Universal Dependencies and Language Contact Annotation: Experience from Warao refugees signs in Brazil

Dalmo Buzato[1]

[1]Federal University of Minas Gerais, Belo Horizonte, Brazil

buzatodalmo@gmail.com

Abstract. *This article aims to present a work in progress that proposes to describe, from the Universal Dependencies (UD) project, the linguistic contact between the Warao language, Venezuelan Spanish, and Brazilian Portuguese on the Warao refugee signs in Brazil. In the present work, in addition to presenting a brief description of the contact between the three languages in the Venezuelan indigenous migration to Brazil, the description of contact languages using the Universal Dependencies is discussed, in addition to initial reflections on methodological choices motivated by the linguistic phenomena observed in the corpus.*

Resumo. *O objetivo do presente artigo é apresentar um trabalho em andamento que se propõe a descrever, a partir do projeto das Universal Dependencies (UD), o contato linguístico entre a língua warao, o espanhol venezuelano e o português brasileiro nas placas de refugiados Warao no Brasil. No presente trabalho, além de apresentar-se uma breve descrição do contato entre os três idiomas na migração indígena venezuelana para o Brasil, discute-se sobre a descrição de línguas de contato valendo-se das UD, além de reflexões iniciais acerca de escolhas metodológicas motivadas pelos fenômenos linguísticos observados no corpus.*

1. Introduction

This article aims to document the ongoing compilation, transcription, and annotation of a treebank consisting of signs written by Warao refugees from Venezuela in Brazil. For this purpose, we rely on the Universal Dependencies (UD) project [Nivre et al. 2016] to describe the linguistic contact between the Warao language, Venezuelan Spanish, and Brazilian Portuguese.

This article will present a brief introduction to the linguistic and sociopolitical context of the Warao migratory uprising from Venezuela to Brazil. Additionally, we will address the relationship between the annotation of contact languages and phenomena (e.g., code-switching, pidgins, creoles, and mixed languages) and the Universal Dependencies project. Finally, we will discuss some initial reflections on the decisions and specific aspects to be taken into consideration in annotating linguistic phenomena in contact contexts for computational purposes.

The Warao are an indigenous ethnic group inhabiting the northeastern region of Venezuela, known as the Orinoco River Delta, as well as some regions in Guyana and Suriname. The Warao people speak a language with the same name, which is an isolated language with no known linguistic relatives. According to [Romero-Figueroa 2020],

[UNHCR 2021a], and [UNHCR 2021b] the Warao constitute the third-largest indigenous population in Venezuela, with approximately 41,000 individuals, making them one of the most prominent and significant indigenous peoples in the country. Anthropological studies suggest that the Warao might be the oldest inhabitants of present-day Venezuela, as they have been residing in the Orinoco Delta region for at least 8000 years.

The humanitarian crisis of the Warao people seems to predate the migratory flow of Venezuelans to Brazil. As reported in [Buzato and Vital 2023] and [GarcÍa-Castro 2006], since the second half of the 20th century, especially from the 1970s onwards, a scenario of subalternity has been observed among the Warao in Venezuela due to the expansion of extractive, agricultural, and industrial activities in the Orinoco delta region. Nomadism was not the traditional way of life for the Warao; however, following the loss of territories to the mentioned activities and in pursuit of better living conditions, the Warao began to migrate to urban centers within Venezuela.

Due to linguistic and cultural differences, the Warao faced numerous challenges in integrating into Venezuelan cities, consequently encountering various hardships such as poverty, violence, discrimination, underemployment, and low educational attainment. These factors could be understood as rendering them subaltern people in their own country after losing their place of origin. The linguistic contact between Warao and Venezuelan Spanish appears to be more enduring, yet its effects are sparsely documented. [Romero-Figueroa 2020] focuses on the lexical effects of this contact.

With the worsening of the political and economic crisis in Venezuela, the migratory flow to Brazil intensified from 2015 onward. Consequently, there was a flow of Warao immigrants to Brazil, initially concentrated in the northern states of the country, particularly in the capitals of Belém and Rondônia. The Warao people, who were already disadvantaged in their country of origin, perceive this condition to be doubly amplified upon migrating to Brazil, now facing an even more distinct and pronounced language barrier.

As a result of the substantial migratory influx from Venezuela to Brazil, the federal government, in collaboration with third-sector organizations, has initiated the ”Operação Acolhida”. This operation entails the internalization of refugees and migrants from Venezuela who arrive in the northern region of Brazil, aiming to enhance their quality of life, facilitate social integration, and mitigate the urban challenges confronted by municipalities near the border.

Despite government and third-sector assistance, a portion of the refugees finds themselves in precarious situations, often relying on the support of the host communities to survive and acquire essential food and daily necessities. In order to do so, refugees resort to signs bearing requests for aid, seeking contributions, typically in monetary form, from the Brazilian population. An overview of these signs and their communicational, textual, and linguistic features can be found in [Mesquita 2020] and [Buzato and Vital 2023].

2. UD e language contact

Linguistic contact occurs when speakers of different languages interact with each other or become part of the same speech community [Crystal 1987]. It is an extremely productive

linguistic phenomenon that has many possible ramifications in linguistic structure. These range from lexical borrowing and code-switching to the emergence of pidgin or mixed and creole languages. The differentiation and distinction among each of these categories are not uniform and continue to be subjects of various discussions in linguistic studies.

Currently, as stated in the Universal Dependencies project documentation, only one Creole language is documented through the framework: the Naija language (Nigerian Pidgin), a contact language spoken in Nigeria. The corpus was constructed from transcriptions of audio recordings collected in 2017 for the *ANR NaijaSyncor* project [Caron et al. 2019]. This oral corpus is also characterized by occasional code-switching to English, as well as to various native Nigerian languages, including Yoruba, Hausa, and Igbo.

However, regarding code-switching studies, there are currently four treebanks dedicated to documenting this phenomenon: *UD Frisian Dutch-Fame* [Braggaar and van der Goot 2021], *UD Maghrebi Arabic French-Arabizi* [Seddah et al. 2020], *UD Turkish German* [Çetinoğlu and Çöltekin 2019], and *UD Hindi English* [Bhat et al. 2018].

The current article appears to contribute to the state of the art in Universal Dependencies by aiming to describe a language that seems to go beyond the boundaries of code-switching, yet is not a Creole language. The language that emerges from the signs produced by refugees appears to surpass the two aforementioned boundaries, constituting a category of emergent languages (e.g., pidgins or mixed languages), which have not been documented within the UD framework up to the present moment.

The discussion about the nature of the language produced by Warao refugees in Brazil is not within the scope of the present article. We agree that for more robust analyses and a more precise definition, a larger amount of data would be necessary, preferably encompassing other communicational situations besides the use of signs for asking for help. Sociolinguistic information about the process of creating these signs, their authors, as well as more detailed profiling of the communities, and the alternation of linguistic uses (for instance, which language they use to communicate among themselves?) would be of great value for a better understanding of the phenomenon and the nature of this emergent language.

Among the documented contact languages in the Universal Dependencies project, *Hindi English* and *Maghrebi Arabic French* are languages that are predominantly documented in written genres. In the case of *Hindi English*, it is documented through tweets, while *Maghrebi Arabic French* has been documented through news articles and comments on Algerian newspaper web forums. This characteristic of other languages seems to support the documentation of contact phenomena in the written modality of language, similar to the case of the Warao refugee signs described in this study. Just as in our case, the written modality of the other treebanks appears to be influenced by the spoken modality, as we will address in the upcoming sections.

3. Data and Initial Discussions

The initially annotated corpus consists of 21 photographs of signs created by refugees, collected by researchers in two medium and large cities in the southeastern region of

Brazil during the months of May 2022 and May 2023. An example of the collected signs can be observed in Figure 1.

Figure 1. Example of a Sign created by refugees

When selecting written texts as the object of analysis, we must consider the specificities of this modality compared to spoken language. However, in our case, the signs are written by speakers who mostly have very low levels of education and formal instruction, as indicated by demographic profiling. We must also take into account the transposition of strategies and phenomena from the oral modality of language into the refugees' textual production. These phenomena are not necessarily derived from linguistic contact, but they certainly influence it in some way. Therefore, choices in transcription related to orality and the absence of formal instruction impact the visualization and understanding of the contact phenomena that we will discuss further.

For example, in Figure 1, despite the absence of uppercase and lowercase letters and spacing between words on the sign, we have chosen to transcribe them separately according to the words present in Brazilian Portuguese. In cases where speakers write words with orthographic deviations, we have decided to retain them due to the possibility of containing contact phenomena. Additionally, if speakers write incomprehensible or nonexistent words in Portuguese, we have chosen to preserve them in that form for potential lexical parallels in the other languages involved in the contact. Lastly, most signs lack any graphical punctuation marks. As we believe that the absence of punctuation usage reveals much about the migrants' formal education level and textual production, we have chosen not to add any punctuation marks to the signs, leaving them with only the possible punctuation marks that each speaker used.

Therefore, the sign we saw above in Figure 1, for instance, was transcribed in our treebank as follows:

(1) boa tarde irmao sou da venezuela preciso ajuda dinheiro amigo para paga luga para comprar roupa gazisa
good afternoon buddy ∅ am from venezuela ∅ need help money friend for payrent for buy clothes cooking gas

If we were to strictly adhere to the rules of grammatical norms, we should reasonably transcribe it as follows:

(2) Boa tarde, irmão! Sou da Venezuela, e preciso de uma ajuda, dinheiro para pagar o aluguel, comprar roupa e gás.
Good afternoon, buddy! I'm from Venezuela, and I need help – money to pay for rent, buy clothes, and get cooking gas.

The content of the photographs was transcribed into a .txt file and automatically annotated using the UDpipe tool [Straka et al. 2016], with a Portuguese language model based on Bosque-UD v. 2.6 [Rademaker et al. 2017]. Afterward, the generated CONLLU files were imported into the annotation tool Arborator-Grew-NILC (https://arborator.icmc.usp). For human annotation and review, we followed the general UD guidelines, as well as the ICMC/USP annotation manual for the Portuguese language [Duran 2021]. In the following sections, we will discuss certain aspects related to the transcription of the signs and the methodological choices made during the process of human review of the content automatically annotated by the model.

The automatic annotation proved to be less effective due to the nature of the input text, along with another factor that significantly influences the written production of refugees: low education levels. From an orthographic perspective, the vast majority of signs lack any punctuation, spacing, or graphical accents. As we are aware, all these factors impact automatic annotation, which certainly explains the low performance of the model with the texts tested here. Furthermore, the writing, strongly influenced by oral aspects, along with anomalous constructions due to the context of linguistic contact and low education levels, also certainly accounts for the substantial need for manual human review and annotation.

For instance, when observing the automatic annotation generated by the model for the sentence provided in (1), shown in Figure 2 below, we notice that the absence of punctuation, combined with certain morphosyntactic characteristics of the signs, led to erroneous annotation by the model. The model mistook the verbal conjugation of the first person singular in the present indicative of the verb precisar (preciso *"I need"*) for the homonymous adjective. This is likely due to the absence of punctuation and the first-person singular pronoun (Eu) before the verb.

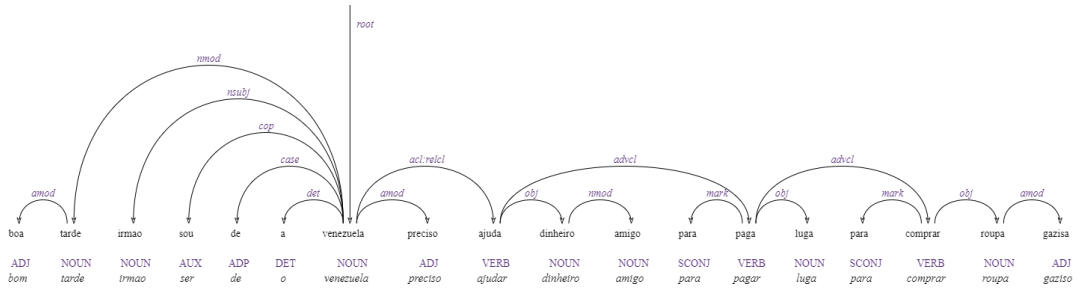

Figure 2. Automatic annotation of transcription (1) generated by the Bosque-UD v. 2.6 model

The model used to automatically annotate the transcriptions was trained on written texts from the journalistic genre, which possibly explains some of the model's difficulties

in annotating more oral-productive phenomena. Examples in transcription (1) can be observed in the model's inadequate annotation of the greeting "boa tarde" (*good afternoon*), an interjection that should be connected to the root with the *discourse* deprel, and the vocative "irmao" (*buddy*), to whom the message is addressed. Written journalistic genres rarely include vocatives or greeting interjections, which could explain the model's struggle in representing them. However, this greeting structure is highly productive in our data, found in almost all signs, along with a high recurrence of structures for farewells and expressions of gratitude (deus abencoe obrigado *"God bless you thanks"*), as seen in Figure 3. Therefore, it is a recurring demand for human correction in our experience.

Figure 3. Example of a Sign created by refugees

Other inconsistencies that can be observed in the automatic annotation of transcription (1) include difficulties in annotating the connections between segments using the *parataxis* and *conj* deprels. Our explanation for this occurrence is that the way elements are connected through these deprels in our transcriptions differs from what is common in Brazilian Portuguese. This is likely influenced by the morphosyntax of the Warao language, as we will discuss later on. Lastly, in using the vocative twice in the passage, seemingly disrupting the linearity of the explanation, as with the word "amigo" (*friend*) between "dinheiro" (*money*) and "para paga" (*to pay*), a phenomenon possibly more common in spontaneous speech, the model annotated "amigo" as a nominal modifier of the word "dinheiro", which lacks semantic consistency. A first revised version of transcription (1) can be found in Figure 4 below.

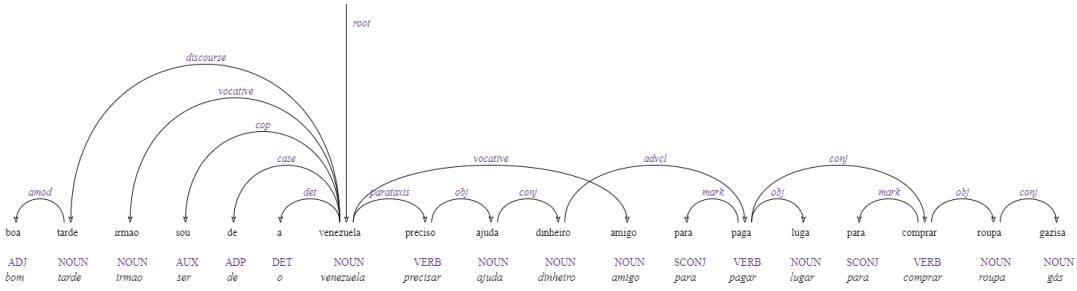

Figure 4. Revised annotation of transcription (1) generated by the Bosque-UD v. 2.6 model

As we can observe in Figure 4, the annotated relationship between the verb "preciso" (*I need*) and the complement "ajuda" (*help*) was labeled with the deprel *obj* (direct

object). According to [Luft 2010], the verb ”precisar” assumes different classifications of verbal transitivity depending on the complement. If the complement is a noun or pronoun, it is common to use a preposition, as exemplified in Figure 5. On the contrary, when the complement is an infinitive verb, the use of a preposition is not required, as in Figure 6.

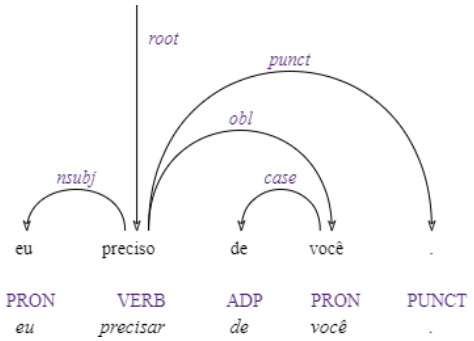

Figure 5. Sentence with the verb ”precisar” classified as an indirect transitive verb (deprel obl)

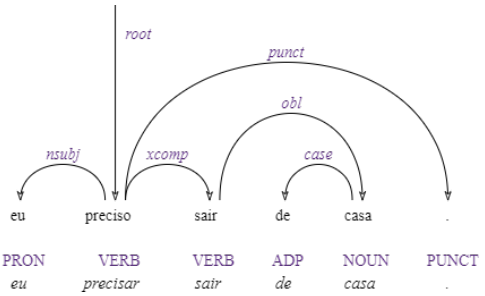

Figure 6. Sentence with the verb ”precisar” without the mandatory use of a preposition (deprel xcomp)

In Brazilian Portuguese, the verb ”precisar” usually requires a preposition in constructions like that in transcription (1), which makes it an indirect transitive verb and leads to its annotation with the deprel *obl* (oblique nominal). In the illustrated example, the refugees use ”precisar” without the preposition, which resulted in the segment being annotated as *obj*, contrary to the pattern observed in Brazilian Portuguese for similar sentences.

This phenomenon seems to have its origin in the transposition of the argument structure of the Spanish verb *necesitar* (to need) into Brazilian Portuguese. Therefore, it is a contact phenomenon. As we can observe in Figure 7 below (translation: ”I need help to buy chicken”), automatically annotated using the AnCora-UD 2.6 model for Spanish [Taulé et al. 2008], the verb *necesito* does not require a preposition in the construction *necesito ayuda*. Due to the proximity between the Spanish language and Brazilian Portuguese, what appears to occur is that speakers transpose the argument structure of the Spanish verb *necesito* onto the Brazilian Portuguese verb *preciso*.

Another phenomenon observed quite productively in the signs was typical phonological phenomena of spoken Brazilian Portuguese, especially in informal speech. As we can see in Figure 8 below, the orthographically proper way in Brazilian Portuguese would be ”preciso de ajuda” (*I need help*) and ”pedimos dinheiro” (*we ask for money*) for

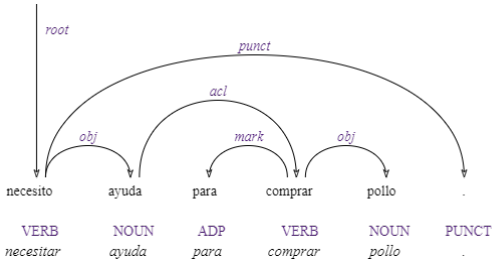

Figure 7. Sentence annotated with the AnCora v. 2.6 model in Spanish

the content in the third line of the sentence. In contrast, we observe the occurrence of "diajuda", where the absence of tonic accent on the preposition forms a single phonetic group [di.a.ˈʒu.dɐ], and it was equally represented as a single word orthographically. Furthermore, there is vowel harmony in "p[i]dimo ∼ p[e]dimo[s]", with the elision of the [s] sound. The elision of the [s], representing the plural mark, also occurs in the fifth line of the sign in the segment "02 menino 03 minina" (*two boy three girl*), which also features vowel harmony in "m[e]nina ∼ m[i]nina". Additionally, we observe diphthongization in the verb "ser" (to be) conjugated in the first person plural in the present indicative tense in the first line of the sign (somos ∼ so[u]mos).

The presence of these occurrences, typical of the oral modality of Brazilian Portuguese, represents an absorption of the language by refugees and especially reveals the specific Portuguese modality to which they have access. They likely do not have contact with the orthographically appropriate written modality or formal instruction in the Portuguese language, which is why they transpose speech phenomena to writing. This transposition also occurs in various textual genres among Portuguese speakers, especially in informal contexts such as daily speech or tweets. These phenomena have proven to be particularly challenging for automatic annotation, and we believe that this difficulty is likely to be encountered when annotating texts exclusively in Brazilian Portuguese, but from text genres that are predominantly informal, as mentioned above.

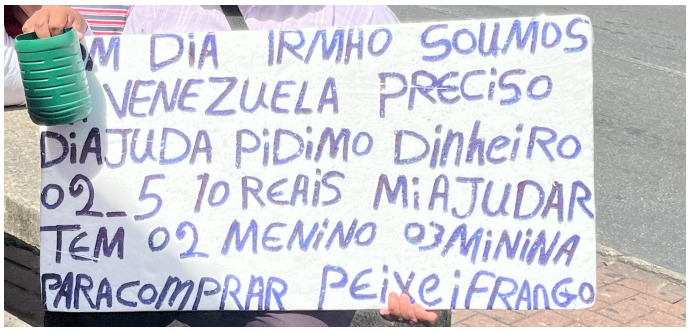

Figure 8. Example of a Sign created by refugees

Furthermore, certain aspects of the morphosyntax of the Warao language, an isolated agglutinative language, appear to be of paramount importance in explaining certain recurring phenomena that we observe in the data. In all the examples presented throughout this article, we notice the absence of conjunctions, for instance. This is possibly due to the lack of this grammatical class in the Warao language, as described by [Romero-Figueroa 1997]. Conversely, according to the author, the connection between

clauses in the language is achieved through parataxis. As a result, in the transcriptions and annotations of the texts, we observe that the majority of connections between clauses in the corpus do not occur through structural marking.

Other significant structural aspects of the Warao language that possibly influence the described contact include the constituent order, which, as per [Romero-Figueroa 1985], is an OSV language; the alternation between facultative use and absence of a copula verb in sentences; the absence of prepositions in the language, instead employing postpositions; the lack of marking for personal pronouns (zero morpheme) for various cases of the second and third person in singular and plural. Finally, the usage conditions and the organization of the pronominal and adposition systems of the Warao language appear consistently divergent from Romance languages. For instance, in Warao, the equivalents of the prepositions ”to” and ”for” for transitive verbal actions can occur through the dative case markers -ma and -to, or through the postposition ”saba”. The equivalents of the preposition ”of” in Warao are the postpositions ”a” and ”abitu”, solely expressing possession.

4. Final remarks

The objective of the current article was to report on the ongoing documentation of linguistic contact among the Warao language, Venezuelan Spanish, and Brazilian Portuguese resulting from Warao migration to Brazil, utilizing the UD project. Additionally, the article discussed language annotation and contact phenomena through UD, along with theoretical and methodological insights made up to the present moment during the annotation and transcription of the collected data.

The upcoming steps involve obtaining additional photographs to enhance the volume of data and facilitate more refined analyses. This will be followed by a stage of reviewing the annotations carried out by experienced researchers in UD. Furthermore, the plan is to also incorporate photographs and recorded interviews available on the web to increase the quantitative data in the treebank. This approach will not only expand the dataset but also offer indications of contact in other contexts and communicative modalities.

Lastly, numerous other theoretical considerations arise from the annotations, such as: Is the emerging language a pidgin or a jargon? Taking into account the reflection by [Holm 2000] on linguistic relatedness, does the linguistic proximity between Venezuelan Spanish and Brazilian Portuguese, even in contact with the Warao language, hinder the emergence of a pidgin? What is the influence of the Warao language's morphosyntax on the emergence of this language, considering linguistic attitudes and the linguistic landscape in which the refugees are situated?

References

Bhat, I., Bhat, R. A., Shrivastava, M., and Sharma, D. (2018). Universal dependency parsing for hindi-english code-switching. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, page 987–998.

Braggaar, A. and van der Goot, R. (2021). Challenges in annotating and parsing spoken, code-switched, Frisian-Dutch data. In *Proceedings of the Second Workshop on Do-*

main Adaptation for NLP, pages 50–58, Kyiv, Ukraine. Association for Computational Linguistics.

Buzato, D. and Vital, Á. (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.

Caron, B., Courtin, M., Gerdes, K., and Kahane, S. (2019). A surface-syntactic ud treebank for naija. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 13–24. Association for Computational Linguistics.

Çetinoğlu, Ö. and Çöltekin, Ç. (2019). Challenges of annotating a code-switching treebank. In *Proceedings of the 18th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2019)*, pages 82–90, Paris, France. Association for Computational Linguistics.

Crystal, D. (1987). The cambridge encyclopedia of language. *UK: Cambridge University*.

Duran, M. S. (2021). Manual de anotação de relações de dependência: Orientações para anotação de relações de dependência sintática em língua portuguesa, seguindo as diretrizes da abordagem universal dependencies (ud). Technical Report ICMC 435, Instituto de Ciências Matemáticas e de Computação, Universidade de São Paulo, São Carlos-SP.

GarcÍa-Castro, Á. (2006). Migración de indígenas warao para formar barrios marginales en la periferia de ciudades de guayana, venezuela. *De Quito a Burgos: migraciones y ciudadanía. Burgos: Gran Vía*.

Holm, J. (2000). *An introduction to pidgins and creoles*. Cambridge University Press.

Luft, C. P. (2010). *Dicionário Prático de Regência Verbal: Nova Ortografia*. Ática, São Paulo, 9 edition.

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.

Nivre, J., De Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajic, J., Manning, C. D., McDonald, R., Petrov, S., Pyysalo, S., Silveira, N., 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.

Rademaker, A., Chalub, F., Real, L., Freitas, C., Bick, E., and De Paiva, V. (2017). Universal dependencies for portuguese. In *Proceedings of the fourth international conference on dependency linguistics (Depling 2017)*, pages 197–206.

Romero-Figueroa, A. (1985). Osv as the basic order in warao. *Lingua*, 66:115–134.

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

Romero-Figueroa, A. (2020). *El contacto warao-espanol: Consideraciones sobre el proceso de aculturacion lexica de la lengua nativa del delta del Orinoco*. Editorial Académica Española.

Seddah, D., Essaidi, F., Fethi, A., Futeral, M., Muller, B., Suarez, P. O., Sagot, B., and Srivastava, A. (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.

Straka, M., Hajic, J., and Straková, J. (2016). Udpipe: trainable pipeline for processing conll-u files performing tokenization, morphological analysis, pos tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297.

Taulé, M., Martí, M. A., and Recasens, M. (2008). Ancora: Multilevel annotated corpora for catalan and spanish. In *Lrec*, volume 2008, pages 96–101.

UNHCR (2021a). Os warao no brasil - contribuições da antropologia para a proteção de indígenas refugiados e migrantes. Technical report, Brasília.

UNHCR (2021b). Perfil socioeconômico da população indígena refugiada e migrante abrigada em roraima. Technical report, Brasília.