Creating datasets for emergent contact languages preservation

Dalmo Buzato
Faculty of Letters
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
dalmobuzato@ufmg.br

Átila Vital
Faculty of Letters
Universidade Federal de Minas Gerais
Belo Horizonte, Brazil
atilavital@ufmg.br

Abstract

The Venezuelan socioeconomic crisis increased the immigration process in Latin America. In Brazil, the Warao ethnic group, from Northeast Venezuela, has arrived in search of jobs and better social conditions, speaking a homonymous isolated language and mostly having Spanish as a second language. The communities have contact with the Brazilian Portuguese, intensifying the possibility for the appearance of an emergent contact language. This paper presents a dataset for the description and preservation of that emergent language. Based on previous works about multimodal data compilation, the dataset will be fed with written and spoken texts, sociolinguistic information, and morphosyntactic annotation. As soon as possible, it will be freely available for web consultation, following precepts of the Open Science Framework and the Digital Humanities paradigm.

1 Introduction

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.

Not all language contact situations lead to the emergence of contact languages. In some cases, there is linguistic borrowing between the languages involved, and changes accumulated over several generations. There are also cases in which a language is used as a mediator of contact but is not necessarily a contact language, in which case we have what linguists call a *lingua franca*. For a systematic discussion of the various linguistic possibilities in a language contact situation and the differences between them, we cite Holm (2000) and Matras (2020).

Much has been discussed about the preservation and ecology of contact languages (Mufwene, 2003). Many contact languages have an unstable status of existence, becoming extinct when the contact situation between speakers of different languages ends (such as business situations, migrations, etc.). In addition, when there is a creole language, i.e. when the contact language is the mother tongue of a generation of individuals in a contact ecology, 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).

In addition to the extremely productive dialogue between corpus linguistics and contact linguistics (Nagy, 2011; Mello, 2014; Adamou, 2016; Léglise and Alby, 2016), relevant discussions have emerged about the creation of corpora and the use of the web for language preservation and documentation. According to Cunha (2020), "in the face of the effective threat of disappearance that thousands of languages around the world are currently suffering, all instruments for the conservation of linguistic diversity must be explored[1]". For the author, the internet has a paradoxical role in this context, because while it contributes to the dissemination of majority languages to more individuals and communities, it can also help to amplify the voice of minority language speakers.

Intending to document and preserve emerging contact languages, this study reports on the ongoing development of a dataset with spoken and written data produced by Venezuelan refugees in Brazil. Most of the data was produced by indigenous refugees of the Warao ethnic group, as we will detail later in the text.

We believe that this work falls within the field of Digital Humanities because it promotes the documentation and maintenance of a language through

[1] Original text: "[...] diante da efetiva ameaça de desaparecimento que sofrem, na atualidade, milhares de línguas ao redor do mundo, todos os instrumentos para a conservação da diversidade linguística devem ser explorados."

digital resources. Much more than simply storing audio and text files in a digital database, this paper uses data collected on the web (mostly photographs and video interviews available on the internet produced by the news media) to document the linguistic variety that emerges when Venezuelan Warao refugees arrive in Brazil.

An almost countless amount of data is produced every day on the internet, whether in media outlets, on social networks, or on websites. This data, even though some of it is currently produced with the help of artificial intelligence, is extremely valuable to linguists because it allows access to an exorbitant amount of data full of linguistic phenomena in an accessible and relatively simple way.

Another justification for the development of this paper lies in the very nature of contact languages. Many of them, including extinct ones, have little documentation as they are primarily transmitted orally and have an unstable survival status, such as pidgins, which emerge as emergency languages. Furthermore, if there is significant pressure for social integration, succeeding generations of contact language speakers may abandon it or incorporate various elements from the prestigious language in a process of language planning.

The subsequent sections will be organized as follows: in the upcoming section, we will provide a concise introduction to the Warao migration to Brazil and the language contact resulting from this migration. Sections 3 and 4 will elucidate the methodological details of the dataset, encompassing the nature and processes involved in storing, transcribing, and annotating both the written and spoken data. The former will include a brief description of the Universal Dependencies framework and its use for annotating linguistic phenomena. The oral data section will present a brief description of the C-ORAL-BRASIL's (Raso and Mello, 2012) transcription criteria, positioned before the audio section of this work. Section 5 will outline potential linguistic phenomena discerned during the previous analysis, while Section 6 will serve as the concluding remarks on the future of the dataset.

2 Warao migration to Brazil

Since the migration of Venezuelans to Brazil began in mid-2017, this phenomenon has been documented by researchers in law, anthropology, sociology, and linguistics. Since the first records, the presence of indigenous refugees has been noted, mainly from the Warao ethnic group.

Research in linguistics has emerged since the beginning of the migration and has mostly been concentrated in the field of applied linguistics, such as in the areas of language policy and foreign language teaching and learning.

Research into language contact has emerged very recently, mainly analyzing the written productions of Venezuelan refugees asking for help, examples of research taking this approach are Mesquita (2020); Buzato and Vital (2023); Buzato (2023).

Points of relevance for research into language contact in the case of Venezuelan indigenous migration to Brazil is the fact that the Warao are speakers of a homonymous native language with no known linguistic relatives as L1, and are speakers of Spanish as L2 at different levels of proficiency, with a significant percentage of migrants having a low level of schooling and literacy.

In addition, according to anthropological studies (UNHCR, 2021; Soneghetti, 2017), the Warao were not a people with nomadic characteristics before their growing status of subalternity, which began with the loss of land for extractive activities in Venezuela, and with their migration to Brazil.

The refugees have not just stayed in the border regions between Brazil and Venezuela, or concentrated in the north of the country, which is closest to the neighboring country. On the contrary, they've moved inland and made long, independent journeys through towns and cities, always with the help of local citizens, to reach regions they believe are best for them to settle in.

For example, the distance between the city of Boa Vista, the capital of the Brazilian state of Roraima (the main initial concentration of refugees after leaving Venezuela), and the city of Belo Horizonte (where some of the photographs were taken) is over 3,000 kilometers, a route traveled independently by the refugees with their families and belongings.

3 Written signs and the dataset

This section will discuss some linguistic aspects of the written signs produced by the refugees, to ask the Brazilian population for help. As will be described below, the written dataset is of a mixed nature, with a percentage of photographs collected from news sites on the internet, and the other part of the photographs of the signs were collected by the researchers, since March 2022, in a fieldwork carried out in the city of Belo Horizonte and metropolitan region, in the state of Minas Gerais, located in the southeast region of Brazilian territory.

Figure 1 below represents an example of a sign written by indigenous refugees. Although the signs collected in the city of Belo Horizonte in almost two years of fieldwork represent an important part of the data, we believe that the photographs collected from the web represent greater quality and representativeness of the phenomenon, since we have reported signs from 2018 to the present year 2024, and collected in several Brazilian cities of different population sizes and regions.

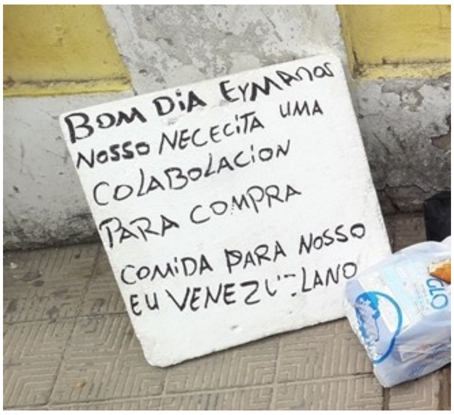

Figure 1: Example of a sign in the dataset

Currently, the photographs of the written signs, whether they come from the web or the researchers' fieldwork, are also transcribed and annotated according to the criteria of the Universal Dependencies (UD) project (Nivre et al., 2016). The transcriptions were made in a standard txt file, and the morphosyntactic annotations are in CoNLL-U format, the standard format of the UD project[2], as can be better elucidated in Buzato (2023). The choice of the UD framework for the written signs is based on its typological proposal and its growing use for annotating non-Indo-European and minority languages, in the spoken and written modalities of language.

Below is an example of how the sign shown in Figure 1 was transcribed. As we can see, no spelling or linguistic corrections have been made to the text produced since they can contain contact phenomena. Furthermore, due to the textual and writing context of the signs, as well as the socio-economic variables of the refugees, most of our signs do not have any graphic punctuation marks. We also decided not to include them, as the absence of punctuation is an important aspect for our research.

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

3.1 Universal dependencies (UD) and language contact phenomena

The UD framework is increasingly developing treebanks to document contact languages and varieties. Currently, it has treebanks of Creole languages and varieties of spoken and written code-switching, derived from diverse texts such as comments on websites (Seddah et al., 2020) or radio interviews (Braggaar and van der Goot, 2021). However, the documentation of pidgin or mixed languages is still underdeveloped. A proposal was recently presented by Buzato (2023) whose annotations will form part of the dataset described here.

The presence of minority/low-resource languages is essential for any typological project, which certainly includes varieties emerging from language contact, especially in initiatives that promote the use of computational tools for typological analyses and the use of large amounts of data from different languages to improve models and tasks in computational linguistics and natural language processing. For this reason, documenting the variety presented here employing UD also contributes to the framework's objective and explores its potential for morphosyntactic annotation.

As can be seen in the annotation guidelines of UD for phenomena of foreign expressions and code-switching[3], it essentially covers lexical borrowing and code-switching phenomena. These phenomena are typically considered to emerge from language contacts of lesser intensity between two or more communities. The annotation of these phenomena depends on the nature of the corpus (if it is a code-switched corpus or a monolingual one), and which specific phenomenon is under consideration. In such instances, multilingual material is annotated in features like `Lang` (language), `Foreign`, and `OrigLang` (Original language).

Since the annotation methods mentioned above do not encompass the phenomena found in our corpus, we have decided not to fully adopt them. In the example in Figure 1, for instance, we have

[2] https://universaldependencies.org/format.html

[3] https://universaldependencies.org/foreign.html

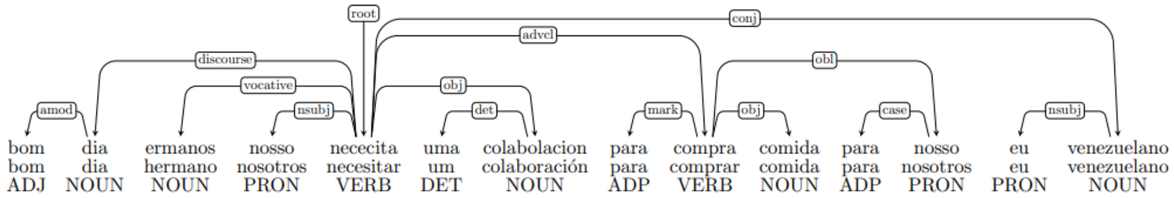

Figure 2: Example of a passage present in the written corpus annotated according to the UD framework

words like ’ermanos’ borrowed from the Spanish ’*hermanos*’, ’colabolacion’ borrowed from ’*colaboración*’, and ’nosso’ borrowed for ’*nosotros*’. As a regular occurrence in mixed languages, there is a more productive blend between the repertoires of the languages in contact. Unlike borrowing, an element is never precisely derived from a language, there are often innovations (addition or loss) present in the linguistic element. Taking into account the aspects presented above, Figure 2 illustrates how the content presented in Figure 1 was annotated following the guidelines of UD.

3.2 Future steps on written section

We know, for example, of the existence of computer tools that allow multimodal texts to be annotated synchronously with the image, which is widely used in research into historical linguistics, such as the TEITOK tool (Janssen, 2016). Furthermore, the use of this tool is increasingly common for documenting minority or low-resourced languages such as Judaeo-Spanish (Quintana, 2020) and Galician (Blanco and Seoane, 2020), or varieties of languages spoken by second language (L2) speakers. This is a work in progress and certainly, one of the next steps in the project, which involves annotating and documenting the written texts produced by the refugees, most of which are multimodal texts - recorded by photographs, includes the use of TEITOK to unify the photograph and the linguistic annotation made by us.

4 The audio section

As a multimodal dataset, there are planned 20 recordings and transcriptions of spontaneous speech. In the first step of the audio compilations, it is important to perform tests to elaborate specific transcription criteria. The basic methodology to be applied is the same used in the C-ORAL-BRASIL’s project (Raso and Mello, 2012), adapted by essential changes that will be elaborated in terms of the potential grammaticalization phenomena to be represented in the dataset.

4.1 A brief look to the C-ORAL-BRASIL’s criteria

One of the most important aspects when dealing with the speech is the package of information conveyed by the prosody (Izre’el et al., 2020). Unlike written texts, the orality does not use punctuation to integrate the discourse into its morphosyntax parameters. Much more than that, through the combination of the fundamental frequency, length, and intensity, the spoken discourse integrates form and function, indicating the way the morphosyntax and the semantic/pragmatics relationship work across the sequence of words (González Ledesma et al., 2004; Raso and Mello, 2012).

Based on prosodic parameters, the transcriptions criteria used in the C-ORAL’s corpora follow the segmentation of speech flow in terms of utterances and tonal units. In other words, the utterance is considered the minimal unit of the speech that conveys a complete communicative function (Izre’el et al., 2020). Along with that, two types of prosodic brakes (terminal and non-terminal brakes) give us perceptual clues about the compositionality and the non-compositionality of linguistic sequences.

The transcription criteria were adapted to the spontaneous spoken Brazilian Portuguese based on the C-ORAL-ROM’s (Cresti and Moneglia, 2005). To provide consistent guidelines for the transcription crew, the authors organized several pilot studies. Those studies helped the establishment of a series of semi-orthographic criteria capable of capturing cliticizations, apheretic forms, erasing of verbal morphology, new pronominal paradigms, disfluencies, and many others. If the transcriptions followed purely orthographic criteria, many relevant lexical and grammatical phenomena would be hidden for future research.

In 1, there is an example of utterance recorded by C-ORAL-BRASIL I. The double bars “//” indicate the end of an utterance. Simple bars “/” indicate intonational units that do not convey a complete communicative function.

Example 1 (bfammn06)

> JOR: aonde a gente tem muito poblema de liquidez / até em empresas que têm / &he / formação de família / na segunda pa terceira geração / já começa a dar poblema e &f [/2] e [/1] e fecha //

Considering the prescriptive orthography, the utterance in 1 has lots of particularities. In a brief look at it, we can identify errors of pronunciation in the word "poblema", which would be written "problema" according to the grammatical prescriptivism. The choice to represent faithfully the way the speaker spoke is important for studies in variational linguistics that have been done with Brazilian Portuguese. Additionally, the phonetic contraction of the preposition "pra" (pronounced and transcribed as "pa") can reveal the complex topic of prepositions and their forms in romance languages.

Just like the mispronunciation, the transcription developed for the C-ORAL-BRASIL represents disfluencies (self-corrections and fragmented words) and time-taking units (entire conversational turns with only "he" and "uhn"). The letter "&" represents a filled pause (&he) or an incomplete word (&f). The mark "[/n]", in which "n" is a natural number, represents a retracting, a very common disfluency in spontaneous speech, a.k.a. self-correction; it happens when the speaker produces a word or a part of it and immediately corrects himself. The number inside the brackets means the number of canceled words (Raso and Mello, 2012).

4.2 First application of the transcriptions criteria to the immigrants' spoken language

The first applications of the transcriptions give us important inferences about the richness of linguistic phenomena presented in a new spoken dataset. The goal of this subsection is to validate the conventions created for C-ORAL-BRASIL to the application in the emergent immigrant's language. To do this, there were transcribed some audio parts collected from Warao's documentary available on YouTube.

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 // porque / indígena não [/1] não mora nas cidades / não mora na montanha // sim // então / lá não não chega médico / não [/1] &n não é possible / que [/1] que médico chegue lá //

With the transcription, there will be available the header's file, which compile possible sociolinguistic information, comments about the transcription, and conventionalized forms. In some moments of 2, we find Portuguese and Spanish lexical combinations ("si" and "possible"). It was considered important to represent those words in the way they were spoken with the appropriate comments in the header, as follows. The layout was inspired in the C-ORAL-BRASIL's corpus as well.

> @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.

During the audio compilation, we will value high-quality recordings. That makes possible Phonetic and Prosodic investigations. The example 3 shows an utterance with glottalization and particular morphosyntax.

Example 3 (documentary_AAA)

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

The figure 3 shows the waveform and the spectrogram of 3. The high acoustic quality is rare to be found in emergent language descriptions.

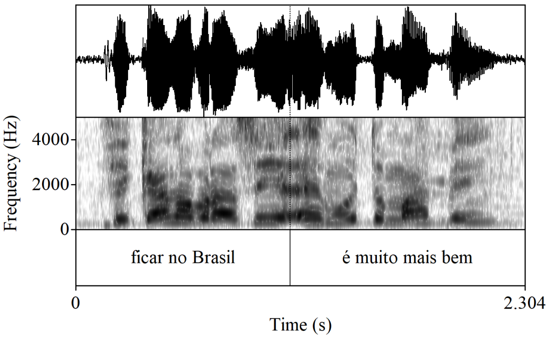

Figure 3: Waveform and spectrogram of example 3.

When building the dataset, all audio files and their respective text-sound alignments will be made available, in addition to the speakers' metadata txt files.

5 Potential linguistic phenomena

Preliminary descriptions of the Warao's signs have been made in previous works (Buzato and Vital, 2023; Buzato, 2023). Even with the objection that writing is not the primary form of emergence of a linguistic system, the proposals initially outlined aim to reflect on the structural particularities of the language used by immigrants.

In addition to the characteristics of the written representation - which seems to resemble speech (Figure 1) - and the constant borrowings from Spanish and Portuguese, some occurrences catch our attention. Firstly, there is a recurring confusion between the use of the adjective related to Venezuela (Venezuelan) and the name of the country itself. It is not uncommon to find data like "eu sou da venezuelano" ("I am from Venezuelan") or "eu (sou) venezuela" ("I (am) Venezuela"). Until now, there is no sufficiently structured data to verify the co-occurrence of these structures with prepositions or specific syntactic positions.

Another syntactically important phenomenon to be punctuated is the use of the copula. Romero-Figueroa (1997) points out that the Warao language allows the optional use of the verb copula in the expression of properties of nominal entities. The reuse of Warao's syntactic structure is what may explain the absence of a linking verb in Figure 1, given the juxtaposition between the pronoun "eu" ("I") and the adjective "venezuelano" ("Venezuelan"). On the contrary, reflections of the confusion between the linguistic structures of the Spanish language and the Portuguese language are also found on the signs. An example is the use of an accusative pronoun postposed to the verb, as in "ajuda me" ("help me"), a less frequent form in Brazilian Portuguese, which favors the preposition "me ajuda" ("help me").

These are just some initial structural notes from the studies carried out with the signs. In the case of the audio files, we hope to confirm the data coming from the signs and describe even more contact phenomena.

6 Conclusions and future and the dataset

This is preliminary work towards the construction and availability of a dataset of an emerging contact language. 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.

The linguistic description through immigrant signs is not common to be found in literature. Still, together with the collection of spontaneous speech data from Warao refugees, the data that will be accumulated and publicized could open the way for new methodologies in the study of emerging languages. We believe that, in the case of languages that emerge during migration crises, signs and writings asking for help may represent the only registers of the emerging forms. Both the development of methodologies and databases are welcome at this time.

At an opportune moment, when we have the first spoken and written data transcribed, annotated, and reviewed, the multimodal dataset will be freely available for web consultation.

Our goal is to document other emerging contact languages through the above protocols, using spoken and written data, mainly in low-resourced varieties in the global south. Furthermore, already extinct contact varieties, such as pidgin or mixed languages, can be transcribed and annotated using the same protocols, thus providing the creation of a set of multilingual datasets of emerging contact languages.

References

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

Rosario Álvarez Blanco and Ernesto Xosé González Seoane. 2020. *Calen barbas, falen cartas: A escrita en galego na Idade Moderna*. Consello da Cultura Galega.

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.

Emanuela Cresti and Massimo Moneglia. 2005. *C-ORAL-ROM: integrated reference corpora for spoken romance languages*. John Benjamins Publishing.

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.

Ana González Ledesma, Guillermo De la Madrid, Manuel Alcántara Plá, R De la Torre, and Antonio Moreno-Sandoval. 2004. Orality and difficulties in the transcription of spoken corpora. In *Proceedings of the Workshop on Compiling and Processing Spoken Language Corpora, LREC*.

John Holm. 2000. *An introduction to pidgins and creoles*. Cambridge University Press.

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.

Yaron Matras. 2020. *Language contact*. Cambridge University Press.

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.

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

Salikoko S Mufwene. 2003. Language endangerment: What have pride and prestige got to do with it. *When languages collide: Perspectives on language conflict, language competition, and language coexistence*, pages 324–346.

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

Joakim Nivre, Marie-Catherine De Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, Christopher D Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, et al. 2016. Universal dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666.

Aldina Quintana. 2020. CoDiAJe–the Annotated Diachronic Corpus of Judeo-spanish. *Scriptum digital. Revista de corpus diacrònics i edició digital en Llengües iberoromàniques*, (9):209–236.

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

Andrés Romero-Figueroa. 1997. *A Reference Grammar of Warao*. Lincom Europa, München.

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.

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

UNHCR. 2021. Os Warao no Brasil - Contribuições da antropologia para a proteção de indígenas refugiados e migrantes. Technical report, Brasília.