Talian corpus:

um corpus de dados escritos do talian (vêneto brasileiro)

Guilherme D. Garcia e Natália B. Guzzo

Université Laval

gdgarcia.ca • nataliaguzzo.github.io

Abralin em Cena 17

Junho, 2023

CRBLM

Centre for Research on Brain, Language and Music

Introdução

- **Objetivo:** Apresentar o *Talian corpus*, um corpus de dados escritos de vêneto brasileiro
  - Exemplo de aplicação: **metafonia**
- **Por quê:** embora saibamos que o processo é variável, pouco sabemos sobre como essa variação é estruturada

Vêneto brasileiro

- Imigrantes italianos se estabelecem no Brasil a partir de ~1850
- Diversas áreas são ocupadas, especialmente nas regiões sul e sudeste

- Origem da maioria dos imigrantes: norte da Itália (Vêneto)

Vêneto brasileiro

- Poucas comunidades de imigrantes falavam apenas uma língua (Frosi and Mioranza 2009)
- Contato entre variedades + contato reduzido com o português

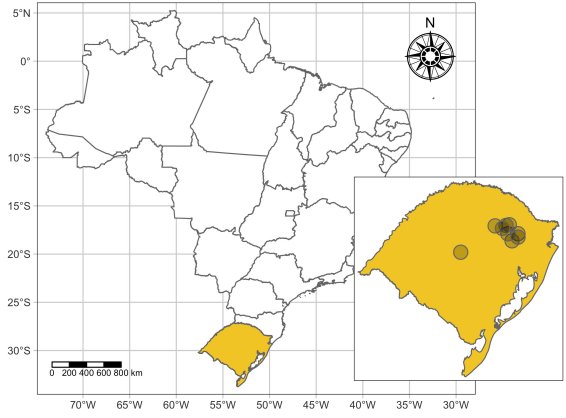
☞ Situação contribuiu para o desenvolvimento de um dialeto baseado em vêneto: VB ou **talian**)

Região de origem	**%**
Vêneto	54.0
Lombardia	33.0
Trentino-Alto Adige	7.0
Friuli Venezia-Giulia	4.5
Outras	1.5

Dados adaptados de Frosi and Mioranza (2009)

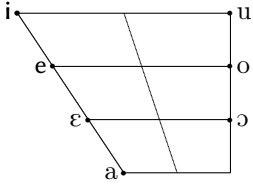
Vêneto brasileiro

- No Rio Grande do Sul, imigrantes receberam terra em áreas onde não havia virtualmente nenhum outro grupo populacional, o que contribuiu para o desenvolvimento de uma variedade de base vêneta

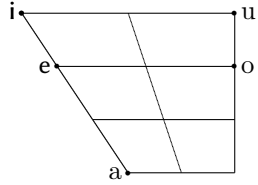


Vêneto brasileiro

- VB: bastante similar ao vêneto central (Frosi and Mioranza 1983; Belloni 2009; Guzzo 2022)
  - um dialeto de vêneto falado na Itália (e.g., Pádua, Verona)
- Ambos exibem uma janela trissilábica para o acento
- Ambos compartilham uma grande proporção do léxico



Posição tônica



Posição átona (pretônica)

Metafonia em vêneto (brasileiro, central)

- Em vêneto, /e, o/ acentuados elevam variavelmente para [i, u] quando seguidos por /i/ átono (Zamboni 1974; Walker 2005, 2010; Belloni 2009)
- O gatilho é geralmente um morfema separado
  - (marcador de plural ou flexão de 2PS)

- O alvo são **todas** as posições acentuadas (exemplos do talian)

1. **Metafonia com acento paroxítono:**

ˈov-i ~ ˈuv-i	‘ovo.PL’
ˈpes-i ~ ˈpis-i	‘peixe.PL’
ˈbev-i ~ ˈbiv-i	‘beber.2PS’
ˈkor-i ~ ˈkur-i	‘correr.2PS’

Metafonia em vêneto (brasileiro, central)

2. **Metafonia com acento proparoxítono:**

- A vogal átona na penúltima sílaba também eleva

  ˈzovɐ̯n-i ~ ˈzuvɪ̯n-i — ‘jovem.PL’

- Também pode haver elevação sem alvo em posição tônica

  ˈɔmen-i ~ ˈɔmin-i — ‘homem.PL’

3. **Metafonia com acento final:**

   faˈzo-i ~ faˈzu-i — ‘feijão.PL’ (sing. /fazol/)

   niˈso-i ~ niˈsu-i — ‘lençol.PL’ (sing. /nisol/)

Esta apresentação

- Examinamos metafonia a partir de um corpus de dados escritos:
  `the Talian Corpus` (Garcia and Guzzo 2021)

Por que dados escritos?

1. O talian não possui **ortografia nem gramática padronizadas**, então a variação na escrita *pode* refletir ao menos em parte a variação da língua falada pelos autores
2. Apesar de não possuir ortografia oficial, os autores são consistentes em seus textos, e o mapeamento grafema-fonema é relativamente constante (e.g., letra *u* = [u], letra *o* = [o, ɔ])

Métodos

Talian corpus

- Pouco material digitalizado
- OCR[1] usando Tesseract (treinado com dados do italiano padrão) (Smith 2007)
- **Materiais:**
  - Excertos de livros e artigos de jornais
  - Jornais: *Correio Riograndense* (fundado em 1909); *O Florense* (fundado em 1986) também faz parte do corpus, mas seus artigos podem ser acessados online (escaneados, não digitalizados)
  - Todos os materiais, até onde nos consta, são de autores do sul do Brasil, muitos dos quais da Região de Colonização Italiana do RS

[1] Optical Character Recognition.

Métodos

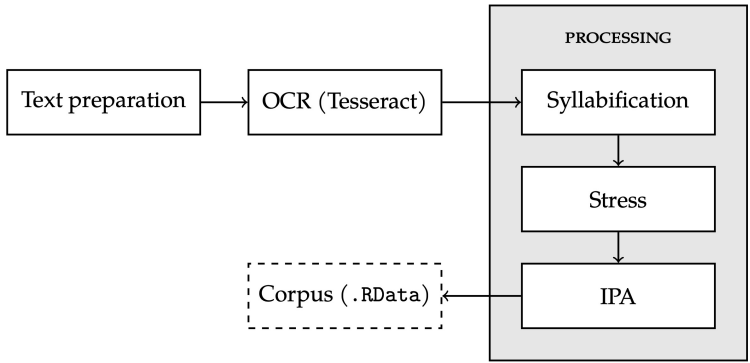
Talian corpus

1. Preparação de dados
   - >1 artigo por página
   - figuras
   - múltiplas colunas
   - quebras de linha
   - impressão frequentemente fraca (excertos de livros)
2. OCR
   - checagem textual
   - correções gerais
3. Compilação do corpus
   - R (R Core Team 2022)

Exemplo de artigos de jornal

Métodos

Talian corpus – disponível no Open Science Framework



Métodos

Talian corpus – disponível no Open Science Framework

- Formato: arquivo `RData` com dados em estilo *tidy data* (Wickham 2014)
- Tamanho: **237.774 palavras**

> ☞ Disponível em `nataliaguzzo.github.io/talian`

Métodos

Talian corpus

- Atualmente, 25 variáveis:

line	logFreq	nSyl	v_3	onset_1
sentence	author	syl_3	coda_3	v_1
wd	title	syl_2	onset_2	coda_1
sLength	year	syl_1	v_2	stressed_V
freq	IPA	onset_3	coda_2	stress

Métodos

Análise sobre metafonia

- Extração de todas as palavras que poderiam sofrer metafonia (**itens-alvo**):
  - palavras polissilábicas que terminam em /i/ átono
  - com vogal subjacente /e, o/ em posição tônica ($n = 3088$)
  - checagem manual de altura vocálica (vogais médias) em todas as palavras

- Palavras extraídas com script em R

- **Exemplos** (forma ortográfica):
  - senti 'sentir.2PS' (sem metafonia)
  - amori 'amor.PL'
  - dóveni 'jovem.PL'
  - curri 'correr.2PS' (com metafonia)
  - cagniti 'cachorro.DIM.PL'
  - fasui 'feijão.PL'

Métodos

Análise sobre metafonia

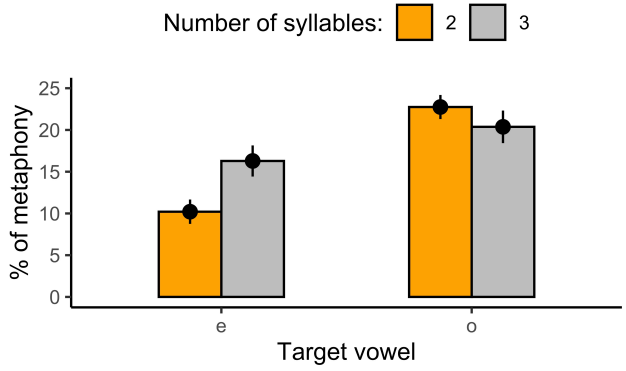
- Dada a distribuição de tokens na amostra, focamos em:

1. paroxítonas
2. de 3 e 2 sílabas

- Número de itens: 2.095 ($n = 490$ types)

- Itens codificados para aplicação/não aplicação de metafonia (variável resposta)
- Preditores:
  - **vogal alvo**, **número de sílabas**, **morfologia**, qualidade de onset, qualidade de coda

Resultados & análise

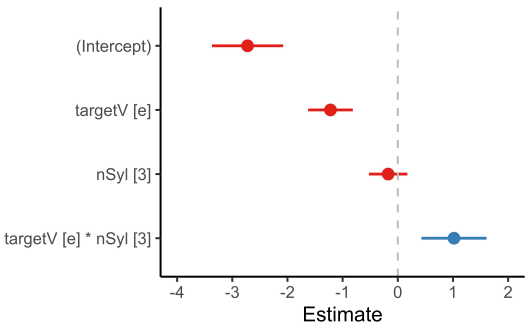


- **Assimetria:** mais metafonia com /o/ do que com /e/ (interação aparente)
- Preditores relacionados ao perfil fonotático não tiveram efeito claro

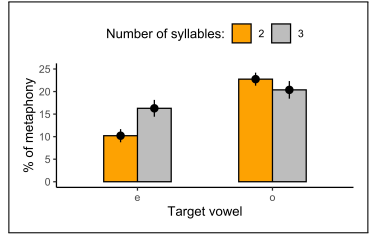
Resultados & análise

- Regressão logística hierárquica

```
metaphony ~ targetV * nSyl + (1 | author)
```

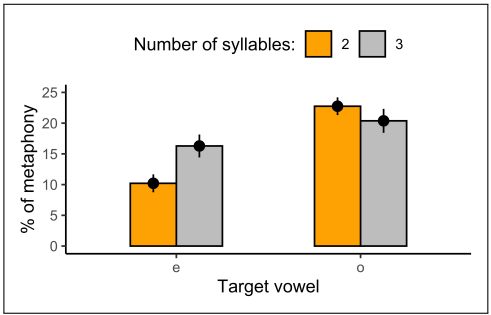


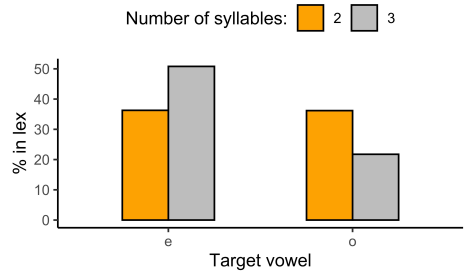
Coeficientes do modelo ($\hat{\beta}$), dados em log-odds



Discussão

- O que poderia estar causando a assimetria em questão?
- Um fator plausível: **estatística lexical**

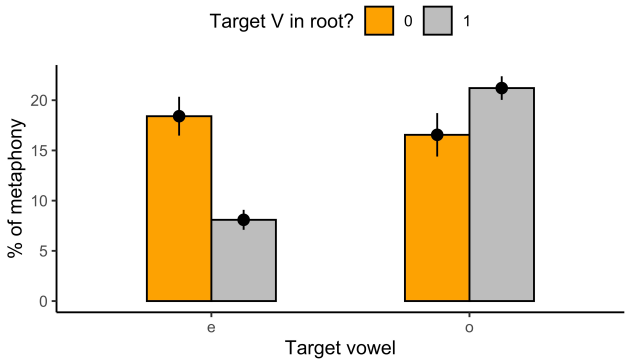




Corpus inteiro

Discussão

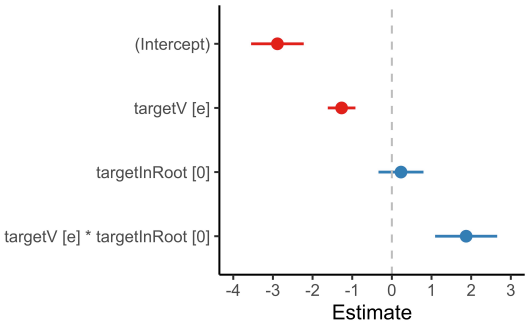
- Efeito morfológico potencial: novamente, um padrão assimétrico emerge

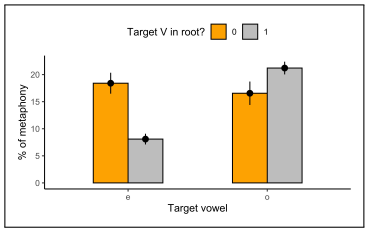


Discussão

- Regressão logística hierárquica

  `metaphony ~ targetV * targetInRoot + (1 | author)`





Model estimates ($\hat{\beta}$), given in log-odds

Discussão

	/**e**/	/**o**/
Corpus inteiro	mais /e/s em 3-σ	mais /o/s em 2-σ
Itens alvo	mais metafonia com /e/ em 3-σ	mais metafonia com /o/ em 2-σ
$\sqrt{\mathbf{V}}$	menos metafonia	mais metafonia

- Para consultar a formalização desses dados e padrões, ver Garcia and Guzzo (2023)

Considerações finais

- Análise da metafonia do talian usando um corpus escrito nos permite concluir que:
- Há uma assimetria entre /e/ e /o/ em palavras de 2- e 3-σ
- Padrões refletem o que vemos no corpus (léxico)
- Assimetria também é encontrada quando consideramos efeitos de morfologia

- Até o momento: dados escritos como *proxy* para a gramática do talian
- **Próximo passo:** coletar dados empíricos para avaliar nossas observações

Referências I

Belloni, S. (2009). *Grammatica veneta*. Esedra, Padova.

Frosi, V. M. and Mioranza, C. (1983). *Dialetos italianos: um perfil linguístico dos ítalo-brasileiros do nordeste do Rio Grande do Sul. [Italian dialects: a linguistic profile of the Italian-Brazilians in the northeast of Rio Grande do Sul]*. EDUCS, Caxias do Sul, Brazil.

Frosi, V. M. and Mioranza, C. (2009). *Imigração italiana no nordeste do Rio Grande do Sul*. EDUCS, Caxias do Sul, 2nd edition.

Garcia, G. D. and Guzzo, N. B. (2021). Talian corpus: a written corpus of Brazilian Veneto.

Garcia, G. D. and Guzzo, N. B. (2023). A corpus-based approach to map target vowel asymmetry in Brazilian Veneto metaphony. *To appear in Italian Journal of Linguistics*.

Guzzo, N. B. (2022). Brazilian Veneto (Talian). *Journal of the International Phonetic Association*, page 1–15.

Referências II

R Core Team (2022). *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna, Austria.

Smith, R. (2007). An overview of the tesseract OCR engine. In *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)*, volume 2, pages 629–633.

Walker, R. (2005). Weak triggers in vowel harmony. *Natural Language and Linguistic Theory*, 23:917–989.

Walker, R. (2010). Nonmyopic harmony and the nature of derivations. *Linguistic Inquiry*, 41:169–179.

Wickham, H. (2014). Tidy data. *Journal of Statistical Software*, 59(10):1–23.

Zamboni, A. (1974). *I dialetti del Veneto*. Pacini, Pisa.

Obrigad$^{o}_{a}$ • Gràssie

Agradecimentos:

- **Bolsistas:** Émilie Dubé, Alexandra Lancaster, Ray Marks, Fabian McCarthy, Lamia Oudni, Carolyn Rathgeber, e Jovia Wong (McGill); Jenna Gramlich (Ball State); Gabriel Frazer-Mckee (Laval)
- Dr. Valdemir Guzzo