with intercept being the general probability for the reference)

Results

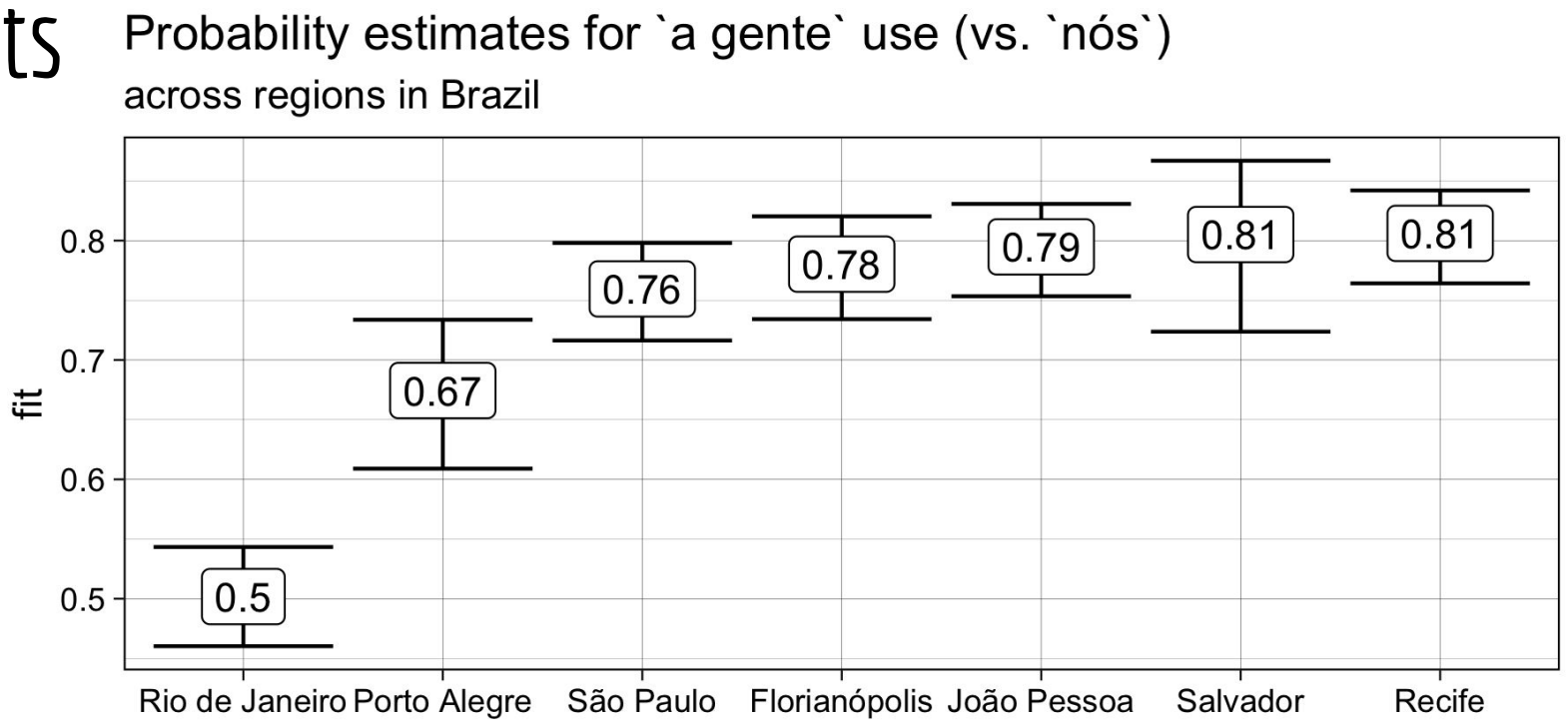

Figure 3: Probability estimates for *a gente* use for first person plural subject pronoun (vs. *nós*) across metropolitan regions in Brazil

Results - first person (we)

	Our Results	Silva Pacheco, 2018
João Pessoa	.79	.79
Florianópolis	.78	.72
Porto Alegre	.67	.69
Rio de Janeiro	.50	.76

Table 2: Comparison of our current results for first person plural pronoun variation, showing probability fit for *a gente* use across regions compared to average percentage use per region reported in Silva Pacheco, 2018

Results

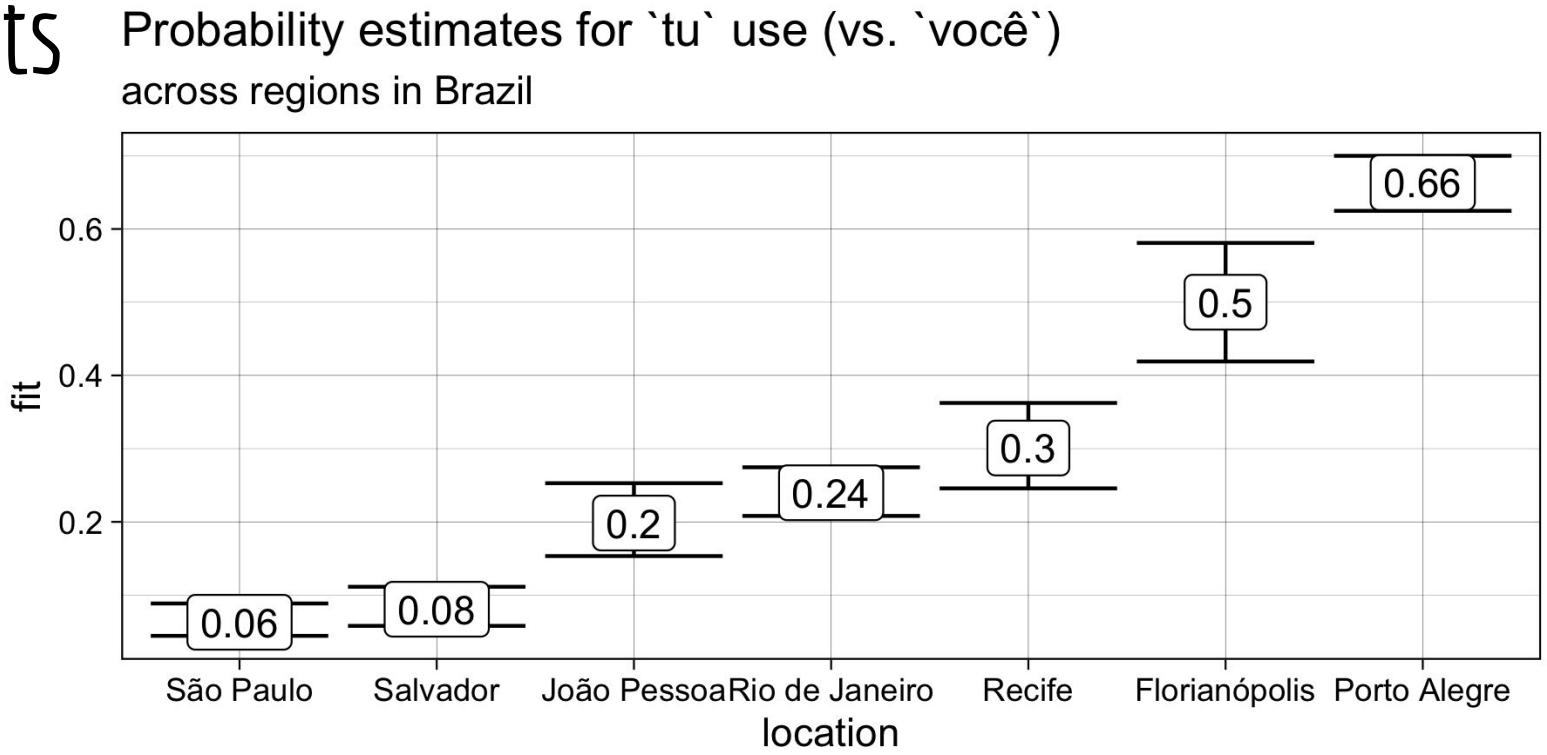

Figure 4: Probability estimates for *tu* use for second person singular subject pronoun (vs. *você*) across metropolitan regions in Brazil

Results - second person singular (you)

Table 3: Comparison of our current results for second person singular, showing probability fit for *tu* (you) use across regions compared to average percentage use per region according to previous literature (Scherre, Andrade, & Catão, 2020; Scherre, Yacovenco, & de Paiva, 2019).

	Our Results	Scherre et al. 2021; 2019
Porto Alegre	.66	.91
Florianópolis	.50	.76
Rio de Janeiro	.24	.27
Recife	.30	.14
João Pessoa	.20	.04
Salvador	.08	.01
São Paulo	.06	.00

Discussion + Next Steps

- Our results for subject pronoun variation in Brazilian Portuguese parallel previous findings
- The permanence of *nós* in Tweets in Rio de Janeiro is going to be addressed in another study
- Linguistic variables such as use of internet language, average length of words per tweet, pragmatic function are being coded and will be analyzed in a multivariate analysis in future studies
- Our hand-annotated corpus of tweets as ground truth for the automatic parsing of the data for subject pronoun variation study purposes

Thank you! Questions?