Brazilian Indigenous Languages Revitalization At Scale: Reducing Cost And Development Time for Low-Resource Language Courses

Gustavo Polleti[1] , **Fabricio Gerardi**[2] , **Carolina Aragon**[3] , **Fernando Bororo**[4] , **Mariel Kujiboekureu**[1] , **Wayali Salomã**[5] , **Fabio Cozman**[1]
[1]Universidade de São Paulo (USP), Brazil
[2]União das Faculdades Católicas de Mato Grosso (UNIFACC), Brazil
[3]Universidade Federal da Paraíba (UFPB), Brazil
[4]Universidade Federal de Rondonópolis (UFR), Brazil
[5]Universidade do Estado do Mato Grosso (UNEMAT), Brazil
gustavo.polleti@usp.br, gerardifabricio@gmail.com, f.kudoro@aluno.ufr.edu.br, mariel.kujiboekureu@edu.mtgov.br, fgcozman@usp.br

Abstract

Brazil is home to about 180 Indigenous languages, which span a wide range of sociolinguistic circumstances, from critically endangered to relatively stable. Across this continuum, Indigenous communities often lack pedagogical resources and are underserved by existing language-learning technologies, which are typically designed for high-resource languages and assume solid connectivity and large datasets. This research project proposes AI-assisted tools and workflows for the collection and annotation of textual and speech data that substantially reduce the time and cost required to produce engaging language-learning game apps. Our goal is to implement a language-learning game app that Indigenous students can use to practice their reading, writing and speaking skills at home. We propose novel speech processing models for low-resource Indigenous languages and offline support in low-connectivity environments. Our project adopts a co-creation model that actively foster collaboration between Indigenous educators, linguists, and youth, is adapted to the their context, and complies with ethical guidelines. We outline an implementation plan with Bororo and Enawene Nawe communities to test our methods and, potentially, produce an AI-driven platform for Indigenous language education that is applicable across diverse sociolinguistic contexts in Brazil and beyond.

1 Problem Statement and Goals

This research project aims to scale up Indigenous language education by reducing the time and cost required to produce engaging language-learning game apps that students can use to practice at home.

Game language-learning apps play a key role in Indigenous language education by supporting literacy, vocabulary acquisition, and oral practice [Thomason, 2015]. LARA [Akhlaghi *et al.*, 2019] is an example with key contributions towards preservation and educational initiatives [Rayner and Wilmoth, 2023]. However, most apps only offer support to high-resource languages and require substantial effort to support languages with few written or spoken data. As a result, course development remains slow and difficult to scale. For instance, "7000 Languages" [1], a non-profit focused on producing online courses for endangered languages, typically needs 1–2 years to develop a single course. Such timelines are incompatible with the linguistic diversity and urgency observed in contexts like Brazil, where dozens of Indigenous languages may simultaneously require educational materials adapted to different levels of vitality and transmission. Brazil hosts approximately 180 Indigenous languages [glo, 2024], collectively referred to here as Brazilian Indigenous Languages (BILs); however, these languages often display very limited transmission to younger generations.

In many Indigenous communities, opportunities for language use are concentrated within formal school settings. While students may practice their Indigenous language with a teacher during class, they often lack interlocutors at home, particularly when parents and caregivers lack fluency or where Portuguese predominates in everyday communication. Language practice restricted to the classroom environment is insufficient to maintain proficiency or confidence, especially where the other national language is dominant in daily life [Hinton, 2011; Grenoble and Whaley, 2006]. Tools that enable home-based practice, strengthen learning, normalize the presence of the language in everyday routines, and support students even when there are few or no fluent speakers at home [King, 2001; Nicholas, 2019]. Our goal is to produce an interactive language app that the student can use to practice reading, writing and speaking in their native language at home.

AI research can help producing apps with embedded speech processing capabilities, enabling students to practice their speech at home. Industry leaders on language learn-

[1]https://www.7000.org/

ing apps (Duolingo [2] and Busuu [3]) already successfully employ speech recognition models to create exercises that foster speaking fluency. However, such apps are restricted to high-resource languages, require stable internet connectivity, and do not meet privacy requirements [Polleti, 2024].

In this paper, we outline our plan to produce a language learning game app with speaking exercises for Brazilian Indigenous languages. First, in Section 3, we describe the technical challenges that currently prevent this solution to exist, notably the high cost to produce reliable language resources and the limitations of current on-device speech technology. We propose a novel AI-assisted process that address the most expensive data collection bottlenecks by enabling Indigenous speakers to annotate the data themselves while automating part of the manual work typically done by linguistic experts; see Section 4.1. We will employ the produced multimodal datasets to develop lightweight speech and language models that can fit in accessible smartphones; see Sections 4.2 and 4.3. We will use these models to power dynamic reading, writing and speaking exercises within a game app; see Section 4.4. We will prove our proposal by producing novel multimodal datasets, with texts and audio recordings, for two Brazilian Indigenous languages, Bororo and Enawene Nawe; see Section 5. The game app and their components will be evaluated in both simulated and real-world scenarios, as detailed in Section 6. We aim at releasing the game app as a tool that Indigenous professors can use to enhance the learning experience of their students. Section 7 discusses ethical considerations and limitations of our research and, finally, Section 8 outline the implementation timeline.

2 Target SDGs and the Societal Impact

This research project aligns with three UN Agenda 2030 Sustainable Development Goals (SDGs) [4]: 4 (Quality Education), 10 (Reduced Inequalities), and 16 (Peace, Justice, and Strong Institutions). It also advances the aims of the UNESCO International Decade of Indigenous Languages (2022–2032), which prioritizes the preservation and promotion of Indigenous languages. Additionally, the UN Permanent Forum on Indigenous Issues (PFII) emphasizes that "... protecting indigenous languages and safeguarding traditional knowledge is crucial for addressing climate change and other global challenges." [5]

3 Related Work

3.1 Brazilian Indigenous Languages Resources

Creating pedagogical material for Indigenous languages is particularly challenging because available spoken and written resources may be both scarce and dispersed [Pinhanez *et al.*, 2023]. To illustrate, in Brazil, for some indigenous languages, such as Guajajara, Asurini, and Bororo, dictionaries are only recently available [Harrison and Harrison, 2013; Ferraz Gerardi,]. For other languages, treebanks are available through the Universal Dependencies Project (UD) [Nivre *et al.*, 2020], though they vary in length and quality. Some languages, however, have only a handful of miscellaneous resources [Monserrat, 2000]. Building a course for languages with few spoken and written resources typically requires significant effort on (often manual) data collection.

Early stage initiatives to consolidate Brazilian Indigenous language resources, TuLeD [Gerardi *et al.*, 2022b], a lexical database, the TuDeT [Gerardi *et al.*, 2022a] dependency treebanks, and the associated linked-data BILGraph [Polleti *et al.*, 2024] have already played a key role in enabling language-learning applications for Indigenous communities [Polleti, 2024]. The release of the Bororo Corpus [Ferraz Gerardi and Toribio Serrano, 2025], aligned with its corresponding UD treebank [Ferraz Gerardi, 2024], has further supported the development of computational tools for educational and online materials. However, all the aforementioned databases were assembled from heterogeneous secondary sources rather than through a structured data-collection process, except for the Bororo language. As a result, the existing datasets are highly incomplete, particularly in terms of dependency tree coverage paired with Portuguese translations. Because much of the source material used in TuDeT originates from foreign research groups, most translations are available only in English. This lack of Portuguese translations limits their applicability for pedagogical applications intended for actual Brazilian users. A further limitation lies in the scarcity of multimodal resources – speech recordings on Indigenous languages are even more scarce than written resources. Existing approaches frequently rely on Christian religious texts, such as the New Testament [Pratap *et al.*, 2024]. While these sources are readily available, the use of religious content, often produced by external institutions, fails to reflect Indigenous epistemologies, worldviews, and culture. The reliance on foreign religious materials underscores the urgent need for speech datasets that are created, curated, and governed by Indigenous communities themselves.

We argue that the current data collection and annotation process is limited largely because existing tools for audio and text annotation are designed for linguistic professionals and focus only on formal description of language resources, such as lexical databases and dependency treebanks [Brugman and Russel, 2004]. While such tools are key for properly documenting languages, they remain costly and slow in producing annotated corpora due to their dependency on scarcely available experts. Furthermore, ethical and practical concerns arise from the fact that experts operating annotation tools are often external to the Indigenous communities whose languages they are working on [Pinhanez *et al.*, 2023]. This creates substantial challenges in ensuring that data collection and annotation procedures are compliant with established ethical frameworks [Lewis *et al.*, 2020], such the Los Pinos Declaration [6]. It is also hard to ensure that linguistic annotations are properly reviewed by actual Indigenous speakers.

[2] https://www.duolingo.com/

[3] https://www.busuu.com/

[4] https://www.un.org/sustainabledevelopment/

[5] https://sdg.iisd.org/news/indigenous-peoples-have-a-crucial-role-in-implementing-sdg-16-concludes-permanent-forum/

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

3.2 On-Device Speech Models

Typical AI-assisted speaking exercises consist of asking the student to speak a given word or sentence in their native language and then use automated analysis of the recorded audio to provide immediate feedback. Even if we assume that we are able to collect and annotate enough data to finetune and train multilingual speech processing models for Indigenous languages, the current architectures would still be difficult to integrate into a mobile app that can be properly used in practice [Sharma *et al.*, 2025]. Although there are speech processing models with broad multilingual support, for both speech-to-text (STT) and text-to-speech (TTS) tasks [Pratap *et al.*, 2024], they are poorly suited to many Indigenous contexts because they assume stable connectivity and sufficient computational resources [Pinhanez *et al.*, 2023]. Communities located in remote or forested regions often lack reliable Internet access and consistent electricity supply, making cloud-dependent applications unfeasible in practice. Most existing AI tools cannot operate offline, as they rely on large models that require server-side processing and cannot be executed on low-power mobile devices commonly available in these settings. Even versions of SOTA speech recognition models designed for on device inference have more than 150MB in size [Radford *et al.*, 2023], which is impractical for lite mobile apps in low-end smartphones.

We plan to employ techniques to optimize model size so we can have multilingual speech processing models that are suited for lite mobile apps, size under 10MB. We focus on three complementary strategies: model pruning and quantization [Jacob *et al.*, 2017]. Recent work on compressed STT models shows that combined techniques can reduce model size while maintaining accuracy [Sainath *et al.*, 2020].

4 Proposed Strategies and Methods

In this section we describe the components of our game app.

4.1 AI-assisted Speech Collection and Annotation

We propose that next-generation tools should be speech-oriented and designed for use by lay Indigenous speakers to accelerate the data collection and annotation process. We claim that AI-assisted tools can simplify the annotation process so that a native speaker without expert linguistic knowledge can annotate words with their own speech, perform translations, and associate morphemes to word tokens. The goal is to enable the Indigenous communities to produce and govern their own language datasets, removing bottlenecks associated with linguistic experts, manual time-consuming work, and the lack of dataset standards. Preliminary work shows that autocompletion by simple speech-to-text models can speed up the data annotation process by 29.7% [Holdt *et al.*, 2025]. We plan on evolving speech-to-text and text-to-speech models for low-resource languages, autocompletion, and other methods to produce AI-assisted tooling for data annotation and collection. Our preliminary proposal, based on previous works [Holdt *et al.*, 2025], consists of a website/mobile-app where the native speaker is requested to annotate a word or sentence, one unit at a time. For each word or sentence, the speaker should record the utterance and written form. For example, the tool might display a written word in Portuguese and ask for its translation in the native language, or vice-versa. We propose autocompletion tools where we already display the fields prefilled and only require the native speaker to review and correct orthographies or record a new audio if needed. Our initial proposal relies on the engagement of community members, but we will actively search for funding to improve the projects vitality.

4.2 Offline Speech Processing Models

Our proposed method to reduce model size consists of: (1) applying quantization in an existing speech processing model, (2) pruning the model while optimizing for the speech exercise task. Consider we want to use a STT model to correct speaking exercises. We have collected and annotated audio recordings for several words in an Indigenous language. We may have enough data to train or finetune a STT model for the given Indigenous language, but the resulting model may be too heavy to run on device. One naive way to implement speech exercises is to transcribe student utterances using a STT model and checking if the transcript matches a target word. Unlike an actual speech recognition task, our speaking exercise is a closed world task because we know the target word beforehand. In a sense, our speaking exercise is close to a similarity problem as it evaluates whether a student utterance for a target word is similar enough to its respective gold truth audio recording. In that context, we are free to choose the best representation to perform the comparison. In our naive example, this representation is the audio transcript provided by the STT model, however, this may not even capture subtle speech attributes such as pronunciation [El Kheir *et al.*, 2023]. Embeddings from intermediary layers of the STT model can produce latent representation spaces more suited for our speaking exercises. Figure 1 illustrates our proposal. First, we prune a STT model Θ up to a predefined layer N,whether or not the model has been fine-tuned for the target Indigenous language. We then run inference with the pruned model to obtain embedding vectors O and $\hat{O}$ for the gold truth and student recording, respectively. Finally, we compute a similarity score between these embeddings using a predefined metric m, and classify the student's response as correct or incorrect based on whether this score exceeds a selected threshold ϵ. The hyperparameters Θ, N, m and ϵ can be tuned using our gold truth recordings. Our pruning proposal can be effective in producing speech processing models small enough to enable offline speech exercises. It is also worth noting that our approach is language-agnostic, so that a pruned STT model can be reused for languages it has never been trained or finetuned for. As we propose to simply compare audio latent representations instead of actual transcripts, the same model can potentially be reused for previously unseen languages. In this research project, we are committed to deliver: (1) an open source python package to implement our pruning method, including the tuning of its hyperparameters, (2) pruned models ready to be used for STT and TTS tasks, (3) a comprehensive evaluation on a real Brazilian Indigenous languages benchmark.

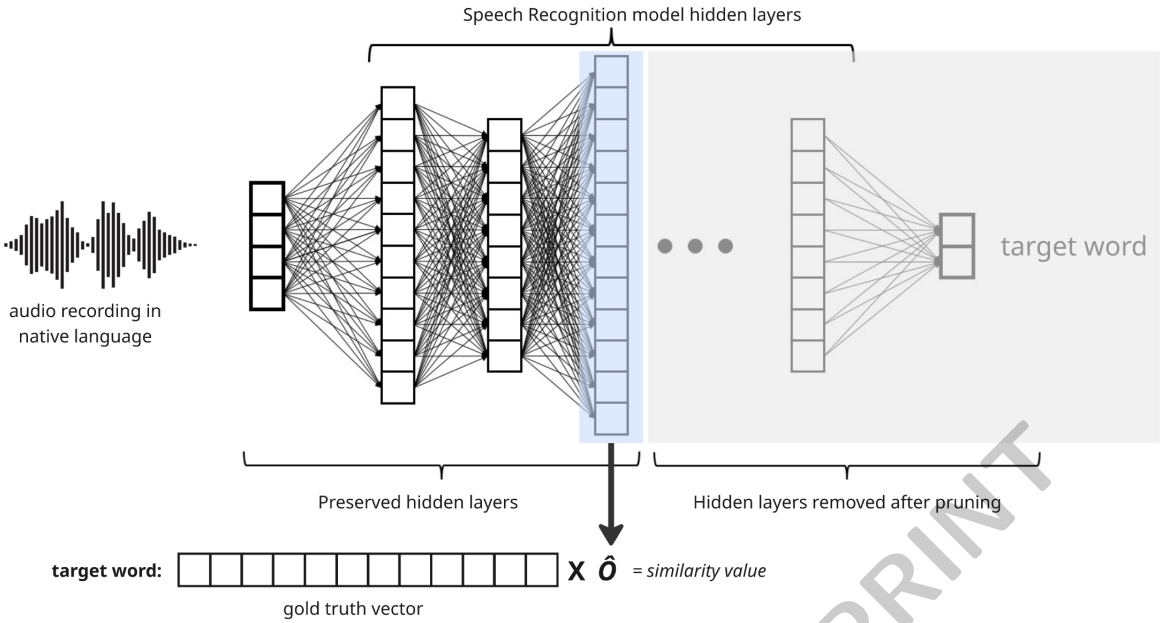


Figure 1: Pruning method proposal for Speaking exercises. Original Speech-to-Text model is pruned up until an intermediary layer where similarity is measured between gold truth and student recording latent representations.

4.3 Micro Language Models

This project investigates the development of "micro" language models to operate under severe data scarcity while supporting pedagogical content generation for Indigenous language education. Unlike general-purpose language understanding, these models are optimized for grammatical control. In practice, these micro language models adopt lightweight architectures augmented with explicit linguistic constraints in order to remain data-efficient, interpretable, and deployable in low-resource educational contexts. Unlike large language models trained on web-scale data, these micro language models are trained on carefully curated resources and are constrained by grammatical objectives, since it would be unrealistic to proceed otherwise.

The micro language models are trained on the datasets produced by our proposed collection and annotation workflow. Audio recordings are transcribed and normalized, after which textual data are systematically annotated using the Universal Dependencies (UD) framework [Nivre *et al.*, 2016; de Marneffe *et al.*, 2021]. UD annotations provide rich grammatical representations, such as dependency structure and part-of-speech categories, that can potentially compensate for limited corpus size.

In addition to annotated corpora, the training pipeline integrates structured lexical resources derived from community-developed dictionaries. Lexical entries supply canonical forms, morphological segmentation and usage examples, which are aligned with corpus instances. This lexicon–corpus alignment supports more robust generalization by exposing the model to paradigmatic relationships that are otherwise underrepresented in small datasets. These structured lexical and grammatical resources can be viewed as forming a lightweight, language-specific knowledge graph that complements the micro language model by encoding explicit relations among forms, meanings, and grammatical functions.

The resulting training data thus combine surface text, UD syntactic structure, and lexically grounded morphological information. We argue these representations can introduce inductive biases capable of guiding model learning toward linguistically valid patterns even in low-resource NLP scenarios [Bisazza and others, 2020]. As new annotated materials become available, we can iterate and evolve the model in line with human-in-the-loop and incremental learning paradigms [Settles, 2012].

Once trained, the micro language models are used for controlled, grammar-aware generation of pedagogical materials [Reiter and Dale, 2000; Wiseman *et al.*, 2018]. Given an input sentence, the model can generate diverse variants by modifying tense or person marking while preserving grammatical well-formedness. This capability enables the automatic exercise creation, thus increasing diversity and supporting adaptive learning paths.

Generated materials follow a grammar-aware sequencing strategy, in which grammatical constructions are introduced incrementally and reused systematically in accordance with the instructional plan. Content generation is further constrained by culturally informed vocabularies and narrative patterns defined by Indigenous educators. Generated materials are sequenced according to a grammar-aware curriculum, in which constructions are introduced gradually and in alignment with instructional goals. This design enables adaptive curriculum generation, in which the selection and ordering of generated materials respond to learners' progress while respecting the grammar-aware sequencing.

4.4 Bringing It All Together: The Game App

The proposed language resources and speech modeling techniques will be integrated into a language learning game app for Indigenous students to practice reading, writing, and speaking in their native language at home. The research to produce this game app is inherently interdisciplinary and will require iterative co-design sessions engaging various stakeholders such as Indigenous professors, linguistic experts, engineers, Indigenous students and community leaders. The purpose of our game app is to complement educational activities within and outside school. The game app will be fully operated and governed by the Indigenous professors. The professor will define the curriculum, decide which exercises will be presented, their content and all other aspects related to the student learning journey. Our goal is to empower the professor with an additional tool they can use to engage their students to practice Indigenous languages during classes and at home. We will conduct this research closely with Indigenous communities to co-create the game app and deliver the most appropriate experience for their specific context.

We adopt a gamified language-learning structure based on Bilingo [Polleti, 2024], inspired by industry applications such as Duolingo. Students progress linearly through the course by completing lessons, each containing three exercise types: speaking (SP), concept match (CM), and sentence translation (TS), as illustrated in Figure 2. In TS exercises (Figure 2c), students receive a sentence in Portuguese and must select the correct sequence of Indigenous-language tokens to form the translation. CM exercises (Figure 2b) present an Indigenous-language word alongside images representing candidate concepts, and the learner selects the matching image. In SP exercises, students speak a target word, and our proposed speech processing model evaluates whether the utterance corresponds to the expected pronunciation. We reinstate this is a preliminary proposed design, which is subject to modifications as we iterate with Indigenous professors during the research.

5 Bororo and Enawene Nawe: Case Studies

5.1 Bororo

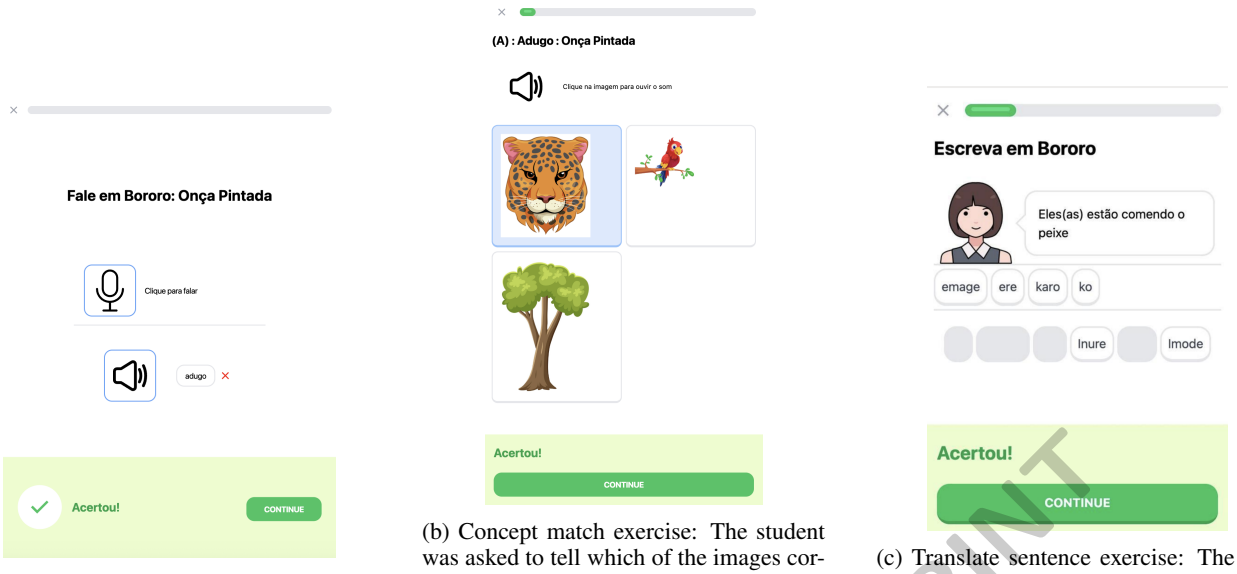
Most existing linguistic documentation of Bororo was produced by Christian missionaries rather than academically trained linguists. The Salesian Mission of Mato Grosso has maintained continuous contact with Bororo communities since 1901, producing dictionaries, grammatical descriptions, and pedagogical materials that remain foundational for contemporary research [Colbacchini, 1925; Cibaeikare and Etua, 1990; Albisetti and Ravagnani, 1992; Camargo, 2011]. Additional materials were produced in the late 1970s by missionaries associated with the Summer Institute of Linguistics (SIL). Despite their overall quality and importance, little new linguistic work of comparable scope was produced in subsequent decades.

This project relationship with the Bororo community began approximately three years ago in Meruri and has since expanded to the Córrego Grande and Tadarimana territories. It has already produced a publicly accessible annotated corpus, an interactive multimedia dictionary, and a Bororo dependency treebank. All resources are developed collaboratively with Bororo educators and community members, with Indigenous authorship and participation as central principles. The Bororo population is estimated at approximately 3,500 individuals, but intergenerational transmission has been severely disrupted. Fluent speakers are predominantly over the age of 40, and daily use of Bororo is limited; nevertheless, the language remains active in ritual contexts, formal education, and some family settings, making revitalization efforts both urgent and feasible. In partnership with the Mato Grosso State Department of Education (SEDUC-MT), the project has already conducted multiple fully funded training workshops for Bororo teachers from different territories, fostering technical capacity-building, inter-community exchange, and a collaborative network of educators engaged in language revitalization. We plan on expanding the scope of this research with the Bororo community to increase the annotated corpora volume, add audio recordings and, thus, using these resources to produce a functional language learning game app for the Bororo language. Our game app prototypes received positive feedback from Bororo professors but require accuracy improvements, notably for speaking exercises, before they can be effective tested in a real educational setting.

5.2 Enawene Nawe

The Enawene Nawe are a recent contact group that lacks both a grammatical description and a formal orthography. The writing system currently in use is neither consensual nor sufficiently representative of the language's phonemic inventory. Existing materials on Paresi represent an important baseline reference due to the close connection between the two languages. These resources have been key for starting the construction of the Enawene Nawe treebank, which is expected to be incorporated into the Universal Dependencies in an upcoming release as part of the data collection and annotation described in this research project. The Enawene Nawe people often only speak fluently their native language and display limited Portuguese proficiency, which is mostly concentrated among men. This configuration contrasts with the Bororo context and requires distinct strategies.

Our research project draws on the experience accumulated in the Bororo case, particularly regarding collaborative corpus construction, community participation, and the integration of language documentation with educational technologies. However, it is recognized that the Enawene Nawe context requires substantial adaptations, both methodological and ethical. Given the near absence of prior linguistic materials, the current effort represents the first systematic documentation of the language. The Enawene Nawe case is a sensitive and consequential example in which linguistics and computational tool development converge from the beggining. Our approach emphasizes participatory processes and respect for local sociocultural dynamics, which can become a reference for future works for recently contacted Indigenous people. We hope, our pioneering work helps to bring visibility, autonomy, and value to a language that has been largely overlooked in previous research.



(a) Speech exercise: The student was asked to speak "jaguar" in Bororo.

(b) Concept match exercise: The student was asked to tell which of the images corresponded to the Bororo word "adugo" that means jaguar.

(c) Translate sentence exercise: The student was asked to sort the words in the correct order to build the sentence.

Figure 2: Exercise types in our learning language tool. The images are simple toy examples to illustrate the tool format. For actual usage, we will select images that are better suited for the Indigenous context.

6 Expected Results and Their Evaluation

The main results and contributions of this research include but are not limited to: (1) the most comprehensive and high quality annotated corpora for Bororo and Enawene Nawe languages, (2) accessible AI-assisted tooling for data collection and annotation, (3) novel techniques and open source models for speech processing of low-resource languages under power and connectivity constraints, (4) game app for students to practice at home and, of course, (5) methods, processes and tools that can be replicated for other endangered languages in Brazil and abroad.

We hope our research can turn the tide in favour of endangered language revitalization worldwide. The Brazilian indigenous language landscape is challenging and diverse, sovwe believe that our study can help communities (with similar constraints in terms of resources scarcity, poor connectivity, electrical supply and sensitive ethical concerns) to develop successful revitalization programs. We are committed to deliver an accessible, open source platform that will reduce both cost and time to produce a language learning game app where students can practice reading, writing and speaking skills. We also expect that our research will reduce power and connectivity requirements for language models. The models and techniques developed here can lead to on-device language models that will power novel AI applications that benefit from offline support. Furthermore, we expect to promote multilingual research by proposing language-agnostic techniques. Finally, we expect to be a reference on interdisciplinary AI research co-created in partnership with Indigenous communities. We hope the our collaboration mechanisms, learnings and guidelines can build trust between the AI research and Indigenous communities so that we can strength our partnership and mutually benefit from it.

To evaluate our proposed methods we consider both objective and qualitative metrics. First, regarding the data collection and annotation process, we will assess the dataset coverage and quality. Since annotated corpora may be incomplete or unevenly distributed, we will measure coverage as the fraction of words and sentences missing one or more of the expected annotation types (e.g., audio, written text, translation). To assess annotation quality, we employ two complementary strategies. First, we conduct inter-annotator agreement, or consensus, analyses. The consensus between annotators is measured using traditional metrics k [Cohen, 1960]. Second, we perform linguistic expert manual inspection on samples, for example ones with disagreement or too few annotations.

Our AI-assisted annotation autocompletion evaluation will measure: (1) the dataset creation speed up and (2) the potential bias induction. First, we will randomly turn off autocompletion process for a fraction of our annotations so we can have a comparison baseline to measure the speed up uplift [Holdt *et al.*, 2025]. Second, we antecipate our autocompletion process may induce biases toward the suggested audio pronunciations and orthographies. For example, if we always autocomplete with a specific orthography or audio, we may discourage annotations on regional orthographies and accents. As we aim at a diverse and comprehensive dataset, it is relevant to assess the bias we are introducing with the autocompletion tools. For that purpose, we will randomly insert intentional autocompletion errors to assess whether the annotators are actively reviewing suggestions or over relying on them.

In addition to dataset creation metrics, we will evaluate the accuracy of our speech-processing models through traditional Word Error Rate (WER) on our Indigenous languages bench-

marks [Radford *et al.*, 2023]. We will measure the trade-off between WER accuracy and model size to evaluate the efficiency of our quantization and pruning proposals. To measure how accessible our speech processing models are, we will measure the power consumption and device coverage. Finally, we will evaluate our speech processing models in simulated speaking exercises using ROC-AUC and PR-AUC metrics. We will also perform qualitative evaluations measuring the Mean Opinion Score (MOS) for synthesized audios and transcripts samples following the standardized CrowdMOS framework [Ribeiro *et al.*, 2011].

Finally, we will run structured interviews and usability studies with community members in the Bororo and Enawene Nawe Indigenous communities. We will assess perceptions regarding trust, privacy, data governance, and the exercises quality.

7 Challenges, Limitations and Ethical Considerations

Indigenous languages often lack standardization; while some languages have many competing orthographies, others do not have a formal orthography at all. For instance, the Bororo orthography currently used by many community members originates from missionary work and does not accurately reflect key phonological features [Colbacchini, 1925].

Because Indigenous territories are often fragmented, pronunciation can differ considerably across regions. These differences sometimes become socially stigmatized, particularly in communities where the ancestral language is less widely spoken, creating perceptions that certain pronunciations are "incorrect." This poses an important ethical challenge, as it may cause communities to feel that a new orthography favours certain regional pronunciations in detriment of others. The potential risk is even greater due to the reliance on automatic speech processing models, which may implicitly favor accents depending on the available training data.

The main challenge for the success of this research project lies in the data collection and annotation component. We are heavily dependent on Indigenous community adoption and engagement to use our tools and to generate the datasets required for developing the methods we propose. Despite our interdisciplinary network involving Indigenous professors and leaders, AI researchers, and linguistic experts, sustained participation by community members and local educators is essential for meaningful impact.

It is worth highlighting that the tools and methods produced by this research will not be perfect in any sense, so automatic exercises may ocasionally fail to correct classify student utterances. Indigenous professors should be trained on AI literacy to understand the non-deterministic nature of this technology and their limitations. We foresee the need for continuous training sessions and workshops to increase AI literacy among Indigenous communities.

The research will comply with Brazilian National Health Council (CNS) Resolutions No. 510/2016 and No. 466/2012, respecting the ethical principles of autonomy, beneficence, non-maleficence, and justice. Research involving Indigenous peoples follows the principles of Free, Prior and Informed Consent (FPIC), as established by the International Labour Organization's Convention No. 169, as well as the rights to self-determination, cultural integrity, and control over traditional knowledge, as affirmed in the United Nations Declaration on the Rights of Indigenous Peoples (UNDRIP). The project considers both individual and collective consent, and adheres to the CARE principles for data governance, ensuring collective benefit, community authority over data, and the ethical use of linguistic and cultural data produced. Indigenous communities should always govern their language, so controls compliant with the Brazilian General Law for Data Proctection (LGPD) [BRASIL, 2018] should be developed to ensure the language resources produced by this research remain private, protected and under the exclusive control of their own Indigenous communities. This research project is commited to avoid any data breaches that could risk private language resources being made publicly available without the explicit consent of the Indigenous communities involved.

8 Project Implementation Plan

Our work plan comprises 2 years and consists of the tasks below:

1. By 2026 April, once the ethics protocols are established, the team will present the project to the community, so as to reach community approval.
2. Up until 2026 October, we will conduct training workshops with Bororo teachers to onboard them with the data collection and annotation tool. During the same time we will work on establishing linguistic groundings with the Enawene Nawe.
3. Once we have enough data to work with, we will evaluate existing speech models and develop our proposed method for model pruning. We will run experiments to develop speech exercises prototype. We expect to have a fully functioning on-device speech model for both Bororo language by October 2026 and by January 2027 for Enawene Nawe.
4. In parallel, we will develop and evaluate micro language models for writing and reading exercises on the benchmarks for both Indigenous communities. Prototypes should be ready by the same time as the speech models.
5. Finally, we will integrate our models into an offline mobile gaming app. We will run training sessions with selected Indigenous professors and members of the community to run pilots in their schools and other contexts. Tests should start for Bororo by July 2027 and Enawene Nawe by October 2027.
6. We will continuously keep evaluating our tools and incorporating feedback. This will be an ongoing effort.

Economic sustainability will rely on non-profit funding and government partnerships for integration into educational programs. All software and models will be released as open source, enabling external collaboration and reducing long-term maintenance costs. This strategy ensures that the tools can be continuously improved and extended to additional Indigenous communities in Brazil, such as the Makurap people, with whom the project is already underway, and abroad.

A Curricula Vitae

Gustavo Polleti
Universidade de São Paulo
Gustavo Polleti is an engineer, researcher, and PhD candidate at the Escola Politécnica of the University of São Paulo, specializing in Natural Language Processing with a focus on Indigenous languages. He holds both a bachelor's degree in engineering and a master's degree from the same institution. Gustavo has over six years of professional experience as a Machine Learning and AI engineer, having worked at major technology companies, including in the mobile gaming industry. He leads the Bilingo project and is responsible for its technical development and implementation.

Fabricio Ferraz Gerardi
União das Faculdades Católicas de Mato Grosso (Brazil)
Fabrício Ferraz Gerardi holds a PhD in Linguistics, a Master's degree in Computational Linguistics, a Master's degree in Ancient Hebrew, and a Bachelor's degree in Classical Languages. He has extensive experience teaching language structure, with a particular focus on Brazilian Indigenous languages. He serves as the official linguist of Brazil's National Foundation for Indigenous Peoples (FUNAI) for the recently contacted Enawenê-Nawê people and is Director of the Plataformas Indígenas project. He is the linguist expert for the project and responsible for the Bororo case of study.

Carolina Aragon
Universidade Federal da Paraíba (UFPB)
Carolina Coelho Aragon is an adjunct professor at the Federal University of Paraíba. Her work focuses on Brazilian Indigenous languages, especially on endangered languages and language maintenance. She received her Ph.D. in Linguistics from the University of Hawai'i at Manoa and her master's from the University of Brasília.

Fernando Kudoro Bororo
Universidade Federal de Rondonópolis (Brazil)
Fernando Kudoro Bororo is an Indigenous Bóe-Bororo educator and researcher with extensive experience in Indigenous schooling and teacher education. Since 2004, he has worked as a teacher at the Korogedo Paru Indigenous State School, located in the Tereza Cristina Indigenous Territory. He holds a degree in Pedagogy from the Intercultural Indigenous Faculty of the State University of Mato Grosso (UNEMAT), completed in 2016. He is currently a graduate researcher in the Master's Program in Education (PPGEDU) at the Federal University of Rondonópolis. His research focuses on the representations of Indigenous peoples in Brazilian state-produced textbooks used in the early years of primary education. He is responsible for co-creating the project's pedagogical content and guiding the Bororo case of study.

Mariel M. B. Kujiboekureu
University of São Paulo
Mariel M. B. Kujiboekureu holds a Bachelor's degree in Mathematics and Natural Sciences Teaching and a Master's degree in Science. His academic trajectory is dedicated to integrating scientific knowledge with the traditional knowledge of the Bororo people, using formal education as a means to strengthen Indigenous identity and cultural continuity. Based in the Meruri village, in the municipality of General Carneiro, Mato Grosso, Brazil, he plays an important role in the preservation of his community's intangible cultural heritage. His work focuses on the transmission of Bororo cultural knowledge and the empowerment of younger generations, contributing to the continuity of ancestral knowledge both within the community and in educational contexts. He is responsible for co-creating the project's pedagogical content and guiding the Bororo case of study.

Wayali Iholalare Kahokase Salomã
State University of Mato Grosso (Brazil)
Wayali Salokmã is a native Enawene Nawe educator currently completing a Bachelor's degree in Pedagogy at the State University of Mato Grosso. He is actively engaged as a teacher at the local Enawene Nawe school, where he teaches the Enawene Nawe language, cultural knowledge, and English. He will help conducting the cases of study.

Fabio Cozman
Universidade de São Paulo
Prof. Fábio Gagliardi Cozman is a Full Professor at the Escola Politécnica of the University of São Paulo (USP), Director of USP's Center for Artificial Intelligence (C4AI), and a CNPq Researcher Level 1C. He is responsible for supervising the technical development and implementation.

Acknowledgments

The last 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.

References

[Akhlaghi *et al.*, 2019] Elham Akhlaghi, Branislav Bédi, Matt Butterweck, Cathy Chua, Johanna Gerlach, Hanieh Habibi, Junta Ikeda, Manny Rayner, Sabina Sestigiani, and Ghil'ad Zuckermann. Overview of LARA: A Learning and Reading Assistant. In *Proc. 8th ISCA Workshop on Speech and Language Technology in Education (SLaTE 2019)*, pages 99–103, 2019.

[Albisetti and Ravagnani, 1992] Cesare Albisetti and Oswaldo Martins Ravagnani. A aldeia bororo. *Perspectivas: Revista de Ciências Sociais*, 15, 1992.

[Bisazza and others, 2020] Arianna Bisazza et al. Low-resource machine translation: A survey. *Computational Linguistics*, 2020.

[BRASIL, 2018] BRASIL. Lei nº 13.709, de 14 de agosto de 2018, 2018. Lei Geral de Proteção de Dados Pessoais (LGPD).

[Brugman and Russel, 2004] Hennie Brugman and Albert Russel. Annotating multi-media/multi-modal resources

with ELAN. In Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa, and Raquel Silva, editors, *Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04)*, Lisbon, Portugal, May 2004. European Language Resources Association (ELRA).

[Camargo, 2011] Gonçalo Ochoa Camargo. Boe ewadaru–a língua bororo: breve histórico e elementos de gramática. *Campo Grande: UCDB*, 2011.

[Cibaeikare and Etua, 1990] C G O Cibaeikare and F C T Etua. *História mítica Bororo, vol. I*. Missão Salesiana do Mato Grosso, 1990.

[Cohen, 1960] J. Cohen. A Coefficient of Agreement for Nominal Scales. *Educational and Psychological Measurement*, 20(1):37, 1960.

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

[de Marneffe *et al.*, 2021] Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. Universal dependencies. *Computational Linguistics*, 47(2):255–308, 2021.

[El Kheir *et al.*, 2023] Yassine El Kheir, Ahmed Ali, and Shammur Absar Chowdhury. Automatic pronunciation assessment - a review. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 8304–8324, Singapore, December 2023. Association for Computational Linguistics.

[Ferraz Gerardi,] Fabrício Ferraz Gerardi. *Bororo Dictionary*. Forthcoming. Available upon request.

[Ferraz Gerardi and Toribio Serrano, 2025] Fabrício Ferraz Gerardi and Luis Toribio Serrano. Corpus bororo (corbo), 2025.

[Ferraz Gerardi, 2024] Fabrício Ferraz Gerardi. Universal dependencies ud bororo-bdt, 2024.

[Gerardi *et al.*, 2022a] Fabrício Ferraz Gerardi, Stanislav Reichert, Carolina Aragon, Lorena Martín-Rodríguez, Gustavo Godoy, and Tatiana Merzhevich. *TuDeT: Tupían Dependency Treebank*. Zenodo, May 2022.

[Gerardi *et al.*, 2022b] Fabrício Ferraz Gerardi, Stanislav Reichert, Carolina Aragon, Tim Wientzek, Johann-Mattis List, and Robert Forkel. *TuLeD. Tupían Lexical Database*. Zenodo, May 2022.

[glo, 2024] Glottolog 5.1, 2024. Accessed: 2024-12-15.

[Grenoble and Whaley, 2006] Lenore A. Grenoble and Lindsay J. Whaley. *Saving Languages: An Introduction to Language Revitalization*. Cambridge University Press, Cambridge, 2006.

[Harrison and Harrison, 2013] Carl Harrison and Carole Harrison. *Dicionário Guajajara-Português*. SIL, 2013.

[Hinton, 2011] Leanne Hinton. Language revitalization and language pedagogy: New teaching and learning strategies. *Language and Education*, 25(4):307–318, 2011.

[Holdt *et al.*, 2025] Špela Arhar Holdt, Nikolai Ilinykh, Barbara Scalvini, Micaella Bruton, Iben Nyholm Debess, and Crina Madalina Tudor, editors. *Proceedings of the Third Workshop on Resources and Representations for Under-Resourced Languages and Domains (RESOURCEFUL-2025)*, Tallinn, Estonia, March 2025. University of Tartu Library, Estonia.

[Jacob *et al.*, 2017] Benoit Jacob, Skirmantas Kligys, Bo Chen, Menglong Zhu, Matthew Tang, Andrew G. Howard, Hartwig Adam, and Dmitry Kalenichenko. Quantization and training of neural networks for efficient integer-arithmetic-only inference. *CoRR*, abs/1712.05877, 2017.

[King, 2001] Kendall A. King. *Language Revitalization Processes and Prospects: Quichua in the Ecuadorian Andes*. Multilingual Matters, Clevedon, 2001.

[Lewis *et al.*, 2020] 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. Indigenous protocol and artificial intelligence position paper. Project Report 10.11573/spectrum.library.concordia.ca.00986506, Aboriginal Territories in Cyberspace, Honolulu, HI, 2020. 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/.

[Monserrat, 2000] Ruth Fonini Monserrat. *Vocabulário Amondawa-Português, Vocabulário e frases em Arara e Português, Vocabulário Gavião-Português, Vocabulário e frases em Karipuna e Português, Vocabulário e frases em Makurap e Português, Vocabulário e frases em Suruí e Português, Pequeno dicionário em Tupari e Português*. Universidade do Caixas do Sul, 2000.

[Nicholas, 2019] Sheilah E. Nicholas. Indigenous language education and literacy practices. In Kenneth L. Rehg and Lyle Campbell, editors, *The Routledge Handbook of Language Revitalization*, pages 346–360. Routledge, London, 2019.

[Nivre *et al.*, 2016] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajic, et al. Universal dependencies v1: A multilingual treebank collection. *Proceedings of LREC*, 2016.

[Nivre *et al.*, 2020] Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. Universal Dependencies v2: An evergrowing multilingual treebank collection. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri,

Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France, May 2020. European Language Resources Association.

[Pinhanez *et al.*, 2023] Claudio S. Pinhanez, Paulo Cavalin, Marisa Vasconcelos, and Julio Nogima. Balancing social impact, opportunities, and ethical constraints of using ai in the documentation and vitalization of indigenous languages. In Edith Elkind, editor, *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence, IJCAI-23*, pages 6174–6182. International Joint Conferences on Artificial Intelligence Organization, 8 2023. AI for Good.

[Polleti *et al.*, 2024] Gustavo Polleti, Fabio Cozman, and Fabrício Gerardi. 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, 2024. SBC.

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

[Pratap *et al.*, 2024] Vineel Pratap, Andros Tjandra, Bowen Shi, Paden Tomasello, Arun Babu, Sayani Kundu, Ali Elkahky, Zhaoheng Ni, Apoorv Vyas, Maryam Fazel-Zarandi, Alexei Baevski, Yossi Adi, Xiaohui Zhang, Wei-Ning Hsu, Alexis Conneau, and Michael Auli. Scaling speech technology to 1,000+ languages. *J. Mach. Learn. Res.*, 25(1), January 2024.

[Radford *et al.*, 2023] Alec Radford, Jong Wook Kim, Tao Xu, Greg Brockman, Christine McLeavey, and Ilya Sutskever. Robust speech recognition via large-scale weak supervision. In *Proceedings of the 40th International Conference on Machine Learning*, ICML'23. JMLR.org, 2023.

[Rayner and Wilmoth, 2023] Manny Rayner and Sasha Wilmoth. Using LARA to rescue a legacy Pitjantjatjara course. In Atticus Harrigan, Aditi Chaudhary, Shruti Rijhwani, Sarah Moeller, Antti Arppe, Alexis Palmer, Ryan Henke, and Daisy Rosenblum, editors, *Proceedings of the Sixth Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pages 13–18, Remote, March 2023. Association for Computational Linguistics.

[Reiter and Dale, 2000] Ehud Reiter and Robert Dale. Building natural language generation systems. 2000.

[Ribeiro *et al.*, 2011] Flávio Ribeiro, Dinei Florêncio, Cha Zhang, and Michael Seltzer. Crowdmos: An approach for crowdsourcing mean opinion score studies. In *2011 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2416–2419, 2011.

[Sainath *et al.*, 2020] Tara N. Sainath, Yanzhang He, Bo Li, Arun Narayanan, Ruoming Pang, Antoine Bruguier, Shuo-yiin Chang, Wei Li, Raziel Alvarez, Zhifeng Chen, Chung-Cheng Chiu, David Garcia, Alex Gruenstein, Ke Hu, Anjuli Kannan, Qiao Liang, Ian McGraw, Cal Peyser, Rohit Prabhavalkar, Golan Pundak, David Rybach, Yuan Shangguan, Yash Sheth, Trevor Strohman, Mirkó Visontai, Yonghui Wu, Yu Zhang, and Ding Zhao. A streaming on-device end-to-end model surpassing server-side conventional model quality and latency. In *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6059–6063, 2020.

[Settles, 2012] Burr Settles. Active learning. *Synthesis Lectures on Artificial Intelligence*, 2012.

[Sharma *et al.*, 2025] Avinash Kumar Sharma, Manas Pandya, and Arpit Shukla. Fine-tuning whisper tiny for swahili asr: Challenges and recommendations for low-resource speech recognition. In *Proceedings of the Sixth Workshop on African Natural Language Processing (AfricaNLP 2025)*, pages 74–81, 2025.

[Thomason, 2015] S. G. Thomason. *Endangered Languages: An Introduction*. Cambridge: Cambridge University Press, 2015.

[Wiseman *et al.*, 2018] Sam Wiseman, Stuart Shieber, and Alexander Rush. Learning neural templates for text generation. In *Proceedings of EMNLP*, 2018.