



OPEN ACCESS

EDITADO POR

- Maria Filomena Spatti Sandalo
(Unicamp)

AVALIADO POR

- Bruna Franchetto (UFRJ)
- Evandro Bonfim (UFSCAR)

SOBRE OS AUTORES

- Thomas Daniel Finbow
Conceptualização, Curadoria de Dados, Investigação, Metodologia, Administração, Supervisão, Visualização, Escrita – rascunho original e Escrita – análise e edição.
- Luciana Raccanello Storto
Conceptualização, Curadoria de Dados, Investigação, Metodologia, Administração, Supervisão, Visualização, Escrita – rascunho original e Escrita – análise e edição.
- Joana Dworecka Autuori
Conceptualização, Curadoria de Dados, Investigação, Metodologia, Administração, Supervisão, Visualização, Escrita – rascunho original e Escrita – análise e edição.
- Marina Geiger dos Santos
Conceptualização, Curadoria de Dados, Investigação, Metodologia, Administração, Supervisão, Visualização, Escrita – rascunho original e Escrita – análise e edição.

DATAS

- Recebido: 15/10/2025
- Aceito: 28/04/2026
- Publicado: 16/06/2026

COMO CITAR

FINBOW, T. D. et al.. (2026).
Processos e Produtos do Projeto
"Processamento Ético de
Línguas Indígenas" do Centro de
Inteligência Artificial da USP
(2022-2025). *Revista da Abralín*,
v. 24, n. 2, p. 89-115, 2025.

RELATÓRIO DE EXPERIÊNCIA

Processos e Produtos do Projeto “Processamento Ético de Línguas Indígenas” do Centro de Inteligência Artificial da USP (2022-2025)

Thomas Daniel FINBOW

Universidade de São Paulo (USP)

Luciana Raccanello STORTO

Universidade de São Paulo (USP)

Joana Dworecka AUTUORI

Universidade de São Paulo (USP)

Marina GEIGER

Universidade de São Paulo (USP)

RESUMO

O PROLIND (Processamento Ético de Línguas Indígenas Brasileiras) é um projeto realizado a partir de 2022, com o propósito de apoiar a documentação e o fortalecimento de línguas indígenas brasileiras, por meio do uso ético de tecnologias e inteligência artificial de ponta. Coordenado por pesquisadores vinculados ao *Centre for Artificial Intelligence*/Centro para Inteligência Artificial (C4AI) da Universidade de São Paulo e à IBM-Brasil e por professores pesquisadores do Departamento de Linguística da Faculdade de Filosofia, Letras e Ciências Humanas da USP, o projeto busca desenvolver assistentes de escrita, tradutores automáticos e bancos de dados linguísticos voltados para comunidades falantes de línguas indígenas, como Nheengatu e Guarani Nhandewa (do povo autodenominado “Tupi-

Guarani”). A atuação do PROLIND prima pela soberania linguística e cultural dos povos indígenas, sendo todas as colaborações formalizadas mediante acordos e consentimentos éticos, em parceria com as autoridades competentes. Entre os principais resultados, a criação de ferramentas de tradução e escrita em Nheengatu, implementadas em comunidades Baré do Alto Rio Negro; e oficinas de revitalização linguística com os autodenominados Tupi-Guarani, no litoral paulista. O projeto prioriza o uso responsável de dados públicos e a criação de tecnologias acessíveis às comunidades. O artigo relata como a tecnologia pode ser ressignificada a partir das necessidades e perspectivas das próprias comunidades indígenas, tornando-se uma ferramenta de resistência e continuidade cultural. Discute também o papel da linguística na mediação entre saberes tradicionais e inovação tecnológica e como processos colaborativos podem gerar soluções éticas e sustentáveis para o ensino e o uso das línguas no ambiente digital.

ABSTRACT

PROLIND (*Processamento Ético de Línguas Indígenas Brasileiras*/Ethical Processing of Brazilian Indigenous Languages) is a project that began in 2022 with the purpose of supporting the documentation and strengthening of Brazilian indigenous languages through the ethical use of cutting-edge technologies and artificial intelligence. Coordinated by researchers affiliated with the Centre for Artificial Intelligence (C4AI) at the University of São Paulo and IBM-Brazil, and by research professors from the Linguistics Department of the Faculty of Philosophy, Letters and Human Sciences at USP, the project seeks to develop writing assistants, automatic translators, and linguistic databases aimed at communities that speak indigenous languages such as Nheengatu and Guarani Nhandewa (who call themselves ‘Tupi-Guarani’). PROLIND’s work prioritises the linguistic and cultural sovereignty of indigenous peoples, with all collaborations formalised through ethical agreements and consents, in partnership with the competent authorities. Among the main results are the creation of translation and writing tools in Nheengatu, implemented in Baré communities of the Upper Rio Negro; and linguistic revitalisation workshops with two communities of the Tupi-Guarani people on the coast of São Paulo. The project prioritises the responsible use of public data and the creation of technologies accessible to communities. The article reports how technology can be reinterpreted based on the needs and perspectives of the indigenous communities themselves, becoming a tool for resistance and cultural

continuity. It also discusses the role of linguistics in mediating between traditional knowledge and technological innovation, and how collaborative processes can generate ethical and sustainable solutions for the teaching and use of languages in the digital environment.

PALAVRAS-CHAVE

Documentação linguística. Fortalecimento linguística. Inteligência artificial. Ética na pesquisa.

KEYWORDS

Language documentation. Language strengthening. Artificial intelligence. Research ethics.

RESUMO PARA NÃO ESPECIALISTAS

Este artigo apresenta o projeto Processamento Ético de Línguas Indígenas (PROLIND), coordenado por Claudio Pinhanez (coordenador do Centro de Inteligência Artificial da USP) e Luciana Storto (professora do Departamento de Linguística da USP) e composto por uma equipe multidisciplinar. O projeto teve como objetivo criar assistentes de escrita para línguas de baixo recursos utilizando a inteligência artificial em dois contextos: com a língua Nheengatu, apoiando escolas da etnia Baré na região do Rio Negro (AM); com Tupi Antigo em dois projetos de Mestrado que estão criando assistentes de escrita focados em tradução. Há ainda uma terceira frente: o projeto de revitalização da língua Guarani Nhandewa em duas aldeias de Mongaguá/SP. Esta ala do PROLIND derivou de uma etapa de trabalho com os Guarani Mbya que foi descontinuada e não envolve inteligência artificial. Desde 2023 o PROLIND atuou com a língua Nheengatu falada em duas comunidades da etnia Baré no alto Rio Negro (AM) desenvolvendo assistentes de escrita, sobretudo tradutores, a fim de possibilitar aos falantes o acesso a serviços jurídicos, de saúde, educacionais e outros serviços governamentais. Em 2024, o projeto passou a atuar com escolas indígenas da etnia Baré em comunidades próximas à cidade São Gabriel da Cachoeira (AM), com a finalidade de trabalhar na criação de co-design e co-desenvolvimento de ferramentas de IA para contribuir com os esforços da comunidade para que sua língua seja usada na vida cotidiana e, cada vez mais, no ambiente digital. Como parte do acordo firmado em 2025 entre a USP e a FOIRN (Federação das Organizações Indígenas do Rio Negro), o PROLIND implantou duas salas de aula digitais totalmente equipadas para apoiar essas atividades. No projeto de revitalização da língua Guarani

Nhandewa em duas aldeias de Mongaguá (SP), onde a transmissão intergeracional da língua foi interrompida são realizados dois encontros mensais com os professores e os alunos das escolas indígenas para trabalhar a retomada da língua por meio da documentação, da elaboração de material de ensino da língua e do apoio aos professores nas aulas de língua indígena. No trabalho com Tupi Antigo, o PROLIND atua por meio de dois projetos de Mestrado que almejam utilizar a inteligência artificial para criar chatbots capazes de se comunicar nessa língua. Essa ferramenta pode ser usada em revitalização da língua ou para o ensino da língua em outros contextos. Um dos projetos, focado em criar tradutores, está utilizando tanto tradução direta quanto tradução pivotada por língua proximal, em que o Guaraní Paraguaio atua como intermediário devido à sua proximidade linguística com o Tupi Antigo. O outro projeto está criando uma gramática computacional do Tupi Antigo, com o intuito, além da tradução, de encontrar um processo repetível com o qual seja possível desenvolver tradutores para qualquer língua com poucos recursos.

Introdução

O projeto Processamento Ético de Línguas Indígenas Brasileiras (PROLIND) envolve linguística, tecnologia e inteligência artificial e nasce, em meados de 2022, com o objetivo de apoiar a documentação e o fortalecimento de línguas indígenas brasileiras, por meio de ferramentas tecnológicas, como assistentes de escrita (completadores de palavras e sentenças e tradutores) e bancos de dados linguísticos. O uso da inteligência artificial é desafiador para línguas de recursos limitados, isto é, línguas sobre as quais há poucos dados disponíveis. No entanto, modelos de linguagem para línguas com grandes recursos como o inglês, alemão e português puderam ser usados e foram adaptados para línguas com baixos recursos através de um ajuste fino.

A equipe do projeto é interdisciplinar, envolvendo 2 cientistas pesquisadores e um engenheiro de software da IBM, 3 pesquisadores do Departamento de Linguística da USP, uma mestranda da área de relações internacionais da USP, duas pós doutorandas, uma em linguística e uma em antropologia, um bolsista técnico, e estagiários de graduação. Trata-se de um projeto idealizado inicialmente por Claudio Pinhanez, que é vice-diretor do Centro de Inteligências Artificiais da USP (C4AI). O C4AI é um centro com verba da FAPESP, USP e do Laboratório de Pesquisa no Brasil da IBM. A linguista Luciana Storto atuou junto com Cláudio Pinhanez como coordenadora do projeto, montando a equipe e auxiliando através de sua rede de contatos, para que as parcerias com as comunidades se estabelecessem. No último ano, juntou-se à equipe um professor do Instituto de

Matemática e Estatística (IME) da USP, Marcelo Finger, que coordenou um projeto de tradução automática do Tupi Antigo.

Desde o início, o projeto foi concebido de forma ética e colaborativa, buscando responder a demandas reais apresentadas pelas comunidades indígenas, sempre em respeito à soberania linguística e cultural dos povos envolvidos, respeitando as resoluções da conferência de Los Piños e da Década das Línguas Indígenas da UNESCO. Foi nesse espírito que se estabeleceram os primeiros diálogos com a comunidade Guarani Mbyá da aldeia *Tenonde Porã*, localizada no distrito de Parelheiros, na cidade de São Paulo, bem como com falantes de Nheengatu do município de São Gabriel da Cachoeira (AM), e com falantes de Terena na região de Bauru, SP. Foram visitadas também as aldeias *Tekoa Mirĩ* em Praia Grande e *Aguapeú* em Mongaguá, no litoral de São Paulo.

Quando as parcerias foram firmadas, acordos foram assinados estabelecendo os termos das colaborações com os falantes de Nheengatu e *Nhandewa*. Não houve coleta de dados durante o projeto, sendo que todos os dados usados vieram de publicações já existentes, que foram utilizadas com o consentimento dos autores.

A principais áreas de atuação do projeto PROLIND até meados de 2025 foram 1) a informatização da documentação da língua; 2) o uso de ferramentas educacionais digitais; 3) o suporte ao uso digital da língua indígena; 4) a promoção da acessibilidade linguística; 5) fomentar a apropriação tecnológica por parte da comunidade. Para tanto, foram desenvolvidos assistentes de escrita para a produção de texto e uso de mídias sociais baseados em modelos indígenas de linguagem. Foi explorado o desenvolvimento de aplicativos Whatsapp e Android para fornecer melhor suporte e acesso a recursos digitais às comunidades. Foram, ainda, desenvolvidos tradutores de línguas indígenas brasileiras para o português por meio de um ajuste fino de tradutores de Modelos de Linguagem para línguas com grandes bancos de dados (LLM).

1. Projeto Guarani Mbyá

Após a realização de algumas reuniões na Terra Indígena *Tenonde Porã*, localizada ao sul da cidade de São Paulo, por decisão da comunidade falante de Guarani Mbyá foi iniciada uma atuação do projeto em oficinas semanais dirigidas ao ensino médio durante um semestre (14 oficinas), onde foi apresentado um protótipo online do tradutor e do completador de palavras e sentenças para os jovens manipularem em 12 notebooks. Também foram apresentadas algumas ferramentas de documentação linguística, como os programas ELAN e FLEX. Arquivos de vídeo gravados pelos alunos foram inseridos no ELAN como exercício em uma das oficinas.

O trabalho com a língua Guarani Mbyá, no entanto, foi descontinuado, pois as lideranças ficaram desconfortáveis com o fato de sua língua estar sendo objeto de tradução automática. Nosso entendimento foi que essa preocupação era influenciada em primeiro lugar pelo clima de especulação receosa a respeito da inteligência artificial, que estava generalizado entre o grande público após o lançamento do ChatGPT. Além disso, a apreensão parecia decorrer também de um receio de que

partidos inescrupulosos não específicos pudessem eventualmente hackear a tecnologia e torná-la contra a comunidade de alguma maneira, por exemplo, acessando e divulgando certos conteúdos, p. ex., os religiosos e políticos, que a comunidade desejava manter protegidos.

O tradutor Guarani Mbyá foi descartado e foi iniciado um tradutor Nheengatu a pedido de duas comunidades de etnia Baré falantes de Nheengatu no Alto Rio Negro. Membros das aldeias Itaóca Tupi e Barigui, localizadas em Mongaguá/SP, que estavam presentes na reunião inicial com os Mbyá, solicitaram ao projeto um apoio na revitalização da língua Guarani *Nhandewa* nessas localidades.

1.1 Metodologia de IA

Os dados existentes das línguas indígenas para as quais foram elaborados tradutores automáticos de IA foram dicionários, materiais acadêmicos, materiais educativos e dados públicos da internet. Eles alimentaram um Modelo de Linguagem público do português, inglês, alemão, etc. (LLAMA, BERT, WMT19), e depois passaram por um processo de adaptação para a língua em questão. Isso gerou um modelo computacional da língua indígena, e da lá foram criados os produtos, a saber, tradutores, dicionários, corretores e geradores de texto, todos de propriedade das comunidades de falantes (CAVALIN, P. *et al.*, 2023, 2024a, 2024b, 2025, PINHANEZ *et al.*, 2023, 2024).

Ficou acordado que as questões relacionadas ao acesso e uso dessas ferramentas e aplicativos seriam decididas pelas comunidades. No futuro, uma nova parceria poderá ser estabelecida na qual os falantes poderão avaliar as ferramentas criadas e as correções inseridas poderão ser usadas para realimentar o processo de adaptação do modelo.

O assistente de escrita foi desenvolvido com base em ferramentas da IBM Research e adaptadas para uso com o Guarani Mbyá, utilizando dados de dicionários e livros online (CAVALIN, P. *et al.*, 2024a). Em oficinas com alunos do ensino médio da escola indígena na aldeia *Tenonde Porã*, ideias e protótipos foram explorados pelos alunos. A partir da participação dos alunos e professores (críticas, comentários e observação de uso), foram desenvolvidos novos protótipos do assistente de redação em Guarani Mbyá.

Na segunda fase do projeto, está sendo desenvolvido um assistente de escrita para a língua Nheengatu, que amplia os conceitos e funcionalidades iniciais. A nova versão será testada e refinada com falantes da língua Nheengatu no futuro. O novo assistente de escrita inclui um dicionário eletrônico, um completador de palavras, um assistente que sugere a próxima palavra e um tradutor de sentenças. Detalhes do projeto com a língua Nheengatu serão apresentados abaixo, numa seção dedicada a este projeto.

2. Projeto *Nhandewa*

O chamado das comunidades de Mongaguá foi fundamental para orientar a direção prática do projeto. As lideranças locais manifestaram preocupação em garantir que as novas gerações tivessem acesso a recursos de qualidade para aprender e manter viva a língua Guarani *Nhandewa*, autodenominada ‘Tupi-Guarani’ ou ‘Tupi’. A grafia com <w> é uma convenção de D’Angelis (2024, c.p.) para diferenciar os ‘Nhandeva’, que moram no interior e que são mais parecidos linguisticamente aos Guarani Kaiowá, do grupo diaspórico que habita o litoral de São Paulo e que é mais próximo dos Guarani Mbyá, depois de um convívio de quase um século.

O processo de negociação com a comunidade de Itaóca Tupi foi marcado por múltiplas visitas, nas quais a equipe teve a oportunidade de conhecer de perto a dinâmica da aldeia. Uma dessas ocasiões incluiu uma visita com mais tempo hábil ao território, permitindo compreender melhor o espaço de vivência comunitária e colher alguns registros para usufruto da comunidade. As tratativas tiveram início em reuniões realizadas na escola local, com a presença de todas as lideranças da comunidade, cujo estilo de organização política se caracteriza por uma liderança compartilhada entre membros influentes. Nessas reuniões, foram feitas diversas exposições sobre os objetivos e metodologias do PROLIND, ao mesmo tempo em que professores e lideranças expuseram suas necessidades, expectativas e desafios. O diálogo foi central, as propostas iniciais apresentadas pela equipe foram discutidas ponto a ponto, aceitas em alguns casos e reformuladas em outros, de acordo com os interesses e demandas.

Havia uma clara necessidade de produzir materiais adequados a diferentes faixas etárias, que pudessem ser utilizados tanto em sala de aula como em contextos informais e que fossem construídos em diálogo com professores, estudantes e falantes mais velhos. A partir dessas conversas, estabeleceu-se um acordo de colaboração: o PROLIND se comprometeria a desenvolver jogos didáticos voltados a crianças menores, do ensino infantil e fundamental 1, protótipos de material de ensino de segunda língua (L2) para adolescentes, do fundamental 2, e oficinas de revitalização linguística, sempre com a participação ativa da comunidade. A proposta não era entregar produtos prontos, mas construir coletivamente ferramentas que fizessem sentido no contexto cotidiano da aldeia.

Ao final desse processo de escuta e negociação, os termos de colaboração foram estabelecidos de forma consensual. Ficou acordado que a prioridade seria a implementação de oficinas de revitalização linguística, associadas à produção de jogos e materiais de ensino da língua, de modo a atender tanto às crianças quanto aos adolescentes. A partir desse momento, as oficinas tiveram início, consolidando a parceria entre a comunidade e o projeto. Essa dinâmica reforçou o caráter participativo do PROLIND, demonstrando que os resultados obtidos em Mongaguá são fruto direto de um processo coletivo, em que a comunidade define suas próprias prioridades e participa ativamente da construção das soluções propostas.

Os primeiros passos do trabalho em Mongaguá incluíram a criação de jogos linguísticos para crianças, estimulando a aprendizagem por meio de associações entre palavras, imagens e narrativas da tradição Tupi-Guarani (autodenominação étnica dos membros das duas aldeias), além da

elaboração de um protótipo de material de ensino de segunda língua (L2). Esse material, que está em fase de elaboração, foca no método comunicativo e utiliza dados obtidos em primeira mão com falantes nativos e dados de outros materiais de ensino de línguas Tupi, como *Curso de Língua Geral (Nheengatu ou Tupi moderno)*, do professor Eduardo de Almeida Navarro (2016), *Lições de Gramática Nhandewa-Guarani vol. 1*, coordenado por professores da Unicamp e por uma equipe de consultores falantes da língua (MARCOLINO *et al.*, 2016) e *Lições de Gramática Nhandewa/Tupi-Guarani vol. 2* (DJATSY *et al.*, 2018), estes focados em temas gramaticais para falantes da língua, portanto, materiais de ensino de primeira língua (L1). Um material didático sobre sentenças complexas em Mbyá (orações adverbiais e complemento) foi elaborado por duas bolsistas de IC baseado no trabalho do linguista Guillaume Thomas. Um falante de Mbyá e Nhandewa teve acesso a este material e ficou de procurar usos para sua aplicação.

Em paralelo, oficinas de revitalização linguística passaram a ser realizadas com a participação de falantes mais velhos da própria comunidade ou de outras aldeias, professores e estudantes, estimulando o engajamento comunitário. Além disso, está sendo criada uma base de dados da língua Guarani Nhandewa no software FLE_x (*Fieldworks Language Explorer*, SIL) com dados de materiais pré-existent, como os anteriormente citados (MARCOLINO *et al.*, 2016), (DJATSY *et al.*, 2018). Essa base de dados sistematiza os dados dos materiais de ensino de língua e ao mesmo tempo fornece suporte para a criação de novos materiais.

Atualmente, os desdobramentos do projeto em Mongaguá têm mostrado resultados positivos. Professores relatam maior interesse das crianças nas atividades escolares quando os jogos e materiais produzidos são utilizados, e as oficinas têm se tornado momentos importantes de encontro comunitário, em que diferentes gerações se reúnem para pensar a língua como parte da vida cotidiana. Além disso, o projeto está fornecendo bases metodológicas e teóricas para a prática dos professores em sala de aula e nas práticas escolares de modo geral.

Fatores externos e internos à comunidade podem fazer com que esta deixe de falar sua língua e passe a falar somente a língua da sociedade dominante. Conforme relato dos moradores, no caso das aldeias Itaóca e Barigui, a substituição do Guarani Nhandewa pelo português se deu em grande medida como uma forma de se proteger da estigmatização dos falantes pela sociedade não indígena, ao se verem envolvidos por uma sociedade anti-indígena em decorrência, entre outros fatores, de disputas territoriais. Assim, mascarar a identidade indígena foi uma estratégia de sobrevivência. Isso englobava não ensinar a língua materna para as crianças. Além disso, casamentos interétnicos, de indígenas com não indígenas, e uma maior abertura dos Guarani Nhandewa à sociedade não indígena, em comparação com os Guarani Mbyá, por exemplo, acarretaram o interrompimento da transmissão da língua às próximas gerações. Atualmente, nas aldeias Itaóca Tupi e Barigui há somente falantes com mais de 60 anos (duas falantes em Itaóca e um falante em Barigui).

Em relação às experiências atuais de revitalização de línguas ao redor do mundo, o projeto de revitalização da língua Guarani Nhandewa das aldeias Itaóca Tupi e Barigui precisa de adequações para garantir sustentabilidade e atingir um maior alcance e eficácia. Para que a revitalização linguística seja levada a cabo, foram adotados os apontamentos de McCarty (2018) e Amaral (2021) sobre a

metodologia que designam 'Planejamento Linguístico de Base Comunitária'. Este embasamento deve ser constituído por um tripé: (i) planejamento situacional, (ii) planejamento do *corpus* e (iii) planejamento da aquisição/transmissão. Em resumo, no planejamento situacional, a comunidade define quais serão os contextos de uso, isto é, em que situações a língua será falada (na escola, dentro das casas, nos espaços religiosos/espirituais, nas reuniões comunitárias, por exemplo) e quem serão os atores envolvidos em cada situação, bem como os responsáveis por liderar a revitalização em cada contexto. No planejamento do *corpus*, define-se os conjuntos de léxico e estruturas gramaticais necessários para os contextos em que a língua será usada. No item planejamento da aquisição/transmissão, são definidas as formas como a língua será aprendida/transmitida.

Um planejamento linguístico de base comunitária, como o nome sugere, deve ser protagonizado por membros da comunidade. Naturalmente a presença de consultores externos, como falantes da língua, linguistas e antropólogos, por exemplo, é bem-vinda. No entanto, como diz Amaral (2021) a atuação do linguista nesse tipo de projeto é limitada. A revitalização nas aldeias Itaóca Tupi e Barigui é promissora e está avançando precisamente porque a iniciativa partiu da vontade das comunidades e não foi imposta por pessoas de fora. Estamos partindo de bases corretas, mas temos ciência das limitações do projeto até o momento, sobretudo no fato de estar restrito a encontros quinzenais em que a linguista e um falante nativo de uma aldeia próxima conduzem uma aula de língua Guarani Nhandewa com os professores da escola e seus alunos. Nesse modelo, o projeto está centrado na linguista e dependendo das atividades que ela desenvolve junto com o falante.

O projeto se desdobra nas aulas regulares de língua indígena da escola e aos poucos vem ganhando espaço fora da sala de aula. Nas aulas de língua indígena e nas aulas para as turmas do infantil, os professores se baseiam nas atividades feitas nos encontros quinzenais para criar atividades para seus alunos, que também têm participado dos encontros. Na escola, recentemente uma professora encabeçou um pequeno projeto para colocar placas com o nome de cada sala da escola (diretoria, cozinha, refeitório, almoxarifado, sanitários). Esse projeto envolveu a aula de língua indígena e a aula de artes para confeccionar os letreiros para serem colocados na porta das salas. Outro projeto realizado por essa professora foi a produção de um vídeo em que os alunos ensinam a fazer uma receita tradicional na cozinha da escola.

Algumas propostas para engajar mais pessoas da comunidade e criar espaços de uso para além da sala de aula foram sugeridas e tiveram simpatia dos participantes do projeto, mas ainda não houve disponibilidade e iniciativa dos mesmos para levar as propostas a cabo. Nesta fase, visa-se trabalhar a revitalização por meio de projetos que durem aproximadamente dois meses e envolvam não somente alunos e professores, mas também as famílias e outros membros das aldeias. Cada projeto resultaria em um produto concreto que utilize a língua. Alguns exemplos de projeto são: vídeos (de culinária, encenações), teatros, jogos de equipes e jogos de cartas. Para a realização de cada projeto, inicialmente faríamos o planejamento linguístico de acordo com o planejamento linguístico de base comunitária (McCARTY, 2018, AMARAL, 2021), abarcando os três pilares de planejamento situacional, em que são detalhadas as etapas necessárias para a realização do produto final, a pessoa responsável

e os atores envolvidos em cada etapa, planejamento do *corpus*, e planejamento da aquisição/transmissão.

São muitas as vantagens de trabalhar a revitalização por meio de projetos. A maior delas é o incentivo para que os professores e outros participantes se tornem protagonistas dos processos de aquisição, saindo da posição de meros alunos das aulas de língua que acontecem nos encontros quinzenais, e tornando-se os produtores/realizadores de seu próprio processo de aquisição. Também aumentam as possibilidades de engajar mais pessoas da comunidade, uma vez que produtos como vídeos e teatros podem ser exibidos para todos e isso pode gerar interesse em participar do próximo projeto. Ainda, é possível pensar em projetos que necessitem do envolvimento de pessoas que não têm relação direta com a escola, sobre plantas medicinais, horta, música, artesanato, por exemplo.

Até o momento, o incentivo para que os professores se tornem protagonistas do processo de revitalização tem ocorrido durante os encontros quinzenais, durante as aulas e atividades de língua, quando são chamados a adaptar a atividade para seus alunos.

3. Projeto ‘Yêgatu Digital’

Desde 2023, o projeto PROLIND trabalha com a língua Nheengatu (escrito *yêgatu*, na ortografia usada atualmente no Rio Negro). Essa língua é falada nativamente principalmente no nordeste do estado de Amazonas (Alto e Médio Rio Negro) e, residualmente, por algumas famílias do povo Mawé, na região de Parintins (AM), além de estar em revitalização em três áreas diferentes da Amazônia (baixo Tapajós, baixo Amazonas, baixo Madeira/Solimões), e em algumas comunidades no Nordeste do Brasil. No caso do alto Rio Negro, a língua é usada por vários povos e etnias, incluindo casos como o povo Baré em que a língua substituiu a língua original, tornando-se um marcador importante da etnia. Desde meados de 2023, temos nos envolvido com muitas lideranças e grupos relacionados à língua Nheengatu no Brasil. Isso foi feito paralelamente aos esforços de construção de modelos iniciais para os tradutores de e para o português brasileiro e ao desenvolvimento de protótipos iniciais do assistente de escrita para essa língua indígena. Após esse envolvimento inicial com essas comunidades, trabalhamos em três linhas de trabalho básicas.

Primeiramente, fechamos uma parceria com um grupo de estudantes indígenas, em sua maioria da etnia Baré, da Universidade Estadual de Campinas. Eles buscam métodos mais fáceis para escrever em sua própria língua em computadores e celulares. Essa colaboração envolveu oficinas de escrita estruturadas usando uma versão nova e aprimorada da ferramenta de assistente de escrita Nheengatu que estamos desenvolvendo, visando a levar, em última análise, ao desenvolvimento de materiais educacionais e culturais. Duas oficinas realizadas no segundo semestre de 2024 com o auxílio do assistente de escrita indicaram que o uso é bastante diferente entre indivíduos que sabem ler e escrever e aqueles que não sabem. Este fluxo de trabalho foi descontinuado no fim de 2024, devido à

dificuldade de envolver os alunos falantes de Nheengatu daquela comunidade estudantil de forma consistente, principalmente por causa das outras demandas em seu tempo.

A segunda linha de trabalho alinha-se à necessidade de serviços e materiais de tradução mais abrangentes. A disponibilidade de serviços de tradução, especialmente de línguas dominantes para línguas indígenas, é importante para aumentar a acessibilidade a serviços jurídicos, de saúde, educacionais e outros serviços governamentais. Ao mesmo tempo, é fundamental que os indígenas possam se comunicar em sua própria língua para denunciar crimes, solicitar serviços e participar de debates políticos e sociais, tornando a tradução de suas línguas para uma língua dominante também essencial para a cidadania. Vale lembrar aqui que no município de São Gabriel da Cachoeira (AM), o Nheengatu é uma língua cooficial, junto com o Tukano, o Baniwa, e o Português, porém, tal estatuto oficial raramente sai do papel.

Em ambos os cenários, o objetivo não é substituir os tradutores indígenas, mas sim tornar seu trabalho mais produtivo e confiável, permitindo que um número, muitas vezes pequeno, de tradutores atenda a um volume maior de demandas e tarefas. Atualmente, estamos estabelecendo uma colaboração com uma equipe de tradutores e escritores profissionais na língua Nheengatu, que recentemente realizou a tradução da Constituição Brasileira para o Nheengatu.

O objetivo é desenvolver uma versão do assistente de escrita adaptada para tarefas de tradução e, possivelmente, com foco em textos jurídicos e de saúde, incorporando tradutores automáticos de alta qualidade como forma de aumentar a produtividade e a qualidade. Para isso, temos desenvolvido uma versão do assistente de escrita mais adequada para trabalhos de escrita e tradução, com capacidade para lidar com textos longos, e investigando sua incorporação em editores de texto profissionais ou de código aberto. A continuidade deste trabalho foi adiada para o segundo semestre de 2025 devido às necessidades do terceiro trabalho.

O terceiro, e mais importante trabalho analisa o uso de ferramentas de IA no contexto de escolas indígenas de ensino fundamental e médio da etnia Baré em comunidades próximas à cidade São Gabriel da Cachoeira (AM). Esse município é uma das áreas com maior diversidade linguística da Amazônia e com o maior número de indígenas declaradas no censo do IBGE. O processo teve início em abril de 2024 com uma apresentação sobre o projeto durante uma assembleia de uma organização local, quando a comunidade aprovou a ideia de explorar uma parceria e a entrada dos pesquisadores e técnicos na terra indígena.

Com o apoio da Federação das Organizações Indígenas do Rio Negro (FOIRN), em setembro de 2024, visitamos duas dessas comunidades. Trata-se de aldeias indígenas típicas às margens do Rio Negro, acessíveis apenas por pequenas embarcações, a cerca de 5 horas de viagem de São Gabriel da Cachoeira. A primeira comunidade que visitamos abriga cerca de 20 famílias, possui uma escola local de ensino fundamental e médio, e só recentemente obteve acesso à internet de alta velocidade por meio de um provedor de satélite barato. Antes, eles só podiam usar a internet por rádio, em baixas velocidades. A energia elétrica ainda é um problema na região, pois a eletricidade gerada por geradores a diesel ou a gasolina é proibitivamente cara, já que o preço do combustível é muito alto,

devido aos custos de transporte. A eletricidade gerada por energia solar ainda é limitada, devido aos custos de instalação e ao baixo desempenho durante a estação chuvosa.

Tivemos três reuniões com líderes e professores dessa comunidade, cerca de 15 pessoas no total, nas quais fomos informados sobre as dificuldades enfrentadas pelos alunos para aprender Nheengatu na escola local, principalmente a falta de livros didáticos. Também nos falaram sobre as preocupações da comunidade com o uso excessivo da internet pelos jovens e como o fortalecimento do uso da língua é essencial para criar um forte senso de identidade. Em determinado momento, mostramos a eles um vídeo do assistente de escrita que estávamos desenvolvendo, com base em dados linguísticos, e na discussão que se seguiu, a maioria das pessoas concordou que poderia ser uma ferramenta a ser testada na escola com os alunos, como forma de fazê-los escrever mais e melhor. Sugerimos a realização de workshops semanais explorando o uso de ferramentas de escrita digital como parte das aulas de idiomas na escola, para os quais concordaram se também poderíamos fornecer a infraestrutura.

A segunda comunidade que visitamos é maior que a primeira, com cerca de 100 famílias, e dispõe de um programa escolar bilíngue de ensino fundamental e médio bem estabelecido. Fomos recebidos na 'oca comunitária', tradicional local de encontros mais formais, onde fomos recebidos por professores e pela liderança local e recebemos demonstrações das habilidades dos alunos na escrita da língua Nheengatu. Em seguida, participamos de uma série de reuniões, nas quais observamos, pela primeira vez, as preocupações dos professores locais com o impacto da chegada da internet via satélite e como eles estão desligando a internet às 22h para coibir o uso excessivo, especialmente por jovens.

Esta é uma comunidade onde, até a década de 1970, crianças eram enviadas para escolas missionárias, onde, com frequência, sofriam punições se falassem em língua indígena. Aprendemos com sobreviventes dessa época como o uso da língua é, ao mesmo tempo, motivo de orgulho e medo para eles. Também conversamos com alunos do ensino médio, que confirmaram a percepção dos professores de que, embora usem o nheengatu como a língua preferida para interações orais na comunidade, nunca o utilizam na internet, nem mesmo em segmentos falados nas mídias sociais. A proposta de estabelecer uma oficina semanal com alunos do ensino fundamental e médio também foi bem recebida pelos professores, que queriam facilitar o uso do Nheengatu na internet e na criação de conteúdo no idioma para atrair e motivar os jovens. No entanto, isso deve ser feito na ortografia específica da língua Nheengatu usada na região, o que motivou modificações importantes nas ferramentas em desenvolvimento.

Em maio de 2025, firmamos uma parceria entre a Universidade de São Paulo (USP), a FOIRN e essas duas comunidades que visitamos para estabelecer um programa de oficinas semanais para apoiar o uso do assistente de escrita, utilizando tablets e internet via satélite, inicialmente nas escolas dessas duas comunidades e, posteriormente, em outras escolas que, como essas, estão profundamente enraizadas na floresta amazônica. É importante ressaltar que nosso plano é trabalhar com essas comunidades no *co-design* e codesenvolvimento dessas ferramentas baseadas em IA para Nheengatu e no estudo de seu uso em contextos reais, como forma de contribuir para os esforços

da comunidade para garantir que sua língua seja usada por jovens e adultos na vida cotidiana e, cada vez mais, no ambiente digital. O programa também busca explorar, com professores e outros produtores de conteúdo, como usar essas e outras ferramentas semelhantes para produzir mais e melhor material escrito na língua em formato digital.

Como parte do acordo, implantamos em julho de 2025 duas salas de aula digitais totalmente equipadas, uma em cada escola comunitária, para apoiar essas atividades. O equipamento foi doado por um doador privado, aprovado pela FOIRN e pela USP, e compreende, para cada sala de aula, 20 tablets para os alunos, 1 laptop para os professores, 1 laptop para controlar os equipamentos da sala de aula, 1 monitor de 40 polegadas, 1 câmera de videoconferência, 1 sistema Starlink e um kit de painel solar de 250W para suportar o carregamento e o uso de todos esses equipamentos.

Nossa principal esperança é que este projeto mostre que é possível desenvolver ferramentas de apoio à escrita para línguas indígenas ameaçadas de extinção, em seus contextos típicos de recursos de dados extremamente baixos, e que essas ferramentas facilitem a transição da escrita do papel para o formato digital. As lideranças dessas comunidades enfatizaram diversas vezes sua preocupação com a chegada da internet rápida e, embora cientes das muitas novas ferramentas e facilidades que ela traz, como a internet pode se tornar mais uma ameaça à continuidade de suas línguas nativas. Nesse contexto, isso é particularmente relevante porque muitas pessoas consideram o uso de uma língua indígena a chave para sua identidade como pessoa indígena, de modo que fortalecer seu uso se torna uma forma não apenas de preservar a língua, mas também de garantir a continuidade de suas práticas culturais.

Vemos o design, o conteúdo e os recursos atuais de nossas ferramentas simplesmente como um passo inicial que permite um uso mínimo pela comunidade para impulsionar o processo de codesenvolvimento. Esperamos que mais dados sejam necessários para tornar as ferramentas úteis em um número maior de contextos. Desde meados de abril de 2025, temos colaborado remotamente com dois educadores das comunidades para aprimorar o assistente de escrita e oferecer melhor suporte ao uso escrito de suas línguas indígenas. Nosso envolvimento com eles seguiu uma abordagem participativa e iterativa, estruturada em três atividades principais:

1. Revisão do Dicionário e Verificação do Significado das Palavras:

Apoiamos e observamos o processo enquanto os professores revisavam o dicionário e corrigiam os significados das palavras, ajudando a garantir a precisão linguística e a adequação cultural dentro do aplicativo;

2. Criação de Texto dentro do Aplicativo:

Os professores foram convidados a escrever textos diretamente usando o aplicativo. Essa atividade nos permitiu identificar desafios práticos de usabilidade e coletar suas sugestões para melhorar a interface e a funcionalidade com base em cenários de escrita do mundo real. Os participantes também ajudaram a identificar palavras que podem ser

inadequadas para uso em um ambiente escolar, apoiando a criação de filtros alinhados às normas culturais e educacionais específicas da comunidade;

3. *Co-design* de ferramentas:

Paralelamente, para melhor compreender como as práticas de *co-design* intercultural têm sido aplicadas no desenvolvimento de ferramentas digitais com comunidades indígenas, realizamos uma revisão sistemática da literatura. O objetivo principal foi identificar e analisar estudos de caso que descrevessem explicitamente metodologias de *co-design*, particularmente aquelas que envolvem educação e tecnologias digitais em contextos indígenas. Essa revisão nos ajudou a refletir sobre as melhores práticas e a subsidiar possíveis adaptações para nosso trabalho em andamento com comunidades amazônicas.

4. Projeto Tupi Antigo

O tupi antigo pode ser considerado ‘a língua indígena clássica do Brasil’ (NAVARRO, 2008, 2013). Essa denominação é justificada no sentido de que o que é chamado de ‘Tupi Antigo’ ou de ‘Tupinambá’ por linguistas e antropólogos, foi o maior agrupamento linguístico no litoral atlântico quando os europeus começaram a chegar na região que futuramente seria o Brasil, no século XVI. Eram variedades dessa língua que os missionários, especialmente os da Companhia de Jesus, aprendiam, com o intuito de “convencer” os povos indígenas a se converterem ao cristianismo. A empreitada descritiva de uma série de missionários, desenvolvida ao longo da segunda metade do século XVI, culminou na publicação de algumas obras de referência: as gramáticas de Anchieta (1595) e Figueira (1621), o catecismo de Araújo (1618), e o anônimo *Vocabulário na língua brasílica* (1621). Esses textos constituem as maiores fontes do nosso conhecimento da ‘língua mais falada na costa do Brasil’, como Anchieta a denominou.

Ao mesmo tempo em que os jesuítas intitulavam seu objeto de estudo ‘língua brasílica’ nos livros que publicavam, ou seja, ‘língua indígena (do Brasil)’ (FINBOW, 2025), outro termo começou a circular para se referir ao conjunto de línguas estreitamente aparentadas que denominamos atualmente ‘tupi-guarani’ entre as quais contava a língua ‘mais falada na costa do Brasil’ descrita pelos jesuítas, a saber, ‘língua geral’. É provável que os portugueses tenham tomado esse termo dos espanhóis, que se referiam assim às línguas originárias empregadas oficialmente como línguas francas para a administração e evangelização nos extensos territórios que dominavam. Algumas dessas *lenguas generales*, como Náuatle e Quíchua, já estavam estabelecidas como línguas francas antes da conquista europeia. Outras, tais como Aimara, Guarani, Mapuche e Muísca, entre diversas outras, foram constituídas como tal, conforme as estipulações de três grandes concílios eclesiásticos em Lima, Peru, de 1551, 1567/8, e 1582/3.

No caso do Império Português, não havia nenhum estado indígena com a extensão geográfica do Império Inca ou Asteca e nenhuma outra língua era usada como uma língua franca. No entanto, conforme a noção de que uma língua geral deveria ser a de maior extensão geográfica ou demográfica na região em que seria usada, uma língua se destacava naturalmente. Foi aquela língua que é chamada de o ‘Tupi Antigo’ ou ‘Tupinambá’ hoje e que os jesuítas tinham batizado ‘a’ Língua Brasileira, por ser a maior geográfica e demograficamente. Ela ocupava uma extensa faixa do litoral, desde o estado atual da Paraíba até o sul do estado de São Paulo, onde transicionava para o Guarani. Vale notar que existiam também várias regiões na costa em que outras línguas indígenas dominavam, como o Goitacá, no norte do atual estado do Rio de Janeiro, e o Aimoré entre o Espírito Santo e o sul da Bahia.

Além da ampla presença no Estado do Brasil, quando os Portugueses iniciaram a conquista do Maranhão a partir de 1615, eles encontraram uma imensa área habitada por povos indígenas que falavam línguas tupi-guarani ao longo da margem sul do baixo Rio Amazonas e em seus grandes afluentes meridionais, o Tocantins, o Araguaia, e o Xingu. Havia também algumas nações, principalmente Tupinambá e Tobajara, que tinham emigrado do litoral do Nordeste na segunda metade do século XVI, para se refugiar dos crescentes conflitos com os invasores. Por azar, eles escolheram se fixar nas localidades amazônicas em que os portugueses viriam a se estabelecer no começo do século XVII.

Portanto, a ampla presença de povos que falavam línguas parecidíssimas com a norma codificada aprendida pelos missionários, que os jesuítas chamavam de *língua brasileira* em seus textos, e as variedades faladas nativamente pelos descendentes dos portugueses com mães indígenas das regiões litorâneas, criava a impressão de que em ambos os estados portugueses, do Brasil e do Maranhão, continham um mar de variedades linguísticas que se pareciam muito. Foi dessa situação que surgiu o conceito de ‘índios de língua geral’, que os missionários aplicavam de forma indiscriminada aos povos de língua tupi-guarani que eles encontravam e procuravam converter ao catolicismo por meio da catequese na variedade codificada.

Deste modo, ao contrário do que se tem afirmado com frequência, a distinção terminológica entre ‘Língua Brasileira’ e ‘Língua Geral’ não se refere a línguas diferentes ou sequer de fases diacrônicas diferentes da mesma língua, mas antes duas visões distintas do mesmo fenômeno: a proliferação de grupos falantes de línguas tupi-guarani encontrados pelos portugueses. ‘Língua Brasileira’ corresponde ao contato inicial no litoral, em que a maior língua indígena se tornou por metonímia “a” língua indígena por excelência, e a versão normatizada pelos missionários passou a ser o veículo da evangelização. Por outro lado, o termo ‘Língua Geral’, tomado emprestado dos espanhóis, foca na vasta extensão territorial de línguas e variedades aparentadas e com um alto grau de inteligibilidade mútua. Ou seja, ‘Língua Geral’ não é uma variedade falada por certo segmento da população da colônia, tal como Rodrigues propõe (1986, 1996). Também não há evidências a favor da criouliização pelo modelo tradicional (LEE, 2005; ARGOLO, 2011a/b, 2012a/b, 2016; OLIVEIRA; ZANETTI; MÓDOLO, 2019; ver também FINBOW, 2022, 2023, 2025).

Mais tarde, na Amazônia, no século XVIII, a Língua Geral assumiu outra função. Além de ser a língua materna da maioria dos integrantes da colônia, ela também servia como uma língua franca mesmo, no sentido de ser uma língua auxiliar falada por muitos, ao lado de sua(s) língua(s) materna(s), que não eram línguas tupi-guarani (*Tapuia*, *Nheengaíba*). Gradualmente, muitas comunidades sofriram uma troca linguística, em que a Língua Geral se tornou o veículo de comunicação principal e acabou substituindo a língua ancestral da comunidade. Assim, a palavra *Tapuia*, que antes designava indígenas que não falassem uma variedade da Língua Geral, ou seja, os não falantes de línguas Tupi-Guarani) passou a ser associada à base da pirâmide da sociedade amazônica, de tal modo que, no século XIX, esse termo se refere exclusivamente ao falante prototípico de Língua Geral.

Ao longo dos séculos, o uso das variedades do tupi do litoral ia diminuindo, a ponto de não ser transmitido entre as gerações. Não é claro, contudo, quando a ruptura na transmissão intergeracional ocorreu em cada lugar. Ainda havia pessoas que se identificavam como Tupiniquim no século XIX, como os relatos de viagem de Sainte-Hilaire, Wied-Neuwied, e D. Pedro II, entre outros, constatam, por exemplo, e nessas comunidades, algumas pessoas ainda falava uma variedade contemporânea da língua. No caso dos Potiguara da Baía da Traição (PB), relatos pessoais sugerem que é possível que algumas pessoas ainda tenham falado em algum grau na primeiras décadas do século XX, embora as investigações oficiais não as tenham registrado (SILVA & ARAGON, 2024).

5. A.I. nas revitalizações do Tupi Antigo

Hoje, três comunidades remanescentes das nações tupi do litoral, os Potiguara (PB, RN), os Tupinambá de Olivença (BA), e os Tupiniquim de Aracruz (ES) estão revitalizando sua língua ancestral. O principal apoio didático à revitalização tem sido os cursos de Tupi Antigo de Navarro (2008) e Barbosa (1956) e materiais desenvolvidos por professores das comunidades a partir de investigações próprias.

O professor de Tupi na Universidade de São Paulo, Eduardo de Almeida Navarro, tem visitado as comunidades Potiguara na Baía da Traição (PB) diversas vezes ao longo das últimas duas décadas e tem contribuído, com seus orientandos, para a produção de material didático-pedagógico e oficinas de Tupi como segunda língua para professores indígenas. Os professores locais estudam e ensinam a língua nas escolas indígenas nos municípios onde as comunidades estão localizadas.

Na USP, atualmente existem dois projetos de Mestrado que almejam utilizar a inteligência artificial para criar *chatbots* capazes de se comunicar em Tupi Antigo que podem ser usados nas revitalizações ou para o ensino da língua em outros contextos, como ocorre atualmente na USP. Um está encerrando seu Mestrado em Ciências da Computação, 'Tupi Antigo: desenvolvimento de ferramentas computacionais para tradução e preservação cultural. Tradução direta e via pivotação tradução por língua proximal', que será defendido em breve, no Instituto de Matemática e Estatística (IME). O outro está prestes a realizar o exame de qualificação do seu Mestrado em Linguística, que está sendo desenvolvido no Departamento de Linguística da Faculdade de Filosofia, Letras e Ciências Humanas. Os dois referidos projetos de pesquisa integram o Projeto descrito neste artigo, intitulado

Processamento de Línguas Indígenas/*Strengthening Brazilian Indigenous Languages with A.I.* (PRO-LIND), vinculado ao Centro para Inteligência Artificial e Aprendizado de Máquinas (CIAAM) e Centro de Inteligência Artificial (C4AI), uma colaboração entre a USP, a FAPESP e a IBM Brasil.

No projeto ‘Tupi Antigo: Desenvolvimento de Ferramentas Computacionais para Tradução e Preservação Cultural’, propõe-se o desenvolvimento de ferramentas computacionais para a tradução automática e a preservação cultural do Tupi Antigo. A pesquisa aprofundada sobre a tradução automática entre o Português e o Tupi Antigo foca em estratégias inovadoras para superar os desafios inerentes à escassez de dados paralelos, à complexidade morfológica e à baixa representatividade computacional de línguas indígenas.

Dado a ausência de suporte nativo para o Tupi Antigo em grandes modelos multilíngues, Rezende propõe uma abordagem metodológica original que ele baseia em duas técnicas complementares: a tradução direta e a tradução pivotada por *língua proximal*, em que o Guaraní Paraguaio atua como intermediário devido à sua proximidade linguística com o Tupi Antigo. Adicionalmente, aplica a camuflagem linguística (*language aliasing*), que consistiu em mascarar o Tupi sob rótulos reconhecíveis pelo modelo NLLB-200 (UTIYAMA; ISAHARA, 2007, TAN et al., 2023).

A pesquisa desenvolveu um pipeline completo que inclui curadoria de corpus, treinamento e refino supervisionado com validação cruzada (*K-Fold*), avaliação por métricas automáticas (*BLEU*, *chrF*, *TER*) e validação qualitativa com análises frase a frase. No caminho Português a Tupi Antigo, observou-se melhora significativa após o refinamento, com aumento da fidelidade semântica e fluidez textual, fato que motivou a substituição do modelo anterior pela nova versão na plataforma pública de tradução *Tupitranslator*. No entanto, no sentido Tupi Antigo a Português, os resultados apontaram para uma estagnação ou até regressão após o refinamento, indicando a necessidade de novos critérios de seleção e ampliação do corpus qualificado.

A dissertação também levanta hipóteses linguísticas para as dificuldades encontradas, como a alta polissemia do Tupi Antigo, sua estrutura aglutinante e as limitações da tokenização baseada em subpalavras para representar morfemas complexos. Além disso, argumenta-se que a preservação de línguas indígenas exige uma abordagem sociotécnica, na qual a tecnologia deve ser sensível ao contexto histórico, cultural e epistemológico das comunidades envolvidas (ZUCKERMANN; WALSH, 2013). Como contribuições teóricas e práticas, este trabalho oferece um repositório de dados, modelos treinados e uma interface pública, além de abrir caminho para aplicações futuras em PLN de baixo recurso, revitalização linguística e tradução mediada por línguas tipologicamente próximas.

No projeto de Mestrado, ‘Computer Translator for Tupi Antigo’, observa-se que as respostas que um LLM moderno como o ChatGPT dá quando solicitado a traduzir o Tupi Antigo tardio são, na melhor das hipóteses, absurdas. Atualmente, se quisermos utilizar o poder dos LLMs em nossos estudos de línguas mortas, moribundas, ou ameaçadas, estamos perdidos quando se trata de tecnologias prontas para uso. Supõe-se que, para ver tecnologia como essa disponível para línguas como o Tupi Antigo, precisamos adaptar os modelos e treiná-los usando nossos próprios recursos, já que uma empresa como o Google ou a OpenAI não certamente não apoiará uma língua sem falantes nativos.

O intuito deste estudo não é apenas de aprimorar as ferramentas disponíveis para pesquisadores de línguas indígenas com poucos recursos linguísticos, mas também se mostrará valiosa para comunidades que estão assumindo um papel ativo na revitalização de suas línguas ancestrais. Entendemos por 'línguas com poucos recursos', aquelas que não têm um corpus amplo para treinar e/ou não têm pessoas competentes suficientes para rotular os corpora existentes. O Tupi Antigo é um excelente exemplo de uma língua com poucos recursos.

A primeira etapa envolveu a digitalização da principal fonte semântica, o *Dicionário do Tupi Antigo* (NAVARRO, 2013). Isso ocorreu em 2023 com a ajuda do próprio Professor Navarro, que forneceu o texto bruto aos autores deste estudo. Certos métodos de limpeza de dados foram aplicados ao arquivo .doc do MS-Word de origem para padronizar sua codificação no estilo Windows com o padrão Unicode mais adequado. Essa etapa foi necessária porque a ortografia de Navarro utiliza símbolos pouco padronizados, como *ý ý ã ã ã ã*, o que torna as buscas em seu documento PDF apenas semifuncional, mesmo com um teclado especializado. Em seu estado atual, o dicionário em formato JSON está pelo menos 98% livre de erros de ingestão e codificação. O texto de origem é editado periodicamente, tal como os algoritmos de processamento de dados para regenerar o JSON quando erros são encontrados pela comunidade.

Com o acesso mais fácil, este dicionário tem sido mais utilizado do que nunca em sua forma digitalizada (SHEIK, 2023a). A linguística computacional em relação ao Tupi Antigo é muito mais fácil com este texto disponível em formato JSON. Os usuários atuais do dicionário também perceberam um valor renovado no texto, permitindo pesquisas em Tupi Antigo e Português em segundos, em vez de folhear páginas fisicamente ou tentar pesquisar em um PDF corrompido e grande. Esses dados fornecerão a base para o aspecto semântico do nosso tradutor, sendo capaz de fornecer traduções e frases de exemplo para todas as palavras conhecidas. Antes de podermos ver o aspecto semântico se concretizar, precisamos expor a estrutura profunda do Tupi Antigo a um LLM existente. Esses dados também nos fornecem insights fonéticos e fonêmicos valiosos, que já influenciaram fortemente a maneira como vemos a língua hoje.

Utilizando as diversas gramáticas do Tupi Antigo escritas ao longo da história, iniciamos o desenvolvimento de uma biblioteca computacional (SHEIK, 2024b) que implementa as transformações gramaticais e fonéticas presentes na língua. Essas gramáticas incluem, entre outras, (ANCHIETA, 1595), (FIGUEIRA, 1621/1686), (BETTENDORF, 1686), (BARBOSA, 1951) e, principalmente, (NAVARRO, 2008), que é uma síntese das fontes mencionadas e onde nossa ortografia base foi definida. Essas obras foram cruciais para nossa compreensão da língua, desde sua utilização ao longo do tempo até seu desuso.

Outra obra importante que também está sendo considerada é *Tupi Portuguesa kuapa* (SILVA & ARAGON, 2024), que nos mostra como os próprios povos indígenas utilizam a língua Tupi Antigo hoje e no futuro. A variedade moderna e revivalista do Tupi Antigo falada na Baía da Traição será definida neste artigo, doravante, como Tupi Potiguara.

A gramática computacional que começamos a desenvolver em 2023 já é capaz de conjugar quase qualquer verbo da língua em um sintagma verbal simples. Este foi o primeiro passo exploratório do

processo. Hoje, ela também consegue aglutinar e flexionar raízes em diferentes classes de substantivos, levando em consideração mudanças fonológicas que ocorrem como resultado de caso, tempo verbal, modo, negação, etc.

Embora complexas, todas essas regras são baseadas diretamente nas atestações e lições dos recursos mencionados sobre a língua. Esta biblioteca tem a capacidade de produzir dados sintéticos perfeitamente anotados na língua. O objetivo é esgotar todos os casos e regras expressos nessas gramáticas de origem para ter uma gramática abrangente do Tupi Antigo, escrita inteiramente em Python.

Ao longo do processo de desenvolvimento da gramática computacional do Tupi Antigo, estamos documentando e descrevendo essas características e processos gramaticais em inglês e online (SHEIK, 2024a). Estamos implementando o próprio código que realiza essas transformações, incorporado e executado no lado do cliente dentro da página web, de forma que linguistas e especialistas em línguas Tupi sem competência computacional possam usá-lo como um recurso.

Quando um erro é encontrado dentro desta gramática, o erro é corrigido a partir da biblioteca computacional. Após a recompilação da gramática inglesa, quaisquer erros serão atualizados em linha com base na biblioteca computacional Python, sem necessidade de revisão. Este aspecto manterá os autores responsáveis perante a comunidade acadêmica linguística em geral e tornará o processo de revisão por pares mais acessível a pesquisadores que não leem Python.

A principal utilidade desta gramática computacional em relação ao nosso objetivo de desenvolver um tradutor para o Tupi Antigo é a geração de dados sintéticos. Esta gramática computacional nos fornece as ferramentas necessárias para gerar uma quantidade indefinida de dados sintéticos de acordo com a sintaxe do Tupi Antigo. Todos esses dados serão completamente rotulados. Dado que outro objetivo deste estudo é encontrar um processo repetível com o qual possamos desenvolver tradutores para qualquer língua com poucos recursos, o primeiro problema que tentaremos abordar é criar recursos artificiais que exponham as regras e a estrutura profunda da língua na forma de big data.

Depois de produzirmos uma quantidade imensa de dados sintéticos semanticamente duvidosos, mas sintaticamente sólidos, treinaremos vários modelos de indução gramatical de última geração com base nesses dados sintéticos. No momento, pequenos modelos LLM, como BERT e GPT2, têm se saído muito bem nos casos de conjugações verbais para rotular os dados que nossa biblioteca computacional já é capaz de sintetizar. Os resultados são muito promissores. Se esse método não for escalonado eficientemente com o comprimento do texto de entrada e/ou a complexidade gramatical, temos uma abundância de técnicas não relacionadas à IA à nossa disposição com as quais podemos testar nossos casos de uso específicos.

Ao criar anotações semissupervisionadas, daremos aos nossos modelos as ferramentas necessárias para receber entrada na forma de Tupi Antigo, que segue a ortografia de (Navarro, 2008), a qual é, por sua vez, uma adaptação da grafia empregada em Barbosa (1951), e então decifrar o texto em sua estrutura profunda. Este formato de saída, com cada classe gramatical e tradução semântica

das raízes claramente identificadas, será muito mais fácil para um LLM intermediário compreender do que a estrutura superficial com a qual as pessoas estão tentando treiná-lo.

Antes deste modelo, anotar o Tupi Antigo era um processo que precisava ser iniciado do zero. Após o desenvolvimento do modelo POS, esse processo seria de validação, e não de produção. Isso reduz significativamente a pressão sobre o linguista humano, eliminando grande parte da atividade repetitiva. Esses resultados da correção humana podem então ser realimentados no processo de treinamento do modelo, corrigindo-o onde ele se desviou.

A gramática computacional também permitiu aos autores criar ferramentas voltadas para o treinamento de linguistas em Tupi Antigo. Um exemplo disso pode ser visto no teste de conjugação de verbos (SHEIK, 2023), no qual o usuário é apresentado a casos aleatórios de verbos conjugados do dicionário (NAVARRO, 2013). Este é um exemplo de recurso didático que utiliza a gramática computacional para facilitar os processos de transferência de conhecimento do livro para o cérebro, sem explicações linguísticas, em um ambiente formal de sala de aula. Acreditamos que este método seja muito mais eficaz na retenção em comparação com exercícios gramaticais tradicionais no contexto do Tupi Antigo. Aqueles que obtiverem pontuações altas neste teste podem ser recrutados para ajudar a verificar os resultados de nossas anotações geradas com o marcador de classes gramaticais do LLM que treinaremos com os dados sintéticos.

Assim que o modelo POS atingir um nível satisfatório de precisão, nós o executaremos em todos os corpora digitalmente disponíveis do Tupi Antigo. O processo de rotulagem passará então por um processo de rotulagem semântica. Nossa hipótese é que essas informações ricas e profundas sobre a estrutura podem ser inseridas em quase qualquer estado da parte GPT ou LLM para receber uma tradução adequada.

Em um subprojeto relacionado intitulado ‘Desenvolvimento de um Corpus Glosado de Tupi Antigo’, visa-se apoiar o projeto de *chatbots* de Tupi Antigo, criando um corpus da língua glosado morfológica e sintaticamente. As informações fornecidas nas glosas não estão prontamente disponíveis nos textos originais e auxiliarão na identificação de diferenças estruturais e organizacionais entre o Tupi Antigo e as línguas próximas, como o Guaraní Paraguaio e o Nheengatu, que estão sendo usadas para desenvolver os *chatbots*, ou seja, a etapa da “Última Milha” do projeto.

Atualmente, três alunos de graduação estão trabalhando em glosar os textos curtos e elementares e frases de exemplo do *Método Moderno de Tupi Antigo*, de Navarro (3ª ed., São Paulo: Editora Global, 2008), o principal livro didático usado para ensinar Tupi Antigo, como parte do projeto de Iniciação Científica. O objetivo é treinar os alunos em técnicas de glosa linguística ao mesmo tempo em que adquirem um conhecimento mais profundo da gramática da língua antes de passar a glosar os extensos textos originais escritos nos séculos XVI e XVII, como o *Catecismo Brasílico* de Araújo (1ª ed., 1618, 2ª ed., 1687) e as poesias e peças de Anchieta (CARDOSO, 1993a/b, NAVARRO, 1997, 1999), e o *Compêndio da doutrina cristã* de Bettendorff (1686), entre outros, bem como as cartas seiscentistas da família Camarão (NAVARRO, 2022), os únicos textos sobreviventes escritos por indígenas falantes nativos do tupi antigo durante o período colonial.

6. Considerações Finais

As consultas às comunidades foram demoradas, pois se realizaram de acordo com a disponibilidade dos professores e lideranças, que foram mediadas pelo Comitê Inter Aldeias para os Guarani e pela Fundação dos Povos Indígenas do Rio Negro (FOIRN). O processo levou dois anos e gerou inicialmente um pedido de oficina no ensino médio dos Mbyá. Um tradutor automático Mbyá foi iniciado, mas foi descontinuado a pedido de lideranças da comunidade. Em seguida, recebemos um pedido para oficinas de revitalização entre os *Nhandewa* Tupi-Guarani de Itaóca Tupi que foi ampliado posteriormente para abranger a comunidade de Barigui, e o estabelecimento de uma parceria para o desenvolvimento de ferramentas de IA para a língua Nheengatu. Foi então iniciado um tradutor automático para o Nheengatu que obteve bons resultados e um grande apoio de duas comunidades de etnia Baré do ARN.

Pode-se avaliar que o projeto deu certo em duas frentes: (1) uma parceria através da FOIRN para uma ferramenta de tradução e assistentes de escrita da língua Nheengatu em Tabocal dos Pereira e Juruti, no Alto Rio Negro. Nestas localidades, foram montadas salas de aula com equipamento para videoconferência, na qual os alunos podem usar 20 tablets para aprender a utilizar as ferramentas desenvolvidas pelo projeto; (2) uma parceria com a comunidade de Itaoca Tupi em Mongaguá, litoral de São Paulo, onde oficinas quinzenais estão sendo realizadas na escola e materiais de ensino de segunda língua (L2) estão sendo criados junto com a comunidade a partir de materiais já existentes para *Nhandewa* como primeira língua (L1). Para tanto, bases de dados estão sendo criadas e alimentadas, e serão deixadas com a comunidade para a criação de futuros materiais.

Em ambos os casos, a linguística está sendo usada a serviço do fortalecimento das línguas indígenas. Isso porque no litoral de São Paulo o *Nhandewa* já está deixando de ser falado pelas novas gerações, e os professores das escolas de Itaóca Tupi e Barigui sentiram a necessidade de envolver a língua em suas atividades escolares em busca de um processo de retomada. Nas comunidades do Alto Rio Negro o Nheengatu é falado, mas escreve-se em Português. Em um mundo digital, onde antenas Starlink foram instaladas e os jovens passam muito tempo ao seu redor jogando e utilizando as mídias sociais, acreditamos que a língua Nheengatu terá mais chances de sobreviver se puder ser usada de forma escrita nos meios digitais.

A situação dos tradutores de Tupi Antigo é um pouco diferente, por dois motivos: 1) não há comunidades de falantes nativos e todos os dados da língua estão a princípio no domínio público, e 2) por existirem interesses entre grupos em retomada linguística, como os Potiguara e *Tupinakyia* (Tupiniquim), como também entre grupos não indígenas que estudam e pesquisam essa língua. Em decorrência deste estatuto especial do Tupi Antigo, sem questões a respeito da origem dos materiais de base ou do acesso aos dados primários, o foco dos pesquisadores nos projetos descritos acima tem sido primeiramente no desenvolvimento de aplicativos e programas que funcionem de forma eficiente, para depois passar a adaptá-las às certas necessidades específicas das comunidades de usuários, se for necessário. No entanto, as mesmas diretrizes têm sido aplicadas que constam no documento da década das línguas indígenas resultante do encontro do GT na Aldeia Buriti em 2025.

Temos a seguinte resolução: “Item 1.1 Criar e garantir o acesso a cursos digitais voltados para o fortalecimento das línguas indígenas, utilizando a tecnologia como ferramenta de aprendizagem”.

Que outras ferramentas podemos propor a partir desta demanda dos povos indígenas? Em parcerias com as comunidades chegaremos a novas propostas. Gostaríamos de mencionar apenas duas possibilidades neste momento.

Ferramentas que os linguistas já usam para a documentação, como o FLEX e ELAN, podem e devem tornar-se mais fáceis de utilizar através da IA. O trabalho dos pesquisadores e dos usuários dos arquivos pode ser facilitado. Tanto a digitação dos dados pode ser automatizada quanto o acesso aos arquivos dentro de acervos deve ser feita de uma maneira mais ágil.

Acreditamos, ainda, que temos que nos esforçar para promover iniciativas a fim de que, em um futuro próximo, os falantes de línguas indígenas tenham como trabalhar em plataformas de transcrição e tradução de arquivos de mídia gravados em suas línguas nativas. Assim, haverá um incentivo financeiro para que os falantes continuem trabalhando permanentemente em documentação linguística e cultural. Os museus, em parcerias público-privadas, talvez consigam implementar plataformas deste tipo com a ajuda de profissionais de IA.

Informações complementares

Avaliação e resposta dos autores

Avaliação: <https://doi.org/10.25189/rabralin.v24i2.2387.R>

Editora

Maria Filomena Spatti Sandalo

Afiliação: Universidade Estadual de Campinas

ORCID: <https://orcid.org/0000-0003-4595-7765>

RODADAS DE AVALIAÇÃO

Avaliador 1: Bruna Franchetto

Afiliação: Universidade Federal do Rio de Janeiro

ORCID: <https://orcid.org/0000-0001-7158-3838>

Avaliador 2: Evandro Bonfim

Afiliação: Universidade Federal de São Carlos

ORCID: <https://orcid.org/0000-0002-0459-928X>

AVALIADOR 1

O artigo avaliado representa, sem a menor dúvida, uma contribuição original e relevante por tratar de uma tecnologia novíssima e cada vez mais usada, diria dominante no atual universo digital: Inteligência Artificial (IA). Para além dos avanços trazidos pelo uso da IA e dos pro e contra apontados e discutidos no mundo todo, a sua aplicação no campo das línguas nativas ou minorizadas ou indígenas é um desafio pelas dificuldades de sua "adaptação" e das responsabilidades éticas que envolve. O artigo expõe com muita clareza as experiências em andamento na USP por um grupo de pesquisadores de alto nível, bem como possibilidades e desafios, através de um texto bem estruturado, compreensível e acessível por um público mais amplo. Título, sumário e conteúdo são impecáveis. Recomendando a publicação, tendo apenas duas observações: (1) gostaria que os autores desenvolvessem minimamente o que eles dizem na Seção 2 (Projeto Guarani Mbyá): "O trabalho com a língua Guarani Mbyá, no entanto, foi descontinuado, pois as lideranças ficaram desconfortáveis com o fato de sua língua estar sendo objeto de tradução automática". Foram só as lideranças (liderança é um termo genérico)? O que seria o mencionado "desconforto" com relação à tradução automática? Seria interessante saber um pouco mais acerca desta reação. (2) o artigo precisa ser submetido a uma revisão do português, corrigindo uns poucos erros e uniformizando a sigla IA (não AI).

AVALIADOR 2

O texto é bem claro e detalhado a respeito das experiências discutidas, incluindo as metodologias e objetivos dos projetos, as particularidades de cada caso/língua/povo indígena, com destaque para o cuidado ético na relação com tais populações e o potencial das tecnologias desenvolvidas de acordo com os interesses das comunidades em temas como fortalecimento e revitalização das suas línguas ancestrais.

Gostaríamos de recomendar apenas uma reflexão maior, talvez nas considerações finais, sobre quais as concepções de língua envolvidas nas decisões de cada povo, pois dimensões como o segredo, por exemplo, fazem parte da ontologia da linguagem de muitos povos indígenas, assim como seria igualmente interessante entender como são concebidas as línguas indígenas que não correspondem à língua ancestral (no caso do Nheengatu) nem a primeira língua, implicando em outra forma de se relacionar com as tecnologias. Outro ponto importante também é a questão da relação entre a oralidade e a escrita, complexa desde os meios analógicos e com especificidades próprias nos ambientes digitais, e a diferença entre as experiências que partem de comunidades em que a língua permanece mais ativa na modalidade oral e da experiência do Tupi Antigo que parte sobretudo de textos escritos. Em relação ao Tupi Antigo, apontamos ainda o caso dos Potiguara do sertão do Ceará, que conseguiram co-oficializar a língua e utilizam de uma combinação do método desenvolvido por Eduardo Navarro e aporte do Nheengatu do Alto Rio Negro.

Conflito de Interesse

Os autores não têm conflitos de interesse a declarar.

Link para *Preprint*

DOI: 10.22541/au.176063088.81298563/v1.

Protocolo e Pré-Registro de Pesquisa

Os autores confirmam que avaliaram os roteiros propostos pela Equator Network, consideramos que nenhum deles se mostra relevante para a pesquisa em tela. Também informamos que a pesquisa desenvolvida não foi pré-registrada em repositório institucional.

Declaração de Disponibilidade de Dados

O compartilhamento de dados não é aplicável a este artigo, pois nenhum dado novo foi criado ou analisado neste estudo.

REFERÊNCIAS

AMARAL, L. How can linguists support language revitalization efforts in the Brazilian context. *Linguistica*, v.17, n.2, pp. 177-193, 2021.

ANCHIETA, J. de. **Arte de gramática da língua mais usada na costa do Brasil**. 1. ed. Coimbra: João Álvares, 1595.

ARGOLO, W. Línguas gerais na história social-linguística do Brasil. **PAPIA**, São Paulo, v. 26, n. 1, p. 7-52, 2016.

ARGOLO, W. História linguística do Sul da Bahia: levantando hipóteses e iluminando caminhos. **Entrepalavras**, v.1, n. 2, p. 270-292, 2012a.

ARGOLO, W. A língua geral da Amazônia como um sistema historicamente novo: jesuítas e tapuias na origem do contexto com interrupção de transmissão linguística entre gerações. In: MATTOS E SILVA, R. V.; OLIVEIRA, K.; AMARANTE, J. (org.). **Várias navegações: português arcaico, português brasileiro, cultura escrita no Brasil, outros estudos. Em homenagem a Therezinha Barreto**. Salvador: EdUFBA, p. 479-513, 2012b.

ARGOLO, W. **Introdução à história das línguas gerais no Brasil: processos distintos de formação no período colonial**. Dissertação de mestrado. Universidade Federal da Bahia, 2011a.

ARGOLO, W. Língua geral na Bahia: comarcas de Ilhéus e Porto Seguro. In: CARVALHO, C. dos Santos; ROCHA, F. Aninger de Barros; PARCERO, L. M. de Jesus (orgs.). **Discurso e cultura: diálogos interdisciplinares**. Salvador: EdUNEB, p. 99-109, 2011b.

BARBOSA, A. Lemos. **Curso de Tupi Antigo: Gramática, Exercícios, Temas de Versão, Vocabulário**. 1. ed. Rio de Janeiro: Livros de Portugal, 1951.

BETTENDORF, J. F. **Catecismo na Língua Brasileira**. 1. ed. Lisboa: Miguel Deslandes, 1686.

CARDOSO, A. **José de Anchieta. Catecismo brasílico, I (Doutrina Cristã)**. São Paulo: Edições Loyola, 1993a.

CARDOSO, A. **José de Anchieta. Catecismo brasílico, II (Doutrina Cristã) – Doutrina autógrafa e confessional**. São Paulo: Edições Loyola, 1993b.

CAVALIN, Paulo, H. Domingues, J. Nogima, C. Pinhanez. *Understanding Native Language Identification for Brazilian Indigenous Languages*. **AmericasNLP workshop of ACL'23**. 2023.

CAVALIN, P., C. Pinhanez, J. Nogima. *Human Evaluation of the Usefulness of Fine-Tuned English Translators for the Guarani Mbyá and Nheengatu Indigenous Languages*. **ILLC workshop of Propor'24**. 2024a.

CAVALIN, P., P. Domingues, C. Pinhanez, J. Nogima. *Fixing Rogue Memorization in Many-to-One Multilingual Translators of Extremely-Low-Resource Languages by Rephrasing Training Samples*. **NAACL 24**. 2024b.

CAVALIN, P., P. H. Domingues, C. Pinhanez. *Sentence-level Aggregation of Lexical Metrics Correlates Stronger with Human Judgements than Corpus-level Aggregation*, **AAAI 25**. 2025.

DJATSY, Lenira Dina de Oliveira et al. (2018). **Lições de gramática Nhandewa/Tupi- Guarani**. Vol 2. Caderno de Atividades multidisciplinares. Brasília, DF: FUNAI.

DOMINGUES, P, C. Pinhanez, P. Cavalin, J. Nogima. *Quantifying the Ethical Dilemma of Using Culturally Toxic Training Data in AI Tools for Indigenous Languages*. **SIGUL workshop of LREC-COLING'24**. 2024.

FIGUEIRA, L. de. **Arte da Língua Brasileira**. 1. ed. Lisboa: António Álvares, 1621.

FINBOW, T. D. The emergence and nature of the 'Língua Geral Amazônica' in accordance with Mufwene's Language Ecology model. **Revista do GEL**. v. 19, n. 2, p. 75-112, 2022.

FINBOW, T. D. Uma reanálise de dois casos de mudança estrutural entre o tupi antigo e a língua geral amazônica/nheengatu tidos como exemplos de criouliização e a apresentação de um caso de mudança gramatical no nheengatu do século XIX por contato com o português. In: LIMA-HERNANDEZ, M. C.; SANTOS, M. M. Soares; ADRIANO, P. Soma (orgs.), **Estudos de aquisição, Mudança e Ensino-aprendizagem de línguas**. VIIIº Simpósio Mundial de Língua Portuguesa (SIMELP 8) / IIIº Simpósio Interdisciplinar de Estudos de Linguagem (SINTEL 3), Brasil – Angola, 4 a 7 de outubro de 2022. vol. 6. São Paulo: Estige Editorial, 2023.

FINBOW, T. D. A sociophilological account of the formation and evolution of the term *Língua Geral*, with emphasis on Amazonia. **Cadernos de linguística, no prelo**, 2025.

GERARDI, F. Ferraz. **UD for Tupinambá**. [S. l.: s. n.], 2024. <https://universaldependencies.org/tpn/index.html>. Accessed: 2024-05-22.

GERARDI, F. Ferraz. **TuLaR**. [S. l.: s. n.], 2024. <https://tular.clld.org/>.

GOMES, M., T. Finbow. *Por uma abordagem perspectivista ecológica do contato linguístico entre português e nheengatu*. **Revista Diálogos (Revdia)**, v. 11, p. 1-29, 2023.

- LEE, M. K. **Conversing in Colony: The Brasilica and the Vulgar in Portuguese America 1500-1759**. Ph.D. Dissertation, John Hopkins University, 2005.
- MARCOLINO, C. et al. **Lições de gramática Nhandewa-Guarani**. Vol 1. FUNAI/Campinas, SP: Curt Nimuendaju, 2016.
- MCCARTY, T. Community-based language planning: Perspectives from indigenous language revitalization. In: HINTON, L.; HUSS, L. M.; ROCHE, G. (eds.). **The Routledge Handbook of Language Revitalization**. New York and London: Routledge, chap. 3, 2018, pp. 22-35.
- NAVARRO, E. de Almeida. **José de Anchieta, Poemas. Lírica portuguesa e tupi**. São Paulo: Editora Martins Fontes, 1997.
- NAVARRO, E. de Almeida. **Método moderno de Tupi Antigo. A língua do Brasil dos primeiros séculos**. 3 ed., São Paulo: Editora Global, 2008.
- NAVARRO, E. de Almeida. **Dicionário de Tupi Antigo. A língua indígena clássica do Brasil**. São Paulo: Global Editora, 2013.
- NAVARRO, E. de Almeida. **Curso de Língua Geral (Nheengatu ou Tupi Moderno) A Língua das Origens da Civilização Amazônica**. 2ª edição (corrigida e aperfeiçoada). São Paulo: Centro Angel Rama da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, 2016.
- NAVARRO, E. de Almeida. Transcrição e tradução integral anotada das cartas dos índios Camarões, escritas em 1645 em tupi antigo. **Boletim do Museu Paraense Emílio Goeldi, Ciências Humanas**, Belém, v. 17, n. 3, 2022.
- OLIVEIRA, M. S. D. de; ZANOLI, M. de L.; MODOLO, M. O conceito de “Língua Geral do Brasil” revisitado à luz da linguística de contato. **Journal of Ibero-Romance Creoles**, v. 9, n.1, p. 306-333, 2019.
- PINHANEZ, C., P. Cavalin, L. Storto, T. Finbow, A. Cobbinah, J. Nogima, M. Vasconcelos, P. Domingues, P. Mizukami, N. Grell, M. Gongora, I. Gonçalves. Harnessing the power of artificial intelligence to vitalize endangered indigenous languages: technologies and experiences, **ArXiv**. Cornell University. (submetido arxiv 2024).
- PINHANEZ, C., P. Cavalin, M. Vasconcelos, J. Nogima. *Balancing Social Impact, Opportunities, and Ethical Constraints of Using AI in the Documentation and Vitalization of Indigenous Languages*. **AI for Good special track, IJCAI'23**. 2023.
- RODRIGUES, A. Dall'Igna. **Línguas brasileiras. Para o conhecimento das línguas indígenas**. São Paulo: Edições Loyola, 1986.
- RODRIGUES, A. Dall'Igna. As línguas gerais sul-americanas. **PAPIA. Revista brasileira de crioulos e similares**, v. 4, n. 2, p. 4-18, 1996.
- SHEIK, K. **Verb Quiz**. 2023. Available from: <https://kiansheik.io/nhe-engu/quiz/>.
- SHEIK, K. **Tupi Grammar Python**. 2024a. Available from: <https://github.com/kiansheik/nhe-engu/tree/main/tupi/tupi>.
- SHEIK, K. **Tupi Grammar in English**. 2024b. Available from: <https://kiansheik.io/nhe-engu/grammar>.
- SILVA, J. R. Guyraakanga Potiguara Araújo da & ARAGON, C. Coelho (orgs.). **Tupi Potiguara kuapa. Conhecendo a língua Tupi Potiguara**. Mamanguape (PB), 2024..

TAN, X.; REN, Y.; HE, D.; QIN, T.; XIAO, T.; LIU, Z.; ZHANG, Y.; CHEN, P.; LI, J.; ZHOU, M.. Multilingual neural machine translation with knowledge distillation. In: **International Conference on Learning Representations (ICLR)**, 2019.

UTIYAMA, M.; ISAHARA, H. A Comparison of Pivot and Direct Translation Approaches in Statistical Machine Translation. **Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics**, 2007, p. 488-495. Disponível em: <https://aclanthology.org/P07-1061.pdf>

ZUCKERMANN, G.; WALSH, M. Revival Linguistics and the Community: Towards a Sustainable Model for Language Revitalization. **Journal of Language Documentation and Conservation**, v. 7, 2013, p. 28-52.