# A Simple Audio and Text Collection-Annotation Tool Targeted to Brazilian Indigenous Language Native Speakers

**Gustavo Padilha Polleti**
Universidade de São Paulo
gustavo.polleti@gmail.com

**Fabio G. Cozman**
Universidade de São Paulo

**Fabricio Gerardi**
Universität Tübingen

## Abstract

In this paper we present an audio and text annotation tool for indigenous languages with focus on native speakers, initially developed for Brazilian indigenous languages. Our tool simplifies the process of language resource annotation and employs gamefication techniques typically found in language learning games. Then we describe the annotation tool and present preliminary results for the Bororo language. We discuss the limitations of our tool, highlighting ethical and practical implementation concerns.

## 1 Introduction

Audio and text annotation tools are key for documenting and building resources for endangered languages (Brugman and Russel, 2004). Existing tools are mostly designed for linguistic professionals and focus on formal description of language resources, such as dependency treebanks and lexical databases. While such tools are fundamental for properly documenting languages, only linguist experts can operate them, and they remain often unknwon outside academia. Hence, despite the pressing need for annotated corpora, language annotation tools remain costly and dependent on scarcely available experts. Annotation tools, in their current form, can hardly scale to address the 2,680 languages at risk of extinction by the end of this century (Wurm, 2001; Lewis, 2009).

Furthermore, ethical and practical concerns arise when we consider that experts who operate language annotation tools are often not members of the indigenous communities themselves (Pinhanez et al., 2023). It is hard to ensure that data annotation procedures are compliant with ethical guidelines (Lewis et al., 2020), such as the Los Pinos Declaration [1], or even that annotations are validated by actual indigenous speakers.

We argue that next-generation tools should be designed for use by lay indigenous speakers to accelerate the data collection and annotation process. While there are few linguist experts, indigenous communities are large. In particular, Brazil is home to a significant number of languages. For example, the Xavante language population alone represents more than 27,000 people. These languages are collectively referred to here as Brazilian Indigenous Languages (BILs). In spite of the high number of languages spoken in Brazil (estimated around 180, see (glo, 2024)), this number is declining fast as populations age and many languages are not learned by younger generations.

In this work, we propose and implement an initial language annotation tool that can be used directly by native speakers in indigenous communities without expert linguistic knowledge. Our proposal simplifies the annotation process so as to only collect words in audio and written text format. Our tool allows indigenous speakers to annotate words with their own speech, perform translations and associate morphemes to word tokens. The main goal is to achieve a source dataset of paired instances, which doe not require further work to develop dependency treebanks, natural language processing tools, and other resources.

We employ a gamification-based design (Sykes, 2018) to maximize engagement among native speakers, encouraging them to produce a high volume of annotations in the shortest possible time. Recognizing the limited availability of indigenous community members, we prioritize a highly user-friendly interface to ensure accessibility and ease of use.

We guide speakers/users through the annotation process by specifying the target word and direct-

[1] https://unesdoc.unesco.org/ark:/48223/pf0000374030

March 2, 2025 ©2025 Association for Computational Linguistics

ing their input to an internal speech recognition component, which transcribes the audio into written text. This transcription includes preliminary annotations, such as morphological information and translations. Speakers can then review, refine, and confirm the prefilled text and annotations before proceeding to the next word.

To enhance usability and minimize friction, we integrate automated annotation components, such as speech recognition. We also address challenges associated with limited computational resources. In our prototype, we employ lightweight models and heuristics that can run offline in a standard web browser or mobile app. Finally, we present preliminary results for the Bororo language as a proof of concept.

The paper is organized as follows. Section 2 describes our annotation tool design and its development, including data sources and methods. Section 3 presents preliminary results for the Bororo language. Section 4 discusses the challenges and limitations of our prototype and offers concluding remarks.

## 2 Methodology

Our data collection and annotation tool aims to empower native speaker communities to collect and annotate language resources by themselves without requiring expert linguistic knowledge. Our tool takes the form of a game, similar to formats often found in language learning game apps from both industry (e.g. Duolingo) and the literature (Polleti, 2024; von Ahn, 2006; Katinskaia et al., 2017).

The tool follows a linear progression structure, where the user advances by completing units. In order to do so, the user is asked to annotate a series of specific words, similar to language exercises. In the annotation screen, depicted in Figure 1 (top), the user is asked to provide speech audio translation in native language for a given Portuguese word. In the figure, the tool asks for a speech translation of the Portuguese word for jaguar, "Onça Pintada", to the Bororo language. First, the user records their speech in native language. The user should say the given word only once within 10 seconds. After the audio recording finishes, we run a speech recognition model to generate a transcript. In our Bororo language example, we have a Bororo-Portuguese dictionary available (Ferraz Gerardi; Polleti et al., 2024), thus, we know in advance that the target word, or the Bororo translation for "onça ointada", is "adugo". However, there are many alternative ortographies or even regional synonyms that can be absent in our knowledge base. In order to avoid enforcing a specific ortography by presenting the target word beforehand, we allow the user to freely annotate so as to preserve linguistic diversity. Next, we check whether the produced transcript matches the target word from the dictionary entry. If a match is found (Figure 1a), we retrieve an image representing the word concept, the written word in native language and its description from the dictionary entry. Finally, the user can make editions if necessary (such as providing an alternative orthography), confirm changes and move on to the next. On the other hand, if we cannot assert that the transcript matches the target word (Figure 1b), the user is required to fill up the written translation and description manually before moving to the next. The tool may fail to properly identify a match by several reasons; for example, the speech recognition may fail, the dictionary may be incomplete or may not contain all synonyms or simply the user translation may be incorrect. We allow the user to retry recording the speech translation multiple times, so if the speech recognition fails due to background noises, computer glitches or any other intermitent issues, it can succeed in a second attempt. If the matching keeps failing even after multiple retries, users can always fill the written translation and description manually. We provide autocomplete options based on lexical similarity to speed up the manual filling process. Additionally, we also provide an option for the user to skip the current annotation and move to the next. For example, if the user does not know the translation for the given word, we want to save time and avoid incorrect annotations by giving them the option to immediately move on to the next. The whole annotation process is depicted in Figure 2.

Now we focus on the speech recognition model and on the word matching heuristic. We propose to reuse speech recognition models that were trained for other languages to be used for low-resource languages. In our proof of concept for the Bororo language, we employed the Web Speech API's Speech Recognition model for Brazilian Portuguese (pt-BR), which can run offline and is available in most web browsers (e.g. Chrome, Edge, Safari, except Firefox). Back to our example, note that “Adugo” is a romanized word. Most writing systems for brazilian indigenous languages were romanized with strong Portuguese language influence, Bororo language included. We observed that the speech recognition model often produced transcripts of portuguese words that are phonetically similar the original Bororo word. For example, the transcript for the word “Adugo” results in “Adubo”, which is a portuguese word with completely different meaning but phonetically similar to the Bororo word. Since Bororo writing system is romanized, we could perform a lexical similarity search between the portuguese transcript and the known Bororo vocabulary to find good match candidates. Additionally, since we know the target word, we can consider a match if the target word has high lexical similarity to the transcript. In our prototype, we built a similarity score based on levenshtein distance and applied an arbitrary 0.9 threshold as the heuristic criteria to tell whether the speech to text process matches or not the target word. Table 1 presents some examples from our prototype. Despite minor spelling issues, for our few examples, we can observe that the portuguese speech recognition model is able to produce phonetically similar transcripts for Bororo words, which can produce accurate matches when coupled to our heuristic.

We define our similarity score as:

$$1-(distance(a,b)/(length(a)+length(b))+\epsilon),$$

where $a$ and $b$ are the target word and transcript, respectively, $distance$ refers to the weighted Levenshtein distance function, $length$ s returns the total number of characters in a string and $\epsilon$ is a hyperparameter that smoothes the similarity score for small words. We observed that our similarity score is often too strict when comparing small sized strings. To avoid missing potential matches, we introduced $\epsilon$ to smoothen the distance metric for small strings. In our prototype we arbitrarily used $\epsilon = 3$. To illustrate, consider the words “caro” and “karo”: they are both very similar, their Levenshtein distance is only 1, but our similarity score would yield only 0.875 if we did not take $\epsilon$ into account. Additionally, we apply NFD unicode normalization form in the transcript string before calculating the similarity score.

## 3 Results

We still need to evaluate our proposal more broadly with the Bororo indigenous community to measure community adoption and engagement. This will require a more comprehensive evaluation of our processes and methods to measure, for example, how effective the speech recognition model is in speeding up the annotation process. At this point, we ran simulated experiments to get preliminary results on: (1) the speech to text recall, (2) how much time the speech to text saves in the annotation process, i.e. the speed up. First, to measure recall, we sampled 50 words from the Bororo dictionary, generated correct speech audio for them and ran a simulation to evaluate how many instances our speech to text process was able to find a match, the fraction of matches over the total number of instances is what we refer to as recall. We obtained 0.56 recall, 28 matches out of 50 words, as presented in Table 2. Next, we got all the words we were able to find a match and asked a volunteer from our University to use the annotation tool, first with the speech to text support and later without it, filling all the information manually. Given that the volunteer is not a native speaker, he had access to the target words and their descriptions during the experiment. We compared the completion times between the volunteer filling it with and without speech to text support to get preliminary insight into the annotation speed up. The volunteer took 3 minutes and 12 seconds to complete the annotation of 28 words, compared to 4 minutes and 33 seconds without speech to text support. We obtained 29.7% speed up, saving around 1 minute in our experiment setup, as presented in Table 3.

Table 2: Speech to text simulation metrics.

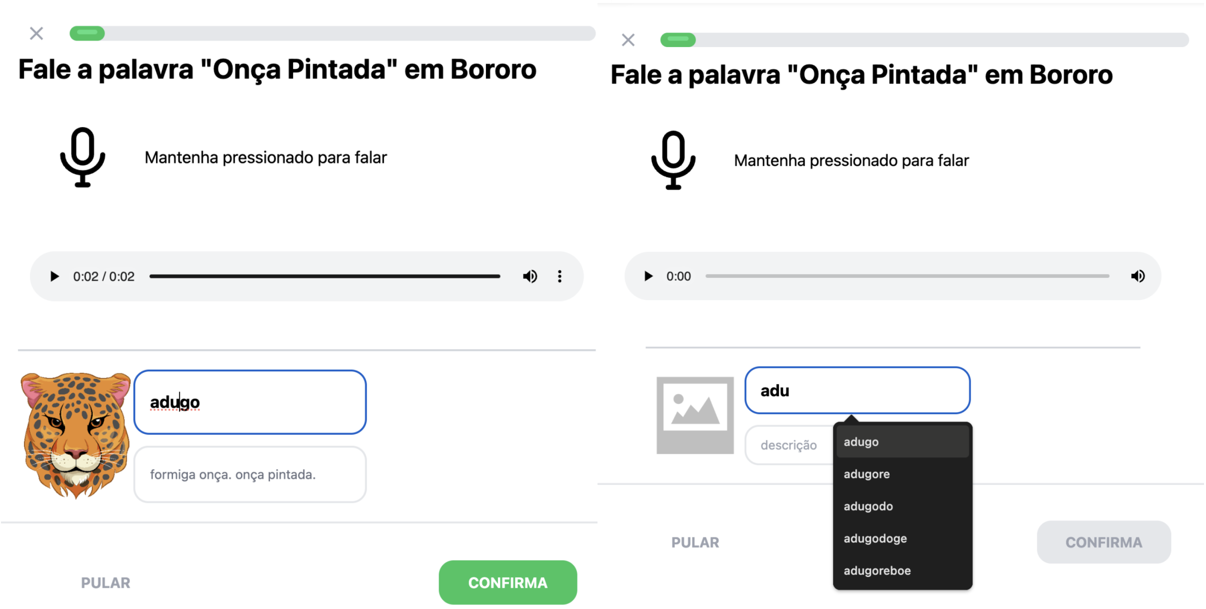
| **Metric** | **Result** |
|---|---|
| Recall | 56% (28 matches out of 50) |
| No transcript | 2% (1 out of 50) |

Table 3: Completion time results.

| **Scenario** | **Total Completion time** |
|---|---|
| without Speech to Text | 273 secs (4 min 33 secs) |
| with Speech to Text | 192 secs (3 min 12 secs) |
| Relative Speed Up | 29.7% |

Table 1: Bororo speech to text examples. The target word is highlighted in the matching candidates.

| Target word (en) | Target word (native) | Transcript (pt) | Match Candidates |
|---|---|---|---|
| jaguar | adugo | adubo | **adugo**, arugo, atugo |
| rain | bubutu | bubu tu | **bubutu** |
| scarlet macaw | nabure | naburi | **nabure** |
| howler monkey | pai | pai | **pai** |
| woman | aredy | aredo | aredo, taredo, **aredy**, arego, arudo, arudu |
| wart | akogo | acogo | **akogo**, apogo, arogo, ecogo |
| fish | karo | caro | **karo**, ocaro, care, caru |
| eye | joku | jogo | jodo, jomo, joto, jugo, joga |
| anteater | apogo | apogo | **apogo**, apogoe, apodo, akogo |
| seed bug | arogo | arrogo | **arogo** |
| potato | tadari | padari | padaro, **tadari** |
| nose | eno | (no transcript) | (no match) |
| dog | arigao | arigato | **arigao** |
| banana | bako | barco | (no match) |
| grandmother | marugo | marugo | **marugo** |


Fale a palavra "Onça Pintada" em Bororo

Mantenha pressionado para falar

adugo

formiga onça. onça pintada.

Fale a palavra "Onça Pintada" em Bororo

Mantenha pressionado para falar

adu

descrição

(a) Successful speech recognition and information retrieval. The transcript identified the word "adugo" and retrieved the associated jaguar image and description.

(b) Failed speech recognition and information retrieval. The transcript failed to indentify a matching word so the user was required to fill manually.

Figure 1: Example of a single session: the user was asked to record the translation in Bororo for the word "jaguar". It depicts autocomplete success and failure scenarios.

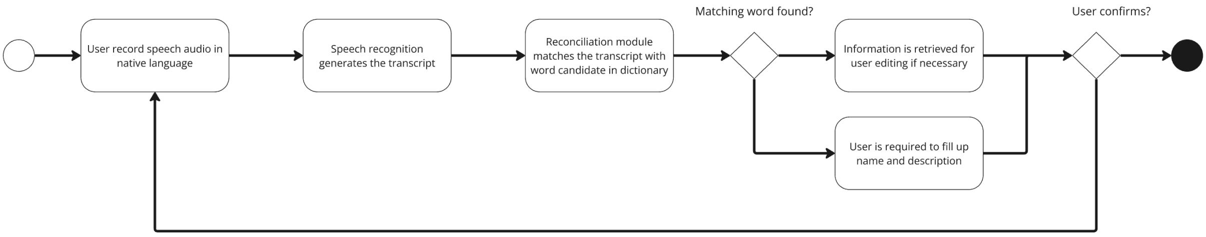

User record speech audio in native language
Speech recognition generates the transcript
Reconciliation module matches the transcript with word candidate in dictionary
Matching word found?
Information is retrieved for user editing if necessary
User is required to fill up name and description
User confirms?

Figure 2: Annotation process diagram.

## 4 Concluding Remarks & Limitations

The annotation tool introduced in this work represents a significant step forward in the advancement of resources for Brazilian indigenous languages. Our proposed design allows native speakers, who do not necessarily require specific linguistic knowledge, to perform annotations in audio and text resources. Our design avoids biases towards specific ortographies by allowing the user to freely annotate their speech and written forms. At the same time, we incorporate speech to text and autocomplete components to speed up the annotation process.

Despite the promising benefits, our prototype falls short in multiple aspects that we now examine. First, our tool currently only supports word annotation. We consider it to be a natural step to evolve our methods to enable sentence annotation. Before we can support sentences, we must require word annotation to be fully functional, which implies better autocomplete and speech recognition capabilities. Additionally, users annotations can vary significantly and we still do not have a proper process to create consensus around them. The orthography currently used by the Bororo people was developed by Catholic missionaries and is not well-suited to their language (see Colbacchini 1925 and Colbacchini 1942). Recent publications have adopted a different orthography, which occasionally leads to minor discrepancies. For example, we have two ortographies for the word "rain" in Bororo, which are "Bubutu" (old) and "Bybyty" (new). If our tool presents "Bubutu" to users, they may be confused as our tool is incentivizing an outdated ortography. Once the Bororo Corpus (Ferraz Gerardi et al., 2024) is completed, this issue is expected to be resolved, as all sources will be unified under a standardized orthography.

One significant issue stems from the fact that Bororo territories are not contiguous, resulting in variations in pronunciation among different regions. These differences can sometimes lead to mockery of speakers from areas where the language is less commonly spoken, as if their way of speaking were "incorrect." This poses an important ethical concern, as it may cause speakers to feel that a new orthography privileges certain pronunciations over others. This concern becomes even more relevant when we consider that our tool employs automatic speech recognition models, which may incentivize specific accents. Given that the speech recognition models were trained in foreign languages, biases towards pronunciation similar to the Portuguese language may occur.

There is still room for improvement in our speech to text process. We considered applying more sophisticated approaches, such as accoustic models (Li et al., 2022, 2020), for zero shot speech recognition in indigenous languages, but models like those require stable internet connectivity as they are too large to run in offline devices. We are currently limited to work with models that can run in the web browser or mobile app so they can be actually used in the field. Future work should conduct evaluate varied speech to text methods and improve their performance.

At this point, we have only implemented a proof of concept for the Bororo language; thus, it is still necessary to assess how well the methods introduced in this work generalize to other languages. Endangered language revitalization requires the development of annotated resources (Miyagawa et al., 2023). We believe that our proposal can be extended to annotate languages beyond Brazilian ones. Similar strategies around phonetical similarities have already been employed in other contexts (Mæhlum and Ivanova, 2023).

Future work should evaluate the effectiveness of our annotation tool in partnership with native speakers and assert its value. We hope our preliminary research can help scaling up data annotation for endangered languages and produce rich data sources to support revitalization initiatives.

## Acknowledgments

The second author was partially supported by CNPq grant 305753/2022-3. We also thank support by CAPES Finance Code 001. The authors of this work would like to thank the Center for Artificial Intelligence (C4AI-USP) and the support from the São Paulo Research Foundation (FAPESP grant 2019/07665-4) and from the IBM Corporation.

The third author is supported by the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (Grant agreement No. 834050).

## References

2024. Glottolog 5.1. Accessed: 2024-12-15.

L. von Ahn. 2006. Games with a purpose. *Computer*, 39(6):92–94.

Hennie Brugman and Albert Russel. 2004. Annotating multi-media/multi-modal resources with ELAN. In *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal. European Language Resources Association (ELRA).

Antonio Colbacchini. 1925. *I Bororos orientali:" Orarimudoge" del Matto Grosso (Brasile)*. Società editrice internazionale.

Antonio Colbacchini. 1942. *Os Boróros orientais*. Companhia Editora Nacional, São Paulo.

Fabrício Ferraz Gerardi. *Bororo Dictionary*. Forthcoming. Available upon request.

Fabrício Ferraz Gerardi, Daniel Sollberger, and Luis Toribio Serrano. 2024. Corpus bororo (corbo) (v0.2).

Anisia Katinskaia, Javad Nouri, and Roman Yangarber. 2017. Revita: a system for language learning and supporting endangered languages. In *Proceedings of the joint workshop on NLP for Computer Assisted Language Learning and NLP for Language Acquisition*, pages 27–35, Gothenburg, Sweden. LiU Electronic Press.

Jason Edward Lewis, Angie Abdilla, Noelani Arista, Kaipulaumakaniolono Baker, Scott Benesiinaabandan, Michelle Brown, Melanie Cheung, Meredith Coleman, Ashley Cordes, Joel Davison, Kūpono Duncan, Sergio Garzon, D. Fox Harrell, Peter-Lucas Jones, Kekuhi Kealiikanakaoleohaililani, Megan Kelleher, Suzanne Kite, Olin Lagon, Jason Leigh, Maroussia Levesque, Keoni Mahelona, Caleb Moses, Isaac ('Ika'aka) Nahuewai, Kari Noe, Danielle Olson, 'Ōiwi Parker Jones, Caroline Running Wolf, Michael Running Wolf, Marlee Silva, Skawennati Fragnito, and Hēmi Whaanga. 2020. Indigenous protocol and artificial intelligence position paper. Project Report 10.11573/spectrum.library.concordia.ca.00986506, Aboriginal Territories in Cyberspace, Honolulu, HI. Edited by Jason Edward Lewis. English Language Version of "Ka?ina Hana ?Ōiwi a me ka Waihona ?Ike Hakuhia Pepa Kūlana" available at: https://spectrum.library.concordia.ca/id/eprint/990094/.

M. Paul Lewis, editor. 2009. *Ethnologue: Languages of the World*, sixteenth edition. SIL International, Dallas, TX, USA.

Xinjian Li, Siddharth Dalmia, Juncheng Li, Matthew Lee, Patrick Littell, Jiali Yao, Antonios Anastasopoulos, David R. Mortensen, Graham Neubig, Alan W Black, and Florian Metze. 2020. Universal phone recognition with a multilingual allophone system. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 8249–8253.

Xinjian Li, Florian Metze, David Mortensen, Shinji Watanabe, and Alan Black. 2022. Zero-shot learning for grapheme to phoneme conversion with language ensemble. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2106–2115, Dublin, Ireland. Association for Computational Linguistics.

Petter Mæhlum and Sardana Ivanova. 2023. Phonotactics as an aid in low resource loan word detection and morphological analysis in sakha. In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 111–120, Tórshavn, the Faroe Islands. Association for Computational Linguistics.

So Miyagawa, Kanji Kato, Miho Zlazli, Salvatore Carlino, and Seira Machida. 2023. Building Okinawan lexicon resource for language reclamation/revitalization and natural language processing tasks such as Universal Dependencies treebanking. In *Proceedings of the Second Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2023)*, pages 86–91, Tórshavn, the Faroe Islands. Association for Computational Linguistics.

Claudio S. Pinhanez, Paulo Cavalin, Marisa Vasconcelos, and Julio Nogima. 2023. Balancing social impact, opportunities, and ethical constraints of using ai in the documentation and vitalization of indigenous languages. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6174–6182. International Joint Conferences on Artificial Intelligence Organization. AI for Good.

Gustavo Polleti. 2024. Building a language-learning game for Brazilian indigenous languages: A case study. Technical report, arXiv:2403.14515.

Gustavo Polleti, Fabio Cozman, and Fabrício Gerardi. 2024. Unified knowledge-graph for brazilian indigenous languages: An educational applications perspective. In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 159–164, Porto Alegre, RS, Brasil. SBC.

Julie M Sykes. 2018. Digital games and language teaching and learning. *Foreign Language Annals*, 51(1):219–224.

S.A. Wurm. 2001. *Atlas of the world's languages in danger of disappearing*. Unesco Pub.