

Língua minoritárias e ferramentas linguísticas: ações do projeto de Iniciação Científica

RESUMO

Este trabalho visa apresentar as ações do projeto de Iniciação Científica (PIBIC) da Universidade X, intitulado X. Dentre as ações conduzidas pelo projeto, este artigo objetiva apresentar o uso de ferramentas linguísticas em associação ao processo de documentação e de descrição de línguas minoritárias. Procuramos relatar ações vinculadas à linguagem CoNLLU na construção de *Treebanks* da *Universal Dependencies* (MARNEFFE et al., 2021) para línguas indígenas, em especial para duas línguas Tupi: Akuntsu e Makurap. Desta forma, partindo da experiência vivenciada no Projeto de PIBIC, apresentamos a eficácia do uso de algumas ferramentas de descrição linguística articuladas à documentação, à descrição, à produção de materiais didáticos e suas relações com o processo de revitalização e fortalecimento de línguas indígenas brasileiras.

PALAVRAS-CHAVE: Ferramenta linguística. Línguas minoritárias. Universal dependencies.

Carolina Coelho Aragon

ca.carolina@ymail.com

Universidade Federal da Paraíba, João Pessoa, Paraíba, Brasil.

Fabrizio Ferraz Gerardi

fabrizio.gerardi@uni-tuebingen.de

Universidade de Tübinga, Baden-Württemberg, Alemanha.

Luana Luiza da Silva Santos

luana.luiza@academico.ufpb.br

Universidade Federal da Paraíba, João Pessoa, Paraíba, Brasil.

Júlia Verônica da Silva

julia.silva@academico.ufpb.br

Universidade Federal da Paraíba, João Pessoa, Paraíba, Brasil.

1. INTRODUÇÃO

As questões de educação e de política de línguas, em especial as línguas minoritárias, são fatores de discussões em diferentes áreas, as quais vêm apontando ações de promoção de diversidade cultural e linguística como vetor de desenvolvimento econômico e social (MORELLO, 2009), principalmente neste primeiro ano da Década Internacional das Línguas Indígenas, fundamentada pela “Declaração de Los Pinos”¹ da Unesco.

Segundo o último censo do Instituto Brasileiro de Geografia e Estatística, há 274 línguas indígenas no Brasil. Esses, porém, são dados questionados por especialistas que consideram o número exagerado. Muitas dessas línguas estão ameaçadas de extinção, seja pelo número reduzido de falantes ou por não mais transmitirem o conhecimento linguístico-cultural para as novas gerações. Na maioria desses casos, estratégias de revitalização e de conservação dessas línguas são imprescindíveis nas comunidades, as quais necessitam de processos contínuos de promoção de políticas de línguas e ações voltadas à educação escolar indígena. A educação escolar indígena no Brasil é reconhecida pela constituição de 1988 no artigo 210, a qual apresenta, de forma geral, um olhar para a diversidade cultural do Brasil e garante o uso das línguas indígenas nas escolas. Com o passar do tempo, outras leis e documentos foram surgindo como: a Lei de Diretrizes e Bases da Educação Nacional (Lei nº 9.394) de 1996; o Referencial Curricular Nacional para as Escolas Indígenas criado em 1998; Plano Nacional de Educação (2001); o Parecer 14/99 do Conselho Nacional de Educação; a Resolução 03/99 do Conselho Nacional de Educação; o Decreto Presidencial 5.051, com a convenção 169 da Organização Internacional do Trabalho – extremamente importante para os povos de recente contato e para os povos isolados; e o Decreto nº 6.861 de maio de 2009.

Com base nesses documentos, fica claro a garantia de uma educação escolar bilíngue, respeitando as diferentes identidades. Porém, várias questões são levantadas durante a execução dessas leis, a começar pela relação entre pedagogia e ensino indígena bilíngue dentro de uma sociedade majoritariamente monolíngue em língua portuguesa. Ao longo dos anos, os povos indígenas têm enfrentado negações de suas línguas, culturas e direitos, sendo a educação escolar indígena, de fato, um desafio (MINDLIN, 2004; D'ANGELIS, 2012). Embora professores indígenas em todo o território nacional ainda enfrentem problemas com a ausência ou a pouca produção de material didático específico para as escolas indígenas, nas últimas décadas, muitos educadores indígenas passaram a ser elaboradores de materiais didáticos em parceria com instituições públicas, organizações não-governamentais e pesquisadores, fortalecendo a educação indígena. Portanto, promover práticas para a promoção e difusão dos saberes indígenas é valorizar e respeitar os povos indígenas, além de ampliar o conhecimento científico em diferentes áreas acadêmicas e setores da sociedade.

Atualmente, com o avanço de tecnologias, ferramentas linguísticas vêm auxiliando não apenas na documentação e na descrição de línguas minoritárias, como também no desenvolvimento de materiais didáticos para as comunidades indígenas. Refletindo sobre essas questões – educação, política de línguas e ferramentas linguísticas –, este trabalho visa apresentar as ações desenvolvidas no Programa Institucional de Bolsas de Iniciação Científica (PIBIC) da Universidade X intitulado “X”, descrevendo as metodologias e as ferramentas linguísticas utilizadas para analisar e para documentar línguas indígenas minoritárias, visando

a produção tanto de materiais didáticos quanto de usos linguísticos variados para as comunidades indígenas e acadêmicas.

2. LÍNGUAS MINORITÁRIAS

O ano de 2022 foi proclamado pela Unesco o início da Década Internacional das Línguas Indígenas (2022-2032) – DILI. Com esses desdobramentos, os povos indígenas estão promovendo ações de trabalhos voltadas à valorização e difusão das línguas indígenas, dentre outras atividades. Segundo o Instituto de Investigação e Desenvolvimento em Política Linguística (IPOL)

“Compreendemos que a DILI é um momento de articular ações que já acontecem em vários lugares, estreitar os laços e colocar essas ações em rede, construindo políticas linguísticas para as línguas indígenas brasileiras. A DILIA é o momento de construirmos um espaço colaborativo, conectar projetos de fortalecimento, revitalização, retomada, tanto aqueles protagonizados pelos próprios povos, quanto aqueles que são construídos em parceria ou dirigidos por organizações não indígenas. Nesse sentido, pretendemos, durante a década, visibilizá-los, apoiá-los e encontrar meios para financiá-los.” (IPOL, 2021)²

Muitas das línguas indígenas da América do Sul estão ameaçadas de extinção com níveis de vitalidade diferenciados, que vão de internacional – amplamente falada, usada internacionalmente – a extinta – não mais usada pelo falante e ninguém mais possui um senso de identidade étnica associado à língua (EBEHARD et al., 2021). No Brasil, as línguas indígenas são consideradas línguas minoritárias, não apenas pelo número de falantes, como por estarem em situação de vulnerabilidade.

Políticas de línguas estão crescendo no Brasil e frentes de conservação e de preservação de línguas minoritárias estão sendo organizadas, como por exemplo, os trabalhos desenvolvidos pelo Instituto de Patrimônio Histórico e Artístico (IPHAN) com a criação do Inventário Nacional da Diversidade Linguística (INDL). O INDL serve, portanto, de “instrumento oficial de identificação, documentação, reconhecimento e valorização das línguas faladas pelos diferentes grupos formadores da sociedade brasileira”³. Até o momento, foram reconhecidas seis línguas indígenas como Referência Cultural Brasileira: Asuriní, Guaraní Mbya, Nahukuá, Matipu, Kuikuro e Kalapalo.

As línguas indígenas brasileiras, segundo Rodrigues (1986), estão classificadas em dois grandes troncos linguísticos (Macro-Jê e Tupi); em famílias menores (e.g. Aruák, Karib, Pano, Tukano, Arawá, Katarina, Nadahup (Maku), Nambikwara, Txapakúra, Yanomani, Bóra, Guaikurú e Múra); em línguas isoladas, que não pertencem a nenhuma agrupamento linguístico (e.g. Aikanã, Kanoê, Kwazá, Irântxe, Mynký, Trumai e Tikuna) (SEIFART; HAMMARSTARÖM, 2018); além de pidgins e línguas crioulas faladas no norte do Brasil, na fronteira com Amapá e Guiana Francesa (Galibí Marworno (Galibí de Uaçá) e o Karipuna do Norte (Karipuna do Amapá))⁴.

Com as frentes de expansão nacional, que ainda reverberam nos dias atuais no Brasil, muitos povos indígenas foram extintos e populações drasticamente reduzidas e integradas forçadamente à sociedade nacional, refletindo não apenas na perda cultural como na linguística. De acordo com dados de Aryon Rodrigues

(2006), no início da chegada da população europeia no Brasil, seriam faladas cerca de 1200 línguas no território brasileiro – atualmente, esse número, muito provavelmente, não ultrapassa 150 línguas. Considerando o número de línguas faladas no Brasil, no início da década de 2000, afirmou-se que cerca de 21% estavam ameaçadas de extinção (MOORE et al., 2008). Muitas comunidades indígenas brasileiras contam com pouquíssimos falantes, sendo que muitas delas, estão em sérios riscos de extinção.

O processo de perda de uma língua ocorre sem que seus membros possam decidir sobre tais perdas, as quais são irreparáveis, pois com elas, perde-se muito, como, por exemplo: a compreensão do valor geral da diversidade linguística; a identidade; o repositório histórico encontrado em cada língua; a soma do conhecimento humano que cada língua adiciona; o direito de existência (CRYSTAL, 2002; NETTLE; ROMAINE, 2000). Essa perda linguística pode ter distintos motivos que vão desde o desaparecimento físico dos falantes – em decorrência de epidemias, extermínio direto, escravização, redução de territórios, destruição das condições de sobrevivência – a aculturação forçada presente na história de um Brasil semeada por um contínuo processo de genocídio dos povos indígenas brasileiros (e.g. RAMOS, 2018).

Sobre o processo de documentação e descrição linguística, é importante observar que, embora se confundam, possuem diferenças fundamentais em suas estruturas. A documentação linguística trata do registro das práticas linguísticas de um povo, dessa forma captura as tradições de uma comunidade, incluindo aspectos culturais, pois registra a prática e funcionamento da língua em diversos contextos. Já a descrição linguística está mais relacionada ao registro de uma língua a partir de seus aspectos gramaticais, entendendo “gramática” como as regras de funcionamento de uma língua independente de sua norma padrão (HIMMELMANN, 2012). Nos dois casos vale ressaltar que se tratam de trabalhos interdisciplinares e que contam com a colaboração da tecnologia para documentação e armazenamento dos dados.

Já sobre a metodologia de pesquisa vinculada às ações de revitalização de línguas, Wilmar D’Angelis (2006) argumenta que é necessário existir um retorno positivo e efetivo para a comunidade estudada, mantendo a comunidade interessada nos resultados da pesquisa, de modo que ela possa participar ativamente de todas as etapas de investigação. Nas palavras de D’Angelis (2016), a pesquisa colaborativa “uma pesquisa cujo escopo foi definido pela própria comunidade que, a partir daí, demandou apoio de linguista(s); ou se trata de uma pesquisa proposta pelo linguista, mas cujo foco e abrangência precisam ser negociados com a comunidade envolvida.” (ibid, p. 60). Pensar em ensino no contexto indígena é, portanto, se adequar às necessidades da comunidade, incentivando e retomando valores culturais da tradição do povo, malhando o aprendizado a partir das experiências territoriais. Isso envolve um conhecimento a partir da língua do povo, valorizando sua identidade para fortalecer suas relações com a língua dominante, o português, e não o contrário, como pontua D’Angelis (2000):

Um grave problema acontece quando as comunidades indígenas são pressionadas ou enganadas para tomar certas decisões e adotar uma certa política linguística, sem saber exatamente onde isso vai dar. Por exemplo, a comunidade pode aceitar e, até apoiar certos programas de ensino que se dizem “bilíngues” que, na verdade, são programas

de ensino que usam a língua indígena na escola, mas que funcionam como desvalorizadores e enfraquecedores da língua (D'Angelis, 2000, p.1).

Com base nessas breves discussões, as ressaltamos como elementos fundantes para a construção de arcações teóricos que motivaram e sustentaram as ações deste projeto vinculadas ao uso de ferramentas linguísticas, as quais descrevemos na próxima seção.

2.1 AKUNTSU E MAKURAP

Dentre as línguas minoritárias brasileiras, o projeto de PIBIC aqui descrito trabalha com duas línguas Tupi (Família Tupari): Akuntsu e Makurap⁵. Atualmente, três mulheres constituem a totalidade do grupo Akuntsu: Pugapia, Aiga e Babawro, as quais vivem na Terra Indígena (TI) Rio Omerê, estado de Rondônia (ARAGON; ALGAYER, 2020). O primeiro contato oficial da Fundação Nacional do Índio (Funai) com os Akuntsu aconteceu no ano de 1995. Nessa época, os Akuntsu estavam reduzidos a sete membros: dois homens, Konibú e Pupák; três mulheres, Pugapia, Aiga e Ururu; e duas crianças, Babakíp e Babakop (SANTOS; ALGAYER, 1995). O histórico desse povo e de outros, como os Kanoé do Omerê, é marcado por fugas e mortes ocasionadas por fazendeiros locais intensificadas no ano de 1985-1986, quando o sudeste do estado de Rondônia começou a ser ocupado por movimentos de “colonização” na região Norte brasileira incentivada pelo Governo Federal⁶. Na década de 1970, a região do sudeste de Rondônia foi destruída etnoambientalmente com a implantação de projetos intensificados na década de 1980 sob um slogan propagandeado pelo então presidente Médici de “uma terra sem homens, para homens sem terra”⁷.

Já os Makurap, possuem uma história distinta dos Akuntsu, mas ao mesmo tempo semelhante quando retratamos as violências sofridas e as enormes perdas populacionais que tiveram ao longo dos anos de exploração do seu território tradicional e dos recursos naturais ali encontrados, incluindo os impactos dos seringais. Os Makurap estão hoje divididos entre as TI Rio Branco e Rio Guaporé em Rondônia, ocupando as margens do lado esquerdo do Rio Branco e do Rio Colorado há anos (MALDI, 1991). De acordo com a autora, essa situação atual é fruto da desterritorialização que ocorreu na década de 1940-1960 com o aumento dos seringais, um processo intenso e marcado nas narrativas dos Makurap, bem como nas de todos os povos do Vale do Guaporé (ver MEZACASA, 2021).

3. O PROJETO PIBIC E SUAS AÇÕES

O projeto de PIBIC X insere-se na temática de educação linguística e etnohistória da Universidade X. Suas ações contemplam o conhecimento linguístico, cultural e histórico dos povos originários do Brasil, objetivando ampliar a documentação e a descrição de línguas indígenas, bem como proporcionar materiais pedagógicos para as comunidades indígenas. O projeto fundamenta-se nos documentos e diretrizes de Educação Indígena, como por exemplo, a Lei de Diretrizes e Bases da Educação Nacional (Lei nº 9.394) de 1996 e o Referencial Curricular Nacional para as Escolas Indígenas criado em 1998, bem como nas discussões sobre letramento e alfabetização em contextos indígenas e nas

discussões teóricas relacionadas à descrição de línguas e ao uso de ferramentas linguísticas. Busca responder às seguintes perguntas: a) Qual a melhor forma de relacionar os conhecimentos linguísticos, culturais e etnohistóricos na produção de materiais didáticos?; b) Como relacionar a análise linguística e histórica com as necessidades atuais de educação escolar indígena nas comunidades foco deste projeto?; e c) Como a relação entre a linguística descritiva e computacional pode ampliar e fortalecer o conhecimento linguístico sobre esses povos?

Justifica-se, desta forma, por promover materiais pedagógicos e acadêmicos que proporcionem a disseminação do conhecimento sobre línguas e culturas indígenas, e por possibilitar aos estudantes de graduação do curso de Letras da Universidade X e demais interessados, experiências na área de análise linguística e ensino-aprendizagem de línguas, fortalecendo o campo da formação de futuros pesquisadores e docentes na área de línguas indígenas.

Com o avanço tecnológico, houve também um desenvolvimento no volume de ferramentas linguísticas para organizar e programar análises quantitativas e qualitativas de um grande número de dados. As ferramentas criadas buscam especificações e padrões que alcancem conformidades relacionadas ao FAIRness data (Findable, Accessible, Interoperable, Reproducible) (WILKINSON, 2016). Com isso, a introdução de técnicas computacionais vem facilitando a criação e o compartilhamento de dados linguísticos em larga escala (big data).

Um exemplo disso, é o desenvolvimento de treebanks – corpus de textos com anotações sintáticas e morfológicas – e materiais didáticos originados de seus resultados são exemplos de ferramentas linguísticas importantes para a documentação e a revitalização de línguas minoritárias, como também para as pesquisas no âmbito cognitivo e psicológico e para as análises comparativas e tipológicas das línguas naturais (faladas e de sinais). Com o repertório de anotações nos treebanks, os dados podem ser facilmente acessados para que materiais pedagógicos possam ser produzidos de uma forma mais rápida, de acordo com a necessidade e as escolhas das comunidades.

Logo, ao buscar estabelecer conexões entre línguas minoritárias e ferramentas linguísticas, este projeto vem buscando desenvolver ações mais específicas que fazem parte de um projeto de pesquisa maior intitulado X. O X é uma base de dados de anotações em treebanks, vinculado ao X, que hoje já possui treebanks de nove línguas Tupi, a saber: Akuntsu, Guajajara, Ka’apor, Karo, Makurap, Munduruku, Guarani antigo, Teko e Tupinambá, cada um em diferentes fases de desenvolvimento. Os treebanks documentados no X são organizados de modo que seja possível a visualização das análises de dependência das frases, dos seus componentes sintático-morfológicos⁸. A tabela 01 apresenta informações sobre as línguas que fazem parte do X atualmente e a quantidade de frases anotadas até o presente momento⁹.

Tabela 1: Status¹⁰ das línguas e quantidade de frases dos treebanks – X

Línguas	Glottocode	Falantes	Status	Frases
Akuntsu	akun1241	3	Quase extinta	243
Guajajara	guaj1255	12000	Forte/Vigorosa	1126
Ka'apor	urb1250	600	Em desenvolvimento	83
Karo	karo1305	200	Forte/Vigorosa	674
Makurap	maku1278	40	Em desaparecimento	31
Munduruku	mund1330	5000	Ameaçada	158
Guarani Antigo	oldp1258	0	Extinta	59
Teko	emer1243	400	Forte/Vigorosa	100
Tupinamba	tupi1273	0	Extinta	546

Esse banco de dados do X envolve anotações de dados linguísticos obtidos de textos e gramáticas já publicados – embora alguns textos na língua Akuntsu sejam inéditos. Uma característica relevante do X é sua terminologia unificada para as anotações morfológicas das línguas Tupi, o que possibilita uma análise diacrônica dessas línguas. Ao consultar distintas descrições de línguas Tupi já publicadas, chegou-se a uma terminologia geral para que as características morfológicas e seus valores sejam os mesmos, tanto quanto possível, para todas essas línguas (RODRÍGUEZ et al., 2022).

Com isso, as ações do PIBIC envolveram diferentes etapas, dentre as que são foco deste artigo, descrevemos a construção dos treebanks da Universal Dependencies (UD) para as línguas Akuntsu e Makurap.

3.1 TREEBANKS

A UD é um programa de anotações morfossintáticas de línguas naturais, baseadas em um sistema arbóreo: os treebanks, os quais são formados a partir do princípio de que a linguagem possui uma relação hierárquica (MARNEFFE et al., 2021). Um aspecto importante, além das anotações, consiste na documentação de cada treebank que inclui descrições sintáticas e morfológicas (NIVRE et al., 2020).

A UD tem como um de seus princípios a descrição do maior número de línguas possíveis. Considerando a diversidade estrutural entre as diversas línguas, a UD descreve, da forma mais universal possível, as classes de palavras, as quais são divididas em oito partes: substantivo, verbo, adjetivo, advérbio, pronome, preposição, conjunção e interjeição. Há ainda as categorias específicas de classes

de palavras, conhecidas por XPOS, as quais são consideradas subclasses das partes do discurso. Portanto, voltando ao princípio de que é possível captar a estrutura da linguagem baseada no núcleo e seus dependentes, as anotações consistem em encontrar essas dependências em cada frase anotada.

Para visualizar melhor a discussão aqui formulada, devemos nos atentar para como funciona, de maneira prática, o sistema de descrição de línguas utilizando a UD. Para isso, dividimos a explicação em duas partes: 1) a descrição de uma palavra a partir de uma frase utilizando um editor de código-fonte – o Sublime Text¹¹; e 2) a formação das dependências a partir do Annotatrix¹² (TYERS et al., 2017).

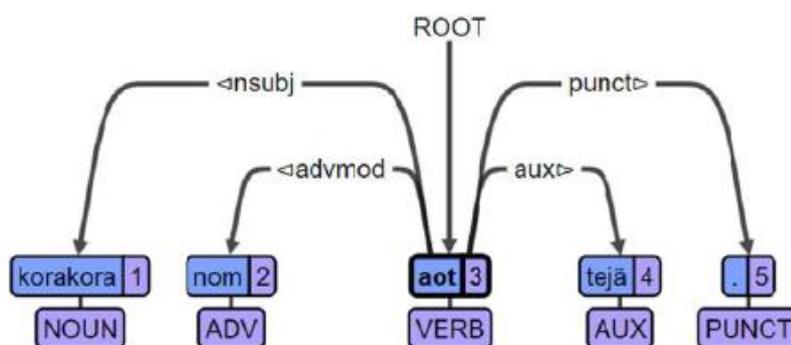
Em um primeiro momento, transferimos as frases de trabalhos publicados da língua Akuntsu (ARAGON, 2014; ARAGON, 2008) e da língua Makurap (BRAGA, 2005) para o editor de texto Sublime Text. Com isso, as frases puderam ser transferidas para o formato CoNLLU, adotado pela UD, que consiste em dez colunas que correspondem aos seguintes campos: Coluna 1. Índice numérico; Coluna 2. O lexema ou morfema como na frase em questão; Coluna 3. Lema (forma base) ou radical da palavra; Coluna 4. Classes de palavras (pré-definidas); Coluna 5. Classe de palavras mais específicas para a língua em questão; Coluna 6. Traços morfológicos e seus possíveis valores; Coluna 7. Núcleo; Coluna 8. Relação de dependência com o núcleo; Coluna 9. Subtipos de relações de dependências; Coluna 10. Qualquer outra informação. A Figura 1 exemplifica uma anotação em formato CoNLLU na língua Akuntsu. A partir daí começamos um trabalho de organização e correção dos erros que o programa, geralmente, gera ao descrever as frases.

Figura 1: Formato CoNLLU: *korakora nom aot tejã*. ‘A galinha não está saindo.’ (ARAGON, 2014, pp. 208).

```
# text = korakora nom aot tejã .
# text_eng = Chicken is not going out (5.19c)
1 korakora korakora NOUN n _ 3 nsubj _ _
2 nom nom ADV adv Advmod 4 advmod _ _
3 aot aot VERB vi _ 4 root
4 tejã jã AUX aux Person=3|Reflex=Yes 3 aux _ _
5 . . PUNCT punct _ 3 punct _ _
```

Após a organização dos textos no *Sublime Text*, as frases são anotadas no *Annotatrix*, onde combinamos abordagens manuais com métodos computacionais para a anotação do corpus linguístico. Começamos anotando manualmente um subconjunto dos dados linguísticos de acordo com a estrutura UD – as características morfossintáticas das frases são codificadas usando aspectos linguísticos, como as: *tags* POS (classes de palavras da UD) e as relações de dependência que envolvem as já encontradas na UD, bem como as categorias específicas de cada língua (as particularidades morfossintáticas que refletem traços tipológicos específicos das línguas Tupi¹³). As relações de dependências correspondentes a essas anotações estão representadas na Figura 2.

Figura 2: Relações de Dependência: *korakora nom aot tejã*. ‘A galinha não está saindo.’ (ARAGON, 2014, pp. 208).



Por último, os dados anotados são vinculados ao Github¹⁴ e publicados pela UD ao passar por um processo de validação, que constitui uma análise de erros das relações de dependências e de traços morfossintáticos anotados para cada treebank.

Após todo este processo de anotações, objetivamos, com o uso dos treebanks, ampliar a documentação e a descrição de línguas minoritárias, buscando a inserção do maior número de textos possíveis. No caso da língua Akuntsu, as anotações procuram sistematizar as descrições já publicadas como também trabalhar com textos inéditos (UD_Akuntsu-treebank)¹⁵, seguindo a metodologia apresentada acima, como por exemplo o texto anotado 'k^wai 011.13', representado parcialmente abaixo. Nesta parte do texto é possível visualizar a estrutura da descrição: sent_id (identificador de sentença), text (texto na língua alvo), text_eng (tradução para o inglês) – as próximas versões incluirão traduções para o português das frases recentemente adicionadas no Github.

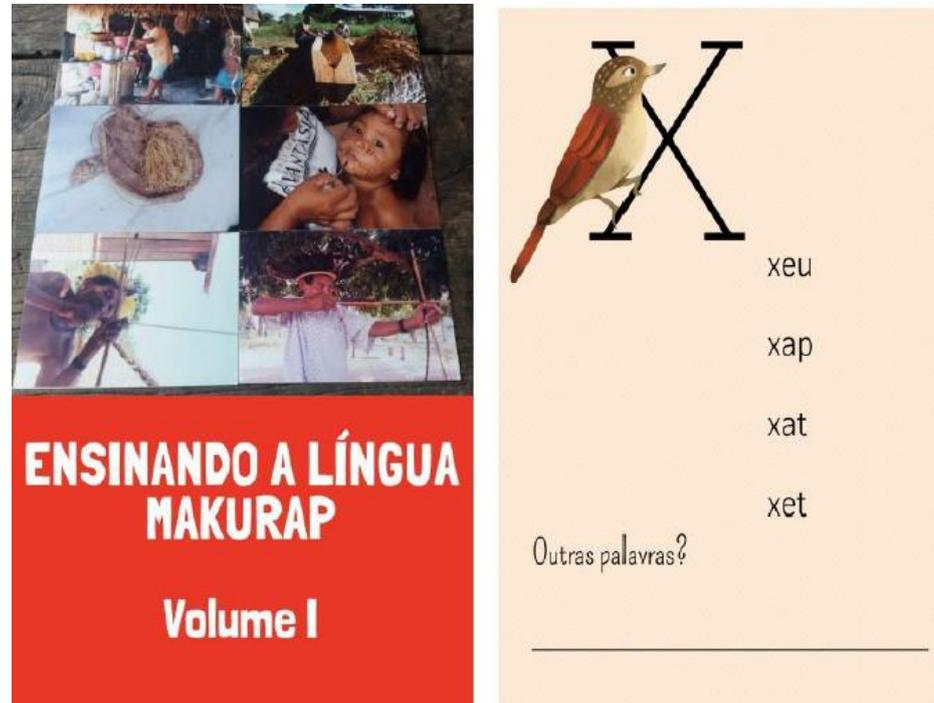
```
# sent_id = 0011.13
# text = ikip āka put . ikip pītka piri put māpara .
# text_eng = Her leg that way hit it. Her leg pierced because of the foot, it is over.
```

1	ikip	kíp	NOUN	n	Person=3	3	obj	-	-
2	āka	āka	PART	pcl	-	3	discourse	-	-
3	put	put	VERB	ideo	-	0	root	-	-
4	.	.	PUNCT	punct	-	3	punct	-	-
5	ikip	kí	NOUN	n	Person=3	6	obj	-	-
6	pītka	pīt	VERB	vt	Trans=Yes	3	parataxis	-	-
7	piri	pi	NOUN	ḥ	Case=Abl	6	obl	-	-
8	put	put	VERB	vi	-	6	parataxis	-	-
9	māpara	māpara	VERB	vt	-	6	parataxis	-	-
10	.	.	PUNCT	punct	-	6	punct	-	-

Além disso, com a ampliação dos *treebanks*, um dos resultados das ações do projeto concernente à língua Makurap culminou na publicação de um material didático – o primeiro livro de alfabetização para as escolas Makurap de autoria do professor Agnaldo Makurap “Ensinando a língua Makurap. Volume I” – para as escolas das TIs Rio Branco e Guaporé (RO), em conjunto com professores e professoras indígenas. Nesse intuito, o projeto voltou-se a um dos seus objetivos: revitalizar e fortalecer línguas minoritárias dentro de uma pesquisa colaborativa. Essa produção na língua Makurap (Foto 1), portanto, trata-se de uma colaboração para o processo de revitalização dessa língua, visto que a partir desse trabalho, professores indígenas poderão ter suportes didáticos para o ensino da língua Makurap para as crianças que

frequentam as escolas dentro das TIs. Dessa forma, este trabalho, além de englobar o aspecto da pesquisa, ampliou seus resultados para a comunidade estudada, envolvendo-a não apenas como o alvo da pesquisa, mas como parte interessada.

Foto 1 - Livro “Ensinando a língua Makurap” (Fonte: X)



Por fim, com a compreensão dos *treebanks* e a descrição da metodologia empregada para as anotações de dados de línguas minoritárias, pode-se imaginar que as ferramentas linguísticas aqui explicadas são demasiadamente complexas para se trabalhar na graduação, visto que para anotar as línguas é preciso ter um conhecimento morfossintático avançado. Acerca da interface do projeto, os alunos de graduação envolvidos relatam que não tiveram dificuldades, visto que este tem uma estrutura organizacional muito lógica e possível de compreender, desde as *tags* POS às explicações do próprio site da UD, que, em conjunto com discussões e explicações práticas durante as reuniões do projeto, o manuseio das ferramentas foi realizado com sucesso pelos alunos. Já com respeito às relações arbóreas no programa *Annotatrix*, os alunos de graduação puderam se basear nas classes de palavras espelhadas pela UPOS, dado que, desta forma, a compreensão do todo passou a ser mais simples, como por exemplo, quando compreenderam noções básicas como o núcleo ser sempre a raiz da frase, e este a base das demais relações de dependências.

4. CONCLUSÃO

Este artigo, ao descrever algumas das ações do Projeto de PIBIC, procurou relacionar a pesquisa linguística às temáticas que envolvem as línguas minoritárias

brasileiras, buscando estabelecer conexões entre documentação, descrição, revitalização de línguas e ferramentas linguísticas. Desta forma, objetivamos associar a importância de estabelecer relações com a linguística computacional em conjunto com

a descritiva como um meio de fortalecer não apenas a descrição de línguas ameaçadas de extinção, como de promover conhecimentos sobre suas comunidades, suas histórias e a importância da diversidade linguística. Pois, ao investigar a diversidade de línguas, conseguimos entender e compreender estruturas gramaticais diferenciadas, importantes para os estudos linguísticos, uma vez que compreender as línguas é também entender a estrutura de seus sistemas linguísticos nos seus mais diversos contextos de uso, os quais são vitais para ampliar os dados tipológicos existentes sobre as línguas Amazônicas.

Em conclusão, destacamos a importância do contato dos alunos de graduação com os tópicos abordados neste artigo, não apenas focando aspectos acadêmicos, como também os sociais, ao proporcionar espaços para a formação profissional de docentes e pesquisadores em formação. Com a continuação deste projeto, esperamos que as atividades propostas possam contribuir para que os alunos continuem a compreender a relevância das línguas indígenas e que os professores indígenas envolvidos possam ter ferramentas necessárias para o fortalecimento e revitalização de suas línguas.

Ao entender a história e a organização das línguas indígenas brasileiras, procura-se também ressaltar os impactos sociais que influenciaram e influenciam a identidade dos povos indígenas e as relações linguísticas intrínsecas à cultura. Espera-se que os participantes deste projeto possam ter uma visão clara sobre a diversidade linguística e sobre como a língua de cada povo indígena está associada aos direitos e às políticas de línguas. Busca-se, desta forma, fortalecer os mecanismos de parceria entre a universidade pública e as demandas dos povos indígenas, tendo em vista a urgência na mitigação dos impactos sociohistóricos impostos às línguas indígenas ao longo dos processos violentos de contato. Além de que, esperamos, com este artigo, possibilitar um olhar amplo para linguística, observando-a como um leque de possibilidades que nos auxilia a compreender os espaços sociais e culturais do Brasil. Quando se trata de línguas indígenas, olhar para elas é restaurar parte de histórias silenciadas, além de nos despertar para o fortalecimento das políticas de línguas dos povos originários como também das políticas públicas de proteção desses povos.

Minority languages and linguistic tools: actions of the Scientific Initiation program

ABSTRACT

This paper aims to present the results of the project X from Institutional Scientific Initiation Scholarship Program (PIBIC) part of the X University. Among the activities carried out by this project, this article focuses on providing a discussion on the usage of linguistic tools in association with the documentation and the description of minority languages. Here we seek to show some tasks regarding the CoNLLU language used in treebanks of the Universal Dependencies (MARNEFFE et al., 2021) for minority languages, especially for two Tupian languages: Akuntsu and Makurap. Thus, based on the PIBIC experiences, we show the efficiency of the use of linguistic tools, computational resources, to the documentation, the description, and the production of pedagogical materials in order to strengthen Brazilian indigenous languages.

KEYWORDS: Linguistic tool. Minority languages. Universal dependencies.

NOTAS

1 - Declaração de Los Pinos. <<https://es.unesco.org/news/declaracion-pinos-chapoltepek-sienta-bases-planificacion-global-del-decenio-internacional>> Acesso em: 25/05/2022.

2 - IPOL.<<http://ipol.org.br/a-decada-internacional-das-linguas-indigenas-no-brasil/>>Acesso em 01 de outubro de 2022.

3 - INDL. <<http://portal.iphan.gov.br/indl>>. Acesso em 01 de outubro de 2022.

4 - Ver também Campbell e Grondona (2012).

5 - As escolhas dessas duas línguas deu-se X.

6 - Para compreender o histórico de contato desses povos e o que acontecia no sudeste de Rondônia nessa época, sugerimos o documentário “Corumbiara” (2009) de Vincent Carelli.

7 - BRASIL. Decreto-Lei Nº 1.106, de 16 de junho de 1970. Cria o Programa de Integração Nacional, altera a legislação do imposto de renda das pessoas jurídicas na parte referente a incentivos fiscais e dá outras providências. Brasília, 1970. <http://www.planalto.gov.br/ccivil_03/decreto-lei/1965-1988/Del1106.htm> Acesso em 10 junho de 2022.

8 - De regra, há também tradução das frases para o inglês, com exceção do treebank Teko, que inclui traduções para o francês e o treebank Guajajara que também conta com traduções para o espanhol.

9 - O número de frases refere-se à versão 2.10, lançada em maio de 2022. A próxima versão, 2.11, será lançada em 15 de novembro de 2022.

10 - De acordo com a classificação de Eberhard et al. (2021).

11 - Editor de código HTML com linguagem Python.

12 - Ferramenta linguística para anotações dos treebanks.

13 - Para maiores detalhes sobre os métodos computacionais utilizados, ver RODRÍGUEZ et al. (2022).

14 - Plataforma gerenciadora de programas open source.

15 - Ver X.

REFERÊNCIAS

ARAGON, C.; ALGAYER, A. A história contada pelos Akuntsú: ocupação territorial e perdas populacionais. **Revista Brasileira de Linguística Antropológica**, [S. l.], v. 12, n. 1, p. 223–234, 2020. Disponível em: <https://periodicos.unb.br/index.php/ling/article/view/29633>. Acesso em: 16 out. 2022.

ARAGON, C. C. **A Grammar of Akuntsú, a Tupian language**. Tese (Doctor of Philosophy in Linguistics) - University of Hawaii at Manoa, Honolulu, 2014. Disponível em: <http://etnolinguistica.wdfiles.com/local--files/tese%3Aaragon2014/CarolinaAragonFinal.pdf>. Acesso em: 29 set. 2022.

ARAGON, C. C. **Fonologia e aspectos morfológicos e sintáticos da língua Akuntsú**. Dissertação (Mestrado em Linguística) - Departamento de Linguística, Português e Línguas Clássicas, Universidade de Brasília, Brasília (DF), 2008. Disponível em: https://repositorio.unb.br/bitstream/10482/5135/1/2008_CarolinaCoelhoAragon.pdf. Acesso em: 29 set. 2022.

BRAGA, A. O. **Aspects morphosyntaxiques de la langue makurap-tupi**. Tese (Doctorat en Sciences du Langage) - Université de Toulouse - Le Mirail, Toulouse, 2005. Disponível em: <http://www.etnolinguistica.org/tese:braga-2005>. Acesso em: 16 out. 2022.

BRASIL. Constituição. **Constituição da República Federativa do Brasil**. Brasília, DF: Senado Federal: Centro Gráfico, 1988.

BRASIL. Lei no 9.394, de 20 de dezembro de 1996. **Lei de Diretrizes e Bases para a educação nacional**.

BRASIL. **Parecer 14/99 do Conselho Nacional de Educação**, de 10 de novembro de 1999.

BRASIL. **Resolução 03/99 do Conselho Nacional de Educação**, de 10 de novembro de 1999.

BRASIL. Lei no 10.172, 9 de janeiro de 2001. **Plano Nacional de Educação**.

BRASIL. **Lei de Diretrizes e Bases da Educação Nacional**, LDB. 9394/1996.

CAMPBELL, L.; GRONDONA, V. **The Indigenous Languages of South America: A Comprehensive Guide**. Berlin, Boston: De Gruyter Mouton, 2012.

CRYSTAL, D. **Language death**. Cambridge University Press, 2002.

D'ANGELIS, W. **Alfabetizando em comunidade indígena**. [S.l.]: Portal Kaingang, 2000. Disponível em: <http://www.portalkaingang.org/alfabetizando.pdf>. Acesso em: 25 abr. 2021.

D'ANGELIS, W. Quando os falantes nativos são os professores, os professores são pesquisadores, e os linguistas são parceiros. **Signo y seña**, n. 29, p. 63-77, 2016.

D'ANGELIS, W. **Aprisionando sonhos: a educação escolar indígena no Brasil**. Editora Curt Nimuendajú, 2012.

DE MARNEFFE, M. C.; MANNING, C. D.; NIVRE, J.; ZEMAN, D. Universal Dependencies. **Computational linguistics**, [s.l.], v. 47, n. 2, p. 255-308, jun, 2021.

ETHNOLOGUE: languages of the World. Disponível em: <https://www.ethnologue.com/>. Acesso em: 16 out. 2022

HIMMELMANN, N. P. Linguistic data types and the interface between language documentation and description. **Language documentation & conservation**, [s.l.], v. 6, p. 187-207, 2012.

JENSET, G. B.; MCGILLIVRAY, B. **Quantitative historical linguistics: A corpus framework**. Oxford University Press, 2017.

MALDI, D. O complexo cultural do Marico: sociedades indígenas dos rios Branco, Colorado e Mequens, afluentes do Médio Guaporé. In: FURTADO, L. G. **Boletim do Museu Paraense Emílio Goeldi**, Série Antropologia, v. 7, n. 2. Belém: Museu Paraense Emílio Goeldi, 1991. p. 209-269.

MEZACASA, R. **Por histórias indígenas: o povo Makurap e o ocupar seringalista na Amazônia**. Tese (Doutorado em História), Universidade de Santa Catarina, Florianópolis, 2021. Disponível em: <https://repositorio.ufsc.br/handle/123456789/226949>. Acesso em: 29 set. 2022

MINDLIN, B. A política educacional indígena no período 1995-2002: algumas reflexões. **Revista de estudos e pesquisas**, FUNAI, Brasília, v. 1, n. 2, p. 101-140, dez., 2004.

MOORE, D.; GALUCIO, A.V.; GABAS JR, N. **O Desafio de Documentar e Preservar as Línguas Amazônicas**. Belém, Museu Paraense Emílio Goeldi, 2008.

MORELLO, R. Diversidade no Brasil: Línguas e Políticas Sociais. **Synergies Brésil**, v. 7, p. 27-36, 2009. Disponível em: <http://gerflint.fr/Base/Bresil7/bresil7.html>. Acesso em: 16 out. 2022

NETTLE, D.; ROMAINE, S. **Vanishing voices: The extinction of the world's languages**. Oxford University Press on Demand, 2000.

NIVRE, J.; DE MARNEFFE, M.; GINTER, F.; HAJIC, J.; MANNING, C. D.; PYYSALO, S.; SCHUSTER, S.; TYERS, F.; ZEMAN, D. Universal dependencies: An evergrowing multilingual treebank collection. **European Language Resources and Evaluation**, Marseille, v. 2, p. 4034-4043, mai., 2020. Disponível em: <https://aclanthology.org/2020.lrec-1.497.pdf>. Acesso em: 26 out. 2022

RAMOS, Alcida R.. Vivos, afinal! Povos indígenas do Brasil enfrentam o genocídio. **Série Antropologia**, v. 461, 2018.

RODRIGUES, Aryon Dall'Igna. **Línguas Brasileiras: para um conhecimento das línguas indígenas**. São Paulo: Edições Loyola, 1986.

RODRIGUES, A. As línguas indígenas no Brasil. In: RICARDO, F.; RICARDO, B. **Povos Indígenas no Brasil**. São Paulo: Instituto Socioambiental, 2006. p. 58-63

RODRIGUES, A. D.; CABRAL, A. S. A. C. Tupían. In: CAMPBELL, L.; GRONDONA, V. **The Indigenous Languages of South America: A Comprehensive Guide**. Berlin, Boston: De Gruyter Mouton, 2012. p. 495-574.

RODRÍGUEZ, L. M.; MERZHEVICH, T.; SILVA, W.; TRESOLDI, T.; ARAGON, C.; GERARDI, F. F. Tupían Language Resources: Data, Tools, Analyses. In: **PROCEEDINGS OF THE 1st ANNUAL MEETING OF THE ELRA/ISCA SPECIAL INTEREST GROUP ON UNDER-RESOURCED LANGUAGES**, 2022, Marseille. Anais de Evento. Paris: European Language Resources Association, 2022. p. 48-58

SANTOS, M. ALGAYER, A. **Índios Isolados do Vale do Corumbiara**. Fundação Nacional do Índio, Brasília. Ms, 1995. Relatório Técnico.

SEIFART, F.; HAMMARSTRÖM, H.. Language isolates in South America. In: CAMPBELL, L. **Language isolates**. New York: Routledge, 2017. p. 260-286.

TYERS, F.; SHEYANOVA, M.; WASHINGTON, J.. In: **PROCEEDINGS OF THE CEEDINGS OF THE 16h INTERNATIONAL WORKSHOP ON TREEBANKS AND**

LINGUISTIC THEORIES. 2018, Prague, Czech Republic. Anais de Evento. Praga: Jan Hajič, 2017 p. 10-17.

WILKINSON, M. D.; DUMONTIER, M.; AALBERSBERG, I. J.; APPLETON, G.; AXTON, M.; BAAK, A.; MONS, B. The FAIR guiding principles for scientific data management and stewardship. **Scientific data**, [s.l.], v. 3, n. 1, p. 1–9, 2016. Disponível em: <https://www.nature.com/articles/sdata201618#citeas>. Acesso em: 17 out. 2022.

Recebido: 23 mar. 2024

Aprovado: 20 nov. 2024

DOI: 10.3895/rl.v26n48.16072

Como citar: ARAGON, C.C.; GERARDI, F.F.; SANTOS, L.L.S.; SILVA, J.V. Linguas minoritárias e ferramentas linguísticas: ações do projeto de Iniciação Científica. *R. Letras*, Curitiba, v. 26, n. 48, p. 121-137, jan./jun. 2024. Disponível em: <<https://periodicos.utfpr.edu.br/rl>>. Acesso em: XXX.

Direito autoral: Este artigo está licenciado sob os termos da Licença Creative Commons-Atribuição 4.0 Internacional.

