FALE

FACULDADE
DE LETRAS

UFMG

Datasets for describing emerging contact languages: the summary of ongoing research into Warao-Spanish-Portuguese contact

Átila Vital & Dalmo Buzato

Faculty of Letters
Universidade Federal de Minas Gerais
Brazil

Outline

- Contact languages
- Warao situation
- Studying written signs
- Spoken data
- Principles
- Next steps

The unstable existence of contact languages

- Contact languages become extinct when the communicative 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 describe these languages?

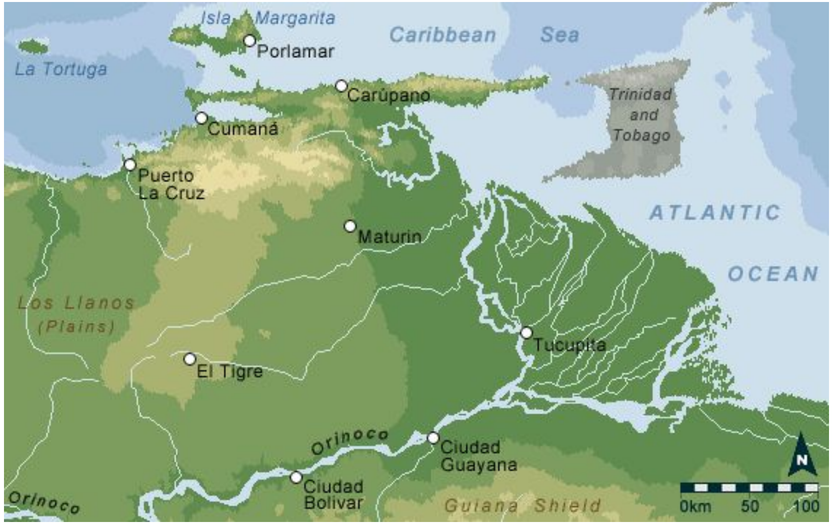
- Creating corpora for contact languages (Nagy, 2011; Mello, 2014; Adamou, 2016; Léglise and Alby, 2016)
  - A very difficult task!
  - Creating treebanks (UD treebanks) is also being a strategy (Seddah et al., 2020; Braggaar and van der Goot, 2021).

- The use of the web for language preservation and documentation (Cunha, 2020)
  - 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.

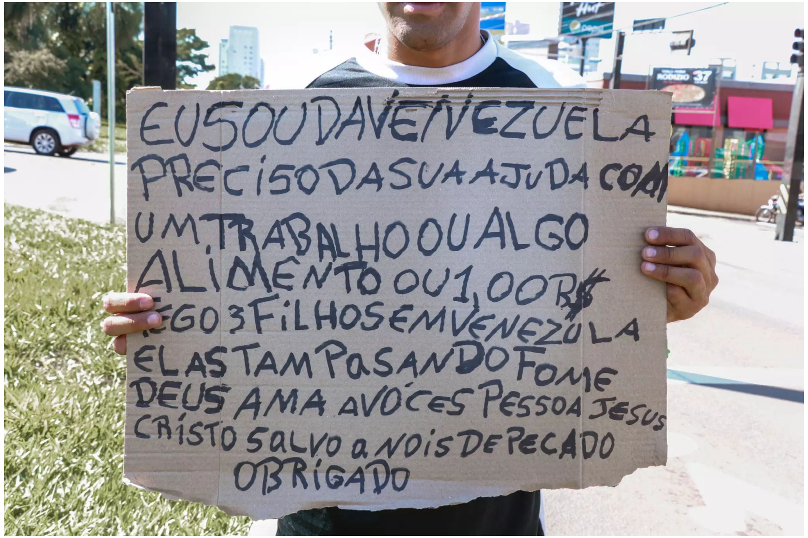
Warao migration to Brazil



- From Orinoco Delta
- +45.000 people in Venezuela
- Speakers of a homonymous native language with no known linguistic relatives (Romero-Figueroa, 1997)
- Some of them are speakers of Spanish as L2
- They were not a people with nomadic characteristics before their growing status of subalternity (Soneghetti, 2017)

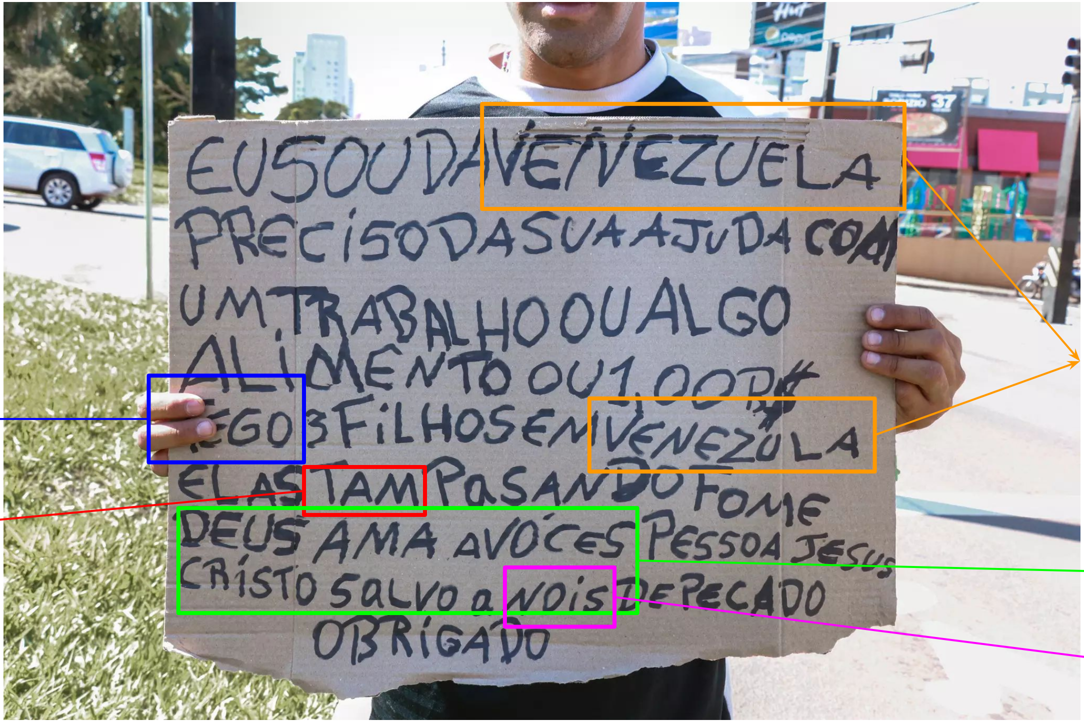
Written signs

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 Mesquita (2020), Buzato & Vital (2023, 2024) and Buzato (2023)
- Presence of diverse linguistic phenomena that go beyond code-switching

Written signs



tego ~ tengo

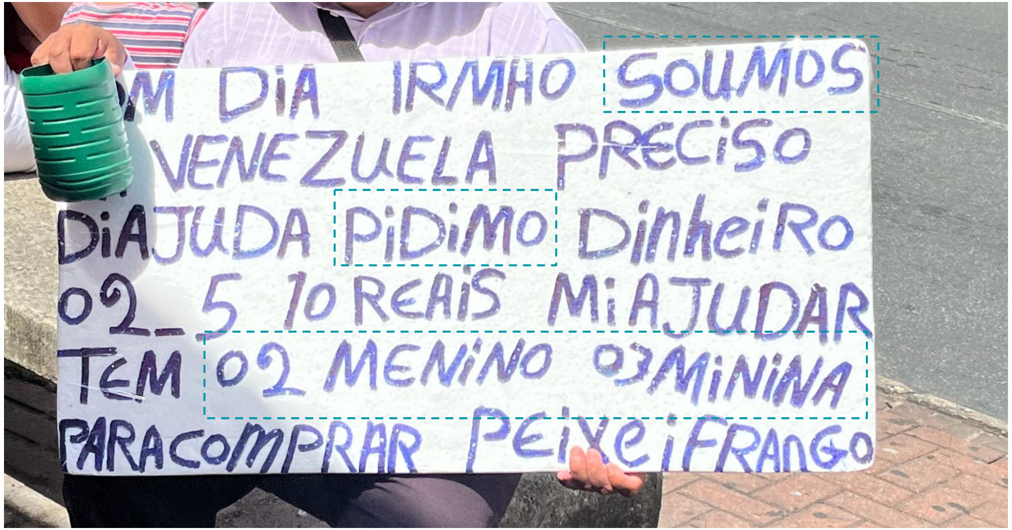
estão -> tão -> tam
estan -> tan -> tam

venezuela ~ venezula

[SN SV SP]

nois ~ nós

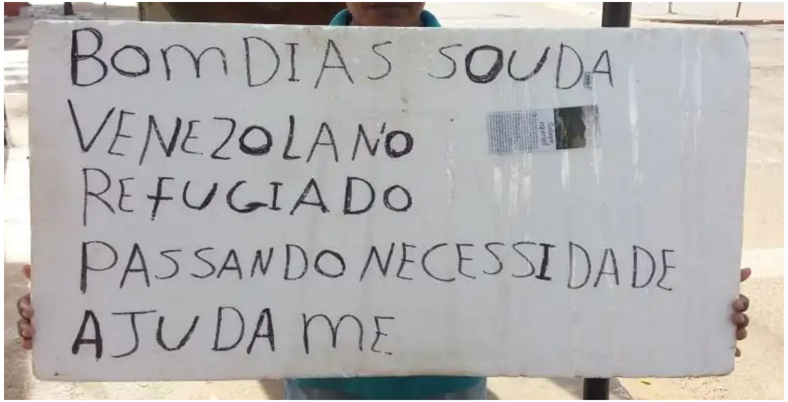
Similarities with PB



M DiA IRMHO SOUMOS
VENEZUELA PRECiSO
DiAJUDA PiDiMO DinheiRO
02 5 10 REAiS MiAJUDAR
TEM 02 MENiNO 03MiNiNA
PARACOMPRAR PEiXEiFRANGO

- Contact with informal spoken BP
- so[u]mos (epenthesis)
- p[i]dimo ~ p[e]dimo
  (vowel harmony)
  (drop of the plural morpheme -s)
- 02 menino 03 minina
  (drop of the plural morpheme -s)
  (vowel harmony em minina)

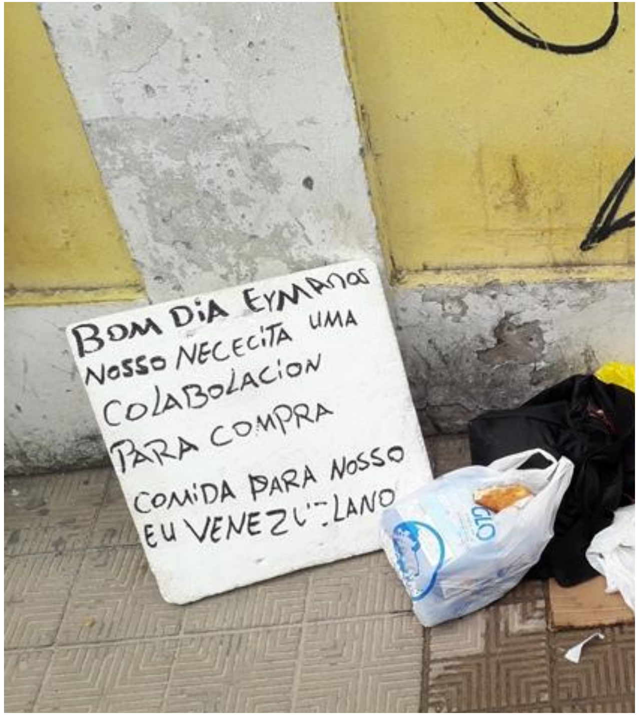
Differences with PB



The proximity between Spanish and Portuguese is a strategy used by refugees.

- bom dias
  (borrowing from Spanish *buenos días*)
- sou da venezolano
  (use of gentile instead of toponym in NP)
  (borrowing from Spanish *venezolano* (pt: venezuelano))
- ajuda me
  (post-verbal clitic instead of pre-verbal as in informal spoken BP)
  (borrowing from Spanish *ayúdame* and the productivity of the pronoun after the verb in Spanish in the imperative mood)

Differences with PB

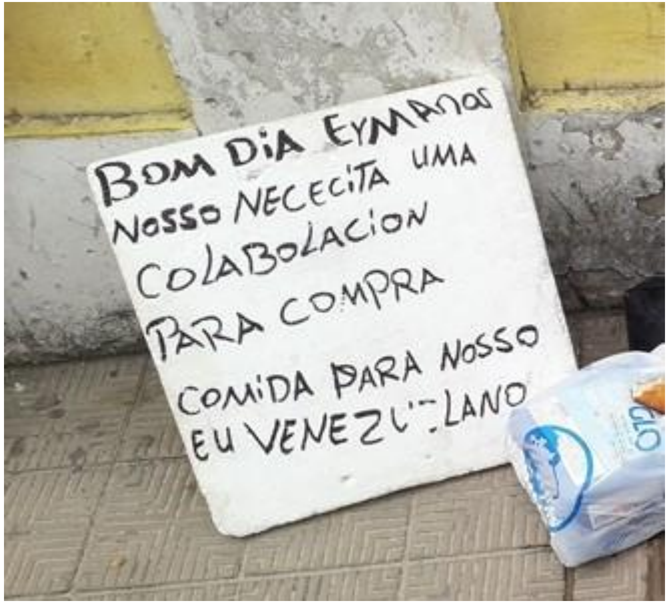


The proximity between Spanish and Portuguese is a strategy used by refugees.

- ermanos
  (borrowing from Spanish *hermanos*)
- colabolacion
  (borrowing from Spanish *colaboración*)
- compra ~ comprar
  (drop of the infinitive morpheme –r)
- nosso
  (truncation of *nosotros*, used as 4P 'nós', PB.)
- eu venezuelano
  (absence of copula verb)
  Romero-Figueroa (1997) describes the absence of the copula verb in Warao

Written signs

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

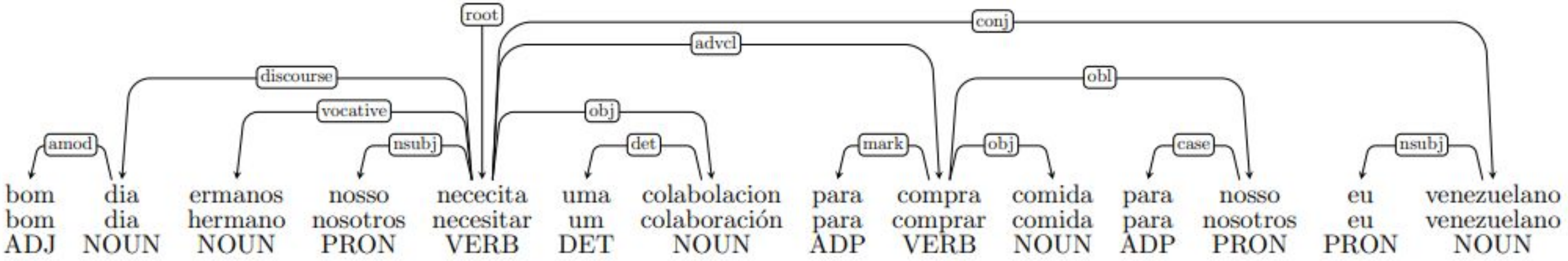


- **Transcription:** bom dia ermanos nosso nececita uma colabolacion para compra comida para nosso eu venezuelano

- 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

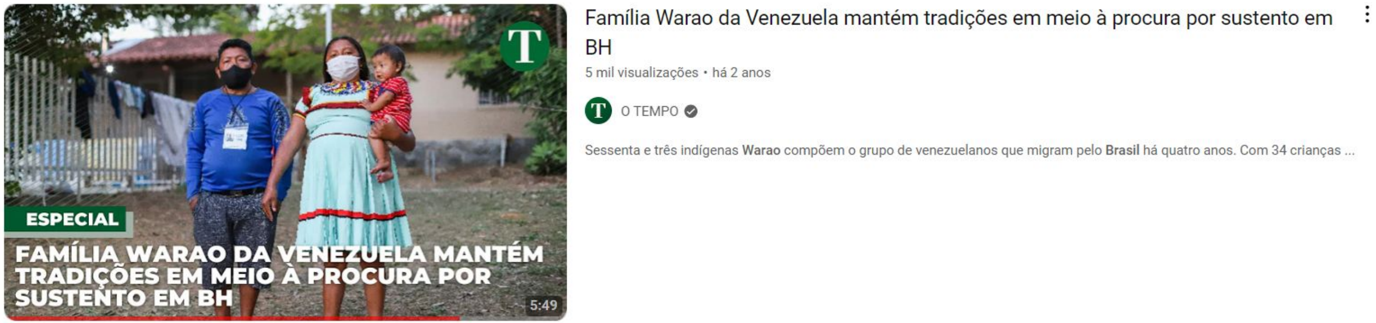


bom	dia	ermanos	nosso	nececita	uma	colabolacion	para	compra	comida	para	nosso	eu	venezuelano
bom	dia	hermano	nosotros	necesitar	um	colaboración	para	comprar	comida	para	nosotros	eu	venezuelano
ADJ	NOUN	NOUN	PRON	VERB	DET	NOUN	ADP	VERB	NOUN	ADP	PRON	PRON	NOUN

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

Spoken data

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



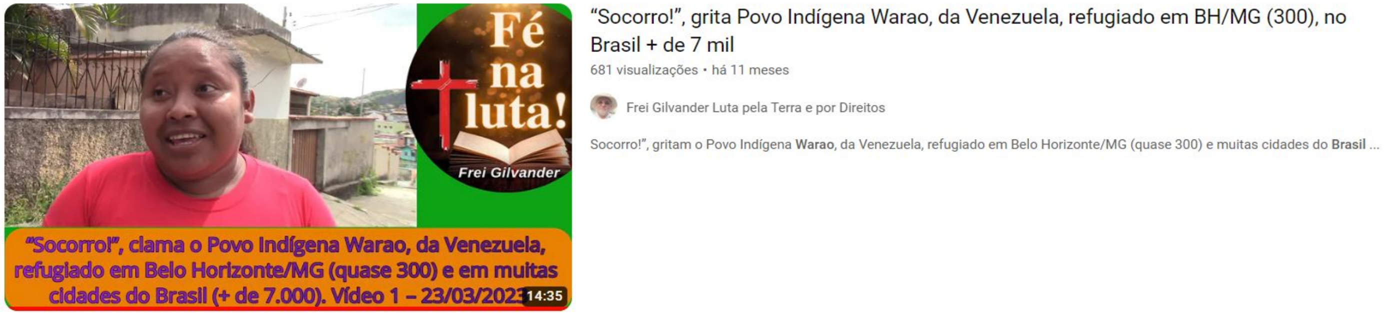
FAMÍLIA WARAO DA VENEZUELA MANTÉM TRADIÇÕES EM MEIO À PROCURA POR SUSTENTO EM BH

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!", clama o Povo Indígena Warao, da Venezuela, refugiado em Belo Horizonte/MG (quase 300) e em muitas cidades do Brasil (+ de 7.000). Vídeo 1 – 23/03/2023

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

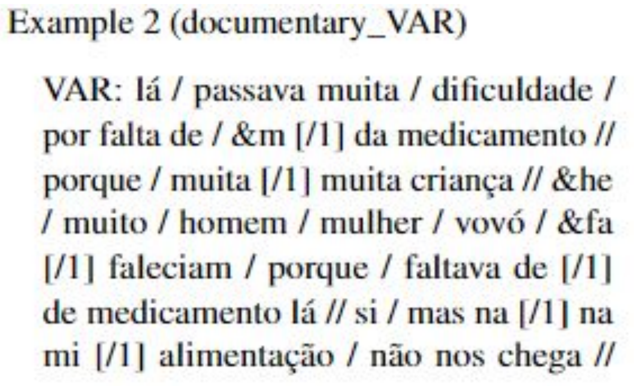
681 visualizações • há 11 meses

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 following 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

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

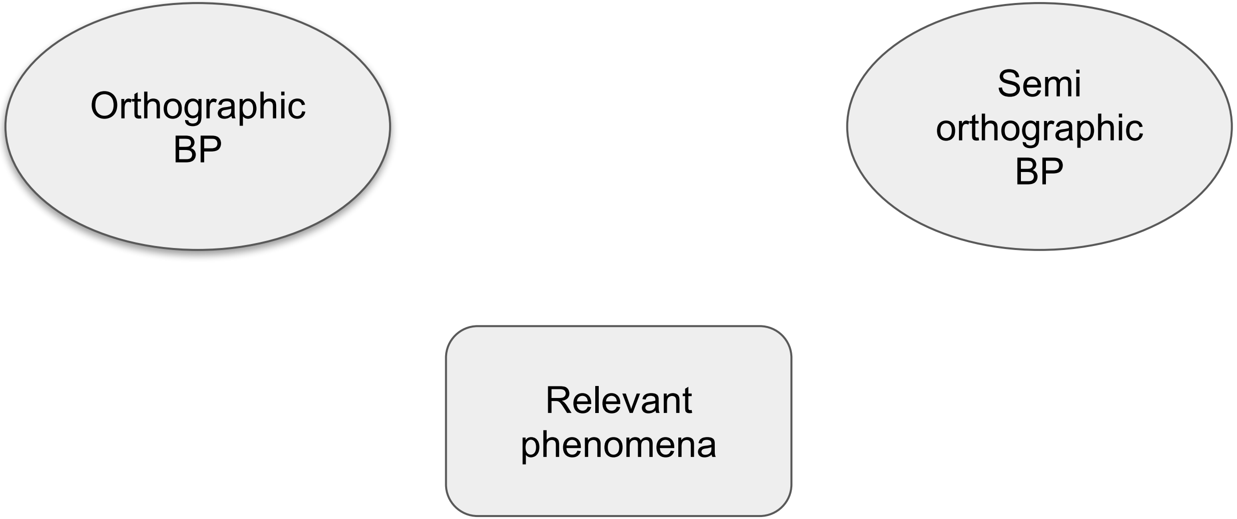
- Transcription based on the Language into Act Theory (L-AcT) (Cresti 2000)
- Utterance is a linguistic unit with prosodic and pragmatic autonomy (Cresti 2000) and conveys a speech act (Austin 1962)

SIL: porque enquanto descansa /=TOP= carrega pedra //=COM=

CAR: pra três pessoas //=COM=

Spoken data

Criteria adaptation to represent relevant phaenomena in contact variety



Orthographic BP

Semi orthographic BP

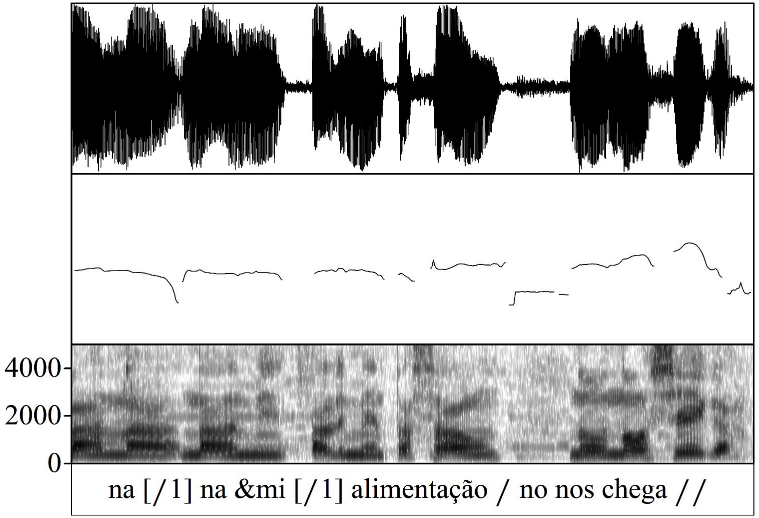
Relevant phenomena

Spoken data

Criteria adaptation to represent relevant phaenomena in contact variety

- The negation

  C-ORAL-BRASIL: não | nũ | n' é não + **<u>non | no</u>**



Spoken data

Criteria adaptation to represent relevant phaenomena in contact variety

- The negation

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

- “Mais”

  C-ORAL-BRASIL: mais + **ma**

- “Sim”

  C-ORAL-BRASIL: sim + **si**

- “Mas”

  C-ORAL-BRASIL: mas + **mai**

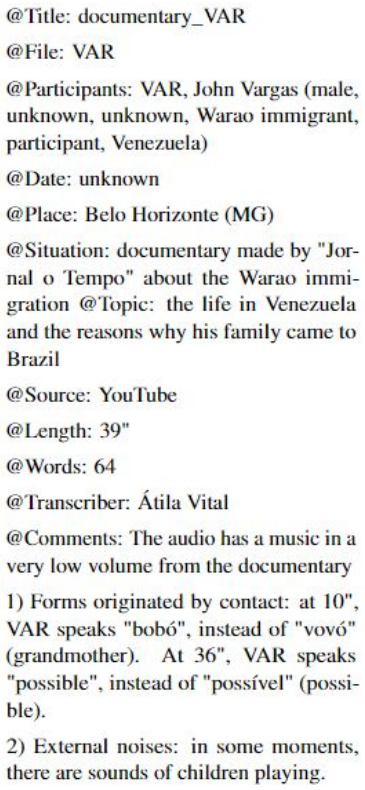
Spoken data

Criteria adaptation to represent relevant phaenomena in contact variety

- Disfluencies

*ANI: família Warao tá em [/1] em **&Ri [/1] &i [/1] Rio** / em / Brasília / tá em São Paulo / ah / quase tudo né / estado //

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

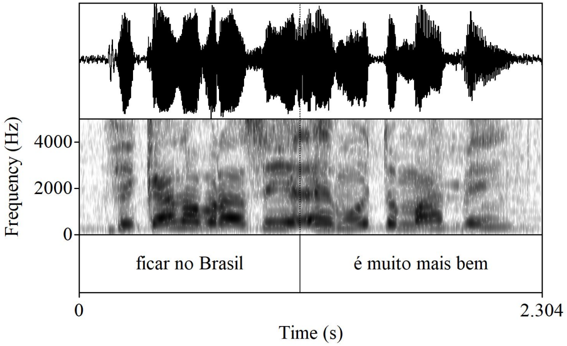
1) Forms originated by contact: at 10", VAR speaks "bobó", instead of "vovó" (grandmother). At 36", VAR speaks "possible", instead of "possível" (possible).

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.

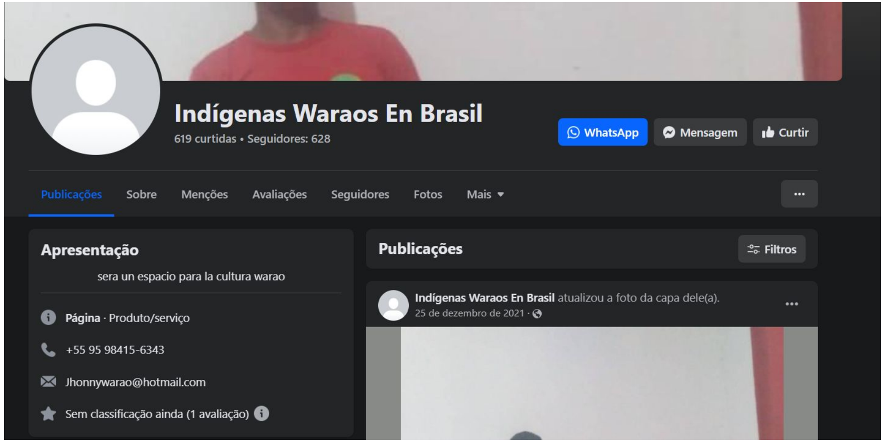
Potential linguistic phenomena



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



Indígenas Waraos En Brasil

619 curtidas • Seguidores: 628

Apresentação

sera un espacio para la cultura warao

Página · Produto/serviço

+55 95 98415-6343

Jhonnywarao@hotmail.com

Sem classificação ainda (1 avaliação)

Publicações

Indígenas Waraos En Brasil atualizou a foto da capa dele(a).

25 de dezembro de 2021

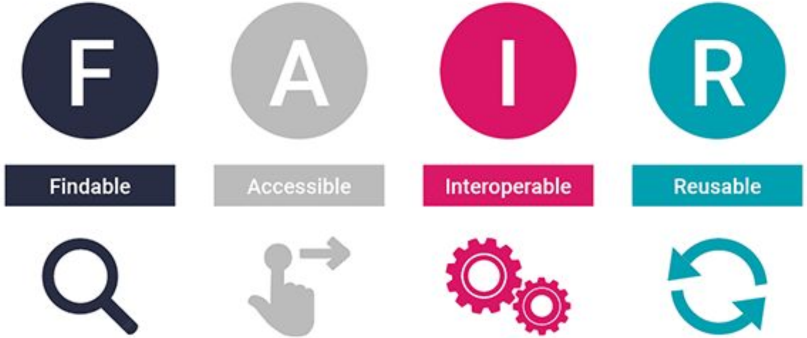
- Our initial objective is to contain around **60 transcribed and annotated signs**, and **20 recordings of spontaneous speech**, totalling approximately 1,500 words. All of them will be transcribed, segmented and aligned;
- Investigate the use of some Warao refugees on digital social networks;
- **Spontaneous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

CARE Principles

Carroll et al. (2020)

- Promote community empowerment through language description work.
- **C**ollective Benefit, **A**uthority to Control, **R**esponsibility, **E**thics
- Fieldwork with indigenous people living in shelters managed by Cáritas Brasileira in the city of Belo Horizonte

FAIR Principles



Wilkinson et al. (2016)

- Data will be openly available in digital repositories.
- The annotations will follow protocols previously developed for comparison by other researchers (e.g. Universal Dependencies & C-ORAL-BRASIL methodology)
- The use license will allow use and citation by researchers worldwide.

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

Dalmo Buzato and Átila Vital. 2024. Creating datasets for emergent contact languages preservation. Proceedings of the 16th International Conference on Computational Processing of Portuguese-Vol. 2.

Evandro L T P Cunha. 2020. A web como ferramenta de suporte à preservação e à revitalização linguística. Cadernos de Linguística, 1(3):01–14.

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

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.

References

Mesquita, R. (2020). Diaria o fixo: fotografias sociolinguísticas de Boa vista–Roraima e as novas perspectivas para as pesquisas do contato linguístico na fronteira. In Cruz, A. and Aleixo, F., editors, Roraima entre línguas: contatos linguísticos no universo da tríplice fronteira do extremo-norte brasileiro. Editora da UFRR.

Wilkinson, Mark D., et al. The FAIR Guiding Principles for scientific data management and stewardship. Scientific data 3.1 (2016): 1-9.

Carroll, Stephanie, et al. The CARE principles for indigenous data governance. Data science journal 19 (2020).

Romero-Figueroa, A. (1997). A Reference Grammar of Warao. Lincom Europa, Munchen.

Soneghetti, Pedro Moutinho Costa. Parecer Técnico acerca da situação dos indígenas das da etnia Warao na cidade de Manaus, provenientes da região do delta do Orinoco, na Venezuela. Procuradoria Geral da República/AM, 2017.

Cresti, E. Corpus di italiano parlato. [S.l.]: Accademia della Crusca, 2000.

Austin, John. How to do things with words. Harvard University Press, 1962.

FALE
FACULDADE
DE LETRAS

UFMG

Muito obrigado! Thank you!

atilavital@ufmg.br
dalmobuzato@ufmg.br

VII Simpósio Internacional de Linguística Funcional (SILF)

Belo Horizonte, Minas Gerais

September 4-6, 2024