UNIVERSIDADE FEDERAL DA PARAÍBA

CENTRO DE CIÊNCIAS HUMANAS LETRAS E ARTES

COORDENAÇÃO DO CURSO DE LETRAS PORTUGUÊS

LUANA LUIZA DA SILVA SANTOS

ANOTAÇÃO DE *CORPUS* EM AKUNTSÚ (TUPÍ, TUPARÍ) E O MODELO DE DEPENDÊNCIAS UNIVERSAIS: ASPECTOS DESCRITIVOS E ETNO-HISTÓRICOS

João Pessoa

2023

LUANA LUIZA DA SILVA SANTOS

ANOTAÇÃO DE *CORPUS* EM AKUNTSÚ (TUPÍ, TUPARÍ) E O MODELO DE DEPENDÊNCIAS UNIVERSAIS: ASPECTOS DESCRITIVOS E ETNO-HISTÓRICOS

Trabalho de Conclusão de Curso apresentado ao Departamento de Língua Portuguesa e Linguística da Universidade Federal da Paraíba, como pré-requisito para a obtenção do título de Licenciada em Letras Português.
Orientadora: Profa. Dra. Carolina Coelho Aragon

João Pessoa

2023

LUANA LUIZA DA SILVA SANTOS

ANOTAÇÃO DE *CORPUS* EM AKUNTSÚ (TUPÍ, TUPARÍ) E O MODELO DE DEPENDÊNCIAS UNIVERSAIS: ASPECTOS DESCRITIVOS E ETNO-HISTÓRICOS

Trabalho de Conclusão de Curso apresentado ao Departamento de Língua Portuguesa e Linguística da Universidade Federal da Paraíba, como pré-requisito para obtenção do título de Licenciada em Letras Português, sob orientação da Profa. Dra. Carolina Coelho Aragon

Aprovado em: __/__/____

Banca examinadora:

_____

Profa. Dra. Carolina Coelho Aragon
(Orientadora/UFPB)

_____

Prof. Dr. Magdiel Medeiros Aragão Neto
(Examinador/UFPB)

_____

Profa. Dra. Rosana Costa de Oliveira
(Examinadora/UFPB)

_____

Prof. Dr. José Wellisten Abreu de Souza
(Suplente/UFPB)

Catalogação na publicação
Seção de Catalogação e Classificação

S237a Santos, Luana Luiza da Silva. Anotação de corpus em Akuntsú (Tupí, Tuparí) e o modelo de dependências universais : aspectos descritivos e etno-históricos. / Luana Luiza da Silva Santos. - João Pessoa, 2023.
48 f. : il.

Orientadora : Carolina Coelho Aragon.
TCC (Graduação) - Universidade Federal da Paraíba/Centro de Ciências Humanas, Letras e Artes, 2023.

1. Descrição linguística. 2. Dependências universais. 3. Treebanks. 4. Línguas indígenas. 5. Akuntsú. I. Aragon, Carolina Coelho. II. Título.

UFPB/CCHLA CDU 81

Elaborado por KARLA MARIA DE OLIVEIRA - CRB-15/485

AGRADECIMENTOS

À minha mãe, Eugênia Silva, que sempre foi minha maior apoiadora durante minha jornada acadêmica, me dando forças e acreditando em mim quando nem eu mesma acreditava.

Ao meu pai, Gilson Barbosa, por sempre enxergar meu potencial e me apoiar com palavras doces e amigas.

À minha irmã, Luiziane Silva, por me apoiar durante esses anos e ser uma amiga com quem pude contar e me aconselhar.

À minha companheira, Jaciára Araújo, por toda paciência durante os longos períodos de leitura e estudos e por me apoiar durante os anos de graduação.

Aos meus colegas, Ana Maria, José Carlos e Layane Ferreira, por tornarem meus dias na UFPB agradáveis e mais leves, bem como por escutarem por horas meus monólogos sobre línguas indígenas e linguística computacional.

À professora Carolina Aragon, por me proporcionar a imersão neste universo encantador que é a descrição de línguas indígenas e me orientar com tanta paciência, compromisso e cuidado.

RESUMO

O presente trabalho descreve as anotações de *corpus* da língua Akuntsú (Tupí, Tuparí) e o modelo de Dependências Universais aplicado à descrição de línguas indígenas. Nesse contexto, estabelecemos relações interdisciplinares entre os estudos linguísticos e a área da Etno-história. As línguas indígenas são marcadas pelo apagamento histórico (Cavalcante 2011), o que reflete, dentre outros aspectos, na perda de diversidade linguística, de identidade e de valores significativos para toda uma comunidade. Este trabalho, portanto, justifica-se na importância do uso de ferramentas tecnológicas utilizando o *Annotatrix* e o formato CoNLL-U das Dependências Universais (De Marneffe, 2021) para a descrição de línguas minoritárias, por meio da construção de *treebanks*. A primeira etapa deste estudo refere-se à discussão de referências bibliográficas de obras que fundamentaram esta pesquisa, como: Aragon (2014), Cunha (1992), Mezacasa (2021), De Alencar (2023) Rodrigues (2015), Aragon e Algayer (2020); Himmelman (1998) e De Marneffe *et al*. (2014). A segunda etapa consistiu no trabalho com as anotações de dados da língua Akuntsú realizadas ao longo da vivência no projeto de PIBIC "Educação, Linguística, História e Comunidades Indígenas" (Edital 2021-2021) executada na Universidade Federal da Paraíba (UFPB). Os resultados mostram a descrição linguística vinculada às anotações morfossintáticas como um dos meios de preservação histórica, de documentar e de compreender histórias orais de um povo. Esperamos, desta forma, abrir caminhos para futuras pesquisas nesta temática, à medida que se demonstra importante para o diálogo interdisciplinar e para possibilidades de estudos de línguas voltados aos povos minoritários.

PALAVRAS-CHAVE: Descrição Linguística; Dependências Universais; *Treebanks*; Línguas Indígenas; Akuntsú.

ABSTRACT

The present study delineates the corpus annotations of the Akuntsú language, a member of the Tupí, Tuparí linguistic family, and the application of the Universal Dependencies model in describing indigenous languages. In this context, it establishes interdisciplinary connections between linguistic studies and the field of Ethno-history. Indigenous languages are characterized by a historical erasure, as noted by Cavalcante in 2011, resulting in the loss of linguistic diversity, cultural identity, and significant values within indigenous communities. The rationale for this work lies in the importance of leveraging technological tools such as Annotatrix and the CoNLL-U format of Universal Dependencies (De Marneffe, 2021) for the documentation of minority languages through the creation of treebanks.The initial stage of this study involves a discussion of relevant bibliographic references that underpin this research, including works by Aragon (2014), Cunha (1992), Mezacasa (2021), De Alencar (2023), Rodrigues (2015), Aragon and Algayer (2020), Himmelman (1998), and De Marneffe et al. (2014). The subsequent stage entails the annotation of data in the Akuntsú language, conducted during the course of the PIBIC project "Educação, Linguística, História e Comunidades Indígenas" (Notice 2021-2021) at the Universidade Federal da Paraíba (UFPB). The findings demonstrate that linguistic description, coupled with morphosyntactic annotations, serves as a means of historical preservation, documenting, and elucidating the oral histories of the Akuntsú people. In this manner, we aspire to pave the way for future research in this field, recognizing its importance in fostering interdisciplinary dialogue and the exploration of language studies concentrated on minority populations.

KEY-WORDS: Linguistic Description; Universal Dependencies; Treebanks; Indigenous Languages; Akuntsú.

LISTA DE ILUSTRAÇÕES

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

AM	Armazenamento de máquina
DEPREL	*Universal Dependency relation*
DU	Dependências Universais
FEATS	*List of morphological features from the universal feature inventory*
PIBIC	Programa Institucional de Bolsas de Iniciação Científica
PLN	Processamento de Línguas Naturais
TI	Terra Indígena
TuDeT	*Tupían Dependency Treebank*
TuLaR	*Tupían Language Resources*
TuLeD	*Tupian Lexical Database*
TuMoD	*Tupian Morphological Database*
TuPAn	*Tupian Plants and Animals*
UPOS	*Universal part-of-speech tag.*
XPOS	*Optional language-specific part-of-speech*

SUMÁRIO

1 INTRODUÇÃO

Embora existam divergências quanto ao número de línguas indígenas faladas no Brasil, estima-se que esse número não ultrapasse 200 línguas. Dentre elas, cerca de 21% estão ameaçadas de extinção (Moore e Galucio 2016), como é o caso da língua Akuntsú, pertencente ao tronco linguístico Tupí, família Tuparí, falada por apenas três mulheres monolíngues (total da população) localizadas no estado de Rondônia, na TI Rio Omerê.

Considerando o uso de ferramentas linguísticas para o contexto de línguas minoritárias, em especial para a língua Akuntsú, este trabalho apresenta o uso do formato CoNLL-U como ferramenta de descrição de línguas indígenas brasileiras. Além disso, demonstra como a anotação morfossintática, utilizando as Dependências Universais (DU), programa de anotações morfossintáticas de línguas naturais baseada em um sistema arbóreo (*treebanks*), pode ser um veículo não apenas de descrição linguística, mas também de preservação de histórias e da cultura de um povo.

Mais especificamente, objetivamos: a) apresentar o modelo da DU e algumas ferramentas linguísticas; b) demonstrar a construção de *treebanks* da língua Akuntsú utilizando dados de Aragon (2014) e dados de campo da autora; c) apresentar o caráter interdisciplinar deste estudo: linguística e etno-história. Para isso, partimos do pressuposto de que uma ferramenta computacional é uma inovação importante para a descrição de línguas minoritárias, assim como uma forma de disponibilizar o acesso de dados para diferentes públicos, viabilizando um sistema lógico de organização desses dados. Ressaltamos, porém, que este trabalho não foca em apresentar análises morfossintáticas da língua Akuntsú e nem em propor discussões aprofundadas relacionadas à Linguística Computacional; mas, sim, demonstrar como a morfossintaxe pode ser trabalhada por meio do uso de ferramentas linguísticas (possíveis metodologias e usos).

O interesse na temática deste trabalho aconteceu após o contato inicial com as línguas indígenas, Akuntsú e Makurap (Tupí, Tuparí), durante as experiências vivenciadas no Programa Institucional de Bolsas de Iniciação Científica (PIBIC), projeto intitulado 'Educação, Linguística, História e Comunidades Indígenas', coordenado pela professora Carolina Aragon. Entre as etapas desenvolvidas, trabalhamos com ações voltadas à revitalização da língua Makurap e à descrição das línguas Makurap e Akuntsú[1]. Durante o projeto, elaboramos materiais didáticos em conjunto com professores indígenas e expandimos as anotações de dados para a construção dos *treebanks* dessas duas línguas. Mesmo após a conclusão do projeto, vigência 2021-2022, os trabalhos prosseguiram com a discussão de textos e com as anotações morfossintáticas. Destacamos que, embora o projeto envolvesse o trabalho com essas duas línguas, optamos por descrever neste trabalho apenas o uso de ferramentas linguísticas voltadas à língua Akuntsú, possibilitando, assim, uma descrição mais detalhada dos objetivos aqui propostos.

As línguas indígenas possuem um histórico de apagamento político e social (Cavalcante 2011). A 'morte' dessas línguas é uma das muitas consequências que marcam a história dos povos indígenas no Brasil. Portanto, este trabalho justifica-se na importância da descrição linguística agregada ao uso de ferramentas tecnológicas com o intuito de ampliar o conhecimento linguístico e etno-histórico dessas línguas.

Para isso, dividimos este trabalho em sete capítulos. O capítulo 1, trata-se desta introdução. No capítulo 2, apresentamos como as disciplinas de História e de Linguística se relacionam no trabalho de resgate linguístico e cultural, ressaltando a importância do estudo interdisciplinar voltado às línguas indígenas. No capítulo 3, demonstramos aspectos relevantes da história e da língua do povo Akuntsú. No capítulo 4, conceitualizamos a documentação e a descrição linguística; no capítulo 5, apontamos descrições teóricas sobre a DU e os *treebanks*; no capítulo 6, explicamos

[1] O trabalho de revitalização linguística não foi realizado para a língua Akuntsú, visto que todas as falantes falam fluentemente a língua (monolíngues).

a metodologia adotada para a construção deste trabalho; e, por fim, no capítulo 7, analisamos alguns dados da língua Akuntsú dentro do formato CoNLL-U e relacionamos essa análise com aspectos etno-históricos.

2 LINGUÍSTICA E ETNO-HISTÓRIA

Por muito tempo houve o mito de que povos indígenas não possuíam história, isso porque, como afirma Cavalcante (2011), os estudos históricos eram centrados apenas nas culturas europeias que possuíam registros escritos[2]. A falsa ideia de que as sociedades indígenas não possuíam história corroborou para o silenciamento das culturas complexas e do dinamismo social que essas comunidades possuíam, reforçando a ideia errônea de que os povos indígenas são atrasados tecnologicamente em relação aos outros povos, ou seja, são povos parados no tempo, embora vivam na contemporaneidade (Fausto e Heckenberger, 2007).

Ainda nos dias atuais é comum que a história indígena seja tratada tendo como ponto de partida a história colonial. Nesse sentido, Cunha (1992) relembra a tendência de excluir o indígena de seu local de protagonismo, de sua própria história, colocando-o, muitas vezes, como vítima do acaso, sem vontades, verdades ou crenças. É importante ressaltar que inserir os povos indígenas como protagonistas de suas histórias não significa culpabilizá-los pela invasão das Américas, tampouco pelo genocídio que sofreram (e ainda sofrem), mas enxergá-los como sujeitos que (re)escrevem as suas histórias de acordo com as suas cosmologias, como pontua a referida autora:

> A percepção de uma política e de uma consequência histórica em que os índios são sujeitos e não apenas vítimas, só é nova eventualmente para nós. Para os índios, ela parece ser costumeira. É significativo que dois eventos fundamentais – a gênese do homem branco e a iniciativa do contato – sejam frequentemente apreendidos nas sociedades indígenas como produto de sua própria ação ou vontade. (Cunha, 1992, p. 18)

[2] Vale ressaltar, de acordo com Cavalcante (2011), que algumas culturas não europeias possuíam seu próprio sistema de escrita, os quais foram suprimidos em detrimento dos sistemas de escrita europeu.

Em 1992, a supracitada autora já chama atenção para o fato de que os nossos livros de história se iniciam em 1500 e, embora atualmente haja avanços em relação aos estudos de culturas originárias, ainda são necessários trabalhos que busquem refletir sobre os povos indígenas e suas histórias. Nesse sentido, a Linguística, em seu viés interdisciplinar, se apresenta como uma importante aliada nessa reflexão.

A relação entre Linguística e Etno-História fica clara, por exemplo, se for levado em consideração o fato de itens lexicais serem permeados de história e de valores culturais. Entendendo a 'documentação linguística' como o registro de práticas linguísticas envolvendo, dentre outros aspectos, tradições orais de um povo, e a 'descrição linguística' como o registro de uma língua explorando seus aspectos gramaticais, como fonética-fonologia, morfologia e sintaxe (bem como os aspectos semânticos-pragmáticos) (Padovani; Miranda; Bastos, 2019), é possível atrelar essas práticas ao modo de compreender a história e a cultura de um povo (falaremos mais sobre documentação e descrição de línguas nos próximos capítulos).

A exemplo do que explicamos, Mezacasa (2021) traz o histórico do termo 'marico', bolsas feitas da folha de tucum (*Astrocaryum vulgare*) — um elemento cultural de povos indígenas localizados no lado direito do rio Guaporé (RO) (Maldi, 1991), incluindo o povo Akuntsú. De acordo com a autora, a palavra 'marico' é uma apropriação indígena de um termo usado pelos seringueiros para designar uma bolsa utilizada para levar os seus utensílios[3]. A partir desse e de outros olhares, a autora remontou a história de desterritorialização do povo Makurap e seus caminhos até o território atual (Terra Indígena (TI) Rio Branco e TI Guaporé, estado de Rondônia). Além disso, a palavra 'marico' é usada por diferentes povos indígenas da região falantes de línguas afiliadas geneticamente ou não, um elemento que indica, portanto,

[3] De acordo com Mezacasa (2021), o 'Seringal' na região do Guaporé possuía uma economia muito mais complexa do que a exploração da borracha, envolvendo outras formas de produção e extração, como, por exemplo, a coleta de Castanha do Brasil (*Bertholletia excelsa*). Essas relações econômicas culminaram no contato forçado com os indígenas da região por meio da escravização e violência.

possíveis empréstimos linguísticos frutos de contatos/relações/encontros de etnias distintas ao longo da história (cf. Crevels; van der Voort, 2008 e Algayer; Aragon; Mezacasa 2022).

No que diz respeito a esses aspectos, a Linguística Computacional, área que "lida com o processamento automático de uma língua" (Freitas, 2022, p. 12), demonstra-se importante para o processo de descrição de línguas, incluindo as minoritárias[4]. Isso porque, em seu lado aplicado, essa área proporciona ferramentas para anotação de *corpus*, como, por exemplo, anotadores morfossintáticos. Eles servem não apenas como meio para armazenamento de dados linguísticos, mas também para conferir aplicação prática a essas anotações, como, por exemplo, a criação de ferramentas automáticas como preditor de palavras, corretor ortográfico e, inclusive, aplicações na Inteligência Artificial (IA).

De acordo com De Alencar (2023, p. 75) "a diferença [da Linguística Computacional] em relação às abordagens linguísticas não computacionais é a completa formalização dos modelos, permitindo uma computação mecânica das representações e regras postuladas". Ou seja, esta linha metodológica visa prover modelos lógicos e estruturais, em níveis de Aprendizado de Máquina (AM), para desenvolver a descrição de elementos linguísticos de uma determinada língua.

Duran *et al*. (2022) afirma que o desenvolvimento tecnológico, desde os anos 1990, revolucionou a Linguística Computacional, exigindo tarefas de processamento de textos. Como o AM demanda que as máquinas aprendam tarefas baseadas em comportamento humano, é necessário ter disponível um grande número de exemplos, *corpora* textuais, para que os algoritmos possam aprender. Desta forma, a anotação de *corpus*, que é uma ciência, passa a exigir diretrizes que "têm por objetivo tornar claro para os anotadores como o conjunto de etiquetas deve ser utilizado, com rica

[4] De acordo com Griep (2021), línguas minoritárias são aquelas pertencentes a grupos desprestigiados social, cultural ou politicamente.

exemplificação que contemple desde casos comuns e frequentes até casos mais raros e difíceis de anotar" (Duran *et al.*, 2022, p. 1614)

Nesse viés, surge o modelo de Dependências Universais (DU) como um dos resultados de avanço do Processamento de Linguagem Natural (PLN). A DU é um esquema de anotação de dependências morfossintáticas, a qual fornece diretrizes comuns a todas as línguas, portanto universais. Essas diretrizes fornecem ferramentas para auxiliar no tratamento e na consulta de dados linguísticos, facilitando a disponibilidade desses dados e construindo possibilidades de aplicabilidade, como já mencionado anteriormente.

Portanto, ressaltamos neste trabalho a relevância em realizar anotações de *corpus* na língua Akuntsú, descrever aspectos morfossintáticos da língua, em conjunto com a tradição oral do povo, remontando sua história a partir do trabalho linguístico, fortalecendo a cultura e a história indígena — a etno-história desses povos.

3 TRÊS MULHERES AKUNTSÚ

É relevante partir do princípio de que a linguagem é a capacidade biológica que todos os seres humanos têm de utilizar um sistema de expressão e comunicação complexo por meio de sons, é um ato de manifestação sonora intencional e representativa, como afirma Mattoso Câmara Jr. (1974). Enquanto a língua, é a manifestação desse traço biológico, visto que é aprendida por meio da interação entre indivíduos e, portanto, é "algo adquirido e convencional" (Saussure, 2012, p. 41).

Todas as línguas estão em constante, lenta e gradual mudança, decorrente de diversos fatores externos e internos (e.g. Labov 2008). Segundo Aryon Rodrigues (2015, p. 3), embora haja mudanças, o fato da língua ser uma experiência compartilhada entre os membros de uma mesma comunidade, dentro de uma mesma cultura, faz com que essa mesma língua permaneça coesa, visto que as inovações são levadas e adaptadas por cada membro da comunidade, causando uma "unidade de uma língua". Essa unidade rompe-se no momento em que as inovações não passam de uma comunidade para outra comunidade e, aos poucos, o que era tão semelhante, se torna ininteligível. Após alguns séculos de alterações, uma língua passa a ser duas ou mais. Sobre a classificação de línguas, ressaltamos:

> Línguas que têm tanto em comum, que são mutuamente inteligíveis formam um grupo dialetal ou o que comumente se chama simplesmente de língua [...]. Já as línguas entre as quais não há mais inteligibilidade mútua, mas cuja separação e diferenciação corresponde aproximadamente à de línguas como o português, o italiano, o francês constituem uma família linguística, neste caso a família românica. Línguas ainda mais remotamente aparentadas, como o português, o alemão, o russo e o híndi formam um tronco linguístico, neste caso o tronco indo-europeu. (Rodrigues, 2015, p. 3)

De acordo com o trabalho desenvolvido por este mesmo autor (Rodrigues, 1986), estima-se que no Brasil as línguas indígenas estão divididas em dois troncos:

Tupí e o tronco Macro-Jê. Além de famílias menores, como a Família Tukano; línguas isoladas, que não pertencem a nenhum agrupamento linguístico, como a língua Kanoé; pidgins e línguas crioulas (Campbell; Grondona, 2012). A língua Akuntsú, é uma língua minoritária, que pertence à família linguística Tuparí, tronco Tupí.

Antes de adentrar na discussão sobre ferramentas de descrição linguística, especialmente as anotações e o modelo da DU, apresentamos neste capítulo alguns aspectos relevantes sobre a língua e a história do povo Akuntsú.

3.1 Língua

A língua Akuntsú é falada atualmente por apenas três mulheres, a totalidade de um povo, localizadas atualmente na TI Rio Omerê, estado de Rondônia. Está severamente ameaçada de extinção em decorrência das perdas populacionais e da decisão da única mulher do grupo em condições físicas de engravidar de não ter filhos. Como afirma Aragon (2014):

> Os falantes não conseguem transmitir sua língua para outra geração, principalmente devido a tabus de parentesco e à recusa em permitir que homens de outros grupos se casem com a mais jovem falante. A sobrevivente mais jovem é uma mulher de [aproximadamente] 30 anos, com outros quatro Akuntsú com mais de 40 anos, sem filhos e sem perspectivas de aumentar seu grupo. Em resumo, assumindo que essas circunstâncias não mudarão no futuro, aceitamos que essa língua está fadada a desaparecer. (Aragon, 2014, p. 2, tradução nossa).[5]

[5] "The speakers cannot pass their language on to another generation, mainly due to kinship taboos and their refusal to allow men from other groups to marry the youngest speaker. The youngest survivor is a woman in her 30s, with four Other Akuntsú over forty, with no children and no prospects for increasing their group. In short, assuming these circumstances will not change in the future, we accept that this language is doomed to disappear."

Quanto aos aspectos gramaticais e tipológicos da língua Akuntsú, destacamos alguns pontos (Aragon 2014):

- A língua é predominantemente aglutinativa, com alguns graus de síntese;
- Há mais sufixos que prefixos;
- As classes abertas de palavras expressam objetos, ações e atributos. São elas: substantivos, verbos, adjetivos e advérbios;
- Já as classes fechadas expressam emoções do falante, distância, posição, aspecto, negação e foco. São elas: posposições, quantificadores/numerais, demonstrativos/dêiticos, partículas e interjeições;
- Os verbos podem ser sintaticamente transitivos ou intransitivos;
- Não há cópula;
- Há subclasses de verbos auxiliares;
- Apresenta morfemas direcionais;
- A negação pode ser expressa por partículas ou sufixos marcados nos nomes ou nos verbos;
- Objetos tendem a preceder o verbo;
- A ordem Sujeito(S)-Objeto(O)-Verbo(V) é a mais frequente, embora relações pragmáticas desencadeiam ordens distintas;
- A combinação de orações pode ocorrer por coordenação ou por subordinação.

3.2 História do povo

Os impactos culturais e físicos que culminaram no estado atual dessa língua estão relacionados ao início da extração da borracha na segunda metade do Século XVIII até a primeira metade do século XX, quando seringueiros começaram a chegar na região do rio Guaporé. Um exemplo disso é que das vinte pessoas mais citadas nas histórias contadas pelas Akuntsú, mais de 50% foram mortas por arma de fogo

(Aragon; Algayer, 2020). Nesse sentido, convém ressaltar alguns aspectos da história desse povo.

Segundo relatos dos Akuntsú documentados por Aragon e Algayer (2020), esse povo viveu durante muito tempo nas proximidades do Rio *Ykytarẽj* e *Ykytxaro* (ver mapa abaixo). Primeiramente, eles foram diretamente afetados pela ocupação dos seringueiros e caucheiros na região onde viviam e, posteriormente, pela liberação de lotes para projetos de "colonização" da região, resultando na derrubada da floresta para extração de madeira e abertura de fazendas.

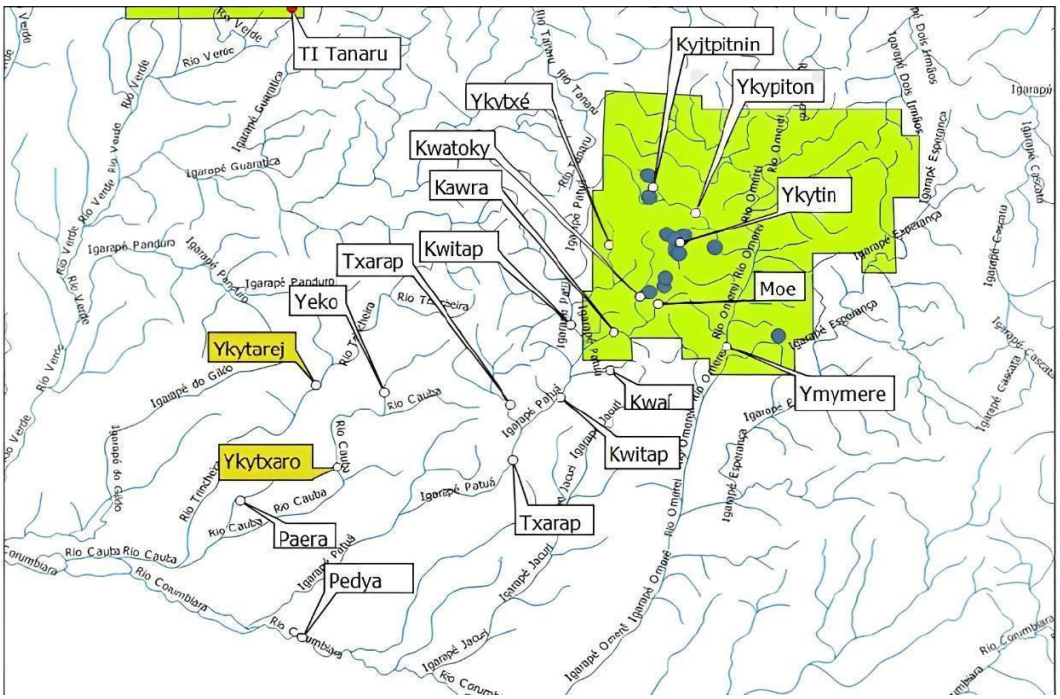

Figura 1: Mapa hidrográfico do território tradicional dos Akuntsú. Nos rios destacados em Amarelo (Ykytarej e Ykytxaro), é possível ver a localização dos rios citados como território Akuntsú; e em verde, região do mapa marcada por um quadrado assimétrico, a TI Rio Omerê; os círculos dentro desse quadrado, em azul, representam os rios que estão dentro dessa TI. Fonte: Algayer e Aragon (2020).

Na década de 1980, essa região foi alvo de projetos de "colonização" orientadas pelo Governo Federal, o que intensificou a ocupação desses territórios por não indígenas, resultando no processo de genocídio de populações indígenas e,

consequentemente, no aumento massivo da destruição ambiental, como afirmam Aragon, Algayer e Mezacasa (2022). Esse projeto consistia, resumidamente, em distribuição de lotes para a ocupação da região por não indígenas, ou seja, na transformação da floresta para construção de áreas agrícolas. Um projeto tão explícito de desmatamento causou debates, cujas pressões resultaram na criação de diretrizes específicas para os povos indígenas e na demarcação de algumas TIs, o que não diminuiu a invasão de madeireiros, tampouco, as doenças virais que atacavam os povos que não tinham imunidade.

O contato oficial da Funai com os Akuntsú ocorreu no ano de 1995. Porém, já em 1984, seus vestígios foram identificados pela Funai, após um dos massacres que ocorreram na região da atual fazenda Yvypitã. Os indígenas que habitavam o local, hoje sabemos que eram os Akuntsú, fugiram e não foi possível continuar com o processo de interdição da área. Como os vestígios da presença indígena tinham desaparecido após esse episódio, não houve meios legais de isolar a área e, portanto, a exploração do local continuou. Porém, mesmo com as evidências desse massacre, indigenistas, com a certeza da existência dos sobreviventes, retornaram para região até conseguirem fazer o contato oficial com os Akuntsú e com os Kanoé do Omerê e, assim, assegurar os direitos e a sobrevivência desses povos (Aragon, 2008, 2014; Carelli, 2009; Tavares, 2020). Na época, os Kanoé eram um grupo de quatro pessoas e os Akuntsú de sete integrantes.

Hoje o povo Akuntsú, indígenas de recente contato[6], são constituídos por apenas três mulheres: Pugapia, Aiga e Babawro; já os Kanoé, são atualmente três adultos: uma mulher, Txiramanty; dois homens, Purá e Bukwá; e o mais novo integrante, o recém-nascido, Tony Kwikay, filho de Bukwá com Celita Aruá.

[6] A Funai (Brasil, 2021) considera povos de recente contato os grupos indígenas que mantêm contato permanente ou não com segmentos da sociedade não indígena, desempenhando, em sua relação, autonomia na incorporação de bens e serviços. Diferentemente dos povos isolados, que não mantêm ou mantêm muito pouca relação com a sociedade não indígena ou com indígenas de outras etnias.

4 DESCRIÇÃO E DOCUMENTAÇÃO LINGUÍSTICA

Como afirmamos na introdução deste trabalho, estima-se que o número de línguas indígenas faladas no Brasil não ultrapasse 200 línguas. Embora haja um número significativo de línguas indígenas, de acordo com Moore e Galucio (2016), 21% dessas línguas faladas no Brasil correm sério risco de extinção, como é o caso do Akuntsú. É importante considerar que, embora os dados do IBGE (2010)[7] indiquem que haja, aproximadamente 305 etnias indígenas no Brasil[8], esse número não deve ser confundido com o número de falantes de uma língua, nem com o número total de línguas.

Nesse aspecto, Padovani, Miranda e Barros (2019) mencionam o fato de que essa confusão fez com que fosse subestimado o alto grau de extinção das línguas indígenas brasileiras. Um dos fatores responsáveis pela vulnerabilidade de línguas são: 1) a perda populacional causada pela exploração ilegal da Floresta Amazônica, como já mencionado; e 2) as pressões que essas línguas sofrem frente a uma língua majoritária, no nosso caso o Português Brasileiro. Aqui vale mencionar que não é apenas a língua que sofre essas pressões, mas também os aspectos socioculturais de um povo. Diante desses fatos, os trabalhos de descrição, documentação e revitalização linguística surgem, não apenas para colaborar com o enriquecimento das discussões na área, mas também para preservação dos aspectos socioculturais e estruturais das línguas ameaçadas de extinção, como afirmam Toledo e Miranda (2021).

É possível entender a documentação e a descrição como tarefas relacionadas, porém essas práticas possuem objetivos diferentes. De acordo com Himmelman

[7] Disponível em: https://indigenas.ibge.gov.br/images/indigenas/estudos/indigena_censo2010.pdf Acesso em: 10 out. 2023.

[8] Embora os primeiros resultados do censo 2022 já estejam disponíveis no site do IBGE e mencionem um aumento no número de indígenas residentes no Brasil — de 896.917 para 1.693.535 —, até a data de escrita deste trabalho, não houve atualizações no número de etnias e de línguas indígenas.

(1998), a documentação é uma prática que consiste na captura de aspectos linguísticos manifestados por meio da oralidade expressa no dia a dia dos falantes de uma comunidade e do conhecimento metalinguístico dos falantes nativos. Ainda para esse autor, essa prática tem como objetivo fornecer um registro compreensível das características linguísticas de uma comunidade de fala. Woodbury (2011) complementa e inclui a divulgação e a preservação de registros como objetivos da documentação.

A documentação é uma prática multidisciplinar, isso porque abrange diversas áreas. A tecnologia é uma grande aliada da documentação, visto que por meio dela é possível o registro de alta qualidade nos trabalhos de campo, além do uso de programas de computadores que possibilitam a criação de acervos digitais. O armazenamento e disponibilidade dos dados é uma parte importante, desse processo, sobre isso Toledo e Miranda (2021), destacam que:

> O resultado da documentação linguística é um registro acessível e de interesse de várias pessoas, sejam linguistas, antropólogos, historiadores, pesquisadores envolvidos na educação e no planejamento de revitalização das línguas e culturas e, claro, inclui-se, essencialmente, os membros da comunidade linguística e seus descendentes. (Toledo; Miranda, 2021, p. 9)

Outro aspecto relevante da documentação, é que, em regra, o interesse deve vir dos membros da comunidade, parte importante nesse processo. É por meio deles que surge o interesse na documentação e na revitalização de suas línguas. Sem o interesse, envolvimento e a disposição dos falantes da língua, seria impossível coletar os materiais de pesquisa, daí justifica-se a importância do envolvimento da comunidade de fala com os pesquisadores e não apenas como objetos de pesquisa.

Sobre aspectos mais técnicos, é importante destacar que os dados coletados em um processo de documentação podem ser primários ou secundários. Os dados primários são aqueles que irão compor o *corpus*: gravações, vídeos, transcrições. Já

os secundários são aqueles que revelam informações sobre a coleta de dados, como: os participantes da gravação, o local e data do registro, os equipamentos utilizados, o público alvo e uma breve descrição do conteúdo que está sendo registrado.

Dentro desse contexto, convém ressaltar as cinco características da documentação expostas por Himmelmann (2006): 1) Sua maior preocupação deve ser a coleta e a análise de dados que devem ser disponibilizados; 2) Os dados primários devem conter evoluções nas análises linguísticas; 3) É necessário haver uma preocupação com o armazenamento desses dados primários a longo prazo; 4) A documentação requer conhecimentos e informações de áreas que vão além da linguística; 5) É importante haver cooperação e envolvimento direto da comunidade de fala, como produtores e pesquisadores.

A descrição linguística, por sua vez, também utiliza do método de captura de um *corpus*, porém, como já falado anteriormente, ela difere da documentação, pois abrange aspectos mais estruturais da língua. Himmelmann (1998) elucida que a descrição, em regra, tem como público-alvo pesquisadores que se dedicam a trabalhar com gramáticas e comparações de línguas. Como bem destaca Padovani, Miranda e Barros (2019, p. 912), "a descrição abrange uma compreensão da língua em níveis mais abstratos, como um sistema de elementos, regras e construções fonológicas, morfossintáticas e semânticas", ou seja, de modo geral, a descrição se dedica a apresentar uma gramática descritiva da língua e/ou um dicionário.

Em síntese, a descrição é voltada para a construção de dicionários e gramáticas descritivas e tem como procedimento a análise fonética, fonológica, morfossintática e semântica de uma língua. Objetiva, portanto, fazer "o registro de uma língua, sendo 'língua' um sistema de objeto de elementos abstratos, construções e regras que constituem a estrutura subjacente invariante dos enunciados observáveis de uma comunidade de fala" (Toledo; Miranda, 2021, p. 12).

Assim como o trabalho de documentação, a descrição é um trabalho interdisciplinar que conta, dentre outras áreas, com a linguística computacional, como o uso do formato CoNLL-U como ferramenta de descrição, como veremos a seguir.

5 ANOTAÇÃO DE *CORPUS* E O MODELO DAS DEPENDÊNCIAS UNIVERSAIS

Como visto, o processo de descrição e documentação linguística requer ferramentas linguísticas. Com o avanço tecnológico, tendências da linguística computacional surgem para definir padrões e especificações que auxiliem na acessibilidade e reprodução dos dados, principalmente os de larga escala, *big data*. Uma dessas tendências é a DU que, por sua vez, "visa oferecer uma representação linguística que seja útil para a pesquisa morfossintática, a interpretação semântica e o processamento prático da linguagem natural em diferentes idiomas" (De Marneffe *et al*., 2021, p 256).

Segundo o site oficial do programa[9], o intuito da DU é "[...] fornecer um inventário universal de categorias e diretrizes que contribuam com a construção de anotações de maneira similar, independente das línguas, permitindo, ao mesmo tempo, extensões próprias de uma língua específica, quando necessário" (tradução nossa)[10]. O modelo DU tem sido usado para anotar *corpus* de distintas línguas, não apenas as línguas minoritárias, inclusive, há um *treebank* para a Língua Portuguesa (Rademaker *et al*., 2017). Quanto às indígenas, além do Akuntsú, há *treebanks* para as línguas: Guajajara, Ka'apor, Karo, Makurap, Munduruku, Guarani Antigo, Teko e Tupinambá — todas incluídas no projeto *Tupían Language Resources* (TuLaR) (descreveremos esse projeto no capítulo de metodologia)[11]. Além disso, outros *treebanks* estão sendo desenvolvidos por diferentes pesquisadores focados em outras

[9] Disponível em: https://universaldependencies.org/introduction.html. Acesso em: 10 mai. 2023

[10] "The general philosophy is to provide a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages, while allowing language-specific extensions when necessary."

[11] Disponível em: https://universaldependencies.org/#current-ud-languages. Acesso em: 10 mai. 2023

línguas indígenas brasileiras: Nheengatu (Tupí Moderno) (De Alencar, 2023); Mbyá Guaraní (Thomas, 2019); Apurinã; e Xavante.

Para formular os *treebanks*, *corpora* morfossintaticamente anotados, é necessário utilizar um programa de anotação de textos, uma linguagem específica de programação, denominada formato CoNLL-U. Esse formato consiste em linhas verticais em que são feitas a descrição de uma palavra. Cada coluna possui o seguinte significado: **Coluna 1**. Índice de palavras (*ID*), contando o número de palavras a partir do 1, aqui, é relevante ressaltar que sinais e pontuação também contam; **Coluna 2**. A palavra em si ou o símbolo da pontuação (*Form*); **Coluna 3**. Lema ou radical da palavra (*Lemma*); **Coluna 4**. Classe de palavras pré-definidas pelo programa de acordo com a *universal part-of-speech tag* (*UPOS*)[12]; **Coluna 5**. A classe de palavras segundo o idioma descrito (*XPOS)*; **Coluna 6**. Lista de recursos morfológicos presentes nas palavras de acordo com a *universal feature inventory* (*FEATS*)[13]; **Coluna 7**. Cabeçalho da palavra (*HEAD*); **Coluna 8**. A relação sintática que as palavras mantêm entre si (*DEPREL*)[14], essa linha é a que denomina a relação das palavras no sistema arbóreo, para visualizá-la e estabelecer essas relações, utilizamos o programa *Annotatrix*, que será abordado posteriormente; **Coluna 9**. Gráfico de dependência aprimorado (*DEPS*); **Coluna 10**. Qualquer outra anotação (*MISC*).

Com base na ideia de que a língua possui uma relação hierárquica (De Marneffe *et al*., 2021) depreende-se uma estrutura baseada em um núcleo e seus dependentes. Desta forma, as anotações consistem em encontrar as dependências nas frases. As dependências morfossintáticas são anotadas no *Annotatrix*[15] — ferramenta de anotação *on-line-off-line* voltada para construção do *treebank*. Essa ferramenta

[12] Disponível em: https://universaldependencies.org/u/pos/index.html. Acesso em: 03 abr. 2023
[13] Disponível em: https://universaldependencies.org/u/feat/index.html. Acesso em: 03 abr. 2023
[14] Disponível em: https://universaldependencies.org/u/dep/index.html. Acesso em: 03 abr. 2023
[15] Disponível em: https://github.com/jonorthwash/ud-annotatrix. Acesso em: 03 abr. 2023

pode ser acessada por meio de um navegador da internet e possui vários recursos, incluindo a alteração e inclusão das relações sintáticas das palavras.

Para a anotação no *Annotatrix*, são utilizadas as *tags* de relações da DU, originalmente descritas no trabalho *Universal Stanford Dependencies: A cross-linguistic typology* (De Marneffe *et al.* 2014) e disponibilizadas, também, no site oficial do programa. Embora as descrições exijam um nível aprofundado de conhecimento sintático e morfológico na língua descrita, os sistemas de *tags* apresentam um modelo lógico a ser seguido, o que facilita a leitura e a descrição.

As análises no formato CoNLL-U ocorrem por meio de uma série de tarefas de programação em linguagem *Python*, que vão desde a separação de pontos e vírgulas até a classificação de palavras em análise morfológica e sintática. Porém, não descreveremos esse processo neste trabalho, visto que nosso enfoque é demonstrar as anotações em seus aspectos estruturais. A descrição detalhada dessa tarefa demandaria uma discussão aprofundada sobre a linguagem *Python*, a qual não teríamos espaço viável, além de não ser o foco deste estudo.

6 METODOLOGIA

Este trabalho foi organizado em duas etapas. A primeira etapa foi o levantamento de referências bibliográficas relacionadas aos temas centrais deste estudo. A segunda etapa consistiu nas anotações do *corpus* da língua Akuntsú, já construídas ao longo da vivência no projeto de PIBIC "Educação, Linguística, História e Comunidades Indígenas" (Edital 2021-2021), executada na Universidade Federal da Paraíba (UFPB).

O *corpus* linguístico deste trabalho foi retirado de Aragon (2014) e de dados inéditos de trabalho de campo da autora na TI Rio Omerê. Quanto ao número de anotações, contabilizamos um total de 368 frases até a publicação da versão 2.11 lançada pela DU em novembro de 2022[16] — essas anotações são descritas no capítulo 7. As anotações são de minha autoria, de minha orientadora e de Fabrício Gerardi, construídas em reuniões regulares realizadas na UFPB no formato *on-line* e presencial.

Após converter as frases com um *script* em linguagem de programação *Python* para o formato CoNLL-U, cada uma delas foi transcrita no *Sublime Text*, editor de texto com visualizações para extensões específicas de arquivo. E, em seguida, inseridas manualmente no *Annotatrix* para que possamos fazer as anotações. Durante esse processo, corrigimos e construímos as relações sintáticas das frases.

De acordo com Santos, Aragon e Gerardi (no prelo)[17], esse trabalho manual é necessário neste momento em que estamos iniciando os *treebanks* do Akuntsú. Porém, quando houver um número alto de frases anotadas, será possível usar ferramentas de anotações automáticas. Tais ferramentas automáticas são programas

[16] Disponível em: https://github.com/UniversalDependencies/UD_Akuntsu-TuDeT/blob/dev/aqz_tudet-ud-test.conllu. Acesso em: 15 out. 2022

[17] SANTOS, L. L. S.; ARAGON, C.; GERARDI, F. Línguas Minoritárias e Anotações Sintáticas de Corpora: experiências de pesquisa na iniciação científica. Revista Letras de Hoje. No prelo.

computacionais cujos algoritmos aprendem a língua (vocabulário, morfologia e dependências) a partir de frases já anotadas e aplicam o aprendizado a novas frases, ainda não anotadas (Rodríguez *et al*. 2022). Os dados anotados são, na sequência, inseridos no repositório DU, presente na plataforma *GitHub*, e publicados no site DU[18]. Nesta plataforma, diferentes desenvolvedores podem adicionar, compartilhar e alterar dados de seus projetos. Todos os dados são de acesso livre.

É importante que alguns conceitos abordados neste trabalho, comuns à área de programação e às Dependências Universais sejam esclarecidos de maneira mais detalhada. Os dados são anotados em formato CoNLL-U, que, como já abordamos, se refere a uma linguagem de programação própria da DU para anotação de dados linguísticos. As dependências são anotadas no *Annotatrix*, uma ferramenta acessada por meio de um navegador *web* para estabelecer as relações de dependências entre as palavras. Os dados anotados são disponibilizados na plataforma *GitHub* — plataforma que, dentre outras utilidades, acomoda projetos de programação.

As atividades aqui descritas fazem parte de um subprojeto, intitulado *Tupían Dependency Treebank*[19] (TuDeT), voltado para a construção de banco de dados de línguas do tronco linguístico Tupí com o objetivo de expandir a descrição dessas línguas (Rodriguez *et al*., 2022). Esse subprojeto, por sua vez, faz parte do *Tupían Language Resources* (TuLaR)[20], um projeto em andamento que está construindo e disponibilizando dados referentes às línguas do tronco linguístico Tupí. Já estão disponíveis alguns *treebanks* (Gerardi, *et al*., 2022a) e um banco de dados lexicais (Gerardi, *et al.*, 2022b). Atualmente, além do TuDeT, outros bancos de dados também compõem o TuLaR, os quais estão em diferentes níveis de desenvolvimentos: TuLeD

[18] Podendo, assim, gerar um *Digital Object Identifier* (DOI).
[19] TuDeT: Tupían Dependency Treebank. Disponível em: https://zenodo.org/record/5655343#.Y0xMHXbMJPZ. Acesso em: 16 out. 2022
[20] TuLaR. Tuían Language Resources. Disponível em: https://tular.clld.org/. Acessso em: 16 out. 2022.

(*Tupian Lexical Database*), TuMoD (*Tupian Morphological Database*), e TuPAn (*Tupian Plants and Animals*).

Antes de seguirmos, retomamos os objetivos deste TCC:

a) Apresentar a DU como um modelo para a descrição de línguas;
b) Descrever os aspectos da anotação de *corpus* da língua Akuntsú utilizando ferramentas linguísticas atuais;
c) Demonstrar a aplicação dessas anotações e do modelo DU na construção dos *treebanks* da língua Akuntsú;
d) Fomentar o diálogo interdisciplinar: Linguística e Etno-História.

Por fim, esclarecemos que não discutiremos todas as frases anotadas no *GitHub*, primeiro pelo espaço aqui destinado e segundo por mantermos o foco na metodologia e na descrição utilizada para anotar dados do Akuntsú. Portanto, no capítulo que segue, separamos quatro anotações de frases com estruturas sintáticas distintas: 1) duas frases com predicados transitivos; 2) uma frase com predicado intransitivo; 3) uma frase com construção genitiva (possessiva). Objetivamos descrever como as anotações foram realizadas em cada uma delas e mostrar a diferença das relações de dependências. Além disso, separamos também um trecho de uma narrativa, sistematizando o diálogo interdisciplinar que propomos identificar neste trabalho.

7 *TREEBANKS* DA LÍNGUA AKUNTSÚ

Neste capítulo serão apresentados, primeiramente, quatro exemplos seguidos do passo a passo de suas construções. No primeiro exemplo, demonstraremos alguns aspectos gerais do formato CoNLL-U e das dependências construídas no *Annotatrix*, discutindo a importância desse procedimento para a descrição linguística; no segundo, terceiro e quarto exemplos, demonstraremos alguns aspectos linguísticos do Akuntsú. Na última anotação demonstraremos um recorte de uma história contada pelos Akuntsú e faremos um resgate da discussão interdisciplinar.

1) *mapi ata kom iko* ‘Ele vai levar a flecha’

```
# sent_id = 0010.1592
# text = mapi ata kom iko .
# text_eng = He is going to take the arrow (5.25)
# text_port = Ele vai levar a flecha
```

1	mapi	mapi	NOUN	n	_	2	obj	_	_
2	ata	at	VERB	vt	Tv=Yes	0	root	_	_
3	kom	kom	PART	pcl	Tense=Fut	2	advmod	_	_
4	iko	ko	AUX	aux	Person=3	2	aux	_	_
5	.	.	PUNCT	punct	_	2	punct	_	_

Tabela 1: Análise da frase *mapi ata kom iko* anotada no formato CoNLL-U. Fonte: Aragon *et al*. (2022).

Na tabela 1, temos um exemplo das colunas no formato CoNLL-U. Cada coluna possui uma representação pertinente à descrição da língua apresentada. Convém, desta forma, retomar o significado de cada coluna, usando como exemplo a representação acima, retirada do *corpus* anotado por Aragon *et al*. (2022). Para ficar ainda mais simples compreender o que cada coluna representa, faremos um recorte de apenas um trecho da anotação acima:

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL	DEPS	MISC
2	ata	at	VERB	vt	Tv=Yes	0	root	_	_

Tabela 2: Análise da palavra *ata* no formato CoNLL-U. Fonte: Aragon *et al.* (2022).

A coluna 1 (*ID*) relaciona-se ao índice de palavras, representado por números, contados a partir do 1. Neste exemplo, a palavra é a de número 2, isto é, a segunda palavra posta na frase. Na coluna 2 (*FORM*) é a palavra em si, *ata*; cada palavra é separada por coluna para que fique mais simples estabelecer suas características. A coluna 3 (*LEMMA*) corresponde ao radical[21] de cada palavra, excluindo sufixos e prefixos; radical -*at*. Nas colunas 4 (*UPOS*) e 5 (*XPOS*), é possível encontrar as categorias morfossintáticas que cada palavra recebe, sendo uma a classe de palavras determinada pela DU (*UPOS*). A DU enumera distintas classes de palavras e selecionamos aqui as que utilizamos para as línguas anotadas no TuDeT: adjetivo, adposição, advérbio, auxiliar, substantivo, verbo, pronome, adposição, conjunção (coordenada e subordina), determinador, numeral, partícula, interjeição e X (quando não é possível determinar a classe de palavra).

A outra classificação refere-se a classe de palavras mais específica do morfema descrito (XPOS). Essas especificações são características morfológicas representadas por etiquetas com valores pré-definidos que podem ser estendidos à medida que uma língua necessite. Desta forma, podemos observar que na quarta coluna a palavra é classificada como um verbo (*VERB)* e na quinta coluna, de acordo com as especificidades da classe, um verbo transitivo (*vt*).

Na coluna 6 (*FEATS*) identifica-se os demais morfemas verbais, neste caso, o morfema -*a*, definido por vogal temática (*TV=YES*) — o YES para a DU representa

[21] Ferrari Neto (2014, p. 59), define o radical de uma palavra como "a parte que está presente em todas as flexões dessa mesma palavra […] também entendida como forma lexical". Ainda segundo esse mesmo autor, o radical de uma palavra é determinado eliminando os seus flexionais.

que, nesta língua, há vogal temática. Segundo Aragon (2014), a vogal temática é uma característica verbal, a qual modifica o verbo fonologicamente e cuja função gramatical não é mais visível sincronicamente.

No que diz respeito às colunas 7 e 8, é possível observar as relações de dependências que as palavras possuem entre si (*DEPREL*). As terminologias da DU utilizadas para anotar tais relações são definidas em um quadro apresentado no site da DU (De Marneffe e*t al*., 2014) e ilustradas abaixo:

	Nominals	Clauses	Modifier words	Function Words
Core arguments	nsubj obj iobj	csubj ccomp xcomp		
Non-core dependents	obl vocative expl dislocated	advcl	advmod* discourse	aux cop mark
Nominal dependents	nmod appos nummod	acl	amod	det clf case
Coordination	MWE	Loose	Special	Other
conj cc	fixed flat compound	list parataxis	orphan goeswith reparandum	punct root dep

Figura 2: Lista das *Universal Dependency relation* (DEPREL). Fonte: *Universal Dependencies*[22]

Para compreendermos melhor as relações de dependência na frase (1), convém observar a Figura 3 abaixo representada no *Annotatrix*.

[22] Disponível em: https://universaldependencies.org/u/dep/index.html. Acesso em: 10 set. 2023

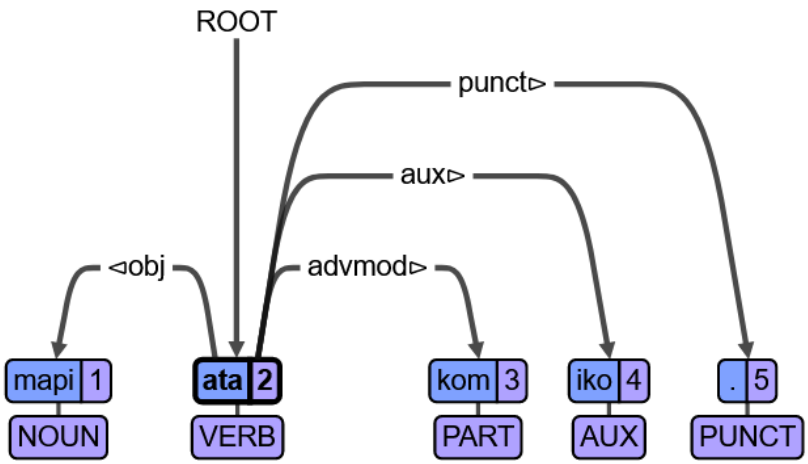

Figura 3: Visão das relações sintáticas da frase *mapi ata kom iko* no *Annotatrix*. Fonte: a autora.

De acordo com De Alencar (2023), os *treebanks* da DU podem ser divididos em dois componentes: 1) os princípios universais que cada língua deve aderir; e 2) os critérios morfossintáticos específicos de uma língua específica. Ambos devem estar em conformidade com o formato CoNLL-U. Posto isto, pode-se observar na Figura 3 que as relações de dependências se iniciam do verbo *ata*, ou seja, o núcleo do predicado (*root*). Se fôssemos fazer uma analogia a uma árvore, a raiz seria o que sustenta a árvore e a formulação dos demais galhos.

Da palavra *ata* ‘pegar’ projetam-se as dependências: a seta para a esquerda determina a relação de *ata* com *mapi* ‘flecha’, objeto (*obj*); três setas para a direita determinam as relações de *ata* com *kom* ‘projetivo’, modificador adverbial (*advmod)*; e *ata* com *iko*, que é de auxiliar (*aux)*; por fim, é estabelecida a relação da raiz com o ponto final (*puntc*), indicando o término da anotação. Nesta frase não há sujeito marcado morfologicamente, pois está subentendido no contexto (quando isso acontece, os falantes tendem a suprimir o sujeito da oração). Outra observação importante é que diferente da Língua Portuguesa que possui o núcleo inicial, ou seja, respeita a ordem verbo-complemento, a Língua Akuntsú possui núcleo final e, por isso, apresenta a ordem complemente-verbo, como pode ser observado na frase (1).

2) *pero õpa Konibu* ‘Konibú bateu na arara’

```
# sent_id = 0010.862
# text = pero õpa Konibu .
# text_eng = Konibú beat the macaw (8.24)
# text_port = Konibú bateu na arara
```

1	pero	pero	NOUN	n	_	2	obj	_	_
2	õpa	õpa	VERB	vt	_	0	root	_	_
3	Konibu	Konibu	PROPN	pn	_	2	nsubj	_	_
4	.	.	PUNCT	punct	_	2	punct	_	_

Tabela 3: Análise da frase *Konibú beat the macaw* anotada no formato CoNLL-U. Fonte: Aragon et al. (2022).

Na tabela 3 convém observarmos alguns aspectos que se manifestam de maneira diferente da frase (1). O sujeito está explícito na oração, estabelecendo uma relação de dependência com o verbo. Como o sujeito da frase é marcado por um nome próprio, a categoria ou *tag* UPOS a qual ele pertence é PROPN. Outras classes morfológicas da DU que comumente assumem posição de sujeito na língua Akuntsú são: substantivo (NOUN) e pronome (PRON). Observe a anotação na Figura 4.

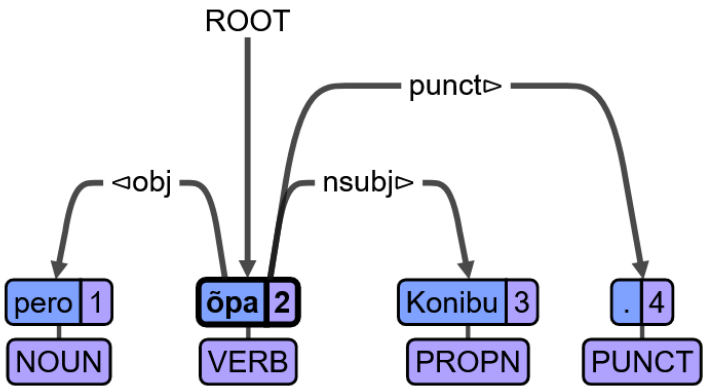

Figura 4: Visão das relações sintáticas da frase *pero õpa Konibu* no *Annotatrix*. Fonte: a autora

3) *kebõ nɨram* ‘Aquela está levantando’

```
# sent_id = 0010.102
# text = kebõ nɨram .
# text_eng = That one is standing up
# text_port = Aquela está levantando
```

1	kebõ	ke	DET	dem	Case=Dat|Deixis=Prox	2	nsubj
2	nɨram	nɨram	VERB	vi		0	root
3	.	.	PUNCT	punct		2	punct

Tabela 4: Análise da frase *kebõ nɨram* anotada no formato CoNLL-U. Fonte: Aragon *et al.* (2022).

Enquanto nos exemplos (1-2) é possível observar a anotação morfossintática em predicados transitivos, no exemplo (3), na tabela 4, podemos observar a relação de dependências de um verbo intransitivo, que, por sua vez, não precisa de nenhum complemento, como é possível observar na Figura 5 abaixo.

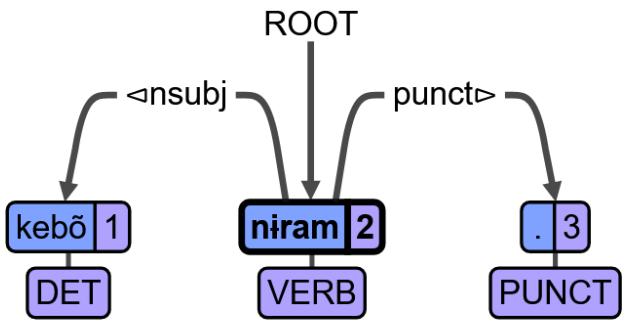

Figura 5: Visão das relações sintáticas da frase pero kebõ nɨram, no *Annotatrix*. Fonte: a autora

4) *tawtʃe tʃogap* ‘Mordida de queixada’

```
# sent_id = 76
# text = tawtʃe tʃogaap
# text_eng = Peccary's bite (4.28e)
# text_port = Mordida de queixada
```

1	tawtʃe	tawtʃe	NOUN	n	_	2	nmod	_	_

2	tʃogaap	tʃoga	NOUN	n	Nomzr=Circ	0	root	_	_

Tabela 5: Análise da frase *tawtʃe tʃogap* anotada no formato CoNLL-U. Fonte: Aragon *et al.* (2022).

No exemplo (4), construção de posse nominal, é possível observar um morfema comum nas línguas Tupí, o nominalizador *-ap* que atribui circunstância aos nomes (Aragon, 2014). As relações de dependências entre determinante (*tʃogap* 'mordida') e determinado (*tawtʃe* 'queixada') são apresentadas na Figura 6.

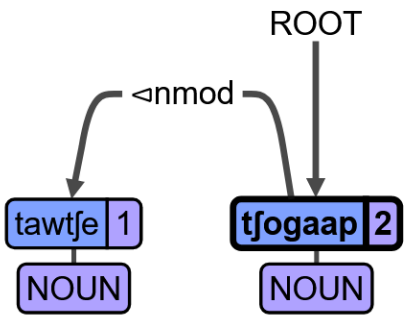

Figura 6: Visão das relações sintáticas da frase *tawtʃe tʃogap* no *Annotatrix*. Fonte: a autora

Como retratado neste trabalho, a DU é um modelo universal que abrange diversas línguas, cada uma com sua especificidade (Duran *et al*. 2022). Portanto, a adoção dos modelos pré-definidos por este programa pode gerar ambiguidades quanto à anotação de uma determinada língua. Como uma forma de lidar com essas ambiguidades, muitas vezes é adotado um manual de anotação, adaptando as regras da DU para uma língua específica. Um exemplo é a partícula de foco *-te* na língua Akuntsú. Inserimos sua descrição com o valor:SIM (Values:Yes), ou seja, especificamos que este morfema se encontra anotado para a língua Akuntsú, como ilustrado na figura 7.

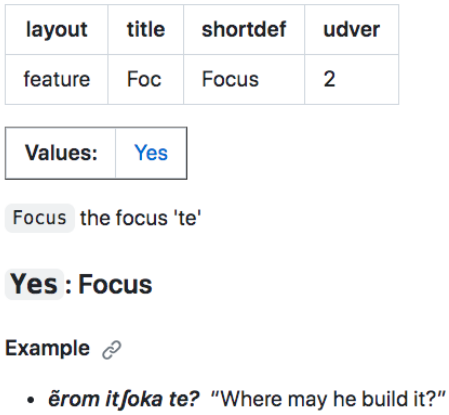

Figura 7: Valor da partícula foco. Fonte: a autora

A DU possui importância, principalmente, em tarefas de treinamento de *parsers* (analisador sintático), visto que viabiliza a anotação de *corpora* anotados sintaticamente e revisados por seres humanos (Duran *et al*. 2022; De Alencar 2023). Além disso, salientamos que o formato de árvore dos *treebanks* não é por acaso, esse formato lógico de dependência, serve, sobretudo, para facilitar o AM. Ademais, um dos princípios da descrição de um *corpus* linguístico é a disponibilidade dos dados, adotar um sistema lógico de anotação e que possibilite a divulgação desses dados *on-line*.

Conectando as anotações com o diálogo interdisciplinar, exemplificamos parte de uma narrativa denominada ‘Pegando mel - Akuntsú e Kanoé’ retirada do banco de dados Akuntsú disponível no *GitHub*:

5) *Ekwitat ko eno , oiat . Txiramanty po ekwita topkora . kojõpi ipa . nom , en nom ekwit pe*. ‘[...] Eu comi a abelha lá, (foi) minha captura. A mão de Txiramanty procurou pela abelha. À noite, retornou. Você não foi, você não foi pelo caminho do mel [...]’

sent_id = 0011.23

text = Ekwitat ko eno , oiat . Txiramanty po ekwitat topkora . kojõpi ipa . nom , en nom ekwit pe .
text_eng = I ate the bee there, my caught (thing). Txiramanty's hand looked for the bee. At night, returned. You didn't, you didn't go to the honey's path.
text_port = Eu comi a abelha lá, (foi) minha captura. A mão de Txiramanty procurou pela abelha. À noite, retornou. Você não foi, você não foi pelo caminho do mel.

1	Ekwitat	Ekwit	NOUN	n	_	2	obj	_	[ɛˈkʷit];mel;honey
2	ko	ko	VERB	vt	_	0	root	_	[ko];comer;eat
3	eno	eno	ADV	adv	_	2	advmod	_	[ˈɛnu];lá;there
4	,	,	PUNCT	punct	_	2	punct	_	_
5	oiat	at	NOUN	vt	Nomzr=Obj|Person[psor]=1	2	appos	_	[at];pegar;to catch
6	.	.	PUNCT	punct	_	5	punct	_	_
7	Txiramanty	Txiramanty	PROPN	n	_	8	nmod	_	_
8	po	po	NOUN	n	_	10	nsubj	_	[pu];mão;hand
9	ekwitat	ekwit	NOUN	n	_	10	obj	_	[ɛˈkʷit];mel;honey
10	topkora	topkora	VERB	vt	_	2	parataxis	_	[topˈkura];procurar;search
11	.	.	PUNCT	punct	_	10	punct	_	_
12	kojõpi	kojõpi	ADV	adv	_	13	advmod	_	[koɲõˈbi];noite;night
13	ipa	ip	VERB	vi	Tv=Yes	10	parataxis	_	[ip];voltar;come.back
14	.	.	PUNCT	punct	_	13	punct	_	_
15	nom	nom	ADV	adv	_	20	discourse	_	[nõm];não;no
16	,	,	PUNCT	punct	_	20	punct	_	_
17	en	en	PRON	pron	Number=Sing|Person=2|PronType=Prs	20	nsubj	_	[ẽn];você;you
18	nom	nom	ADV	adv	_	20	advmod	_	[nõm];não;no
19	ekwit	ekwit	NOUN	n	_	20	nmod	_	[ɛˈkʷit];mel;honey
20	pe	pe	NOUN	posp	_	13	parataxis	_	[pɛ];para;to
21	.	.	PUNCT	punct	_	20	punct	_	_

Tabela 6: Análise de parte do texto 'Pegando mel - Akuntsú e Kanoé' no formato CoNLL-U. Fonte: Aragon *et al.* (2022).

Antes de partirmos para os aspectos históricos contidos nesse recorte, convém observar um aspecto diferencial dessa anotação em relação às anteriores. Na coluna 10, espaço reservado para qualquer outra anotação que o linguista responsável pelo *corpus* achar pertinente, encontra-se a transcrição fonética das palavras. Esse dado está sendo incorporado aos poucos e, por isso, até o momento de escrita deste trabalho, as frases anteriores ainda não contêm essa informação validada na publicação do banco de dados e, por isso, não foi adicionada aqui. Assim como as transcrições fonéticas, as traduções das frases para o português ainda estão em

desenvolvimento. Quanto aos exemplos adicionados neste capítulo, as traduções apresentadas são de nossa autoria e ainda não estão disponíveis na página do *GitHub*.

Consideramos que a Linguística Computacional, Descritiva e a História podem ser inter-relacionadas nas ações voltadas às línguas e culturas indígenas. Uma vez que os povos indígenas possuem uma tradição predominantemente oral, como afirma Cavalcante (2011), a descrição linguística é um dos meios de preservação histórica, uma forma de documentar e compreender histórias orais de um povo. Nesse contexto, a anotação de *corpus* e o modelo da DU, além de favorecer as ferramentas linguísticas, também funcionam como um meio de documentação histórica.

É por meio da língua que as histórias de um povo são narradas, as quais são repletas de conhecimentos culturais (etno-históricos) numa conexão entre povo e território. Para aprofundar essa discussão, observe a narrativa completa do trecho do exemplo (5) acima, contada pela indígena Aiga Akuntsú:

"Comi mel lá, minha captura (o que eu peguei). A mão de Txiramanty (mulher Kanoé) procurou por mel. À noite, voltei. Você não seguiu o caminho do mel. Abelha canudo (sp.), pelo caminho de caça, por ali, casca grossa, árvore do Tucumã (sp.), por ali, perfurou a barriga (o tronco da árvore), abrindo. Foi difícil. Txiramanty empurrou, quebrou. Txiramanty quebrou. A avó derramou (o mel), (e) acabou. Tinha muitos filhotes de abelha, abriu, limpou (tudo de dentro do mel)."

Desta história depreendemos relações entre as Akuntsú e uma mulher Kanoé, Txiramanty. Podemos compreender as relações desses povos com o território na prática de tirar e consumir mel, algo comum entre os povos indígenas, especialmente os Tupí (Cangussu, 2021). O mel faz parte do consumo, mas também das práticas xamânicas do grupo. Os Akuntsú evitam tomar água pura e gostam de misturar água com mel. O mel pode ser passado na ferida aberta como cicatrizante. Há também a menção da árvore Tucumã (*Astrocaryum aculeatum G.Mey.*), comumente encontrada na região norte do Brasil, uma planta de distintos usos para os Akuntsú e os Kanoé. Portanto, partindo dos olhares apresentados neste trabalho e dos valores etno-

históricos depreendidos das frases e das narrativas, reafirmamos a importância da oralidade como meio de registro histórico, como defende Mezacasa (2021). Portanto, acreditamos que a DU fornece ferramentas que auxiliam na construção dos *corpora* não apenas linguístico, mas histórico.

8 CONSIDERAÇÕES FINAIS

Neste trabalho apresentamos ferramentas linguísticas utilizadas na descrição da língua Aluntsú. Além disso, dialogamos sobre o uso das Dependências Universais que surge como uma tendência da Linguística Computacional, viabilizando orientações de anotação de *corpora*. Apresentamos o formato CoNLL-U e o *Annotatrix* como ferramentas computacionais que auxiliam no processo de descrição morfossintática e na construção de *treebanks*.

Este trabalho também buscou ressaltar aspectos, mesmo que breves, sobre a cultura e a história do povo Akuntsú, contextualizando a situação atual de sua língua, elencando alguns aspectos gramaticais (Aragon, 2014). Em seguida, apresentamos a metodologia deste trabalho organizado em duas etapas. A primeira relacionada ao levantamento de referências bibliográficas e a segunda etapa voltada às anotações do *corpus* da língua Akuntsú construídas ao longo da vivência no projeto de PIBIC "Educação, Linguística, História e Comunidades Indígenas" (Edital 2021-2021), executada na Universidade Federal da Paraíba (UFPB).

Demonstramos, ao longo deste estudo, como a anotação morfossintática, utilizando as Dependências Universais, pode ser um veículo não apenas de descrição linguística, mas também de preservação de histórias e da cultura de um povo. Portanto, mostramos como uma ferramenta computacional pode ser uma inovação importante para a descrição de línguas minoritárias, assim como uma forma de disponibilizar o acesso de dados para diferentes públicos dentro de um sistema lógico de organização de dados.

Ainda sobre o que foi apresentado no capítulo 7, descrevemos um recorte dos dados presentes no *treebank* da língua Akuntsú. Para isso utilizamos quatro exemplos: 1) duas frases com predicados transitivos; 2) uma frase com predicado intransitivo; 3) uma frase com construção genitiva (possessiva); 4) uma frase retirada de uma história narrada em Akuntsú. Após cada exemplo, apresentamos as discussões

pertinentes aos aspectos da língua objeto deste estudo, do modelo da DU e da língua como um meio de transmissão histórica e cultural.

Acreditamos, portanto, que este estudo inicial poderá desenvolver novos caminhos e olhares para futuras pesquisas voltadas ao uso de ferramentas linguísticas, como, por exemplo, desenvolver aplicações voltadas a essa e a outras línguas minoritárias. Deste modo, esta temática torna-se relevante para os alunos do Curso de Letras ao demonstrar a interdisciplinaridade e as possibilidades de aplicar os estudos linguísticos.

REFERÊNCIAS

ARAGON, C. 2014. **A Grammar of Akuntsú, a Tupian language**. Tese (Doutorado em Linguística), University of Hawaii. Cidade (Havaí). 2014. Disponível em: http://etnolinguistica.wdfiles.com/local--files/tese%3Aaragon2014/CarolinaAragonFinal.pdf. Acesso em: 29 set. 2022.

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: 28 set. 2022.

ALGAYER, A.; ARAGON, C. C.; MEZACASA, R. Território, Materialidade e Atitude Linguística: ferramentas da Frente de Proteção Etnoambiental Guaporé nos contextos das Terras Indígenas Massaco, Rio Omerê e Tanaru–Rondônia, Amazônia brasileira. **Revista Brasileira de Linguística Antropológica**, v. 14, p. 197-240, 2022.

CAMARA JR. J. M. **Princípios de Linguística Geral**. 4 ed. Rio de Janeiro: Livraria Acadêmica, 1974.

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

CANGUSSU, D. **Manual Indigenista Mateiro**. 2021. Dissertação de Mestrado, INPA.

CAVALCANTE, T. L. V. Etno-história e história indígena: questões sobre conceitos, métodos e relevância da pesquisa. **História (São Paulo)**, v. 30, n. 1, p. 349-371, jan/jun 2011.

CORUMBIARA. 2009. Direção: Vincent Carelli. Produção: Vídeo nas Aldeias. Rondônia. 117 min. Estéreo, colorido.

CREVELS, M.; VOORT, Hein van der. 2008. The Guaporé-Mamoré region. **From linguistic areas to areal linguistics**, v. 90, p. 151-17.

CUNHA. M. C. Introdução a uma história indígena. In: CUNHA, M. C. (org.). **História dos Índios no Brasil**. ed. 2. São Paulo: Companhia das Letras, 1992. p. 9-24.

DE ALENCAR, L. F. Linguística Computacional. In: OTHERO, G. A.; FLORES, V. N. **A linguística hoje: múltiplos domínios**. São Paulo: Contexto, 2023. p. 73-87.

DE ALENCAR, L. F. Yauti: A Tool for Morphosyntactic Analysis of Nheengatu within the Universal Dependencies Framework. In: **SIMPÓSIO BRASILEIRO DE TECNOLOGIA DA INFORMAÇÃO E DA LINGUAGEM HUMANA (STIL)**, 14. 2023, Belo Horizonte/MG. Anais [...]. Porto Alegre: Sociedade Brasileira de Computação, 2023. p. 135-145. DOI: https://doi.org/10.5753/stil.2023.234131.

DE MARNEFFE, M. C.; DOZAT, T.; SILVEIRA, N.; HAVERINEN, K.; GINTER, F.; NIVRE, J.; MANNING, C.D. Universal Stanford dependencies: A cross-linguistic typology. In: **LREC**. 2014. p. 4585-4592.

DE MARNEFFE, M. C.; MANNING, C. D.; NIVRE, J.; ZEMAN, D. 2021. Universal Dependencies. **Computational linguistics**, [s.l], v. 47, n. 2, p. 255-308, jun, 2021.
Disponível em:
https://periodicos.ufsm.br/fragmentum/article/download/23400/13793. Acesso em: 02 jun. 2023.

DURAN, M. S.; NUNES, M. G. V.; LOPES L.; PARDO, T. A. S. Manual de anotação como recurso de Processamento de Linguagem Natural: o modelo Universal Dependencies em língua portuguesa. **Domínios de Lingu@ gem**, v. 16, n. 4, p. 1608-1643, 2022.

ETHNOLOGUE: languages of the World. Disponível em: https://www.ethnologue.com/. Acesso em: 03 jul. 2023.

FAUSTO, C. HECKENBERGER M. 2007. Indigenous History and the History of the Indians. In Fausto, C, Heckenberger, M. (Ed.). **Time and Memory in Indigenous Amazonia**: **Anthropological Perspective**. Florida: University Press of Florida, 2007. p.1-43.

FERRARI NETO, J. Morfologia Flexional. In: Ribeiro, M. G. C. **A morfologia e sua interface com a sintaxe e com o discurso**. 2 ed. João Pessoa: Editora UFPB, 2014

GERARDI, F.; REICHERT, S.; ARAGON, C. MARTÍN-RODDRÍGUEZ, L. GODOY, G. MERZHEVICH, T. 2022a. TuDeT: Tupían Dependency Treebank. 19 May 2022. Zenodo. TuDeT: Tupían Dependency Treebank.

GERARDI, F.; REICHERT, S.; ARAGON, C.; LIST, J.M.; FORKEL, R. 2022b. TuLeD. Tupían Lexical Database. 2022b. Zenodo. TuLeD: Tupían lexical database. Max Planck Institute for Evolutionary Anthropology: Leipzig.

GRIEP, G. W. O que são línguas minoritárias?. **Tesouro Linguístico**, 17 fev. 2021. Disponível em: https://wp.ufpel.edu.br/tesouro-linguistico/2021/02/17/o-que-sao-linguas-minoritarias/. Acesso em: 02 set. 2023.

HIMMELMANN, N. P. Documentary and descriptive linguistics. **Linguistics**. RuhrUniversität Bochum, 1998. v. 36, p. 95-161.

LABOV, William. **Padrões sociolinguísticos**. São Paulo: Parábola Editorial, 2008.

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.

MATTOSO CÂMARA JR. J. **Princípios de Linguística Geral**. 4 ed. Rio de janeiro: Livraria Acadêmica, 1974.

MEZACASA, R. **Por histórias indígenas: o povo Makuráp 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: 10 mar. 2023.

MINDLIN, B. 1986. **Avaliação do componente indígena**. Relatório de andamento. Polonoroeste, Fundação Instituto de Pesquisas Econômicas.

MOORE, D.; GALUCIO, A. V. Perspectives for the documentation of indigenous language in Brazil. In: BÁEZ, G. P.; ROGERS, C.; LABRADA, J. E. R. (org.). **Language Documentation and Revitalization in Latin American Contexts**. 1 ed. Berlin: De Gruyter, v. 295, p. 29-58, 2016.

PADOVANI, B. F. S. L.; MIRANDA, C. C.; BARROS, J. B. A importância da documentação e da descrição linguística para a revitalização de línguas ameaçadas. **Domínios de Lingu@gem**, v. 13, n. 3, 2019. Disponível em:

https://seer.ufu.br/index.php/dominiosdelinguagem/article/download/42062/27302/213080. 03 jul. 2023.

RADEMAKER, A.; REAL, L.; BICK, E.; CHALUB, F.; FREITAS, C.; PAIVA, V.. Universal dependencies for Portuguese. In: **Proceedings of the fourth international conference on dependency linguistics** (Depling 2017). 2017. p. 197-206.

RODRIGUES, A. D. **Línguas Brasileiras**. São Paulo: Loyola, 1986.

RODRIGUES, A. D. Linguística: As línguas indígenas do Brasil. **Fragmentum**, n. 46, p. 289-299, 2015.

RODRÍGUEZ, L.; MERZHEVICH, T.; SILVA, W.; TRESOLDI, T.; ARAGON, C.; GERARDI, 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.

SAUSSURE, F. **Curso de linguística geral**. 28. ed. São Paulo: Cultrix, 2012.

TAVARES, L. K. **Vivendo no "vazio": relações entre os sobreviventes Kanoê e Akuntsú da terra indígena Rio Omerê (RO)**. 2020. Dissertação (Mestrado em Antroplogia) - Programa de Pós-Graduação em Antropologia Social do Departamento de Antropologia da Universidade de Brasília, Brasília, 2020. Disponível em: http://www.realp.unb.br/jspui/handle/10482/39113. Acesso em: 01 ago. 2023.

THOMAS, G. Universal dependencies for mbyá guaraní. In: **Proceedings of the third workshop on universal dependencies** (udw, syntaxfest 2019). 2019. p. 70-77.

TOLEDO, B. F. MIRANDA, C. C. Por que documentar e descrever línguas? A importância desses estudos para revitalização e fortalecimento de línguas indígenas brasileiras. **Articulando e Construindo Saberes**, v. 6, 2021.

WOODBURY, A. C. Language Documentation. In: AUSTIN, P. K.; SALLABANK, J. (org.). **Language documentation and archiving**. New York: Cambridge University Press, 2011. p. 159-186.