FALE
FACULDADE
DE LETRAS

UFMG

Creating datasets for emergent contact languages preservation

Dalmo Buzato & Átila Vital
Faculty of Letters
Universidade Federal de Minas Gerais
Brazil

What is language contact?

- Language contact occurs when speakers of different languages interact with each other in communicative situations.

- Depending on social variables, such as the **intensity of the contact**, the **prestige position of the languages and speakers involved**, and the **need for a mutual means of communication**, a contact language may emerge.

Unstable existence of contact languages

- They become extinct when the contact situation between speakers of different languages ends (such as business situations, migrations, etc).

- In addition, this language usually suffers from low social prestige and is usually not taught in schools, with no other instruments of social stimulation (literature, media use, government use).

How to preserve these languages?

- Creating corpora for contact languages (Nagy, 2011; Mello, 2014; Adamou, 2016; Léglise and Alby, 2016)
  - **A very difficult task!**
  - **Creating treebanks (UD treebanks) is also being a strategy (Seddah et al., 2020; Braggaar and van der Goot, 2021).**

- The use of the web for language preservation and documentation
  - **Digital social networks (Facebook; Instagram; Tiktok; Twitter/X)**
  - **Digital newspapers**

Our dataset

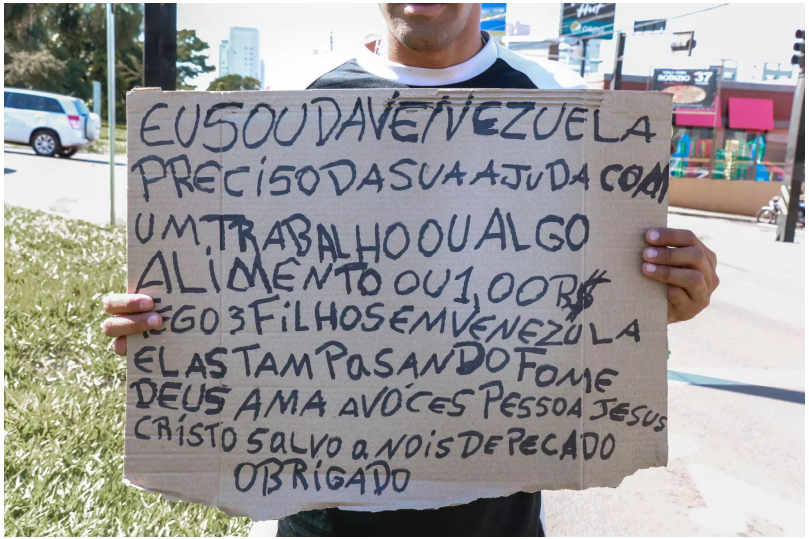
This study reports on the ongoing development of a dataset with **spoken** and **written data** produced by Venezuelan refugees in Brazil. The data was produced by indigenous refugees of the Warao ethnic group.

Warao migration to Brazil

- From Orinoco Delta
- +45.000 people in Venezuela
- Speakers of a homonymous native language with no known linguistic relatives
- Speakers of Spanish as L2
- They were not a people with nomadic characteristics before their growing status of subalternity.

Written signs

Written signs produced by the refugees, to ask the Brazilian population for help.



- **Mixed nature**: photographs collected from news websites (2018 - now) and during a fieldwork carried out in the city of Belo Horizonte (2022- now).
- Initial descriptions made by Buzato & Vital (2023) and Buzato (2023)
- Presence of diverse linguistic phenomena that go beyond code-switching

Written signs

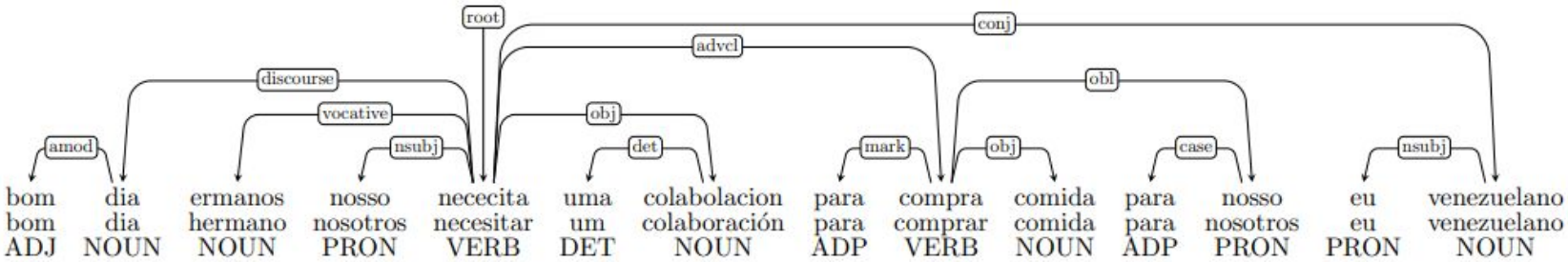
For annotating written signs, we use the Universal Dependencies (UD) framework

- **Transcription:** `bom dia ermanos nosso nececita uma colabolacion para compra comida para nosso eu venezuelano`

- Choice of UD (Nivre et al., 2016) is based on its **typological proposal** and its growing use for annotating non-Indo-European and **minority languages**.

Written signs

For annotating written signs, we use the Universal Dependencies (UD) framework



Example of how the transcription of the slide above has been annotated (morphosyntactically) according to UD guidelines

Spoken data

Main source: spontaneous speech recordings in Belo Horizonte (ongoing)
Pilot source: videos available in internet

Família Warao da Venezuela mantém tradições em meio à procura por sustento em BH

5 mil visualizações • há 2 anos

O TEMPO

Sessenta e três indígenas **Warao** compõem o grupo de venezuelanos que migram pelo **Brasil** há quatro anos. Com 34 crianças ...

"Socorro!", grita Povo Indígena Warao, da Venezuela, refugiado em BH/MG (300), no Brasil + de 7 mil

681 visualizações • há 11 meses

Frei Gilvander Luta pela Terra e por Direitos

Socorro!", gritam o Povo Indígena **Warao**, da Venezuela, refugiado em Belo Horizonte/MG (quase 300) e muitas cidades do **Brasil** ...

Spoken data

Spontaneous speech records from videos available in internet & fieldwork records in Belo Horizonte (currently ongoing)

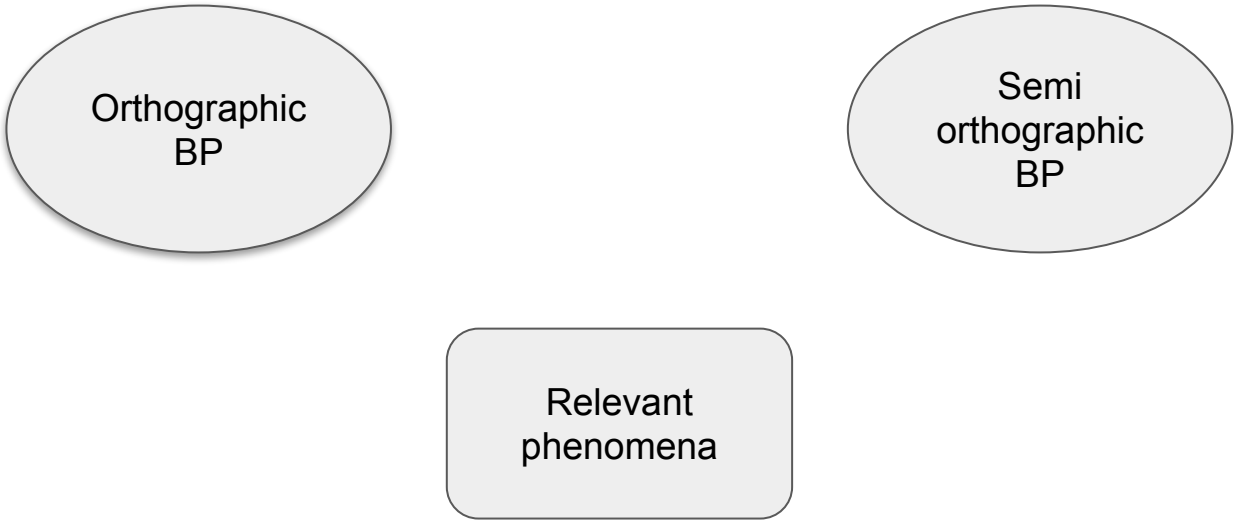
Example 2 (documentary_VAR)

VAR: lá / passava muita / dificuldade / por falta de / &m [/1] da medicamento // porque / muita [/1] muita criança // &he / muito / homem / mulher / vovó / &fa [/1] faleciam / porque / faltava de [/1] de medicamento lá // si / mas na [/1] na mi [/1] alimentação / não nos chega //

- Transcription and annotation adapted from the C-ORAL-BRASIL criteria (Raso and Mello, 2012)
- Package of information conveyed by the prosody (Izre'el et al., 2020)
- Semi-orthographic criteria capable of capturing cliticizations, apheretic forms, erasing of verbal morphology, new pronominal paradigms, disfluencies, and many others.

Spoken data

Criteria adaptation to represent relevant phaenomena in contact variety



Spoken data

Criteria adaptation to represent relevant phaenomena in contact variety

- The negation

  C-ORAL-BRASIL: não | nũ | n’ é não | **non | no**

Spoken data

Spontaneous speech records from videos available in internet & fieldwork records in Belo Horizonte (currently ongoing)

- The transcription criteria will be defined after a better contact with the texts
- Different grammaticalization and lexicalization phenomena
- Code-switching

Sociolinguistic profiling

@Title: documentary_VAR

@File: VAR

@Participants: VAR, John Vargas (male, unknown, unknown, Warao immigrant, participant, Venezuela)

@Date: unknown

@Place: Belo Horizonte (MG)

@Situation: documentary made by "Jornal o Tempo" about the Warao immigration @Topic: the life in Venezuela and the reasons why his family came to Brazil

@Source: YouTube

@Length: 39"

@Words: 64

@Transcriber: Átila Vital

@Comments: The audio has a music in a very low volume from the documentary

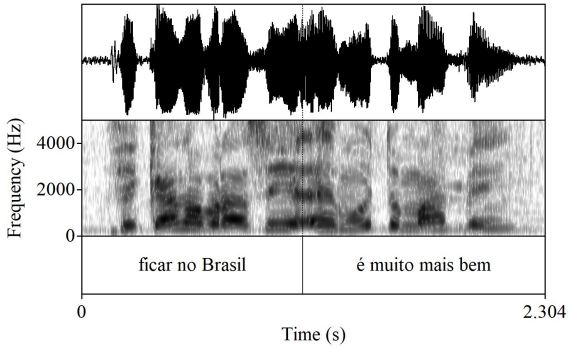
1) Forms originated by contact: at 10", VAR speaks "bobó", instead of "vovó" (grandmother). At 36", VAR speaks "possible", instead of "possível" (possible).

2) External noises: in some moments, there are sounds of children playing.

- Inspired by the C-ORAL-BRASIL model (Raso and Mello, 2012)

- The high acoustic quality is rare to be found in emergent language descriptions. Still, during the audio compilation, we will value high-quality recordings.

Potential linguistic phenomena



AAA: ficar no Brasil / é muito mais bem //

- Constant borrowings from Spanish and (Vernacular) Brazilian Portuguese;
- Recurring confusion between the use of the adjective related to Venezuela (Venezuelan) and the name of the country itself;
- Absence of copula use;
- Use of an accusative pronoun postposed to the verb, as in "ajuda me" ("help me"), a less frequent form in Brazilian Portuguese.

Current and future steps

- Our initial objective is to contain around **60 transcribed and annotated signs**, and **20 recordings of spontaneous speech**, totalling approximately 1,500 words. All of them will be transcribed, segmented and aligned;
- Investigate the use of some Warao refugees on digital social networks;
- Spontaneous speech records with refugees living in Belo Horizonte.

References

Evangelia Adamou. 2016. A corpus-driven approach to language contact: Endangered languages in a comparative perspective, volume 12. Walter de Gruyter GmbH & Co K

Anouck Braggaar and Rob van der Goot. 2021. Challenges in annotating and parsing spoken, code-switched, Frisian-Dutch data. In Proceedings of the Second Workshop on Domain Adaptation for NLP, pages 50–58

Dalmo Buzato. 2023. Universal Dependencies and Language Contact Annotation: Experience from Warao refugees signs in Brazil. In Proceedings of the 2nd Edition of the Universal Dependencies Brazilian Festival, pages 509–519.

Dalmo Buzato and Átila Vital. 2023. O contato linguístico em placas de refugiados venezuelanos em Belo Horizonte e região metropolitana: observações preliminares. In Anais do Congresso Nacional Universidade, EAD e Software Livre, volume 1

Shlomo Izre'el, Tommaso Raso, Alessandro Panunzi, and Heliana Mello. 2020. In search of basic units of spoken language. In Search of Basic Units of Spoken Language, pages 1–452

Maarten Janssen. 2016. Teitok: Text-faithful annotated corpora. In Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16), pages 4037–4043.

Isabelle Léglise and Sophie Alby. 2016. Plurilingual corpora and polylanguaging, where corpus linguistics meets contact linguistics. Sociolinguistic studies, 10(3):357–381.

Heliana Mello. 2014. What Corpus Linguistics can offer Contact Linguistics: the c-oral-brasil corpus experience. PAPIA: Revista Brasileira de Estudos do Contato Linguístico, pages 407–427

Naomi Nagy. 2011. A multilingual corpus to explore variation in language contact situations. RILA : Rassegna Italiana di Linguistica Applicata, pages 65–84.

Tommaso Raso and Heliana Mello. 2012. O Corpus C-ORAL-BRASIL. Editora UFMG, Belo Horizonte.

Djamé Seddah, Farah Essaidi, Amal Fethi, Matthieu Futeral, Benjamin Muller, Pedro Ortiz Suarez, Benoît Sagot, and Abhishek Srivastava. 2020. Building a user-generated content North-African Arabizi treebank: Tackling hell. In Proceedings of the 58th annual meeting of the Association for Computational Linguistics, pages 1139–1150.

FALE
FACULDADE
DE LETRAS

UFMG

Muito obrigado! Thank you!

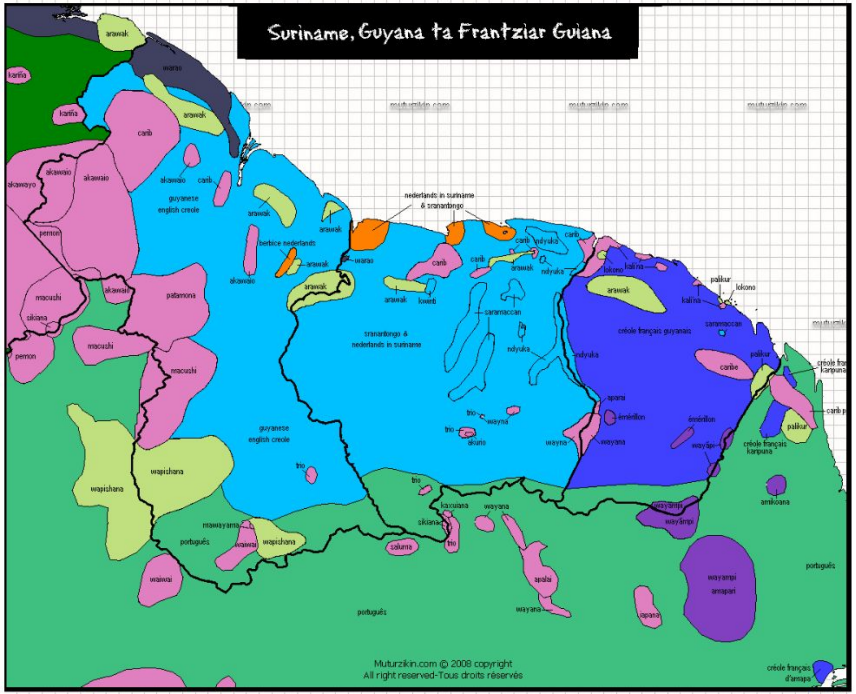
buzatodalmo@gmail.com
4tilavital@gmail.com

Third Workshop on Digital Humanities
and Natural Language Processing

Universidade de Santiago de Compostela

March 12, 2024

“Tenho duas mãos... e o sentimento do mundo”



- Document other emerging contact languages through the same protocols, using spoken and written data, mainly in low-resourced varieties in the global south.
- Examples:
  - **Suriname**: Sarnami-hindi (Surinamese Hindi), Aucano, Saramacano and Sranantongo
  - **Paraguay:** Jopara