OPEN ACCESS

ESCOLA DE HUMANIDADES
PUCRS

LETRAS DE HOJE
Studies and debates in linguistics, literature and Portuguese language
Letras de hoje Porto Alegre, v. 59, n. 1, p. 1-9, jan.-dez. 2024
e-ISSN: 1984-7726 | ISSN-L: 0101-3335

http://dx.doi.org/10.15448/1984-7726.2024.1.44734

SEÇÃO LIVRE

Línguas minoritárias e anotações sintáticas de *corpora*: experiências de pesquisa na iniciação científica

Minorities languages and syntactic annotations of corpora: research experiences in scientific initiation

Lenguas minoritarias y anotaciones sintácticas de corpora: experiencias de investigación en la iniciación científica

Luana Luiza Santos[1]
orcid.org/0000-0002-9843-0206
luana.luiza@academico.ufpb.br

Carolina Coelho Aragon[1]
orcid.org/0000-0001-9459-9939
carolinac.aragon@gmail.com

Fabrício Ferraz Gerardi[2]
orcid.org/0000-0002-1438-7336
fabricio.gerardi@uni-tuebingen.de

Recebido em: 23 mai 2023.
Aprovado em: 22 set 2023.
Publicado em: 11 jan 2024.

Resumo: Muitas das línguas indígenas brasileiras estão ameaçadas de extinção. Na maioria dos casos, estratégias de revitalização e de conservação dessas línguas são imprescindíveis (Crystal, 2002; Harrison, 2007), necessitando de processos contínuos de promoção de políticas linguísticas e de ações voltadas à educação escolar indígena. Este artigo apresenta o uso de ferramentas linguísticas associadas à construção de *treebanks* (*corpus* de textos com anotações sintáticas e morfológicas) e à descrição de duas línguas minoritárias do tronco linguístico Tupí faladas no sudoeste Amazônico. Os *treebanks*, parte das Dependências Universais (De Marneffe *et al.*, 2021; Duran *et al.* 2022), são a base de algumas das atividades do projeto "Educação, Linguística, História e Comunidades Indígenas" vinculado ao Programa Institucional de Bolsas de Iniciação Científica (2021-2022) da Universidade Federal da Paraíba (UFPB). Neste artigo, discutimos a aplicação dessas ferramentas na descrição linguística e exploramos a interseção da linguística computacional com a linguística descritiva.

Palavras-chave: linguística computacional; linguística descritiva; línguas Tupí; dependências universais; *treebanks*.

Abstract: Many of the Brazilian indigenous languages are endangered. In most cases, revitalization and conservation strategies for these languages are essential (Crystal, 2002; Harrison, 2007), requiring continuous processes of promoting language policies and actions focused on indigenous school education. This article presents the use of linguistic tools associated with the construction of treebanks (corpora of texts with syntactic and morphological annotations) and the description of two minority indigenous languages belonging to the Tupían linguistic family spoken in the southwestern Amazon, Brazil. The treebanks, part of the Universal Dependencies project (De Marneffe *et al.*, 2021; Duran *et al.*, 2022), form the basis of experiments conducted in the Institutional Program for Scientific Initiation Scholarships at the Federal University of Paraíba (2021-2022), entitled "Education, Linguistics, History, and Indigenous Communities." We discuss the application of these tools in linguistic description, their relationship with the study of indigenous language typology. Furthermore, we explore the intersection of computational linguistics with descriptive linguistics.

Keywords: computational linguistics; descriptive linguistics; Tupían languages. universal dependencies; treebanks.

Resumen: Muchas de las lenguas indígenas de Brasil están en peligro de extinción. En la mayoría de los casos, las estrategias de revitalización y conservación de estas lenguas son esenciales (Crystal, 2002; Harrison, 2007), requiriendo procesos continuos para promover políticas y acciones lingüísticas dirigidas a la educación escolar indígena. Este artículo presenta el uso de herramientas lingüísticas asociadas a la construcción de treebanks (corpus de textos con anotaciones sintácticas y morfológicas) y a la descripción de dos lenguas minoritarias

[1] Universidade Federal da Paraíba (UFPB), João Pessoa, PB, Brasil.
[2] Eberhard Karls Universität Tübingen, Tübingen, Alemanha.

del tronco lingüístico Tupí habladas en el suroeste de la Amazonía, Brasil. Los bancos de árboles, parte del proyecto Dependencias Universales (De Marneffe *et al*., 2021; Duran *et al*. 2022), son la base de las experiencias desarrolladas en el Programa Institucional de Becas de Iniciación Científica de la Universidad Federal de Paraíba (2021-2022), titulado "Educación, Lingüística, Historia y Comunidades Indígenas". Discutimos la aplicación de estas herramientas en la descripción lingüística, sus relaciones con el estudio de la tipología de lenguas indígenas. Además, exploramos la intersección de la lingüística computacional y la lingüística descriptiva.

Palabras clave: lingüística computacional; lingüística descriptiva; lenguas Tupí; dependencias universales; *treebanks*.

Introdução

A introdução de técnicas computacionais facilita a criação e o compartilhamento de dados linguísticos em larga escala, os quais vão de pesquisas de conteúdos na internet, análise de textos à criação automática de materiais pedagógicos. No que diz respeito a essas discussões, 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, uma vez que fatores de frequência[3] desempenham um papel importante no que diz respeito à estrutura linguística (Hawkins, 2004, 2014). Essa área proporciona ferramentas para anotação de *corpus*, como por exemplo os *treebanks* — sistemas arbóreos que incluem categorias morfossintáticas (Duran *et al*., 2022). 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.

As ferramentas linguísticas, para organizar e para programar análises quantitativas e qualitativas de um grande número de dados, buscam especificações e padrões com acesso e compartilhamento de dados obedecendo ao princípio *FAIRness* de dados: ***F****indable,* ***A****ccessible,* ***I****nteroperable,* ***R****eproducible* (Wilkinson *et al.*, 2016). Um exemplo disso é o desenvolvimento de *treebanks* para as pesquisas no âmbito cognitivo, psicológico e para as análises comparativas e tipológicas das línguas naturais (faladas e de sinais).

Refletindo sobre o uso de ferramentas linguísticas para o contexto de línguas minoritárias, este trabalho visa apresentar o uso do formato CoNLL-U[4] como ferramenta de descrição de línguas indígenas brasileiras a partir de experiências de pesquisas desenvolvidas no Programa Institucional de Bolsas de Iniciação Científica (PIBIC) da Universidade Federal da Paraíba (UFPB) intitulado "Educação, Linguística, História e Comunidades Indígenas". Descreveremos as metodologias e as ferramentas linguísticas utilizadas para analisar línguas indígenas, com foco nas línguas Akuntsú e Makurap, ambas pertencentes à subfamília linguística Tuparí, família (tronco) Tupí.

Este artigo foi organizado da seguinte forma: na primeira seção dialogamos sobre as línguas minoritárias e as situações de vulnerabilidade que as envolvem, bem como determinadas ações políticas em torno delas; na segunda seção, descrevemos o projeto de PIBIC, vigência 2021-2022, e algumas características do plano de trabalho; na terceira seção, apresentamos o modelo de Dependências Universais (*Universal Dependencies*, UD) e as anotações linguísticas desenvolvidas no projeto; e, por fim, na quarta seção, as considerações finais são abordadas.

1 Línguas minoritárias e a situação de vulnerabilidade

Muitas línguas indígenas estão ameaçadas de extinção com níveis de vitalidade diferenciados (Ethnologue, 2023). No Brasil, as línguas indígenas são consideradas minoritárias, não apenas pelo número de falantes como pela situação de vulnerabilidade.

As questões de educação e de política linguística, em especial as referentes às línguas

[3] Grandes quantidades de textos permitem, por exemplo, que se compreenda o porquê de certas estruturas que competem com outras estruturas similares serem preferidas ou evitadas em uma língua. A frequência de certas estruturas possibilita inferências referentes à relação de complexidade, processamento e eficiência (ver Hawkins 2004, 2014).

[4] Formado por uma tabela de 10 colunas com informações associadas aos *tokens* nas linhas. Falaremos mais sobre isso nas próximas seções.

minoritárias, são fatores de discussões em diferentes áreas (Altenhofen, 2013; Maher, 2013), as quais vêm apontando ações de promoção de diversidade cultural e linguística como vetor de desenvolvimento econômico e social (Morello, 2009), principalmente, neste momento da Década Internacional das Línguas Indígenas (DILI) (2022-2032) —, fundamentada pela "Declaração de Los Pinos"[5] da Unesco. Esta declaração reconhece, entre outros aspectos, os fundamentos para a construção de atividades que efetivem a participação dos povos indígenas nas atividades voltadas à valorização e à manutenção das línguas indígenas. Frentes de conservação e de preservação de línguas minoritárias, como os trabalhos desenvolvidos pelo Instituto de Patrimônio Histórico e Artístico (IPHAN) e como o Inventário Nacional da Diversidade Linguística (INDL), servem, por exemplo, de "instrumento oficial de identificação, documentação, reconhecimento e valorização das línguas faladas pelos diferentes grupos formadores da sociedade brasileira" (IPHAN, [2022]).

Aryon Rodrigues (2006) estimou cerca de 1200 línguas indígenas no território brasileiro no início da ocupação europeia — atualmente esse número não chega a 160 línguas (Storto, 2019). No início da década de 2000, cerca de 21% das línguas brasileiras estavam ameaçadas de extinção (Moore *et al*., 2008). A extinção de línguas pode ser desencadeada por distintos fatores que vão desde o desaparecimento físico dos falantes — em decorrência de epidemias; de extermínio direto; de redução de territórios; e de destruição das condições de sobrevivência – à aculturação forçada (Ramos, 2018; Nettle; Romaine, 2000; Crystal, 2002).

Dentre as línguas minoritárias brasileiras, o projeto de PIBIC aqui descrito trabalhou durante o período de 2021 a 2022 com duas línguas da família Tupí: Akuntsú e Makurap, ambas do subgrupo Tuparí. Atualmente o Akuntsú é falado por três mulheres (totalidade do grupo) que vivem na Terra Indígena (TI) Rio Omerê, em Rondônia (Aragon; Algayer, 2020). O primeiro contato oficial da Fundação Nacional dos Povos Indígenas (Funai) com os Akuntsú aconteceu no ano de 1995. Nessa época, esse povo estava reduzido a sete membros (Santos; Algayer, 1995). O histórico desse coletivo e de outros, como os Kanoé do Omerê (os quais dividem o território com os Akuntsú), é marcado por fugas e mortes intensificadas na década de 1980, quando o sudeste do estado de Rondônia começou a ser ocupado por movimentos de frentes expansionistas incentivados pelo governo federal.

A história dos Makurap, assim como os Akuntsú, é caracterizada por violências e perdas populacionais ao longo dos anos de exploração do seu território tradicional e dos recursos naturais ali encontrados. 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, a situação territorial 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 Guaporé (Mezacasa, 2021).

2 O PIBIC

O projeto de PIBIC insere-se na temática de educação, de linguística e de etnohistória, visando: a) disseminar conhecimentos sobre línguas e culturas indígenas; b) viabilizar para os estudantes do curso de Letras da UFPB e demais interessados experiências na área de análise linguística e de ensino-aprendizagem de línguas; e c) fortalecer o campo de formação de futuros pesquisadores e docentes ao produzir materiais pedagógicos e acadêmicos. Já o plano de trabalho, foco deste artigo, desenvolve atividades voltadas à ampliação dos *treebanks* das línguas indígenas Akuntsú e Makurap.

---

[5] DECLARAÇÃO de Los Pinos. *In*: *UNESDOC Digital Library*. França: UNESCO, 2020. Disponível em: https://unesdoc.unesco.org/ark:/48223/pf0000374030. Acesso em: 25 maio 2022.

As ações desse projeto estão incorporadas ao *Tupían Dependencies Treebanks* (TuDeT)[6] (Ferraz Gerardi *et al.*, 2022a) que possui hoje *treebanks* de nove línguas Tupí, a saber: Akuntsú, Guajajara, Ka'apor, Karo, Makurap, Munduruku, Guarani Antigo, Teko e Tupinambá, cada um em diferentes fases de desenvolvimentos, além de banco de dados lexicais (Ferraz Gerardi *et al.*, 2022b). Ademais, outros *treebanks* estão sendo desenvolvidos por diferentes pesquisadores focados em outras línguas indígenas brasileiras: Nheengatu (Tupí Moderno) (De Alencar, 2023); Mbyá Guaraní (Thomas, 2019); Apurinã; e Xavante.[7]

Os *treebanks* documentados no TuDeT são organizados de modo que seja possível a visualização das análises de dependência das frases, dos seus componentes sintático-morfológicos.[8] Uma característica relevante do TuDeT é a sua terminologia unificada para as anotações morfológicas das línguas Tupí, o que, entre outras coisas, possibilita uma análise diacrônica dessas línguas. Ao consultar distintas descrições de línguas Tupí já publicadas chegou-se a uma terminologia geral para a morfologia destas línguas, tanto quanto possível (Rodríguez *et al.*, 2022).

A Tabela 1 apresenta informações sobre as línguas que fazem parte do TuDeT e a quantidade de frases anotadas até o presente momento.[9] Os critérios estabelecidos para os níveis de vitalidade apresentados abaixo foram retirados do Ethnologue (2023), assim como o quantitativo de falantes – com exceção da língua Akuntsú, informados por Aragon (pesquisadora que trabalha diretamente com o grupo).

TABELA 1 – *Status* das línguas e quantidade de frases dos *treebanks* – TuDeT

Línguas	Glottocode	Falantes	Nível de vitalidade	Frases
Akuntsú	akun1241	3	Quase extinta	328
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	913
Tupinambá	tupi1273	0	Extinta	546

Fonte: Os autores (2022).

Pode-se imaginar que as ferramentas linguísticas aqui apresentadas são demasiadamente complexas para se trabalhar na graduação, visto que o processo de anotação requer conhecimentos linguísticos pouco estudados na graduação. Não obstante, os discentes envolvidos relatam que não tiveram dificuldades, já que as anotações possuem uma estrutura organizacional muito lógica e possível de compreender. Além das descrições fornecidas no site da UD, as discussões e explicações práticas durante as reuniões do projeto possibilitaram o manuseio das ferramentas de modo satisfatório.

---

6 O TuDeT faz parte do Tupían Language Resources (TuLaR)

7 O UD é usado para anotar *corpus* de distintas línguas, não apenas as línguas minoritárias, como majoritárias, e.g., o português (Rademaker *et al.,* 2017).

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 2.11 lançada em novembro de 2022.

3 Anotações linguísticas e o modelo da UD

A UD é um programa de anotações morfossintáticas de línguas naturais baseada em um sistema arbóreo, os *treebanks*, formados a partir do princípio de que a linguagem possui uma estrutura hierárquica (De Marneffe *et al.*, 2021; Duran *et al.* 2022). Um aspecto importante da UD consiste na documentação de cada *treebank*, incluindo descrições gramaticais, além de explicações com relação às escolhas do tratamento morfossintático dado às línguas (Nivre *et al.*, 2020). A UD, dessa maneira, visa "fornecer um inventário universal de categorias e de 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" (*Universal Dependencies*, c2014-2022, tradução nossa).[10]

Considerando a diversidade estrutural das línguas, a UD descreve as classes de palavras, as quais são divididas em: adjetivo, adposição, advérbio, auxiliar, substantivo, verbo, pronome, adposição, conjunção (coordenada e subordinada), determinador, numeral, partícula, interjeição e X (quando não é possível determinar a classe de palavra). Há, ainda, as categorias de classes de palavras, conhecidas por XPOS, específicas para cada língua. Características morfológicas são representadas por etiquetas com valores pré-definidos que podem ser estendidos à medida que uma língua necessite. Por fim, há uma série de relações de dependências que capturam a estrutura sintática da frase anotada.

As anotações morfossintáticas (modelo UD) realizadas durante as ações do projeto de PIBIC foram organizadas em três partes principais: a) transcrição das frases utilizando um editor de texto com visualizações para extensões específicas de arquivo – o *Sublime Text*[11] (ver Figuras 1 e 2); b) anotação das dependências no *Annotatrix* (Tyers *et al.*, 2017) (ver Figuras 3 e 4); e c) inserção das anotações no repositório UD no *Github*.

Após converter as frases com um *script* em linguagem de programação *Python* para o formato CoNLL-U, transferimos os dados dos trabalhos publicados da língua Akuntsú (Aragon, 2008; Aragon, 2014) e da língua Makurap (Braga, 2005) para o editor de texto *Sublime Text*.

O formato CoNLL-U 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 opcional. A Figura 1 exemplifica uma anotação em formato CoNLL-U na língua Akuntsú e a Figura 2 na língua Makurap.

Figura 1 – Formato CoNLL-U: *korakora nom aot tejã*. "A galinha não está saindo" (Aragon, 2014)

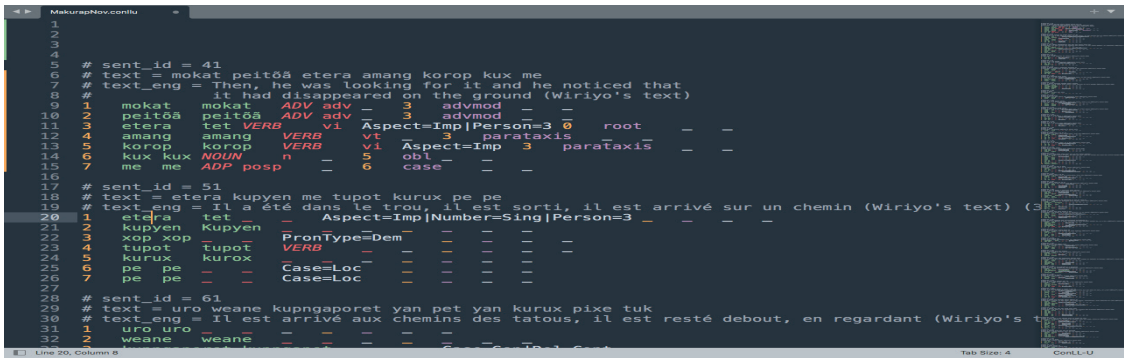
```
# text = korakora nom aot tejã .
# text_eng = Chicken is not going out   (5.19c)
1   korakora    korakora    NOUN    n       _   3   nsubj   _   _
2   nom nom ADV adv Advmod  4   advmod  _   _
3   aot aot VERB vi _   4   root    _   _
4   tejã    jã  AUX aux Person=3|Reflex=Yes 3   aux _   _
5  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

Fonte: Autores (2022).

[10] Do original: 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. Disponível em: https://universaldependencies.org/introduction.html. Acesso em: 10 out. 2022.

[11] Editor de código HTML com linguagem Python.

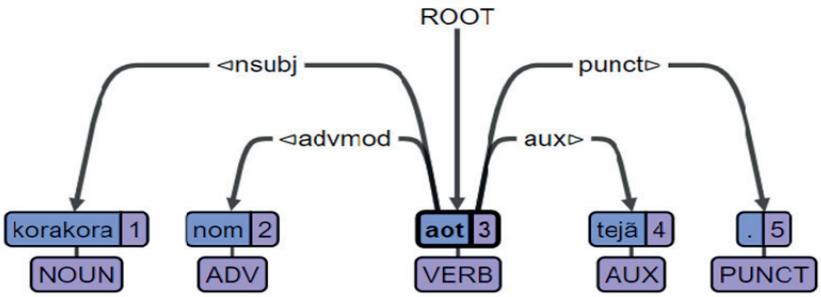
Figura 2 – Formato CoNLL-U: *mokat peitõã etera amang korop kux me.* "Então, ele estava procurando-o e percebeu que havia desaparecido do chão" (Braga, 2005)



Fonte: Autores (2022).

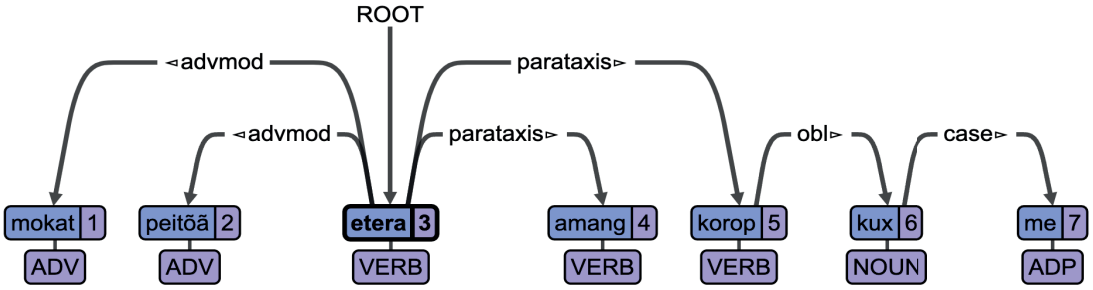
Após converter as frases para o formato CoNLL-U, seguimos para as anotações manuais das dependências no *Annotatrix*. Esse trabalho manual faz-se necessário nesta etapa inicial do projeto, porém, quando houver um número alto de frases anotadas, será possível usar ferramentas de anotações automáticas. Tais ferramentas automáticas são programas 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). Visualização das dependências e da morfologia das frases apresentadas nas Figura 1 e 2 estão ilustradas nas Figuras 3 e 4.[12]

Figura 3 – Relações de dependência: *korakora nom aot tejã*. "A galinha não está saindo"



Fonte: Autores (2022).

Figura 4 – Relações de dependência: *mokat peitõã etera amang korop kux me*. "Então, ele estava procurando-o e percebeu que havia desaparecido do chão" (Braga, 2005)



Fonte: Autores (2022).

[12] ADV - advérbio; ADVMOD - modificador adverbial; AUX - auxiliar; CASE - caso; NOUN - nome; NSUBJ - sujeito nominal; OBL - oblíquo; ROOT - núcleo do predicado; Parataxis - parataxe; PUNCT - pontuação; VERB - verbo.

Por último, os dados anotados são vinculados ao repositório UD no *Github*.[13] Observe no texto representado parcialmente abaixo, um exemplo dos dados Akuntsú anotados no repositório. Na 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) e *text_port* (tradução para o português).

Figura 4 – Texto '*kʷai*'

```
1324    20      mãpara  mãpara  VERB    vi      _       17      parataxis       _       _
1325    21      .       .       PUNCT   punct   _       20      punct   _       _
1326
1327    # sent_id = 0011.13
1328    # text = ikɨp ãka put . ikɨp pɨtka piri put mãpara .
1329    # text_eng = Her leg that way hit it. Her leg pierced because of the foot, it is over.
1330    # text_port = A perna dela, assim, picou. Furou a perna dela, foi por causa do pé, (mas) já acabou.
1331    1       ikɨp    kɨp     NOUN    n       Number=Sing|Person=3    3       obj     _       _
1332    2       ãka     ãka     PART    pcl     _       3       discourse       _       _
1333    3       put     put     VERB    ideo    _       0       root    _       _
1334    4       .       .       PUNCT   punct   _       3       punct   _       _
1335    5       ikɨp    kɨ      NOUN    n       Number=Sing|Person=3    6       obj     _       _
1336    6       pɨtka   pɨt     VERB    vt      Trans=Yes       3       parataxis       _       _
1337    7       piri    pi      NOUN    n       Case=Abl        6       obl     _       _
1338    8       put     put     VERB    vi      _       6       parataxis       _       _
1339    9       mãpara  mãpara  VERB    vt      _       6       parataxis       _       _
1340    10      .       .       PUNCT   punct   _       6       punct   _       _
1341
1342    # sent_id = 0010.442
```

Fonte: Autores (2022).

No caso da língua Akuntsú, as anotações procuram sistematizar as descrições já publicadas como também trabalhar com textos inéditos (UD_Akuntsú-*treebank*),[14] seguindo a metodologia apresentada neste estudo.

Por fim, os dados são publicados pela UD depois de passar por um processo de validação. Esse processo consiste na análise de erros das relações de dependências e de traços morfossintáticos anotados para cada *treebank*. A publicação (*release*) de um a nova versão dos *treebanks* na UD é realizada duas vezes por ano. Essas podem ser citadas e usadas em diversos estudos, já que todos os dados são de livre acesso.

Considerações finais

Este artigo, ao relacionar ferramentas linguísticas e línguas minoritárias, procurou associar o uso de *treebanks* e do formato CoNLL-U à descrição de línguas indígenas. Relacionamos a importância de estabelecer vinculações com a linguística computacional em conjunto com a linguística 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.

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, proporcionando espaços para a formação profissional dos discentes envolvidos no projeto.

Referências

ARAGON, Carolina. *A Grammar of Akuntsú, a Tupian language.* 2014. 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, Carolina. *Fonologia e aspectos morfológicos e sintáticos da língua Akuntsú.* 2008. 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.

ARAGON, Carolina; ALGAYER, Altair. 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.

[13] Plataforma gerenciadora de programas com acesso livre aos dados.
[14] Ver: https://github.com/UniversalDependencies/UD_Akuntsú-TuDeT/blob/dev/aqz_tudet-ud-test.conllu. Acesso em: 8 jun. 2022.

ALTENHOFEN, Cléo V. Bases para uma política linguística das línguas minoritárias no Brasil. *In*: NICOLAIDES, C.; SILVA, K. A.; TÍLIO, R; ROCHA, C. H. (org.). *Política e Políticas Linguísticas*. Campinas: Pontes Editores, 2013. p. 93-116.

BRAGA, Alzerinda. *Aspects morphosyntaxiques de la langue Makurap-tupi.* 2005. 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.

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

DE ALENCAR, Leonel Figueiredo. 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. https://doi.org/10.5753/stil.2023.234131.

DE MARNEFFE, Marie-Catherine; MANNING, Chistopher; NIVRE, Joakim; ZEMAN, Daniel. Universal Dependencies. *Computational linguistics*, [*S. l.*], v. 47, n. 2, p. 255-308, jun. 2021.

DURAN, Magali Sanches *et al.* Manual de anotação como recurso de Processamento de Linguagem Natural: o modelo Universal Dependencies em língua portuguesa. Domínios de Lingu@ gem, [*S. l.*], v. 16, n. 4, p. 1608-1643, 2022.

ETHNOLOGUE. languages vitality count. Ethnologue: Languages of the World, 2023. Disponível em: https://www.ethnologue.com. Acesso em: 16 out. 2022.

FERRAZ GERARDI, Fabrício *et al.*, TuDeT: Tupían Dependency Treebank. [*S. l.*], 19 May 2022. Zenodo. TuDeT: Tupian Dependency Treebank, 2022a.

FERRAZ GERARDI, Fabrício *et al*. TuLeD. Tupían Lexical Database. [*S. l.*], 23 May 2022. Zenodo. TuLeD: Tupían lexical database. Max Planck Institute for Evolutionary Anthropology: Leipzig, 2022b.

FREITAS, Cláudia. *Linguística computacional*. São Paulo: Parábola, 2022.

HARRISON, David. *When languages die:* The extinction of the world's languages and the erosion of human knowledge. Oxford University Press, 2007.

HAWKINS, John A. *Efficiency and complexity in grammars*. OUP Oxford, 2004.

HAWKINS, John A. *Cross-linguistic variation and efficiency*. OUP Oxford, 2014.

IPHAN. Inventário Nacional da Diversidade Linguística. *In*: *Portal Iphan*. [*S. l.*], c2014. Disponível em: http://portal.iphan.gov.br/indl. Acesso em: 1 out. 2022.

MAHER, Terezinha M. Ecos de resistência: políticas linguísticas e línguas minoritárias no Brasil. *In*: NICOLAIDES, C.; SILVA, K. A.; TÍLIO, R; ROCHA, C. H. (org.). *Política e Políticas Linguísticas*. Campinas: Pontes, 2013. p. 117-134.

MALDI, Denise. 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,* Belém, v. 7, n. 2, p. 209-269,1991.

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

MOORE, Denny; GALUCIO, Ana Vilacy; GABAS JR., Nilson. *O desafio de documentar e preservar as línguas amazônicas.* Belém: Museu Paraense Emílio Goeldi, 2008.

MORELLO, Rosângela. Diversidade no Brasil: línguas e políticas sociais. *Synergies Brésil,* [*S. l.*], v. 7, p. 27-36, 2009. Disponível em: http://gerflint.fr/Base/Bresil7/bresil7.html. Acesso em: 16 out. 2022

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

NIVRE, Joakim; DE MARNEFFE, Marie-Cathetine; GINTER, Filip; HAJIC, Jan; MANNING, Chistopher; PYYSALO, Sampo; SCHUSTER, Sebastian; TYERS, Francis; ZEMAN, Daniel. Universal dependencies: An evergrowing multilingual treebank collection. *European Language Resources and Evaluation,* Marseille, v. 2, p. 4034-4043, maio, 2020. Disponível em: https://aclanthology.org/2020.lrec-1.497.pdf. Acesso em: 26 out. 2022.

RADEMAKER, Alexandre *et al*. Universal dependencies for Portuguese. *In*: INTERNATIONAL CONFERENCE ON DEPENDENCY LINGUISTICS, 4., 2017, Pisa. *Proceedings* [...]. Pisa: Linköping University Electronic Press, 2017. p. 197-206.

RAMOS, Alcida. Vivos, afinal! Povos indígenas do Brasil enfrentam o genocídio. *In*: *Série Antropologia*. Brasília: DAN/UnB, 2018. v. 461.

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

RODRÍGUEZ, Lorena.; MERZHEVICH, Tatiana; SILVA, Wellington; TRESOLDI, Tiago; ARAGON, Carolina; GERARDI, Fabrício. Tupían Language Resources: Data, Tools, Analyses. *In*: ANNUAL MEETING OF THE ELRA/ISCA SPECIAL INTEREST GROUP ON UNDER-RESOURCED LANGUAGES, 1., 2022, Marseille. *Anais* [...]. Paris: European Language Resources Association, 2022. p. 48-58.

SANTOS, Marcelo. ALGAYER, Altair. Índios Isolados do Vale do Corumbiara. Brasília: Fundação Nacional do Índio, 1995. (Relatório Técnico).

STORTO, Luciana. R. *Línguas indígenas*: tradição, universais e diversidade, São Paulo: Mercado de Letras, 2019.

THOMAS, Guillaume. Universal dependencies for mbyá guaraní. *In*: WORKSHOP ON UNIVERSAL DEPENDENCIES, 2019, 3., Paris. *Anais* [...]. Paris: The Association for Computational Linguistics, 2019. p. 70-77.

TYERS, Francis; SHEYANOVA, Mariya; WASHINGTON, Jonathan. UD Annotatrix: An annotation tool for Universal Dependencies. *In*: INTERNATIONAL WORKSHOP ON TREEBANKS AND LINGUISTIC THEORIES, 16., 2018, Prague, Czech Republic. *Anais* [...]. Praga: Jan Hajič, 2017. p. 10-17.

WILKINSON, Mark; DUMONTIER, Michel; AALBERSBERG, Ijisbrand; APPLETON, Gabrielle; AXTON, Myles; BAAK, Arie; MONS, Barend. 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.

Luana Luiza Santos

Graduanda em Letras – Português pela Universidade Federal da Paraíba (UFPB), em João Pessoa, PB, Brasil.

Carolina Coelho Aragon

Doutora pela Universidade do Havaí, em Honolulu, HI, Estados Unidos; mestra pela Universidade de Brasília, DF, Brasil. Professora Adjunta do Departamento de Língua Portuguesa e Linguística da Universidade Federal da Paraíba (UFPB), João Pessoa, PB, Brasil.

Fabrício Ferraz Gerardi

Doutor em Linguística e mestre em Linguística Computacional pela Universidade de Tübingen, na Alemanha. Também possui mestrado em língua hebraica pela Universidade de São Paulo (USP), em São Paulo, Brasil. Professor e pesquisador da Universidade de Tübingen, em Tübingen, Alemanha.

Endereços para correspondência

Luana Luiza Santos

Universidade Federal da Paraíba

Centro de Ciências Humanas, Letras e Artes

Jardim Cidade Universitária, 58033-455

João Pessoa, PB, Brasil

Carolina Coelho Aragon

Universidade Federal da Paraíba

Centro de Ciências Humanas, Letras e Artes

Jardim Cidade Universitária, 58033-455

João Pessoa, PB, Brasil

Fabrício Ferraz Gerardi

Eberhard Karls Universität Tübingen

Seminar für Sprachwissenschaft

Keplerstraße 2

72074 Tübingen

Alemanha

Os textos deste artigo foram revisados pela SK Revisões Acadêmicas e submetidos para validação do(s) autor(es) antes da publicação.