Aspectos da construção de um corpus sintaticamente anotado do nheengatu no modelo Dependências Universais

Aspects of the construction of a Universal Dependencies treebank for Nheengatu

Leonel Figueiredo de Alencar *[1]

[1] *Universidade Federal do Ceará, Centro de Humanidades, Fortaleza, CE, Brasil.*

Resumo

O alheamento das tecnologias da linguagem natural constitui fator adicional de enfraquecimento de línguas minoritárias relativamente às línguas majoritárias com as quais convivem. Sobretudo os falantes mais jovens, elos da transmissão linguística, tendem a migrar para a língua favorecida com esses recursos. O nheengatu é uma língua indígena brasileira em perigo de extinção, com índice de suporte digital de apenas 0,07 na escala *Digital Language Support* (DLS), significativamente inferior à pontuação de 0,97 do português, para o qual tem perdido continuamente falantes. O *treebank* do nheengatu da coleção Dependências Universais visa a contribuir para redução dessa deficiência, alimentando o treinamento de um *parser* neural. O *treebank* estreou com 196 sentenças e 2.146 palavras na versão de 15/11/2023 dessa coleção. Este artigo trata da versão mais recente do *treebank*, que, composto de amostras de sentenças extraídas de vinte publicações de diferentes fases históricas do nheengatu, perfazendo 1.470 sentenças e 15.036 palavras, constitui o maior de língua ameríndia da versão de 15/05/2024 da coleção Dependências Universais. A utilização de um analisador automático acelerou o crescimento do *corpus*. Anotadores humanos, porém, revisaram cada anotação automática, assegurando um índice de validação de 100% do *treebank* e concorrendo para a classificação de duas estrelas, a mais alta conferida a *treebanks* de línguas ameríndias da coleção Dependências Universais. A expansão e revisão do *corpus* continuará, visando a abarcar todos os textos em domínio público e alcançar acurácia de *parsing* do estado da arte.

Palavras-chave: Linguística computacional. Processamento de linguagem natural. Tupinologia. Corpus sintaticamente anotado.

Abstract

The alienation of natural language technologies adds up to the weakening of minority languages coexisting with majority languages. Especially younger speakers, who function as links in language transmission, tend to migrate to the language favored by these resources. Nheengatu, an endangered Brazilian indigenous language, has a digital support score of just 0.07 on the Digital Language Support (DLS) scale. This is significantly lower than the 0.97 score for Portuguese, to which Nheengatu has been continually losing speakers. The Nheengatu treebank of the Universal Dependencies collection aims to reduce this deficit by feeding the training of a neural parser. Initially released on 11/15/2023 with 196 sentences and 2,146 words, the latest version, as of 05/15/2024, comprises 1,470 sentences and 15,036 words from twenty publications spanning different historical phases of Nheengatu. This makes it the largest treebank for an Amerindian language in the collection. The use of an automatic analyzer facilitated the rapid expansion of the corpus, while human annotators reviewed each annotation to ensure a 100% validation rate, achieving a two-star rating, the highest for Amerindian language treebanks in the Universal Dependencies collection. The ongoing expansion and revision aim to include all public domain texts and achieve state-of-the-art parsing results.

Keywords: Computational linguistics. Natural language processing. Tupinology. Treebank.

Textolivre
Linguagem e Tecnologia

DOI: 10.1590/1983-3652.2024.52653

Seção:
Artigos

Autor Correspondente:
Leonel Figueiredo de Alencar

Editor de seção:
Daniervelin Pereira
Editor de layout:
Leonel Figueiredo de Alencar

Recebido em:
18 de maio de 2024
Aceito em:
16 de julho de 2024
Publicado em:
20 de agosto de 2024

1 Introdução

É fato conhecido do mundo biológico que espécies medram em ambientes propícios e definham em condições desfavoráveis. Não é diferente no campo das línguas naturais, "organismos vivos que,

*Email: leonel.de.alencar@ufc.br

como outro vivente qualquer, nascem, crescem e se desenvolvem para culminar numa florescência [...] [exuberante] ou estiolar e morrer" (Stradelli, 1929, p. 63). A história tem dado testemunho do desaparecimento de diversas línguas (Rodrigues, 1993). Se uma língua morre sem deixar descendentes[1] e se, concomitantemente, não tiver sido documentada, perde-se para sempre uma peça do quebra-cabeças da linguagem ou mesmo da cognição humana (Rodrigues, 1986; Storto, 2019).

Num país de extrema diversidade linguística como o Brasil, a glototanásia tem vitimado diversas línguas indígenas (Rodrigues, 1986, 1993). Das cerca de 150 línguas indígenas ainda sobreviventes no país (Storto, 2019), um número expressivo está ameaçado de extinção (Eberhard; Simons; Fennig, 2023). Nesse grupo se insere o nheengatu, também conhecido como tupi moderno ou Língua Geral Amazônica (LGA) (Rodrigues, 1996; Freire, 2011; Rodrigues; Cabral, 2011; Moore, 2014). Até meados do século XIX a língua mais falada da região norte (Navarro, 2016), não obstante os estimados 6000 falantes no Brasil e 8000 na Colômbia, o nheengatu ocupa nesses dois países, segundo Eberhard, Simons e Fennig (2023), respectivamente, os níveis 6b e 7 da Escala Graduada Expandida de Interrupção Intergeracional (EGIDS, abreviatura de *Expanded Graded Intergenerational Disruption Scale*). Os níveis 6b e 7 significam que, embora a geração reprodutiva ainda use o idioma, "a transmissão intergeracional está em vias de ser interrompida" (Eberhard; Simons; Fennig, 2023). Na Venezuela, o nheengatu ocupa o nível 8b da escala, indicando a sua quase extinção. O nível 10 da EGIDS é atribuído a línguas extintas, como, por exemplo, ao próprio tupinambá, do qual o nheengatu é descendente.[2]

Conforme a EGIDS, a chave para a vitalidade de uma língua não é tanto o número absoluto de falantes como a transmissão natural de pais para filhos. Com apenas 780 falantes, o jamamadi, por exemplo, classifica-se no nível 5 dessa escala (Eberhard; Simons; Fennig, 2023).

Um dos fatores para a sobrevivência de uma língua nos dias de hoje é a sua presença no mundo digital. As novas gerações, exatamente aquelas responsáveis pela cadeia de transmissão de uma língua, são os usuários mais frequentes das tecnologias digitais, de cujas vidas fazem parte numa proporção dificilmente imaginada na época dos seus pais ou avós. Tradução automática, texto para voz, voz para texto e assistentes virtuais, entre outras ferramentas, representam a culminância de mais de um século de pesquisas em inteligência artificial, processamento de linguagem natural (PLN) e linguística computacional, compondo o ecossistema que determina a vitalidade dos idiomas que delas se beneficiam ou o definhamento daqueles ignorados por esse desenvolvimento científico e tecnológico.

O PLN e suas áreas afins, linguística computacional e linguística de *corpus*, desenvolveram-se muito nos últimos 20 anos no Brasil, focando quase exclusivamente língua portuguesa. Paralelamente, constatamos um crescente interesse da comunidade científica internacional pela construção de recursos para o tratamento computacional de línguas minoritárias, indígenas ou em perigo de extinção. Por exemplo, Ptaszynski, Mukaichi e Momouchi (2013) relatam sobre o desenvolvimento de ferramentas para processamento automático do ainu, língua nativa do norte do Japão quase extinta. Esses esforços, no entanto, praticamente ignoraram a enorme diversidade de línguas indígenas da América do Sul, representada por cerca de 456 línguas, até muito recentemente, como salientam Gerardi, Reichert e Aragon (2021), que propõem uma base de dados lexical da família tupi, segundo eles, especialmente sub-representada, não obstante constituir a maior do continente. Nesse contexto, uma iniciativa digna de nota foi a criação de um *treebank* para shipibo-konibo (Vasquez *et al.*, 2018), aparentemente o primeiro recurso desse tipo para uma língua ameríndia no modelo Dependências Universais (doravante UD) (Nivre *et al.*, 2016; Marneffe *et al.*, 2021).

Felizmente, o ostracismo tecnológico a que se relegaram as línguas indígenas brasileiras começa a reverter-se. Galves *et al.* (2017) e Sandalo e Galves (2023) discutem a elaboração de um esquema de anotação para um *corpus* sintaticamente anotado (*treebank*) do cadiuéu. Rueter *et al.* (2021)

[1] Sobre esse termo, ver nota 2.

[2] Cruz (2011) vale-se do modelo de árvore genealógica para classificar as línguas da família tupi-guarani. No diagrama que propõe, cada nó representa uma língua ou grupo ("ramo") de línguas, conectado a um ou mais nós, de tal modo que se podem estabelecer as diferentes relações de parentesco entre as línguas ou grupos. Por exemplo, o tupinambá constitui o nó "mãe" do nheengatu. Em linguística histórica e comparativa, a árvore genealógica e toda a terminologia associada, como *língua ancestral*, *línguas filhas*, *línguas parentes*, *família linguística etc.*, constituem metáforas, as quais não devem ser tomadas ao pé da letra (Aikhenvald; Dixon, 2001). Como ressalta Faraco (2022), não é a língua que muda, mas os falantes.

relatam sobre as diferentes etapas da construção do primeiro *treebank* do apurinã no modelo UD, visando à implementação de aplicações de PLN de alto nível, v.g., sistemas de resolução de perguntas (QA, do inglês *question answering*), tradução automática *etc.* D'Angelis, Oliveira e Schwade (2021) tratam da tradução do sistema operacional de um fabricante de celulares para caingangue e nheengatu. Alencar (2021) descreve um tradutor automático para predicações qualificativas em nheengatu, língua para a qual Alencar (2023) propõe uma ferramenta denominada Yauti capaz de construir árvores no formato UD. Martín Rodríguez *et al.* (2022) e Santos, Aragon e Gerardi (2024) apresentam diversos recursos computacionais para línguas tupis, incluindo *treebanks* no modelo UD para nove membros dessa família.

O presente trabalho vem somar-se a esses esforços, constituindo mais um passo para dirimir a disparidade de condições de sobrevivência do nheengatu frente ao português.[3] Com a progressiva absorção, a partir do último quartel do século XIX, dos seus falantes à cultura de língua portuguesa, ancorada na escrita, o nheengatu, de tradição essencialmente oral, enfraqueceu-se cada vez mais, perdendo terreno adicional com a televisão e a internet, sobretudo entre os falantes mais jovens (Navarro; Ávila; Trevisan, 2017). A concorrência com a língua portuguesa tornou-se ainda mais desfavorável com a disseminação avassaladora das tecnologias de linguagem natural. Simons, Thomas e White (2022) enfatizam a importância do suporte digital para a sobrevivência de uma língua num mundo mediado digitalmente, fator para o qual propõem cálculo de pontuação de 0 a 1 e classificação numa escala de cinco faixas sequenciadas de forma ascendente (Figura 1).

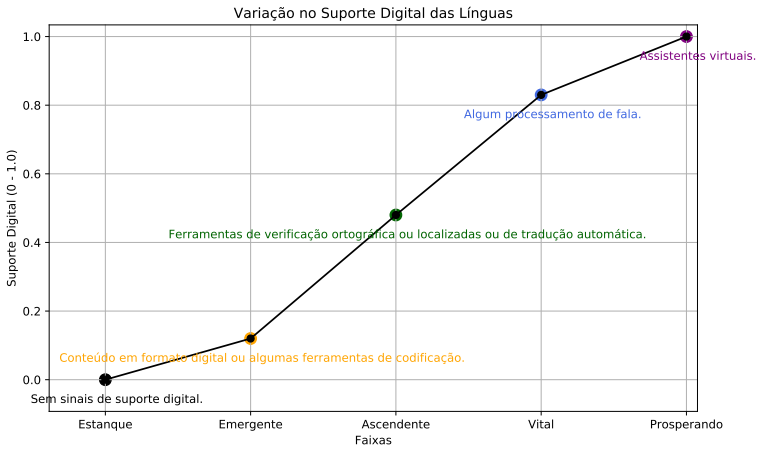


Figura 1. Escala de Suporte Digital às Línguas (DLS, do inglês *Digital Language Support*).

Fonte: Elaboração própria a partir das especificações de Simons, Thomas e White (2022) e Eberhard, Simons e Fennig (2023).

O nheengatu, com base nessa escala, enquadra-se na categoria de línguas com suporte digital emergente, contudo, a pontuação alcançada é de apenas 0,07 (Eberhard; Simons; Fennig, 2023), a uma gigantesca distância de línguas majoritárias como português e inglês, com pontuações de 0,97 e 1,0.

O desenvolvimento de ferramentas robustas de PLN é uma tarefa de engenharia de software não trivial. O que para o usuário final ocorre num passe de mágica oculta nos bastidores um edifício de conhecimentos e técnicas acumulados e aperfeiçoados durante décadas. Trata-se de complexo passível de divisão numa série de componentes encadeados, de tal forma que o desempenho de um componente depende essencialmente daquele que o precede na cadeia. Um dos componentes essenciais é a análise sintática automática (*syntactic parsing*), da qual depende a análise semântica e, consequentemente, a compreensão textual automática, base, por sua vez, para aplicações como sistemas de QA, tradutores

[3] Conforme Eberhard, Simons e Fennig (2023), o espanhol parece pouco usado entre falantes do nheengatu. De um lado, resta, quando muito, um número ínfimo de falantes na Venezuela. De outro lado, na Colômbia, vem ocorrendo uma migração do nheengatu para a língua tucano.

automáticos, assistentes virtuais *etc.*

Este artigo trata do UD_Nheengatu-CompLin, aparentemente o único *treebank* sintático do nheengatu.[4] Distribuído sob a licença CC BY NC SA, expandiu-se de 196 sentenças e 2.146 palavras na versão v2.11 da coleção UD para 1.470 e 15.036 na versão atual (Figura 2). Esse tipo de recurso permite construir um analisador sintático automático (*parser*) por meio de aprendizagem de máquina, utilizando, por exemplo, redes neurais (Straka; Hajič; Straková, 2016). Sua utilidade, porém, transcende o domínio da informática. A linguística de *corpus* passou a constituir a partir da última década um método padrão de levantamento de dados nos mais variados domínios das ciências da linguagem, da dialetologia e sociolinguística à linguística histórica, passando pela teoria gramatical, sanando limitações de dados obtidos por introspecção (Hirschmann, 2019).

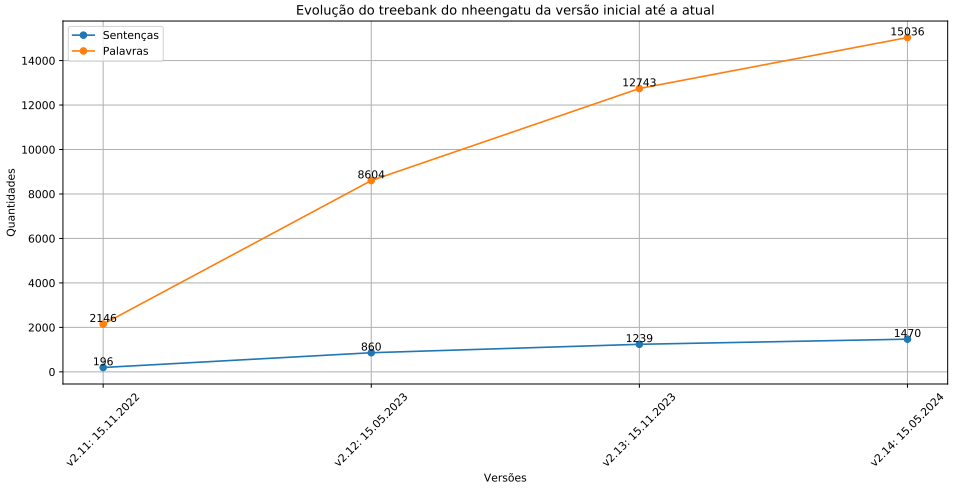


Figura 2. Crescimento do UD_Nheengatu-CompLin.
Fonte: Elaboração própria.

Na próxima seção, expomos o quadro teórico que fundamentou a construção do *treebank*. A seção seguinte trata da metodologia, focando a seleção dos textos, a normalização ortográfica e o fluxograma de anotação e revisão. A seção 4 aborda as diferentes dimensões da anotação. A última seção apresenta as conclusões e oferece sugestões para trabalhos futuros.

2 Aspectos do projeto UD

Mirando tanto investigações linguísticas quanto a interpretação semântica e aplicações de PLN, UD almeja consistência na anotação morfossintática de línguas tipologicamente as mais diversas (Marneffe *et al.*, 2021), no âmbito de um esforço coletivo no espírito do código aberto que resultou numa coleção de 283 *treebanks* de 161 línguas, em contínua expansão desde 2015 (Figura 3).

A Figura 4 exemplifica o formato CoNLL-U do modelo UD. O primeiro componente consiste em uma série de metadados encabeçados por #, enquanto o segundo é uma tabela de dez colunas separadas por tabulações, com uma linha para cada palavra. Essas colunas especificam, para cada palavra, as seguintes informações, nesta ordem: (i) índice de ordem (ID); (ii) forma (FORM); (iii) lema (LEMMA); (iv) etiqueta do conjunto de partes do discurso universais (UPOS) (Tabela 1); (v) classe de palavra de um conjunto XPOS de etiquetas específicas da língua em questão; (vi) traços morfológicos (FEATS); (vii) núcleo regente (HEAD); (viii) relação de dependência (DEPREL); (ix) "dependências aprimoradas" (*enhanced dependencies*) (DEPS) e (x) informações adicionais (MISC). Por exemplo, a forma *yaikú* 'estamos' porta o índice 3, uma vez que é a terceira da sentença. O lema dessa palavra é o radical *ikú* 'estar'. Diferentemente do verbo pleno *purungitá* 'conversar' da

[4] O UD_Nheengatu-CompLin com os respectivos arquivos `stats.xml` e `eval.log` e os principais *scripts* utilizados para processar esses dados encontram-se neste link. Os três primeiros arquivos integram a versão 2.14 da coleção UD (Zeman *et al.*, 2024). O repositório https://github.com/CompLin/nheengatu contém a versão mais atualizada do *treebank* e das respectivas ferramentas.

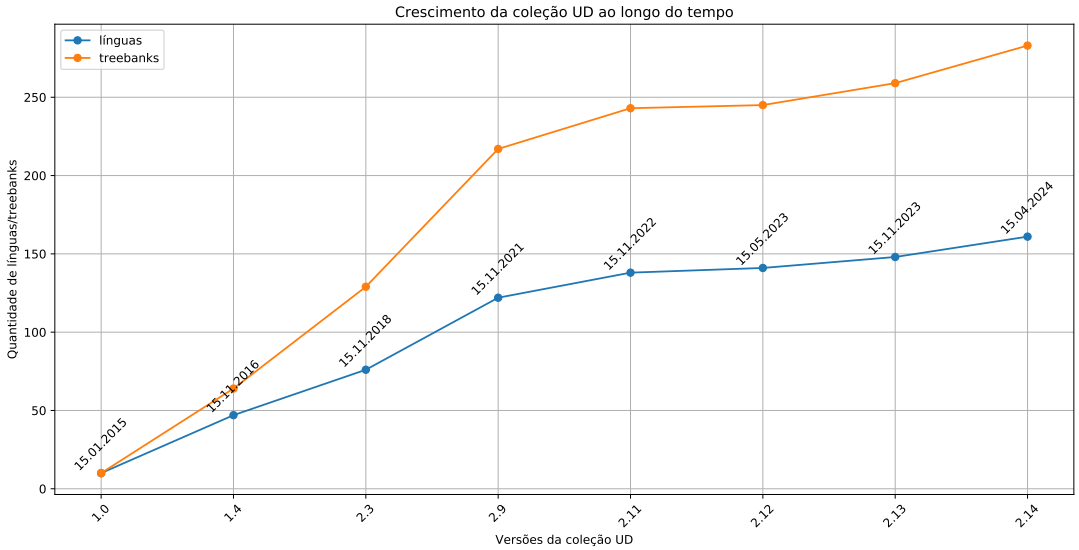


Figura 3. Crescimento da coleção UD do início até a versão atual.
Fonte: Elaboração própria com base nos dados de Marneffe *et al.* (2024c).

```
# sent_id = MooreFP1994:0:0:4
# text = Yandé yapurungitá yaikú nheengatú.
# text_eng = We are talking Nheengatu.
# text_por = Nós estamos falando nheengatu.
# text_source = p. 105
# text_orig = yãndé [[ya-purĩngitá]V [ya-ikú]Aux ye?ẽngatú]VP
# text_annotator = XXXX
# reviewer1 = YYYY
1   Yandé   yandé   PRON    PRON    Number=Plur|Person=1|PronType=Prs   2   nsubj   _   TokenRange=0:5
2   yapurungitá purungitá   VERB    V   Number=Plur|Person=1|VerbForm=Fin   0   root   _   TokenRange=6:17
3   yaikú   ikú AUX AUXFS   Number=Plur|Person=1|VerbForm=Fin   2   aux   _   TokenRange=18:23
4   nheengatú   nheengatú   NOUN    N   Number=Sing 2   obj   _   SpaceAfter=No|TokenRange=24:33
5   .   .   PUNCT   PUNCT   _   2   punct   _   SpaceAfter=No|TokenRange=33:34
```

Figura 4. Análise de sentença do UD_Nheengatu-CompLin no formato CoNLL-U.
Fonte: Elaboração própria.

linha 2, etiquetado como VERB na respectiva coluna (iv), o auxiliar *ikú* 'estar' classifica-se como AUX no modelo UD. Um *treebank* da coleção UD não precisa especificar todos esses campos, mas apenas (i), (iv), (vii) e (viii). A coluna (ix), que abriga as chamadas dependências aprimoradas (Schuster; Manning, 2016), é preenchida apenas por parte dos *treebanks* da coleção UD.

Tabela 1. Etiquetas do conjunto UPOS.

Palavras de classe aberta	**Palavras de classe fechada**	**Outras**
ADJ: adjetivo	ADP: adposição	PUNCT: sinal de pontuação
ADV: advérbio	AUX: auxiliar	SYM: símbolo
INTJ: interjeição	CCONJ: conjunção coordenativa	X: outra
NOUN: nome comum	DET: determinante	
PROPN: nome próprio	NUM: numeral	
VERB: verbo	PART: partícula	
	PRON: pronome	
	SCONJ: conjunção subordinativa	

Fonte: Adaptado de Marneffe *et al.* (2024c).

A Tabela 1 exibe as 17 etiquetas de partes do discurso do conjunto UPOS, postuladas como universais, ou seja, capazes de classificar qualquer palavra de qualquer língua. No entanto, o conjunto de partes do discurso de uma língua particular não precisa coincidir com UPOS, podendo constituir um subconjunto (Marneffe *et al.*, 2021). O modelo UD agrupa as diferentes classes em três macrocategorias, as duas primeiras das quais abrangem a noção tradicional de palavra. A primeira macrocategoria consiste em conjuntos passíveis de contínua expansão, por exemplo, por meio de processos de formação de palavras ou empréstimos de outras línguas, enquanto a segunda macrocategoria abarca as palavras de classe fechada, que constituem conjuntos relativamente fixos, que raramente incorporam novos membros (Jurafsky; Martin, 2009; Lehmann, 2013). Tipicamente, a criação de novos membros das classes fechadas ocorre pela gramaticalização de itens de classe aberta (Lehmann, 2013). Um exemplo do nheengatu é o substantivo *pukusawa* 'comprimento', que passou a funcionar como posposição ('durante') e conjunção ('enquanto') (Cruz, 2011; Avila, 2021).[5]

A terceira macrocategoria consiste de sinais de pontuação, símbolos e qualquer outro tipo de material textual que não se enquadre em nenhuma das demais classes. Sinais de pontuação são tratados no modelo UD como os demais *tokens* que representam palavras num sentido tradicional, ocupando, como estas, nós na árvore sintática, ligados a outros nós por relações de dependência (Figura 5).[6]

Com exceção da classe de particípio, subsumida na classe dos verbos, as duas primeiras macrocategorias da Tabela 1 englobam as taxonomias óctuplas grega e latina antigas (Robins, 1966) e a classificação de palavras de Macambira (1999) e Lima (2011), entre outros, que, na esteira da Nomenclatura Gramatical Brasileira (NGB), consideram substantivo, pronome, artigo, advérbio, verbo, preposição, conjunção, interjeição, adjetivo e numeral suficientes para abarcar todo o léxico do português. Na teoria UD, os artigos integram a classe dos determinantes, que abriga muitos dos itens tratados como pronomes em abordagens mais tradicionais, como indefinidos, interrogativos e demonstrativos. Dado seu viés tipológico, a teoria UD substituiu a categoria preposição pela adposição, que abrange tanto preposições quanto posposições e circumposições. Diferentemente da NGB, a classe numeral do modelo UD, seguindo modelos precedentes de anotação de *treebanks* como o de Santorini (1990), se restringe aos numerais cardinais, excluindo, por exemplo, em línguas como o português, palavras como *segundo* ou *terceiro*, analisadas como adjetivos (Duran, 2021).[7] As outras discrepâncias do modelo UD em relação à NGB são a categoria partícula e a subdivisão de verbos, substantivos

[5] Cumpre ressaltar que os critérios para definição dessas duas macrocategorias variam conforme o arcabouço teórico. Lehmann (2013, p. 27), por exemplo, considera adposições e conjunções como "classes abertas em todas as línguas europeias modernas."

[6] Todas as árvores dependenciais deste artigo foram geradas pelo visualizador https://www.let.rug.nl/kleiweg/conllu/.

[7] Pelo contrário, Francis e Kučera (1979) prescrevem a etiquetagem de cardinais e ordinais como CD e OD.

e conjunções nas classes VERB, AUX, NOUN, PROPN, CCONJ e SCONJ.

A classificação de numerais cardinais e interjeições é controversa. Greenberg (2000) observa que cardinais comportam-se de forma análoga tanto a adjetivos quanto substantivos em várias línguas, com o que Evans (2000) concorda, destacando, por outro lado, a rígida estruturação formal e semântica desses elementos, característica de palavras de classe fechada. Em gramática gerativa, há autores que enxergam uma natureza lexical nos cardinais, classificando-os como substantivos ou adjetivos, enquanto outros os analisam como categoria funcional Num ou Q, que projeta um sintagma numeral (NumP) ou quantificador (QP), respectivamente (Ionin; Matushansky, 2018). Tesnière (1959) classifica o cardinal em *deux livres* 'dois livros' como adjetivo de quantidade, tratando as interjeições como tipo não de palavra, mas de sentença, designando-as *mots-phrases* 'palavras-frase'. Analogamente, Cunha e Cintra (2017, p. 92) excluem a interjeição das diferentes classificações que postulam para palavras e morfemas, dada a sua natureza de "vocábulo-frase" ou "grito com que traduzimos de modo vivo nossas emoções"(Cunha; Cintra, 2017, p. 605). Pelo contrário, Ameka (1992) ressalta o caráter provavelmente universal das interjeições, considerando-as uma classe de palavras. Wilkins (1992, p. 120) contesta a inserção das interjeições "no cesto de lixo dos 'fenômenos paralinguísticos'". Para ele, a interjeição constitui simultaneamente um lexema e um enunciado, relevando à investigação teórica nas mais diferentes subdisciplinas da linguística. No tratamento da interjeição, o modelo UD se soma a uma prática na construção de *corpora* anotados que remonta a Francis e Kučera (1979) e, posteriormente, a Santorini (1990), em cujos esquemas de anotação essa parte do discurso figura sob a etiqueta UH.

O conjunto XPOS do UD_Nheengatu-CompLin constitui-se de 78 etiquetas, incorporando tanto distinções correntes na descrição do nheengatu, como a subclassificação dos pronomes pessoais em primeira e segunda classes (Navarro, 2016; Avila, 2021), ou a subdivisão dos verbos em ativos (dinâmicos) e inativos (estativos) (Cruz, 2011), quanto distinções de granularidade mais fina que visam a capturar particularidades semânticas e/ou morfossintáticas. No UD_Nheengatu-CompLin, via de regra, subdivisões de uma categoria do conjunto UPOS no âmbito do conjunto XPOS são indicadas por sufixos que refletem as propriedades diagnósticas de cada subgrupo. Por exemplo, as etiquetas AUXFS e AUXFR do conjunto XPOS distinguem auxiliares flexionados pós-verbais (Figura 5) e pré-verbais (Figura 6), respectivamente.

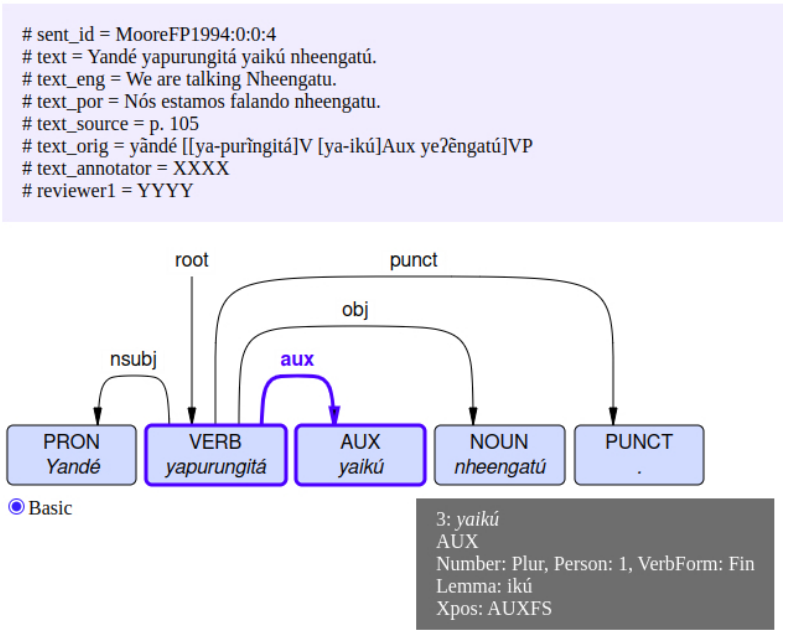


Figura 5. Representação dependencial de sentença do UD_Nheengatu-CompLin destacando os traços da forma *yaikú* 'estamos'.

Fonte: Elaboração própria.

Os traços morfossintáticos da coluna (vi), referida na documentação de UD pela abreviatura

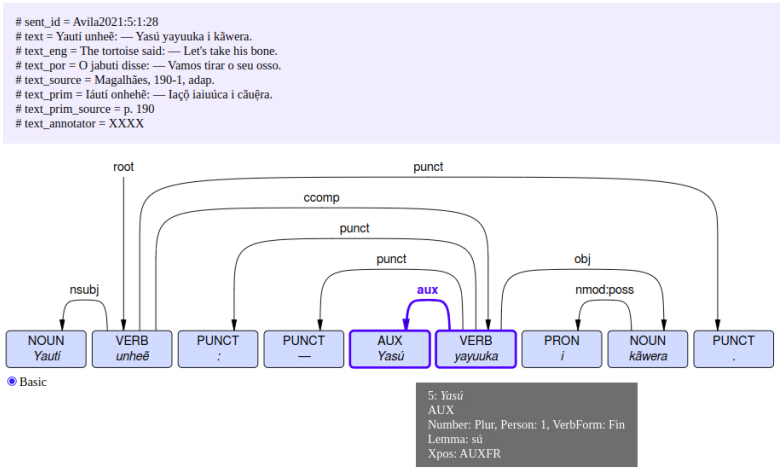


Figura 6. Representação dependencial de sentença do UD_Nheengatu-CompLin com o auxiliar pré-verbal *sú* 'ir'.
Fonte: Elaboração própria.

FEATS, constituem uma lista ordenada alfabeticamente de pares de atributos e valores no formato `ATRIBUTO=VALOR`, separados por |. Por exemplo, na coluna (vi) das linhas 2 e 3 da tabela da Figura 4, os pares `Number=Plur`, `Person=1` e `VerbForm=Fin` especificam que se trata de formas verbais finitas na primeira pessoa do plural. O par `PronType=Prs` na coluna (vi) da linha 1 distingue os pronomes pessoais dos demais tipos de pronomes, como os demonstrativos, os interrogativos *etc.*

As informações das colunas (vii) e (viii) definem a árvore sintática dependencial, que consiste numa série de nós conectados por arcos direcionados de um nó pai regente a um nó filho dependente. Cada arco é rotulado com uma abreviatura de uma das 37 relações de dependência postuladas pela teoria UD (doravante DEPREL), como `nsubj` (sujeito nominal) e `obj` (objeto direto).

O modelo UD adota uma posição lexicalista na análise sintática. Isso significa que a menor unidade linguística do *corpus* é a palavra. Unidades menores, como o prefixo *ya* de primeira pessoa do plural (Figura 5) ou o prefixo relacional *r* de contiguidade (Figura 7), não constituem nós na árvore dependencial. A contribuição desses morfemas para a representação morfossintática da sentença é codificada na estrutura de traços. As árvores sintáticas nesse modelo estabelecem, portanto, relações de dependência entre palavras (Figuras 5 e 6), ao contrário das de teorias constitucionais, que se baseiam na estrutura sintagmática, representando relações de parte e todo (Figura 8).

A raiz da árvore dependencial, identificada pelo índice 0, é o único nó que não representa uma palavra da frase. Trata-se de artifício que confere uma estrutura comum a todo tipo de sentença. Esse nó abstrato domina o nó da palavra hierarquicamente superior da sentença por meio da relação `root`. No caso de predicados verbais, esse nó mais alto é o único verbo pleno da oração (Figura 5) ou o verbo principal (Figura 6). A Figura 7 exibe a estrutura dependencial de sentença em que o nó mais alto constitui um substantivo.

O modelo UD distingue entre palavras de conteúdo, que incluem todas as classes da primeira macrocategoria da Tabela 1, e palavras funcionais, que constituem um subconjunto próprio do conjunto definido pela segunda macrocategoria da Tabela 1, ou seja, toda palavra funcional integra o conjunto de palavras de classe fechada, mas nem toda palavra de classe fechada é funcional (Lehmann, 2023). Esse é o caso, por exemplo, dos numerais e dos pronomes (Lopes; Duran *et al.*, 2022).[8] Via de regra, palavras funcionais figuram apenas como dependentes, ao passo que palavras de conteúdo tanto podem reger outras palavras quanto ser regidas (Marneffe *et al.*, 2024a). Com base nessa distinção, as relações do conjunto DEPREL subdividem-se em dois grupos. O primeiro grupo consiste de relações sintáticas que se estabelecem entre duas palavras de conteúdo, por exemplo, entre dois substantivos, dois verbos ou dois numerais,[9] entre um verbo e um substantivo ou pronome, um substantivo e um

[8] A esse respeito, o modelo UD difere de abordagens como Lehmann (2013), que trata pronomes como palavras gramaticais.
[9] Por exemplo, em expressões como *5 – 6 metros* ou *1920 x 1080 pixels* (Marneffe *et al.*, 2021, p. 285).

```
# sent_id = MooreFP1994:0:0:5
# text = Ixé se ruka upé aikú.
# text_eng = I'm in my house.
# text_por = Eu estou na minha casa.
# text_source = p. 105
# text_orig = išé [se-rúka upé]PP a-ikú
# text_annotator = XXXX
```

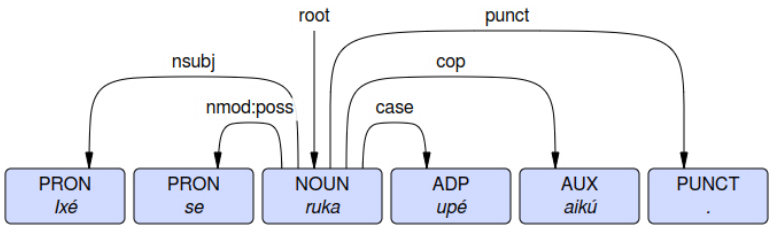


Figura 7. Representação dependencial de sentença do UD_Nheengatu-CompLin com predicado nominal.

Fonte: Elaboração própria.

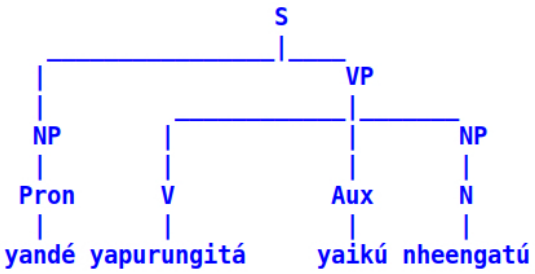


Figura 8. Árvore sintagmática baseada em Moore, Facundes e Pires (1994) para a sentença da Figura 5.

Fonte: Elaboração própria.

adjetivo ou um verbo e um advérbio. Além de `nsubj` e `obj`, esse grupo abarca relações como `iobj` (objeto indireto), `obl` (oblíquo), `csubj` (sujeito oracional), `ccomp` (complemento oracional), `xcomp` (complemento oracional aberto) e `nmod` (modificador nominal). O segundo grupo abrange relações sintáticas entre uma palavra de conteúdo e uma palavra funcional, como `case`, `mark` e `aux`, que se aplicam a adposições, conjunções subordinativas e auxiliares, respectivamente.

A teoria não impõe que toda língua natural possua todas as 37 relações sintáticas. Exige apenas que as relações sintáticas de uma língua constituam um subconjunto dessas relações. Por outro lado, a teoria permite que uma relação sintática seja subtipada, de modo a contemplar especificidades de uma língua. Esses subtipos podem ser criados livremente, mas precisam estar documentados no sistema de UD para serem aceitos pelo *script* de validação. O UD_Nheengatu-CompLin vale-se no momento de três relações subtipadas, a saber `nmod:poss`, `acl:relcl` e `advcl:relcl`, utilizadas por vários outros *treebanks*.

A última coluna da tabela da Figura 4, denominada MISC, abriga quaisquer informações adicionais que construtores de um *treebank* julguem relevantes. Seguindo vários *treebanks* da coleção UD, o UD_Nheengatu-CompLin inclui nesse campo, entre outros, os atributos SpaceAfter e TokenRange. O primeiro indica a ausência de espaço em branco subseguindo o *token*. O segundo atributo indica a posição inicial e final do *token* na cadeia de caracteres (*string*) por meio da notação `n:m`, adotada na linguagem de programação Python, onde `n` é o índice do primeiro caractere do *token* e `m-1`, do último.

3 Metodologia de construção do corpus

3.1 O material linguístico

Nesta seção, tratamos do material linguístico que compõe a versão atual do UD_Nheengatu-CompLin, ao qual continuamente novos exemplos se agregam. No momento, as sentenças do *treebank* provêm de 20 publicações, a maior parte das quais documentam a história do nheengatu de meados do século XIX à segunda década do XXI. Exemplos de outras publicações, como Góes Neto (2015), Avila (2016) e Trevisan (2017) serão igualmente incorporados ao *treebank*. Extrapolaria o âmbito deste artigo discorrer pormenorizadamente sobre cada uma dessas fontes. Para tanto, remetemos o leitor a Avila (2021), que faz um levantamento detalhado dos registros escritos da LGA do século XVIII às duas primeiras décadas do século XXI. Das fontes nominadas na Figura 9, Moore, Facundes e Pires (1994) e Alencar (2021) são as únicas que não integram o corpus histórico-filológico de Avila (2021).

Os *treebanks* da coleção UD de algumas das línguas majoritárias mais ricas em recursos para o PLN, como, por exemplo, alemão, japonês, checo, russo, português e árabe, possuem cada um mais de um milhão de palavras, os três primeiros mais que o dobro disso. No caso de uma língua que, apesar dos mais de 300 anos de história, só mais recentemente iniciou uma tradição escrita própria (os textos anteriores constituindo na maioria das vezes registros de falantes não nativos), um *treebank* à altura dos dessas línguas teria de abranger todos os registros históricos acessíveis e toda ou quase toda a produção dos últimos anos.

Para dar uma ideia da relativa escassez de textos em nheengatu, a tradução do Novo Testamento (Brasil, 2019), publicada pela primeira vez em 1973, constituindo, ao que tudo indica, o texto mais longo do nheengatu do século XX, possui pouco mais de 150.000 palavras. Publicações mais recentes, como o livreto de fábulas contadas por membros da Comunidade de Terra Preta (Bird; Gelbart; McAlister, 2013), com cerca de 2.500 palavras em nheengatu, não alcançam uma fração ínfima dessa quantidade.

Desse modo, o ideal seria incluir no UD_Nheengatu-CompLin todos os textos disponíveis. No entanto, como as ferramentas de processamento computacional do nheengatu ainda não estão suficientemente maduras ao ponto de automatizar de forma mais significativa a anotação do *corpus*, esse objetivo não é realista a um curto ou médio prazo. Um outro fator a ser levado em conta é que os textos de publicações mais recentes são protegidos pelo direito autoral, não podendo ser incorporados irrestritamente ao *treebank*, como é o caso das fábulas referidas, cuja reprodução necessita de autorização por escrito da comunidade responsável. Por outro lado, muitas publicações não estão em formato digital pesquisável, necessitando de transcrição manual, v.g., Costa (1909), enquanto a

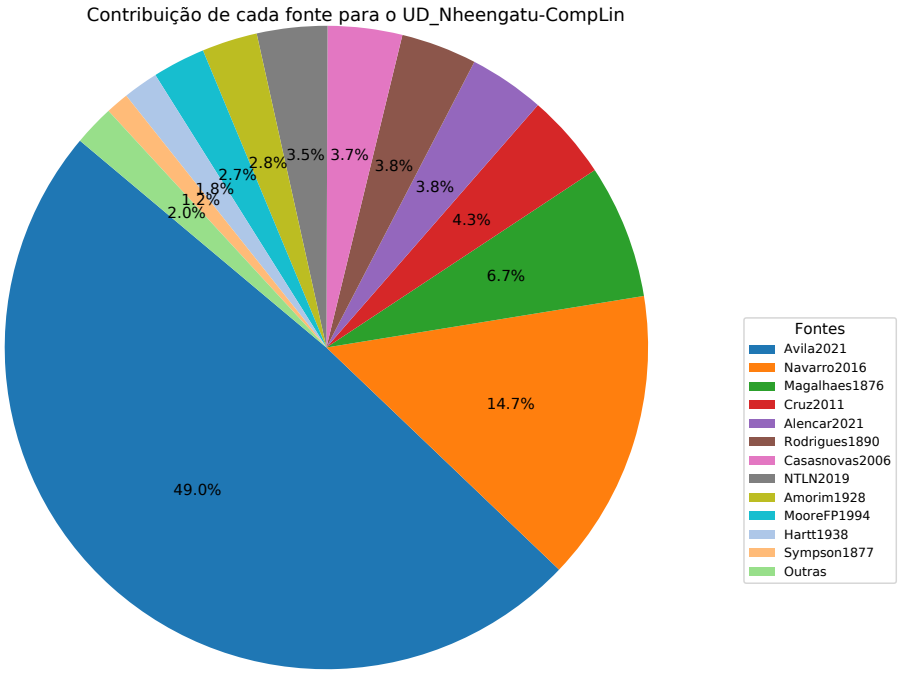


Figura 9. Frequência relativa das principais fontes bibliográficas do UD_Nheengatu-CompLin. A categoria *Outras* engloba fontes com menos de 16 sentenças no *treebank*.
Fonte: Elaboração própria.

maioria das que se encontram nesse formato demandam uma série de intervenções manuais na preparação do texto para anotação (Seção 3.3). Outra dificuldade para a construção de *corpora* anotados e a implementação de ferramentas de processamento é a gigantesca diversidade de ortografias do nheengatu.

Dadas essas limitações, dividimos a construção do UD_Nheengatu-CompLin em diversas etapas. Para a etapa objeto deste artigo, selecionamos primeiramente exemplos que permitissem fornecer uma visão geral da estrutura morfossintática do nheengatu sob a perspectiva do modelo UD, aproveitando as publicações mais acessíveis e atentando para as questões de direito autoral. Outro critério decisivo que permeou essas escolhas foi o grau de facilidade de anotação, determinado por fatores como a disponibilidade e o tipo de versão digital do texto, a qualidade das traduções e o detalhamento da análise lexical e/ou gramatical prévia, principalmente sob a forma de glosas interlineares. Com base nesses critérios, compilamos e anotamos o primeiro grupo de exemplos, extraídos de Moore, Facundes e Pires (1994), Cruz (2011), Navarro (2016), Alencar (2021) e Avila (2021). O segundo grupo de exemplos visou a antecipar os diversos problemas a serem enfrentados nas etapas seguintes. Para tanto, incorporamos progressivamente exemplos de variadas fontes, tanto sentenças individuais quanto passagens mais extensas ou textos inteiros, representativos do tesouro textual do nheengatu de meados do século XIX aos dias de hoje, norteando-nos principalmente pelo abrangente levantamento de fontes que embasam o dicionário de Avila (2021).

A Figura 10 sumaria alguns dos principais dados quantitativos que permitem dimensionar o tamanho atual do UD_Nheengatu-CompLin e compará-lo com os outros cinco maiores *treebanks* de línguas ameríndias na versão v.2.14 da coleção UD.[10] O gráfico evidencia que o UD_Nheengatu-CompLin supera significativamente os demais no número de palavras, consistindo, também, do maior número de sentenças. Quanto ao número de lemas e formas, o UD_Nheengatu-CompLin se aproxima dos *treebanks* que ocupam a primeira posição nessas duas dimensões.

A Figura 11 evidencia que o *treebank* possui uma distribuição diversificada de tipos de sentença em termos de extensão, incluindo tanto sentenças curtas como muitas bastante longas, com 355 sentenças na faixa de 12 a 20 palavras e 126 entre 21 e 49, perfazendo uma média de 10,23 palavras

[10] As Figuras 10, 34, 35 e 36 condensam dados gerados pelo *script* conllu-stats.pl, compilados no arquivo `stats.xml` do repositório de cada *treebank*.

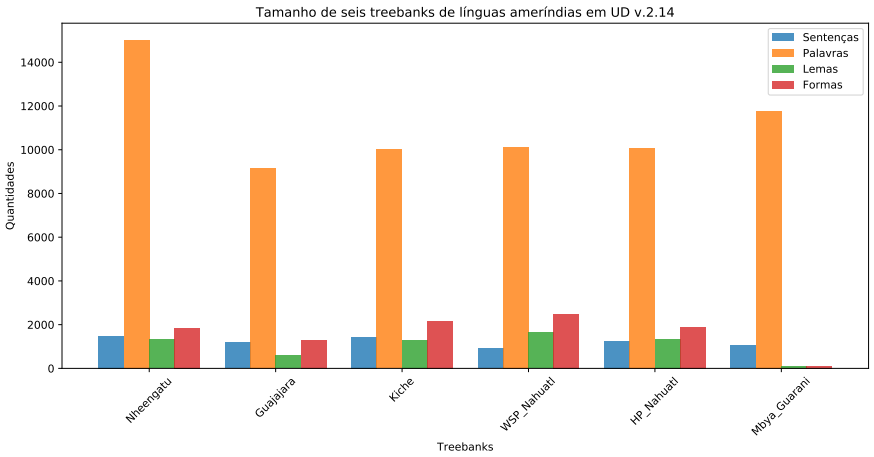


Figura 10. Dados quantitativos dos seis maiores *treebanks* de línguas ameríndias em UD v.2.14.
Fonte: Elaboração própria a partir dos dados de Zeman *et al.* (2024).

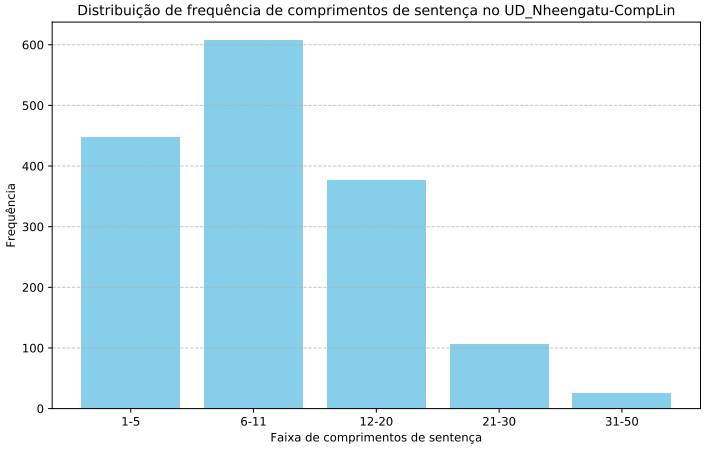


Figura 11. Comprimentos de sentença por quantidade de palavras no UD_Nheengatu-CompLin.
Fonte: Elaboração própria.

por sentença com desvio padrão de 6,93 e uma mediana de 8,0.[11]

Constatamos na Figura 9 que quase metade dos exemplos consiste de abonações dos verbetes de Avila (2021). Totalizando 15% do *treebank*, o segundo maior grupo de exemplos provém de Navarro (2016), um curso dividido em 13 lições, obedecendo a uma progressão gramatical que cobre uma ampla parcela da estrutura morfossintática da língua, exemplificada tanto por trechos ou textos completos adaptados da literatura, como Magalhães (1876), João Barbosa Rodrigues (1890), Stradelli (1929), Amorim (1928) e Cruz (2011), quanto por textos construídos. Um glossário de mais de 800 entradas indica as classes de palavra, propriedades flexionais e acepções do vocabulário desses textos, que podem ser extraídos de arquivo no formato PDF e preparados para anotação automática com poucas intervenções manuais.

Com cerca de 8.000 verbetes, dos quais aproximadamente 6.400 consistem em entradas principais e 1.600 em variantes, Avila (2021) é o mais abrangente dicionário do nheengatu de que se dispõe. Fundamenta-se num conjunto de publicações diatópica e diacronicamente representativo, estendendo-se sobretudo de meados do século XIX à segunda década de XXI, complementado por dados coletados junto a informantes na região do Alto Rio Negro. A microestrutura dos verbetes contempla uma ampla gama de informações, incluindo a etimologia, as marcas de uso, especificações morfossintáticas como classe de palavra e regência, as diferentes acepções e subacepções, os lemas tanto atuais quanto históricos e os registros na literatura antiga com as respectivas variantes ortográficas (Figura 12). Via de regra, não só as lexias simples e complexas, mas também afixos flexionais e derivacionais de todos os exemplos correspondem a lemas do dicionário. Também são consignadas várias centenas de derivados, compostos e locuções de diversas classes de palavras.

Ao todo, abonam os verbetes mais de 4.000 exemplos diferentes, a maioria provenientes da literatura, com apenas cerca de 10% da lavra do próprio dicionarista. As abonações obedecem a uma formatação sistemática, constituída de três partes, a saber, (i) sentença nheengatu, (ii) fonte e (iii) tradução portuguesa, o que facilita a geração automática de parte substantiva dos metadados da anotação. Como podemos constatar na Figura 13, alimentado com uma abonação de Avila (2021), o Yauti[12] automaticamente constrói os valores dos atributos `text`, `text_source` e `text_por`, do último dos quais se vale, via tradutor do Google™, para preencher o atributo `text_eng`.

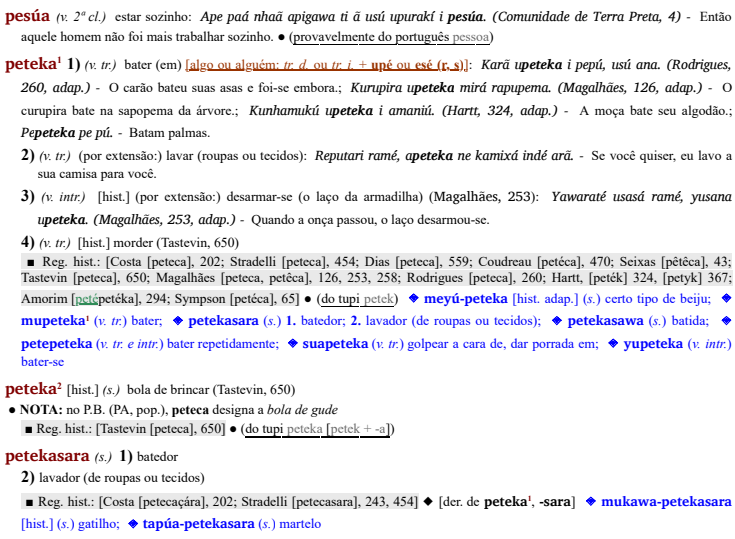
Outro ponto forte do dicionário é que transcende o domínio lexicográfico propriamente dito, uma vez que as informações gramaticais do corpo principal de algumas dezenas de verbetes são complementadas por explicações gramaticais sob a forma de notas, algumas delas bastante extensas. Diversos temas da gramática do nheengatu são aprofundados, tanto do ponto de vista sincrônico quanto diacrônico, em anexos ou no corpo principal da tese que ensejou o dicionário. Todas essas características fazem de Avila (2021) a publicação em nheengatu mais adequada para a rápida construção de um *treebank* conforme a teoria UD.

3.2 A normalização ortográfica

A normalização ortográfica é um procedimento padrão na linguística de *corpus* (Hirschmann, 2019). Num *treebank* do nheengatu, essa questão é crucial, haja vista a inexistência de uma ortografia uniforme. Talvez não seja exagero afirmar que cada publicação em nheengatu adota um padrão próprio, ainda assim, dentro de determinadas obras, deparamo-nos com significativa variação grafêmica (Avila, 2021; D'Angelis, 2023). A discrepância entre as grafias de uma mesma palavra chega a ser tão extrema ao ponto de não compartilharem caractere algum, como, por exemplo, *yã* relativamente a *nhan*, *nhaan* e *naa*, algumas das variantes de *nhaã* 'aquele'. Sem uma normalização ortográfica, seria impraticável a um usuário do *treebank* recuperar todas as ocorrências desse pronome demonstrativo. Uma complicação adicional é que as diferenças ortográficas não se limitam à representação de fonemas, estendendo-se à segmentação de palavras. Esse tipo de variação é pervasivo nos textos nheengatus, afetando diferentes morfemas, escritos ora separados, ora juntos da palavra anfitriã, intermediados ou não por hífen, como o relativizador *waá* (grafado também *ua*, *uaa*, *uá*, *uahá*, *wa etc.*)

[11] Valores computados por meio das funções correspondentes da biblioteca numpy de Python, aplicadas sobre uma lista com as quantidades de palavras sintáticas das 1.470 sentenças do *treebank*.

[12] Disponível no repositório https://github.com/CompLin/nheengatu.



pesúa *(v. 2ª cl.)* estar sozinho: *Ape paá nhaã apigawa ti ã usú upurakí i* ***pesúa****. (Comunidade de Terra Preta, 4)* - Então aquele homem não foi mais trabalhar sozinho. ● (provavelmente do português pessoa)

peteka[1] **1)** *(v. tr.)* bater (em) [algo ou alguém: *tr. d.* ou *tr. i.* + **upé** ou **esé (r, s)**]: *Karã* ***upeteka*** *i pepú, usú ana. (Rodrigues, 260, adap.)* - O carão bateu suas asas e foi-se embora.; *Kurupira* ***upeteka*** *mirá rapupema. (Magalhães, 126, adap.)* - O curupira bate na sapopema da árvore.; *Kunhamukú* ***upeteka*** *i amaniú. (Hartt, 324, adap.)* - A moça bate seu algodão.; *Pe****peteka*** *pe pú.* - Batam palmas.

2) *(v. tr.)* (por extensão:) lavar (roupas ou tecidos): *Reputari ramé,* ***apeteka*** *ne kamixá indé arã.* - Se você quiser, eu lavo a sua camisa para você.

3) *(v. intr.)* [hist.] (por extensão:) desarmar-se (o laço da armadilha) (Magalhães, 253): *Yawaraté usasá ramé, yusana* ***upeteka****. (Magalhães, 253, adap.)* - Quando a onça passou, o laço desarmou-se.

4) *(v. tr.)* [hist.] morder (Tastevin, 650)

■ Reg. hist.: [Costa [peteca], 202; Stradelli [peteca], 454; Dias [peteca], 559; Coudreau [petéca], 470; Seixas [pêtêca], 43; Tastevin [peteca], 650; Magalhães [peteca, petêca], 126, 253, 258; Rodrigues [peteca], 260; Hartt, [peték] 324, [petyk] 367; Amorim [petépetéka], 294; Sympson [petéca], 65] ● (do tupi petek) ◆ **meyú-peteka** [hist. adap.] *(s.)* certo tipo de beiju; ◆ **mupeteka**[1] *(v. tr.)* bater; ◆ **petekasara** *(s.)* **1.** batedor; **2.** lavador (de roupas ou tecidos); ◆ **petekasawa** *(s.)* batida; ◆ **petepeteka** *(v. tr. e intr.)* bater repetidamente; ◆ **suapeteka** *(v. tr.)* golpear a cara de, dar porrada em; ◆ **yupeteka** *(v. intr.)* bater-se

peteka[2] [hist.] *(s.)* bola de brincar (Tastevin, 650)

● **NOTA:** no P.B. (PA, pop.), **peteca** designa a *bola de gude*

■ Reg. hist.: [Tastevin [peteca], 650] ● (do tupi peteka [petek + -a])

petekasara *(s.)* **1)** batedor

2) lavador (de roupas ou tecidos)

■ Reg. hist.: [Costa [petecaçára], 202; Stradelli [petecasara], 243, 454] ◆ [der. de **peteka**[1], **-sara**] ◆ **mukawa-petekasara** [hist.] *(s.)* gatilho; ◆ **tapúa-petekasara** *(s.)* martelo

Figura 12. Excerto de Avila (2021, p. 592).
Fonte: Elaboração própria por meio de recorte de página em PDF.

```
>>> import Yauti
>>> s='''Karã upeteka i/pron2 pepú, usú ana. (Rodrigues, 260, adap.) -  O carão bateu suas asas e foi-se embora.'''
>>> AnnotateConllu.parseExample(s,'Avila2021',0,0,721,annotator='XXXX')
# sent_id = Avila2021:0:0:721
# text = Karã upeteka i pepú, usú ana.
# text_eng = The big guy flapped his wings and went away.
# text_por = O carão bateu suas asas e foi-se embora.
# text_source = Rodrigues, 260, adap.
# text_annotator = XXXX
1       Karã    karã    NOUN    N       Number=Sing     2       nsubj   _       TokenRange=0:4
2       upeteka peteka  VERB    V       Person=3|VerbForm=Fin   0       root    _       TokenRange=5:12
3       i       i       PRON    PRON2   Case=Gen|Number=Sing|Person=3|Poss=Yes|PronType=Prs     4       nmod:poss       _       TokenRange=13:14
4       pepú    pepú    NOUN    N       Number=Sing     2       obj     _       SpaceAfter=No|TokenRange=15:19
5       ,       ,       PUNCT   PUNCT   _       6       punct   _       TokenRange=19:20
6       usú     sú      VERB    V       Person=3|VerbForm=Fin   2       parataxis       _       TokenRange=21:24
7       ana     ana     PART    PFV     Aspect=Perf     6       advmod  _       SpaceAfter=No|TokenRange=25:28
8       .       .       PUNCT   PUNCT   _       2       punct   _       SpaceAfter=No|TokenRange=28:29
```

Figura 13. Geração de metadados e análise morfossintática com o Yauti a partir de exemplo do verbete **peteka**[1] da Figura 12.
Fonte: Elaboração própria.

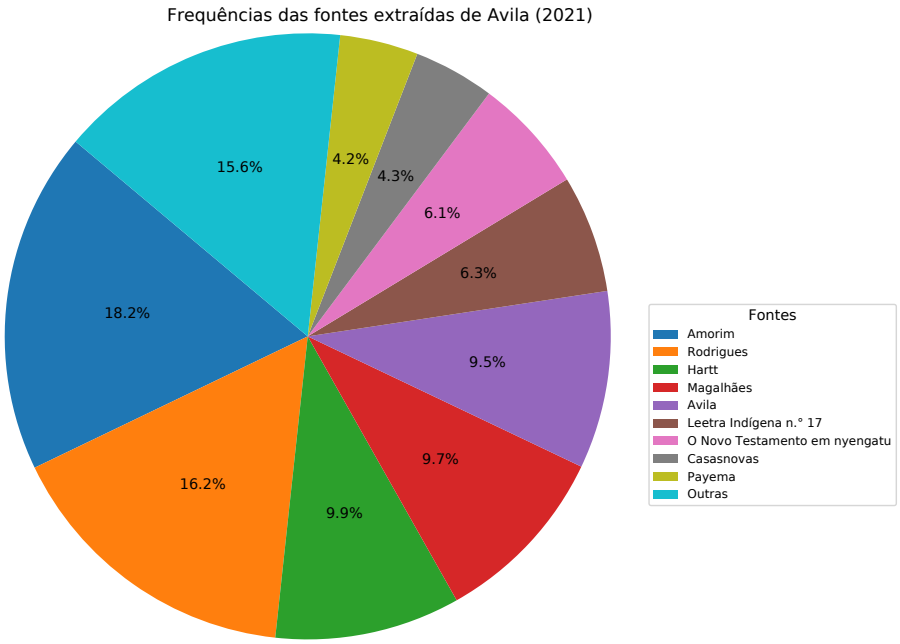


Figura 14. Frequência relativa das fontes bibliográficas de Avila (2021) utilizadas no UD_Nheengatu-CompLin. A categoria *Outras* engloba fontes com menos de 30 sentenças no *treebank*.

Fonte: Elaboração própria.

No UD_Nheengatu-CompLin, `text` assinala a versão normalizada do texto original, preservado *ipsis litteris* como valor do atributo `text_orig` (Figuras 5 e 7). Em exemplos de Avila (2021) provenientes da literatura nheengatu, o atributo `text_prim` reproduz o texto primário (Figuras 6 e 23). A exemplo do anotador morfossintático Yauti (Alencar, 2023), adotamos na normalização ortográfica a proposta de Avila (2021), dada a abrangência diacrônica e sincrônica desse dicionário. Cumpre salientar, no entanto, que essa escolha não representa, absolutamente, um juízo de valor de que essa ortografia é melhor do que as outras. De fato, Avila (2021) propõe esse padrão com a única finalidade de facilitar a consulta ao dicionário. Trata-se de estratégia análoga à ortografia de trabalho definida por D'Angelis, Oliveira e Schwade (2021).

3.3 O processo de anotação e revisão

Nesta subseção, tratamos das diferentes tarefas envolvidas na inclusão de uma nova sentença no UD_Nheengatu-CompLin, as quais se articulam no fluxograma da Figura 15.

Nosso objetivo de longo prazo é abarcar com o *treebank* todo o material textual nheengatu em domínio público, além de textos de cujos proprietários venhamos a obter permissão. Visando a metas de curto prazo de expansão do *treebank* relacionadas às *releases* periódicas da coleção UD, distinguimos duas situações típicas que se deparam no primeiro passo do fluxograma da Figura 15.

A primeira situação configura-se quando se trata de anotar todas as sentenças de uma dada publicação ou de uma ou mais narrativas ou trechos contínuos. Nesse caso, anotamos os exemplos, via de regra, na sequência em que aparecem na respectiva fonte. A segunda situação ocorre quando o objetivo é anotar sentenças com uma determinada palavra ou construção. Por exemplo, as orações relativas constituem uma das construções fundamentais do nheengatu. Desse modo, um *parser* capaz de analisar corretamente os diferentes tipos de estruturas com o relativizador *waá* precisa ser treinado num *treebank* com um grande número de variados exemplos com esse elemento. Para tanto, extraímos sentenças com *waá* de diferentes exposições gramaticais, como as de Seixas (1853), Magalhães (1876), Sympson (1877) e Casasnovas (2006), para somarem-se às sentenças com esse subordinador incluídas pela primeira via.

O Yauti (Alencar, 2023) foi utilizado desde o início para anotar as sentenças do UD_Nheengatu-CompLin e teve seu desenvolvimento guiado pelas necessidades da anotação, crescendo em abrangência paralelamente ao *treebank*. Em ambas as situações delineadas acima, comumente nos deparamos

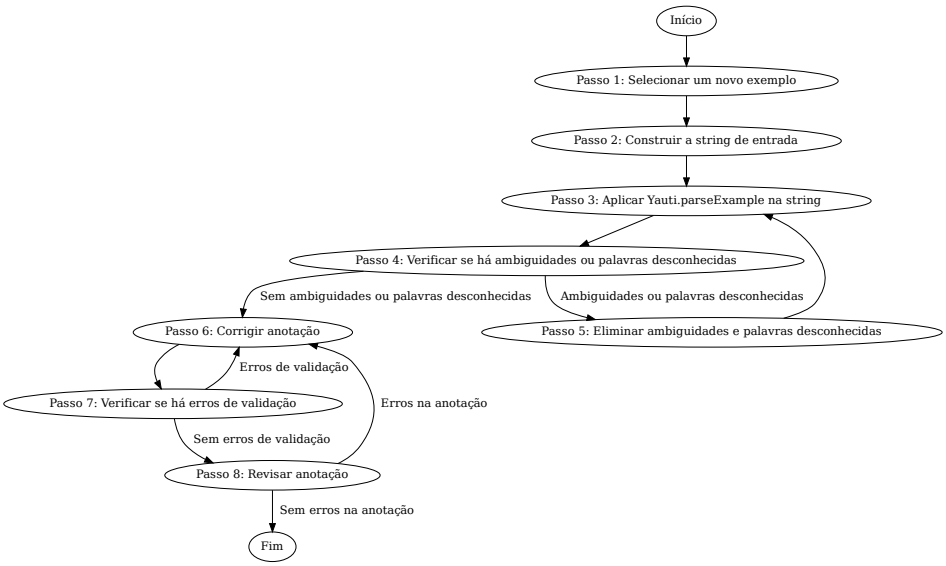


Figura 15. Fluxograma da inclusão de uma nova sentença anotada no UD_Nheengatu-CompLin.
Fonte: Elaboração própria.

com palavras ainda não codificadas no léxico da ferramenta e fenômenos gramaticais ainda não implementados cuja análise sob a perspectiva de UD e codificação em Python implicaram anotar um variado leque de exemplos característicos, de modo também a testar a respectiva implementação. Como o léxico do Yauti ainda é limitado (Alencar, 2023), ao incluir uma nova entrada no glossário da ferramenta para poder analisar um dado exemplo, aproveitamos para anotar o máximo de novos exemplos de Avila (2021) com essa palavra, exemplos esses que, com frequência, continham palavras ainda não implementadas, resultando na anotação de mais sentenças, num efeito em cadeia.

A seguir, explicamos os passos de 2 a 8 da Figura 15 com base num exemplo concreto. No passo 2, preparamos a *string* de Python a ser dada como primeiro argumento da função `parseExample` do Yauti, conforme exemplificamos anteriormente na Figura 13 com abonação de Avila (2021). Em exemplos de Magalhães (1876) e Sympson (1877), como de publicações análogas, atualizamos a ortografia da tradução em português e adaptamos o texto original à proposta ortográfica de Avila (2021), incluindo-o como último componente da referida *string* (Figura 16).

```
>>> import Yauti
>>> s='''— Kurupira upeteka mirá rapupema. (p. 126) — É o Curupira que está batendo nas sapopemas. -
Curupíra opeteca (otucá) mirá rapupêma.'''
>>> Yauti.parseExample(s,'Magalhaes1876','2-19',4,'104-1',annotator='XXXX')
# sent_id = Magalhaes1876:2-19:4:104-1
# text = — Kurupira upeteka mirá rapupema.
# text_eng = "It's Curupira who's hitting the sapopemas."
# text_por = — É o Curupira que está batendo nas sapopemas.
# text_source = p. 126
# text_orig = Curupíra opeteca (otucá) mirá rapupêma.
# text_annotator = XXXX
1       —       —       PUNCT   PUNCT   _       3       punct   _       TokenRange=-1:0
2       Kurupira        kurupira        NOUN    N       Number=Sing     3       nsubj   _       Token
Range=1:9
3       upeteka peteka  VERB    V       Person=3|VerbForm=Fin   0       root    _       TokenRange=10
:17
4       mirá    mirá    NOUN    N       Number=Sing     3       obj     _       TokenRange=18:22
5       rapupema        rapupema        _       _       _       3       _       _       SpaceAfter=No
|TokenRange=23:31
6       .       .       PUNCT   PUNCT   _       3       punct   _       SpaceAfter=No|TokenRange=31:3
2
```

Figura 16. Preparação de exemplo de Magalhães (1876, p. 126) e primeira aplicação da função `parseExample` do Yauti.
Fonte: Elaboração própria.

Nos passos 3 e 4, executamos a função `parseExample`, constatando que o Yauti não reconheceu o quinto *token* da sentença (Figura 16). Isso ocorreu porque o seu léxico ainda não incluía essa palavra, consignada em Avila (2021). Realizada a atualização da ferramenta, conforme o passo 5, reexecutamos a função `parseExample`, gerando a análise da Figura 18. No passo 6, analisamos essa

anotação, recorrendo, geralmente, a uma ferramenta de visualização de árvores no formato CoNLL-U (Figura 19). Nesse exemplo, apenas a tradução automática do português para o inglês necessitou correção. O passo 7 consiste na aplicação do *script* validate.py, ferramenta do projeto UD que verifica se uma análise no formato CoNLL-U obedece a uma série de requisitos do modelo (Marneffe *et al.*, 2024b), como exemplificado na Figura 17 com duas outras sentenças. Como o *script* considera válida a análise da Figura 18, concluímos o ciclo com o passo 8, que consiste na revisão da anotação por um outro anotador. Divergindo anotador inicial e revisor, procedemos à adjudicação das discrepâncias (Hirschmann, 2019).[13]

```
python3 validate.py --lang=yrl --max-err=0 muiratiwa.conllu
[Line 17667 Sent Avila2021:0:0:719]: [L2 Metadata non-unique-sent-id] Non-unique sent_id attribute 'Avila2021:0:0:719'.
[Line 17663 Sent Avila2021:0:0:719 Node 16]: [L3 Syntax rel-upos-advmod] 'advmod' should be 'ADV' but it is 'PRON'
[Line 26614 Sent Aguiar1898:21:300:300]: [L1 Format missing-empty-line] Missing empty line after the last sentence.
Format errors: 1
Metadata errors: 1
Syntax errors: 1
*** FAILED *** with 3 errors
```

Figura 17. Detecção de erros de validação por meio do *script* validate.py.
Fonte: Elaboração própria.

```
>>> Yauti.parseExample(s,'Magalhaes1876','2-19',4,'104-1',annotator='XXXX')
# sent_id = Magalhaes1876:2-19:4:104-1
# text = — Kurupira upeteka mirá rapupema.
# text_eng = "It's Curupira who's hitting the sapopemas."
# text_por = — É o Curupira que está batendo nas sapopemas.
# text_source = p. 126
# text_orig = Curupíra opeteca (otucá) mirá rapupêma.
# text_annotator = XXXX
1       —       —       PUNCT   PUNCT   _       3       punct   _       TokenRange=-1:0
2       Kurupira        kurupira        NOUN    N       Number=Sing     3       nsubj   _       TokenRange=1:9
3       upeteka peteka  VERB    V       Person=3|VerbForm=Fin   0       root    _       TokenRange=10:17
4       mirá    mirá    NOUN    N       Number=Sing     5       nmod:poss       _       TokenRange=18:22
5       rapupema        sapupema        NOUN    N       Number=Sing|Number[psor]=Sing|Person[psor]=3|Rel=NCont
obj     _       SpaceAfter=No|TokenRange=23:31
6       .       .       PUNCT   PUNCT   _       3       punct   _       SpaceAfter=No|TokenRange=31:32
```

Figura 18. Reaplicação da função `parseExample` sobre o exemplo de Magalhães (1876, p. 126) após atualização do léxico da ferramenta.
Fonte: Elaboração própria.

```
# sent_id = Magalhaes1876:2-19:4:104-1
# text = — Kurupira upeteka mirá rapupema.
# text_eng = "It's Curupira who's who's hitting the buttress roots."
# text_eng_ggl = "It's Curupira who's who's beating the sapopemas."
# text_por = — É o Curupira que está batendo nas sapopemas.
# text_source = p. 126
# text_orig = Curupíra opeteca (otucá) mirá rapupêma.
# text_annotator = XXXX
```

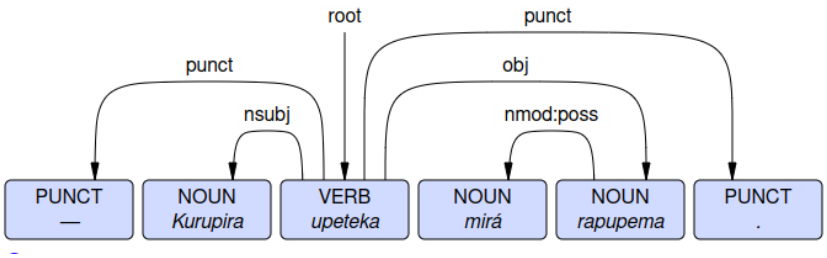


Figura 19. Visualização da análise da Figura 18 com correção manual da tradução do português para o inglês realizada pelo tradutor do Google™.
Fonte: Elaboração própria.

A Figura 20 exemplifica um outro tipo de situação com que o anotador se depara nos passos 4 e 5. Processos produtivos de formação de palavras, abundantes em nheengatu (Sympson, 1877; Stradelli, 1929; Cruz, 2011; Navarro, 2016; Avila, 2021), propiciam a criação de novos lexemas a qualquer momento. Entre esses mecanismos, destacam-se a reduplicação e a sufixação. Enquanto várias dessas formações se lexicalizaram e estão consignadas em Avila (2021), como *purapuranga*

[13] Alencar (2024) pormenoriza a composição da equipe de anotadores e revisores.

'muito bonito', reduplicação parcial de *puranga* 'bonito', e *sesaíma* 'cego', derivada de *sesá* 'olho' por meio do sufixo privativo *ima* 'sem', encontramos nos textos outras derivações ainda não lexicalizadas, como *paya-ima* 'sem pai' (Figura 20).[14] Isso significa que apenas incorporar o inventário integral de lexemas de Avila (2021) não permitiria a um analisador morfossintático automático reconhecer qualquer formação desse tipo. Para lidar com esses casos, o Yauti adota duas estratégias. Em casos completamente previsíveis como a reduplicação total, a ferramenta automaticamente identifica o processo derivacional, preenchendo o campo LEMMA com a base e o campo FEATS com o traço `Red=Yes` (Figura 21). Noutros casos, exemplificado na Figura 20, o anotador precisa inserir uma etiqueta especial prefixada por `/=` indicando o tipo de formação e a classe de palavra resultante, entre outras informações (Alencar, 2023).

```
>>> s='''Taína i/pron2 paya umanú, i/pron2 paya-ima/=prv:x|a. (p. 8) A criança cujo pai morreu é órfã.
- Taý-na i páia hu-manú, i páia ima.'''
>>> Yauti.parseExample(s,'Sympson1877',0,0,90,annotator='XXXX')
# sent_id = Sympson1877:0:0:90
# text = Taína i paya umanú, i paya-ima.
# text_eng = A child whose father has died is an orphan.
# text_por = A criança cujo pai morreu é órfã.
# text_source = p. 8
# text_orig = Taý-na i páia hu-manú, i páia ima.
# text_annotator = XXXX
1       Taína   taína   NOUN    N       Number=Sing     4       nsubj   _       TokenRange=0:5
2       i       i       PRON    PRON2   Case=Gen|Number=Sing|Person=3|Poss=Yes|PronType=Prs     3
nmod:poss       _       TokenRange=6:7
3       paya    paya    NOUN    N       Number=Sing     4       nsubj   _       TokenRange=8:12
4       umanú   manú    VERB    V       Person=3|VerbForm=Fin   0       root    _       SpaceAfter=No|
TokenRange=13:18
5       ,       ,       PUNCT   PUNCT   _       4       punct   _       TokenRange=18:19
6       i       i       PRON    PRON2   Case=Gen|Number=Sing|Person=3|PronType=Prs      4       _
_       TokenRange=20:21
7       paya-ima        paya    ADJ     A       Derivation=Priv 4       _       _       SpaceAfter=No|
TokenRange=22:30
8       .       .       PUNCT   PUNCT   _       4       punct   _       SpaceAfter=No|TokenRange=30:31
```

Figura 20. Preparação de exemplo de Sympson (1877, p. 8) e primeira aplicação da função `parseExample` do Yauti.
Fonte: Elaboração própria.

```
>>> s='''Utuká-tuká ukena. (Example 576 Br) Bateu na porta (repetidamente). - u-tuka-tuka ukena'''
>>> Yauti.parseExample(s,'Cruz2011',0,0,63,annotator='XXXX')
# sent_id = Cruz2011:0:0:63
# text = Utuká-tuká ukena.
# text_eng = He knocked on the door (repeatedly).
# text_por = Bateu na porta (repetidamente).
# text_source = Example 576 Br
# text_orig = u-tuka-tuka ukena
# text_annotator = XXXX
1       Utuká-tuká      tuká    VERB    V       Person=3|Red=Yes|VerbForm=Fin   0       root    _
0:10
2       ukena   ukena   NOUN    N       Number=Sing|Rel=Abs     1       obj     _       SpaceAfter=No|T
:16
3       .       .       PUNCT   PUNCT   _       1       punct   _       SpaceAfter=No|TokenRange=16:17
```

Figura 21. Análise automática da reduplicação total pelo Yauti.
Fonte: Elaboração própria.

Divergimos de Avila (2021) em alguns aspectos relacionados à adaptação dos exemplos antigos, subtarefa do passo 2, adotando uma perspectiva mais conservadora, de modo a possibilitar investigações dialetológicas ou diacrônicas, preservando, via de regra, a pontuação[15] e escolhendo, para cada palavra, a variante histórica mais próxima da forma original, sem inserir nem suprimir palavras, mesmo quando justificável sob a ótica do nheengatu rio-negrino atual.

Comparem-se, na Figura 22, o texto original `text_orig` de João Barbosa Rodrigues (1890), sem o relativizador *waá*, com o texto secundário `text_sec` de Avila (2021) com esse subordinador. O atributo `cross_reference` remete ao exemplo do UD_Nheengatu-CompLin com a versão de Avila (2021), que também lematiza *muirá* 'árvore' como *mirá* e separa com uma vírgula a última oração da sentença. Pelo contrário, limitamos nossas intervenções ao plano ortográfico, mantendo a forma *muirá*, consignada em Avila (2021) como variante histórica.

Assinalamos determinadas idiossincrasias dos textos históricos com os traços `Style=Arch` e `Style=Rare`, indicando, no campo MISC, a forma canônica, como na Figura 23, na esteira de alguns *tre-*

[14] Em Stradelli (1929, p. 276) constam "paiay[m]a" e "maiayma" como acepções de *órfão de pai* e *órfão de mãe*.
[15] Maiusculizamos a primeira letra de sentenças e inserimos a pontuação final faltante, entre outras intervenções mínimas em casos de lapsos ou erros tipográficos óbvios.

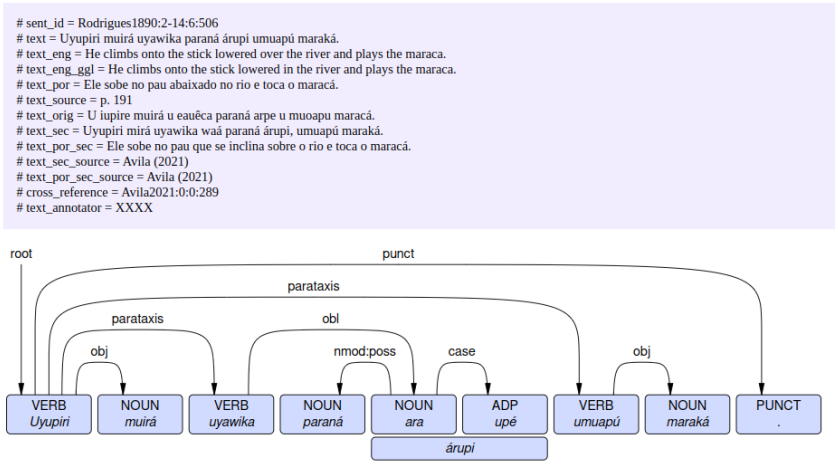


Figura 22. Comparação entre a adaptação textual do UD_Nheengatu-CompLin e de Avila (2021).
Fonte: Elaboração própria.

ebanks da coleção UD. Recorremos a essa estratégia, por exemplo, na anotação de sentenças com pronomes de segunda classe na função de objeto de um verbo sem marca de flexão, fenômeno que Avila (2021) preserva nas suas abonações (Figura 23). No momento, o Yauti analisa esses pronomes, nessa configuração, erroneamente como sujeitos do verbo subsequente. Outras particularidades gramaticais do nheengatu do século XIX registradas nos exemplos de Avila (2021), como o prefixo *e* de imperativo de segunda pessoa do singular e o dativo dos pronomes pessoais, são analisadas automaticamente pelo Yauti. Esses últimos fenômenos não estão marcados no UD_Nheengatu-CompLin com `Style=Arch`, decisão que talvez alteremos no futuro.

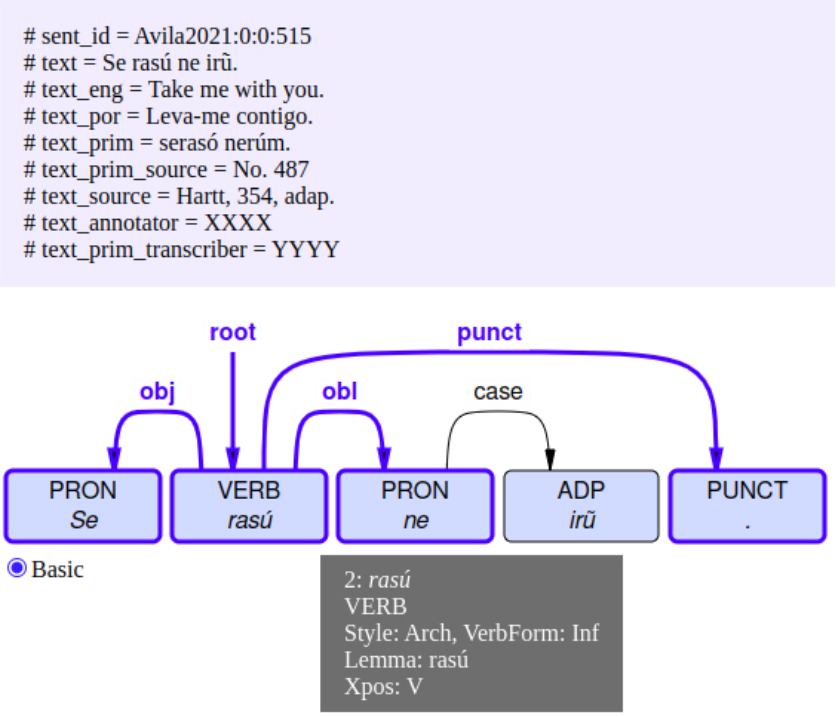


Figura 23. Exemplo com o traço `Style=Arch`.
Fonte: Elaboração própria.

Também discrepamos de Avila (2021) ao seguirmos Cruz (2011) na análise de formas como *taumaã* 'vêem' e *tapurasí* 'dançam' ao invés de desmembrá-las em *ta umaã* e *ta upurasí*, onde *ta* constitui o pronome pessoal de terceira pessoa do plural. O Yauti trata *tokens* desse tipo como formas do

paradigma conjugacional, sem segmentá-los. No entanto, quando se depara com sequências do tipo de *ta umaã*, preserva essa característica do texto, não realizando a junção do pronome e da forma verbal.

4 Aspectos da anotação

Esta seção aborda as principais decisões que nortearam a elaboração das representações no formato CoNLL-U do UD_Nheengatu-CompLin, fornecendo também dados quantitativos nesse domínio. Começamos com a delimitação de sentenças e palavras, para em seguida tratar do preenchimento dos diferentes campos da tabela exemplificada na Figura 4. Finalmente, apresentamos os resultados do UD_Nheengatu-CompLin para as principais métricas de avaliação do projeto UD.

4.1 Segmentação sentencial e vocabular

Antes de proceder à anotação de textos conforme o modelo UD, cumpre determinar as unidades a serem anotadas. Na versão atual do UD_Nheengatu-CompLin, dois tipos de unidades são delimitadas: a sentença e a palavra sintática. O formato CoNLL-U contempla também a segmentação dos textos em parágrafos, unidade que deixamos para incluir numa etapa futura.

Por um lado, a marcação de parágrafos não é relevante para a anotação de grande parte dos materiais linguísticos do nheengatu, especialmente aqueles provenientes de exposições gramaticais, vocabulários e dicionários, como, por exemplo, Sympson (1877) e Cruz (2011). Por outro lado, mesmo em narrativas comumente não se dividem os textos em parágrafos, como no caso das lendas coligidas por Amorim (1928) (Figuras 24 e 25) e Casasnovas (2006). Magalhães (1876), por sua vez, adota uma paragrafação irregular. Na lenda *Como a noite apareceu*, constatamos indentações características de parágrafos, porém, essas marcas tipográficas parecem arbitrárias, não correspondendo aos parágrafos delimitados na tradução portuguesa. Enquanto na relativamente extensa lenda *Jabuti e anta do mato* uma única indentação assinala o início do texto, na lenda *O jabuti e a onça*, de menos de duas páginas, cada sentença é marcada por uma indentação.

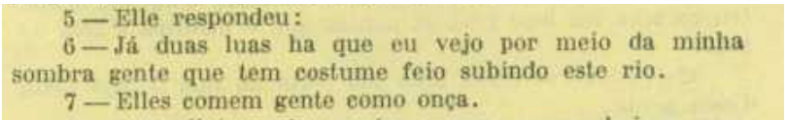

5 — Elle respondeu:
6 — Já duas luas ha que eu vejo por meio da minha sombra gente que tem costume feio subindo este rio.
7 — Elles comem gente como onça.

Figura 24. Tradução de Amorim (1928, p. 301) para o trecho da Figura 25.
Fonte: Elaboração própria de recorte de página em PDF.

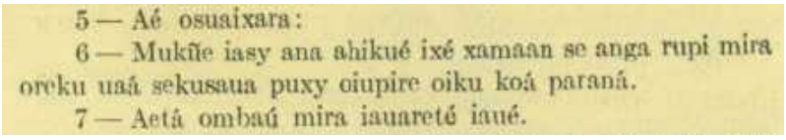

5 — Aé osuaixara:
6 — Mukũe iasy ana ahikué ixé xamaan se anga rupi mira oreku uaá sekusaua puxy oiupire oiku koá paraná.
7 — Aetá ombaú mira iauareté iaué.

Figura 25. Trecho da lenda *Kukuhi* (Amorim, 1928, p. 309).
Fonte: Elaboração própria de recorte de página em PDF.

Via de regra, consideramos sentenças individuais, nos textos em nheengatu, tanto as divisões de parágrafo delimitadas por pontuação final quanto segmentos numerados como o de número 7 da Figura 25. Há casos, contudo, que discrepam dessa situação. Em narrativas, consideramos como partes de uma mesma sentença tanto a fala de personagem quanto o discurso, terminado por dois pontos, que a introduz, ainda que essas partes ocorram em parágrafos ou divisões numeradas distintas, como na Figura 25. No UD_Nheengatu-CompLin, os segmentos cinco e seis constituem uma única sentença, a exemplo da sentença da Figura 6.

A adoção desse critério de segmentação sentencial não é unânime nos *treebanks* da coleção UD. Isso reflete-se na análise sintática da tradução em português, alemão ou inglês do exemplo da Figura 25 por meio do analisador sintático automático UDPipe 2. Quando se aplicam nessas traduções os modelos baseados nos *treebanks* UD_Portuguese-Bosque, UD_German-GSD e UD_English-Atis, todo

o texto é tratado como uma única sentença, ao contrário do que ocorre no caso dos modelos treinados nos treebanks UD_Portuguese-CINTIL, UD_German-HDT e UD_English-ParTUT, que dividem o texto em duas sentenças.

Em sentenças do tipo da Figura 25, a segmentação (toquenização) vocabular constitui uma operação trivial: basta destacar das palavras os sinais de pontuação que as acompanham. Em diversas situações, porém, ocorre um descompasso entre as noções de palavra ortográfica, isto é, delimitada por espaço em branco ou sinal de pontuação, e de palavra sintática. Em nheengatu, esse é o caso, por exemplo, dos auxiliares incorporados (Figura 26) e de clíticos como o alomorfe monossilábico átono da partícula *taá* de interrogativas parciais, o advérbio *ntu* 'somente' ou os alomorfes *pe* e *me* da adposição inessiva *upé* 'em' (Figuras 27 e 28).

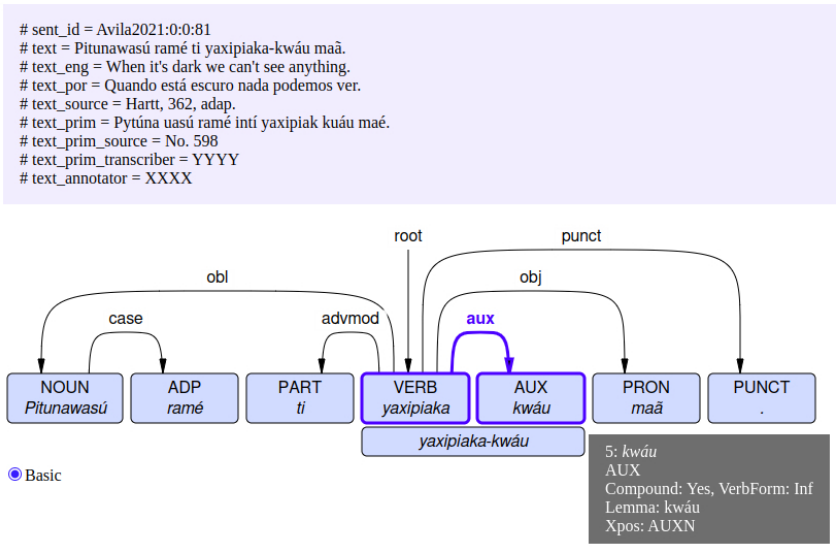


Figura 26. Anotação de auxiliar incorporado.
Fonte: Elaboração própria.

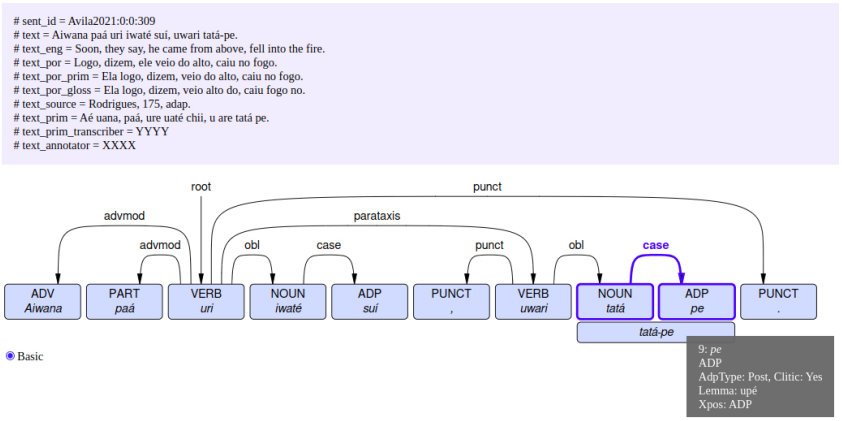


Figura 27. Anotação do alomorfe oral da posposição inessiva enclítica.
Fonte: Elaboração própria.

A análise da Figura 27 reflete a abordagem de Avila (2021, p. 588), que considera *pe* um alomorfe de *upé* 'em', constituindo um clítico com verbete próprio de lema *=pe*.[16] Ao nosso ver, essa análise estende-se naturalmente, no quadro de UD, ao exemplo da Figura 28. Avila (2021, p. 580), pelo contrário, classifica *paraname* como "substantivo locativo", equivalente a *paranã upé* e traduzindo-se como "no rio", constituindo a "forma locativa" de *paraná* 'rio', lema com variante *paranã*. No dicionário, *paraná* e *paraname* encabeçam verbetes principais, ao passo que *paranã* constitui verbete que

[16] O sinal de igualdade indica a natureza clítica do morfema.

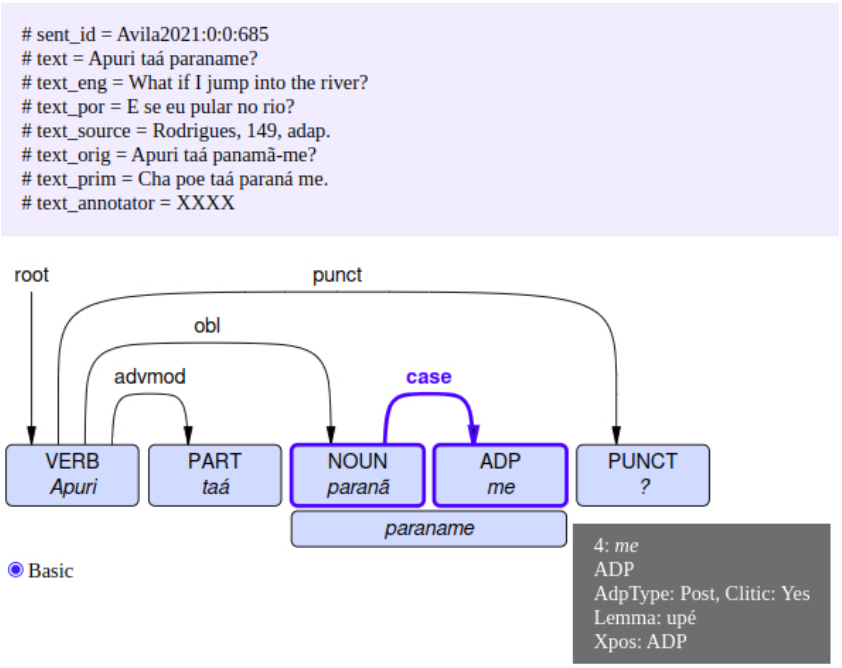


Figura 28. Anotação do alomorfe nasal da posposição inessiva enclítica.
Fonte: Elaboração própria.

meramente remete a *paraná*. Formas locativas análogas constituem verbetes de Avila (2021), como *gantime* 'na proa', igualmente classificadas como substantivos locativos, classe que também engloba diversos lemas terminados em *upi*, como *árupi*, *wírupi* e *pitérupi*, formas locativas, respectivamente, de *ara* 'cima', *wira* 'parte inferior' e *pitera* 'meio'. Navarro (2016) classifica essas últimas formas como posposições, análise que inicialmente abraçamos, antes de adotarmos um tratamento consistente para todas as formas que Avila (2021) considera substantivos locativos.

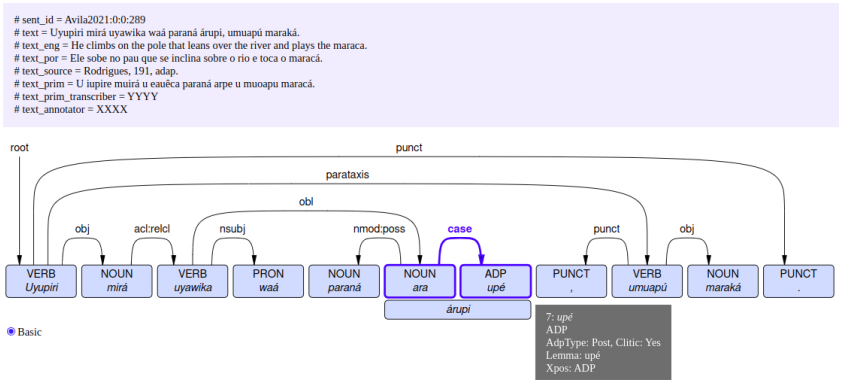


Figura 29. Exemplo de anotação substantivo locativo.
Fonte: Elaboração própria.

No modelo UD, palavras ortográficas que abrangem mais de uma palavra sintática, como nas Figuras 26, 27, 28 e 29, denominam-se palavras fusionadas, cujos componentes se ligam por hífen, conforme a ortografia de Avila (2021), apenas num subconjunto dos casos; comparem-se as Figuras 27 e 28. O Yauti identifica automaticamente os diferentes tipos de palavras fusionadas, segmentando e anotando os dois componentes, o segundo dos quais recebe o traço `Compound=Yes` ou `Clitic=Yes`, conforme se trata de composição (Figura 26) ou cliticização. Na anotação no formato CoNLL-U, uma palavra fusionada é assinalada na coluna 1 por um índice sob a forma de intervalo numérico `n-m`, em que `n` e `m` designam os índices da primeira e da última palavra sintática abrangida. Todos os campos da palavra fusionada são preenchidos por `-`, exceto o campo MISC. A anotação de uma

palavra fusionada precede as anotações das palavras sintáticas que a integram.

4.2 Lematização

Em línguas com morfologia flexional e derivacional como o nheengatu, o valor do campo LEMMA não coincide necessariamente com o campo FORM. Por exemplo, quando se trata de um substantivo uniforme no plural, ou seja, sufixado com *-itá* (Figura 30), ou um verbo da série ativa conjugado (Figura 4), o campo LEMMA abriga a forma de citação de dicionário, que, conforme Avila (2021), consiste no radical nominal e verbal, respectivamente, formas não marcadas de singular e infinitivo.

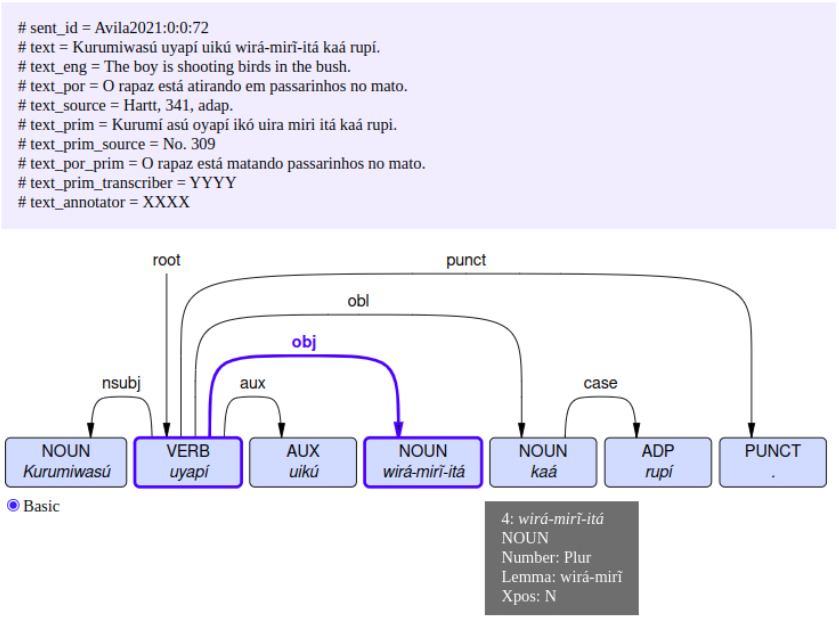


Figura 30. Lematização de substantivo uniforme no plural.
Fonte: Elaboração própria.

O prefixo *yu-* de voz médio-passiva e os sufixos derivacionais modificadores, como os aspectuais, avaliativos, privativos ou de formação de coletivos, também são eliminados no processo de lematização. Nesses casos, as diferentes informações expressas por esses afixos são codificadas sob a forma de traços no campo FEATS (Figura 31). De modo análogo procedemos com formas reduplicadas (Figura 32).

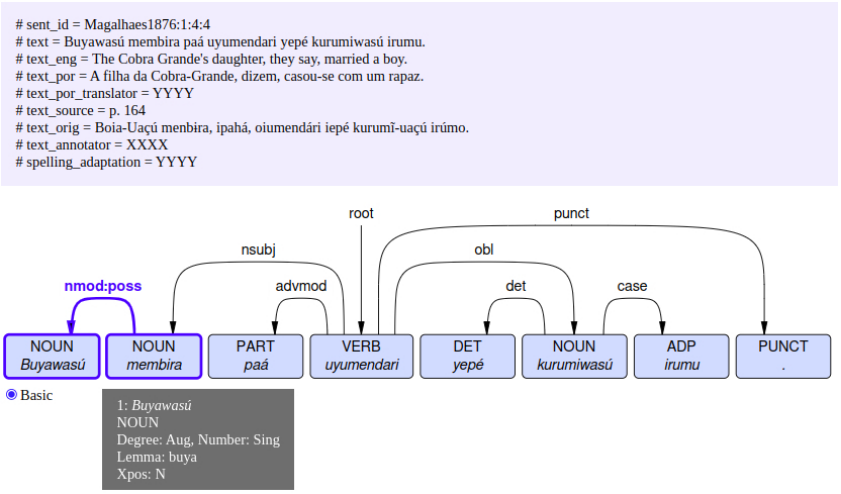


Figura 31. Lematização de substantivo com sufixo aumentativo *-wasú*.
Fonte: Elaboração própria.

O lema de formas lexicalizadas não composicionais, porém, preserva o afixo derivacional; comparem-se as formas *buyawasú* 'cobra grande' e *kurumiwasú* 'rapaz', que lematizamos como *buya* e

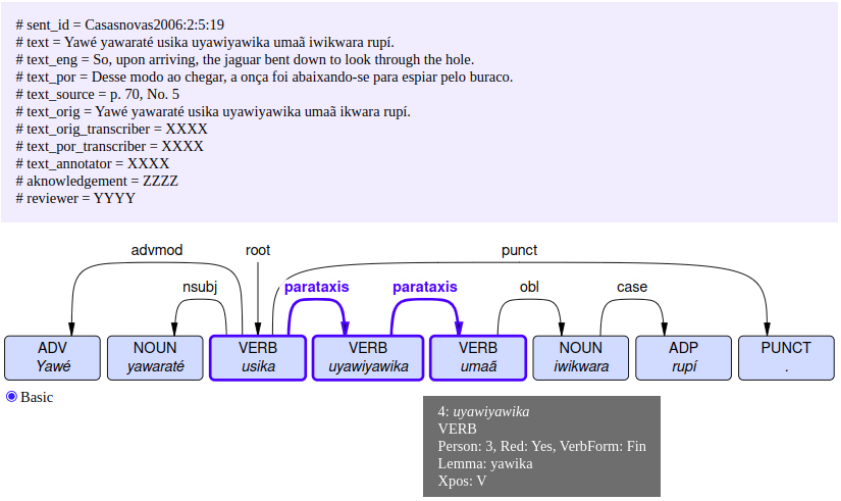


Figura 32. Lematização de verbo reduplicado.
Fonte: Elaboração própria.

kurumiwasú, na esteira de Avila (2021). No UD_Nheengatu-CompLin, porém, nem sempre incluímos no campo LEMMA o lema da forma derivada de Avila (2021), preferindo a decomposição da forma quando composicional. Essa estratégia facilita o levantamento, no *treebank*, dos diferentes processos morfológicos que incidem sobre uma dada palavra, pois basta procurar pelo respectivo lema. Por exemplo, em Avila (2021), consta o verbete principal *yumunhã*, parafraseado como "fazer-se; ser feito", entre outras acepções. Em exemplos do *treebank* com essa acepção, porém, o lema é *munhã* 'fazer'. Dado que lexicalização e composicionalidade são fenômenos graduais, é possível que decisões atuais no *treebank* nesse domínio da lematização sejam revistas à luz de critérios que venham a ser definidos.

De uma maneira geral, adotamos as decisões de lematização de Avila (2021). No entanto, em vez do radical das palavras com prefixos relacionais, preferimos acompanhar Navarro (2016), utilizando como lema a forma absoluta dos substantivos poliformes, v.g., *tetama* 'terra' com prefixo absolutivo *t*, e a forma com prefixo de contiguidade *r* de posposições e verbos de segunda classe, v.g., *ruakí* 'perto de' e *rurí* 'estar feliz'. Cremos que essa decisão se coaduna mais com o espírito lexicalista de UD. De fato, enquanto nos deparamos nos textos com formas como *kwáu* (Figura 26) e *sendú* (Figura 33), uma forma como *urí* em vez de *surí* ou *rurí* não parece possível, pelo menos não encontramos registros em Avila (2021).

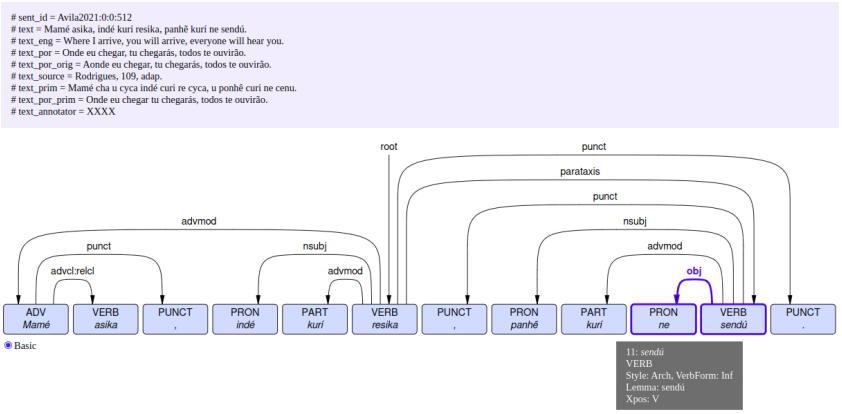


Figura 33. Exemplo com forma verbal no infinitivo.
Fonte: Elaboração própria.

4.3 Classes de palavras, traços morfológicos e relações sintáticas

Língua de gramática alheia a regulamentações oficiais, o nheengatu não dispõe de um inventário de classes de palavras normativo ou consensual. Sympson (1877) e Moore, Facundes e Pires (1994) resumem o vocabulário da língua a conjuntos de sete e oito classes, respectivamente, cuja interseção se limita a cinco elementos. Faltam ao segundo conjunto as conjunções e "sinais" do primeiro, que, por sua vez, carece das partículas, pronomes e demonstrativos daquele, subsumidos noutras classes, como os adjetivos.[17]

Cruz (2011) propõe uma classificação hierárquica com diferentes classes e subclasses. Nesse sistema, as palavras da língua são agrupadas inicialmente em duas macroclasses, a saber, palavras lexicais e palavras gramaticais, perfazendo, ao todo, sete classes, das quais não constam os adjetivos, que inexistiriam no nheengatu. Integram a primeira macroclasse verbos, advérbios, posposições, "índices de pessoa" e nomes, os quais se dividem em substantivos e dêiticos, subclasse que abarca parte dos pronomes de outras abordagens. As palavras gramaticais classificam-se em partículas e clíticos, as primeiras englobando diferentes subclasses, como as conjunções, subordinadores, interjeições e mais de duas dezenas de subtipos de partículas, responsáveis pela expressão da interrogação, negação, tempo, aspecto, modo, modalidade *etc.* Essa taxonomia não esgota o inventário terminológico de que Cruz (2011) se vale para descrever as propriedades morfológicas e sintáticas das palavras nheengatus. Determinante, (artigo) indefinido, (verbo) auxiliar, quantificador e numeral são algumas das outras classes a que se refere, a última das quais sugere integrar a classe dos nomes.

A proposta de Cruz (2011) difere em pontos essenciais da abordagem do modelo UD (Tabela 1). Nessa teoria, a distinção entre clíticos e não clíticos é ortogonal à classificação de palavras, que também não comporta os "índices de pessoa" de Cruz (2011), prefixos de nível subpalavra. Desse modo, no UD_Nheengatu-CompLin, assinalamos com o traço `Clitic=Yes` posposições, advérbios e partículas clíticas. Por outro lado, elementos classificados como morfemas flexionais, como os prefixos da série ativa, não constituem, no *treebank*, nós nas árvores sintáticas. Outra diferença refere-se à análise das partículas, que, conforme a Tabela 1, constitui classe disjunta de conjunções e interjeições. Finalmente, as classes determinante, pronome e numeral de UD não possuem correlatos diretos na proposta de Cruz (2011).

Avila (2021), pelo contrário, opera, na seção *categoria gramatical* da microestrutura dos verbetes do dicionário, com uma classificação em linhas gerais mais próxima das duas primeiras colunas da Tabela 1, das quais discrepa pela ausência das classes auxiliar e determinante e não distinção entre conjunções coordenativas e subordinativas. Na exposição gramatical que precede os verbetes do dicionário, contudo, Avila (2021) identifica diversos tipos de subordinadores. Por outro lado, no corpo das acepções dos verbos *ikú* e *mupika*, refere-se ao emprego deles como auxiliares. A distinção entre determinantes e pronomes da Tabela 1 corresponde, em linhas gerias, aos rótulos pronomes substantivos e pronomes adjetivos de Avila (2021). Avila (2021), na trilha de Moore, Facundes e Pires (1994), entre outros, defende a existência de adjetivos em nheengatu.

A Figura 34 permite comparar as quantidades de etiquetas de classes de palavras, traços morfológicos e relações de dependência do UD_Nheengatu-CompLin com as dos cinco *corpora* que o seguem na lista dos seis maiores *treebanks* de línguas ameríndias da coleção UD. Abstraindo da etiqueta `X`, inusada no UD_Nheengatu-CompLin e no UD_Kiche-IU, e de `ADJ` e `PART`, inexistentes, respectivamente, no UD_Guajajara-TuDeT e nos *treebanks* do Nahuatl, os cinco *corpora* compartilham o mesmo subconjunto das etiquetas da Tabela 1, do qual apenas `SYM` não ocorre. Nas duas outras dimensões, o UD_Nheengatu-CompLin iguala ou supera os demais *treebanks*. Ao todo, o UD_Nheengatu-CompLin contém 36 relações sintáticas, das quais três são subtipadas, e 76 combinações diferentes de atributos e valores, fazendo jus à riqueza gramatical do nheengatu.

A Figura 35 exibe as frequências absolutas das etiquetas de classes de palavras dos três maiores *treebanks* de línguas indígenas sul-americanas e do maior de língua indígena norte-americana na coleção UD. O UD_Nheengatu-CompLin ocupa a primeira ou segunda posição na quantidade de 11 das 16 etiquetas do seu conjunto de partes do discurso. A Figura 36 contrasta os três maiores

[17] Na classe dos "sinais", Sympson (1877) abriga elementos heterogêneos como sufixos derivacionais, interjeições e partículas.

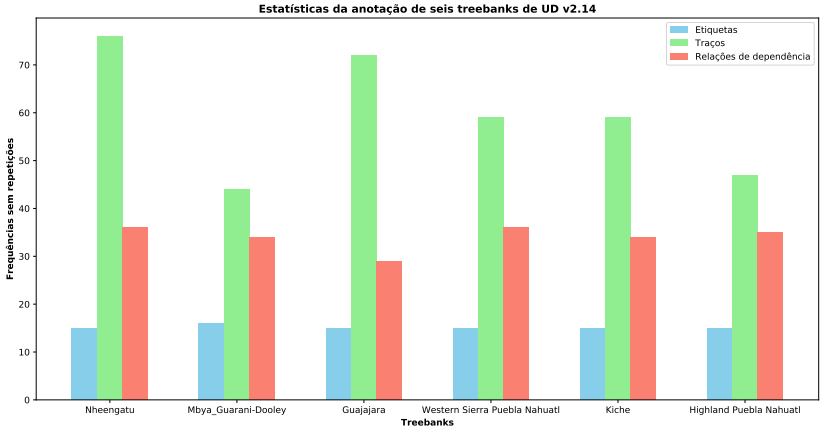


Figura 34. Quantidade sem repetições de etiquetas de classes de palavras, traços morfológicos e relações de dependência dos seis maiores *treebanks* de línguas ameríndias da versão v2.14. da coleção UD.

Fonte: Elaboração própria com base nos dados de Zeman *et al.* (2024).

treebanks de línguas indígenas brasileiras com o maior de língua portuguesa quanto às frequências relativas dessas etiquetas. Constatamos nos dois gráficos que o UD_Nheengatu-CompLin possui um número razoável de ocorrências mesmo daquelas classes de palavras menos frequentes, como `CCONJ`, `NUM` e `INTJ`, diferentemente de outros *treebanks* de línguas indígenas e mesmo do UD_Portuguese-CINTIL, com uma frequência relativa de `INTJ` tão baixa que nem aparece na Figura 36, não obstante o quase meio milhão de palavras desse *treebank*.

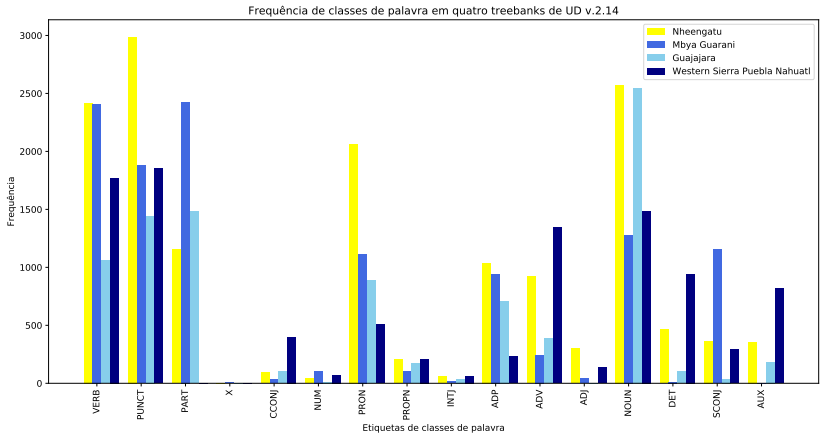


Figura 35. Classes de palavra em quatro *treebanks* de línguas ameríndias na coleção UD.

Fonte: Elaboração própria com base nos dados de Zeman *et al.* (2024).

4.4 Avaliação

Atualmente, o *script* validate.py sanciona integralmente todas as 1.470 sentenças do UD_Nheengatu-CompLin, a anotação de 28.37% das quais foram revistas por um ou dois revisores. Esse *script* constitui o principal crivo para admitir ou rejeitar um *treebank* numa *release* da coleção UD, servindo também para situar os *treebanks* em diferentes faixas de validade. O UD_Nheengatu-CompLin atualmente integra o grupo de 168 *treebanks* (59,36% de 283) de um total de 118 línguas (73,29% de 161) com o status `CURRENT VALID`, o mais alto.

O *script* validate.py, porém, não é abrangente o suficiente para detectar problemas de anotação específicos de uma língua particular ou mesmo análises que parecem implausíveis em qualquer língua, como um verbo regendo dois objetos diretos. Para assegurar uma maior qualidade de anotação, os *treebanks* são submetidos também ao Udapi (Popel; Žabokrtský; Vojtek, 2017), um *framework* que,

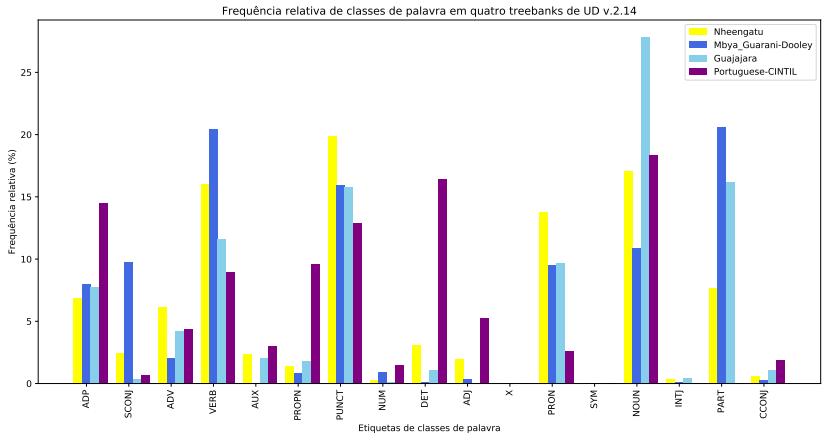


Figura 36. Frequências relativas das etiquetas de classe de palavra dos três maiores *treebanks* de línguas indígenas brasileiras e do maior de língua portuguesa na coleção UD.

Fonte: Elaboração própria com base nos dados de Zeman *et al.* (2024).

entre outras funcionalidades, gera, para um dado *treebank*, um relatório de erros de anotação, como, por exemplo, múltiplos objetos diretos. No exemplo da Figura 37, o Udapi indica que as formas verbais finitas da sentença carecem de indicação de modo verbal. É questionável, porém, se isso realmente constitui erro, uma vez que, no nheengatu, o modo verbal, quando não subentendido, é normalmente indicado por meio de partículas.

```
# sent_id = MooreFP1994:0:0:4
# text = Yandé yapurungitá yaikú nheengatú.
┌─► Yandé yandé PRON PRON Number=Plur|Person=1|PronType=Prs nsubj TokenRange=0:5
├─► yapurungitá purungitá VERB V Number=Plur|Person=1|VerbForm=Fin root Bug=finverb-mood|TokenRange=6:17
├─► yaikú ikú AUX AUXFS Number=Plur|Person=1|VerbForm=Fin aux Bug=finverb-mood|TokenRange=18:23
├─► nheengatú nheengatú NOUN N Number=Sing obj SpaceAfter=No|TokenRange=24:33
└─► . . PUNCT PUNCT _ punct SpaceAfter=No|TokenRange=33:34
```

Figura 37. Relatório de erros de anotação gerado pelo comando `udapy -HAM ud.MarkBugs` para a sentença da Figura 4. Os erros são assinalados pela palavra *Bug*, seguida de uma abreviatura que identifica o tipo de erro.

Fonte: Elaboração própria.

O Udapi computa 2.726 erros de anotação para o UD_Nheengatu-CompLin, 2.580 dos quais são do tipo exemplificado na Figura 37. A abreviatura `fin-verb-mood` indica que a forma verbal finita carece de especificação de modo verbal. A quantidade desses erros compõe com uma série de outros parâmetros um cálculo, registrado no arquivo `eval.log` do ramo mestre do repositório de cada *treebank*, que atribui de zero a cinco estrelas aos *treebanks* da coleção UD. Com duas estrelas, o UD_Nheengatu-CompLin insere-se no grupo de *treebanks* de línguas ameríndias mais bem avaliados, não obstante ultrapassar o patamar estabelecido de um erro por 10 palavras.

5 Considerações finais

Neste trabalho, partimos da premissa de que a inclusão digital é um fator de sobrevivência das línguas minoritárias. Isso é especialmente relevante no caso do nheengatu, dado o seu papel histórico de língua de contato. Como parte significativa dos falantes utiliza cotidianamente o português, com um índice de 0,97 de suporte digital uma das línguas mais privilegiadas do mundo, o nheengatu, com apenas 0,07, enfrenta mais um cenário adverso após mais de 150 anos de progressivo declínio.

Como aporte ao fortalecimento do nheengatu na seara tecnológica, construímos o UD_Nheengatu-CompLin, que reúne sentenças tanto mais curtas quanto mais longas extraídas de um total de 20 publicações de diferentes gêneros e épocas. Com 1,28 a 1,64 vezes mais palavras e 1,02 a 1,62 mais sentenças que os outros cinco maiores *treebanks* de línguas ameríndias na versão v2.14 da coleção UD, o UD_Nheengatu-CompLin iguala ou supera estes na maioria das dimensões aferidas pelo *script* conllu-stats.pl, notadamente no que tange aos traços morfológicos, integrando o grupo

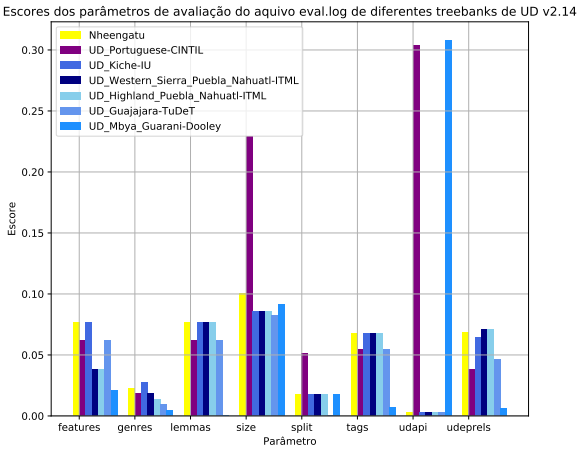


Figura 38. Avaliação de cinco *treebanks* de UD v2.4 em termos de tamanho, diversidade de gêneros textuais, quantidade de erros computados pelo Udapi e diferentes aspectos da anotação.

Fonte: Elaboração própria com base nos dados de Zeman *et al.* (2024).

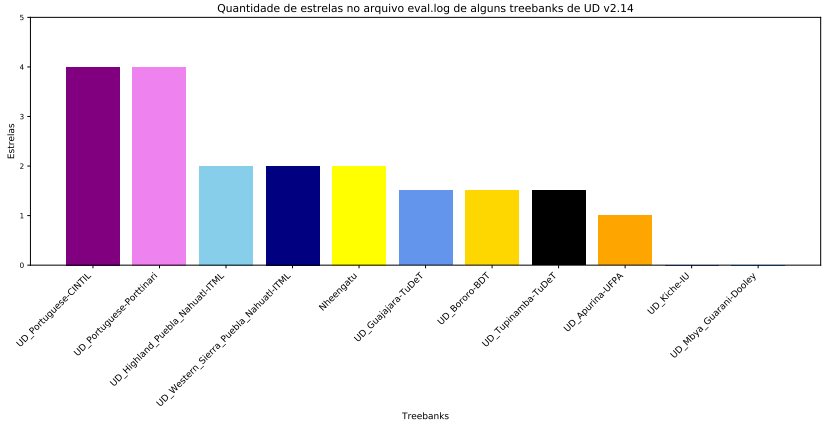


Figura 39. Quantidade de estrelas de onze *treebanks* com base nos escores da Figura 38.

Fonte: Elaboração própria com base nos dados de Zeman *et al.* (2024).

de *treebanks* de línguas ameríndias com a maior quantidade de estrelas. Todas essas características parecem propícias ao treino de um *parser* voltado para anotação das demais sentenças do mesmo conjunto de publicações de onde provêm as sentenças atuais do *treebank*.

Apesar de todos esses avanços, ainda há um longo caminho a percorrer para alcançar um *parser* do nheengatu com desempenho próximo ao índice LAS de 95% que Lopes e Pardo (2024) obtiveram no *parsing* do português a partir de um *treebank* de 8.418 sentenças. Para tanto, precisaríamos pelo menos quintuplicar o tamanho atual do UD_Nheengatu-CompLin, abarcando uma parcela maior de sentenças das fontes listadas na Figura 9. Num experimento de *parsing* utilizando o UDPipe 1.2 com a versão anterior do *treebank*, Alencar (2024) obteve, por meio do método de décupla validação cruzada, índices de LAS de 64,51% ±1,85% e 81,17% ±1,02% para texto cru e texto com etiquetas-ouro, respectivamente. Em sentenças desambiguadas com etiquetas do conjunto XPOS e providas de etiquetas especiais prefixadas por /= para tratamento de palavras desconhecidas, num cenário análogo ao *parsing* com etiquetas-ouro, o Yauti obteve LAS de 73,2% (Alencar, 2023), 7,97 pontos percentuais abaixo do modelo do UDPipe.

Isso sugere integrar um modelo treinado com mais sentenças no fluxograma de anotação da Figura 15. As Figuras 40 e 41 permitem antecipar as vantagens da utilização do UDPipe: o *parser* resolve automaticamente a ambiguidade da palavra *kwá*, que, no exemplo em questão, funciona como advérbio ao invés de pronome ou determinante. No entanto, a Figura 42 evidencia um problema desse *parser* no tratamento de palavras que inexistem no *corpus* de treino e fogem ao padrão morfológico mais geral. De fato, o *parser* não reconheceu o prefixo relacional de contiguidade desse substantivo poliforme, tratando-o como uniforme. Esse tipo de erro poderia passar despercebido a um anotador humano. O Yauti, pelo contrário, ao não gerar análise alguma para palavras desconhecidas, força o anotador a pesquisá-las em Avila (2021), incluindo-as no léxico da ferramenta, ou empregar as etiquetas especiais.

```
>>> s='''Resú-putari será kwá/demx rupí? (p. 413) Queres ir por cá? - Resó putari será cuá rupi?'''
>>> Yauti.parseExample(s,'stradelli1929',0,0,2, annotator='XXXX')
# sent_id = stradelli1929:0:0:2
# text = Resú-putari será kwá rupí?
# text_eng = Do you want to go this way?
# text_por = Queres ir por cá?
# text_source = p. 413
# text_orig = Resó putari será cuá rupi?
# text_annotator = XXXX
1-2     Resú-putari                                                             TokenRange=0:11
1       Resú    sú      VERB    V       Number=Sing|Person=2|VerbForm=Fin       0       root    _       _
2       putari  putari  AUX     AUXN    Compound=Yes|VerbForm=Inf       1       aux     _       _
3       será    será    PART    PQ      PartType=Int    1       advmod  _       TokenRange=12:16
4       kwá     kwá     PRON    DEMX    Deixis=Prox|Number=Sing|PronType=Dem    1       obl     _       TokenRa
17:20
5       rupí    rupí    ADP     ADP     AdpType=Post    4       case    _       SpaceAfter=No|TokenRange=21:25
6       ?       ?       PUNCT   PUNCT   _       1       punct   _       SpaceAfter=No|TokenRange=25:26
```

Figura 40. Análise de uma nova sentença com o Yauti.
Fonte: Elaboração própria.

```
$ echo "Resú-putari será kwá rupí?" | udpipe --tokenize --tag --parse model-10.output
Loading UDPipe model: done.
# newdoc
# newpar
# sent_id = 1
# text = Resú-putari será kwá rupí?
1-2     Resú-putari
1       Resú    sú      VERB    V       Number=Sing|Person=2|VerbForm=Fin       0       root    _       _
2       putari  putari  AUX     AUXN    Compound=Yes|VerbForm=Inf       1       aux     _       _
3       será    será    PART    PQ      PartType=Int    1       advmod  _       _
4       kwá     kwá     ADV     ADVDX   Deixis=Prox|PronType=Dem        1       obl     _       _
5       rupí    rupí    ADP     ADP     _       4       case    _       SpaceAfter=No
6       ?       ?       PUNCT   PUNCT   _       1       punct   _       SpacesAfter=\n
```

Figura 41. Análise da sentença da Figura 40 com um dos modelos do UDPipe 1.2. utilizados no experimento de décupla validação cruzada de Alencar (2024).
Fonte: Elaboração própria.

O UD_Nheengatu-CompLin está em constante expansão e revisão. Uma das tarefas mais urgentes é resolver os erros de anotação apontados pelo Udapi. Esperamos também alcançar 2000 sentenças na próxima *release* de UD e dobrar o número de sentenças revisadas.

```
$ echo "Kurupira upeteka mirá rapupema." | udpipe --tokenize --tag --parse model-10.output
Loading UDPipe model: done.
# newdoc
# newpar
# sent_id = 1
# text = Kurupira upeteka mirá rapupema.
1       Kurupira        kurupira        NOUN    N       Number=Sing     2       nsubj   _       _
2       upeteka peteka  VERB    V       Person=3|VerbForm=Fin   0       root    _       _
3       mirá    mirá    NOUN    N       Number=Sing     4       nmod:poss       _       _
4       rapupema        rapupema        NOUN    N       Number=Sing     2       obj     _       SpaceAfter=No
5       .       .       PUNCT   PUNCT   _       2       punct   _       SpacesAfter=\n
```

Figura 42. Análise do exemplo da Figura 16.
Fonte: Elaboração própria.

Agradecimentos

Agradecemos às diversas pessoas e instituições que contribuíram para o UD_Nheengatu-CompLin. A Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), no âmbito do projeto DACILAT (Processo No. 22/09158-5), e a Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico (Funcap) prestaram auxílio financeiro para o engajamento de estudantes na transcrição, anotação e revisão de sentenças. Nesse contexto, destacamos especialmente as contribuições de Juliana Lopes Gurgel, bolsista de treinamento técnico da FAPESP, e Dominick Maia Alexandre, bolsista de iniciação científica da Funcap. Eduardo de Almeida Navarro gentilmente cedeu os materiais textuais de Navarro (2016). A Editora da Universidade Federal do Amazonas (UFAM), na pessoa do seu diretor, Sérgio Freire, permitiu a utilização dos textos de Casasnovas (2006). Marcel Twardowsky Avila generosamente compartilhou conosco a sua *expertise* filológica em nheengatu, esclarecendo muitas questões lexicais e gramaticais sobre a língua e adaptando o texto da lenda *Como a noite apareceu* (Magalhães, 1876). João Marcos Cardoso, especialista em pesquisa e curador da Biblioteca Brasiliana Guita e José Mindlin da Universidade de São Paulo (USP), franqueou-nos as transcrições das narrativas de Amorim (1928) e João Barbosa Rodrigues (1890), diligentemente realizadas por Gabriela Lourenço Fernandes e Susan Gabriela Huallpa Huanacuni, estagiárias daquela instituição.

A versão final do artigo beneficiou-se imensamente dos comentários e sugestões dos dois revisores anônimos, a quem somos profundamente gratos.

Reconhecemos a utilização do ChatGPT para acelerar a escrita de código em Python e LaTeX. Examinamos, testamos e frequentemente corrigimos cada uma das sugestões dessa ferramenta.

Referências

AIKHENVALD, Alexandra Y.; DIXON, R. M. W. Introduction. *In:* AIKHENVALD, Alexandra Y.; DIXON, R. M. W. (ed.). *Areal diffusion and genetic inheritance: Problems in comparative linguistics*. Oxford: Oxford University Press, 2001. p. 1–26.

ALENCAR, Leonel Figueiredo de. Uma gramática computacional de um fragmento do nheengatu / A computational grammar for a fragment of Nheengatu. *Revista de Estudos da Linguagem*, v. 29, n. 3, p. 1717–1777, 2021. DOI: 10.17851/2237-2083.29.3.1717-1777.

ALENCAR, Leonel Figueiredo de. Yauti: A Tool for Morphosyntactic Analysis of Nheengatu within the Universal Dependencies Framework. *In:* ANAIS do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana. Belo Horizonte/MG: Sbc, 2023. p. 135–145. DOI: 10.5753/stil.2023.234131. Disponível em: https://sol.sbc.org.br/index.php/stil/article/view/25445.

ALENCAR, Leonel Figueiredo de. A Universal Dependencies Treebank for Nheengatu. *In:* GAMALLO, Pablo; CLARO, Daniela; TEIXEIRA, António; REAL, Livy; GARCIA, Marcos; OLIVEIRA, Hugo Gonçalo; AMARO, Raquel (ed.). *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 2*. Santiago de Compostela, Galicia/Spain: Association for Computational Linguistics, mar. 2024. p. 37–54. Disponível em: https://aclanthology.org/2024.propor-2.8.

AMEKA, Felix. Interjections: The Universal Yet Neglected Part of Speech. *Journal of Pragmatics*, v. 18, n. 2, p. 3, 1992.

AMORIM, Antonio Brandão de. Lendas em Nheêngatu e em Portuguez. *Revista do Instituto Historico e Geographico Brasileiro*, Imprensa Nacional, v. 154, n. 100, p. 9–475, 1928. Tomo 100, vol. 154 (2º de 1926).

AVILA, Marcel Twardowsky. *Estudo e prática da tradução da obra infantil A terra dos meninos pelados, de Graciliano Ramos, do português para o Nheengatu*. Mar. 2016. Diss. (Mestrado) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo. Disponível em: https://doi.org/10.11606/D.8.2016.tde-16052016-142700.

AVILA, Marcel Twardowsky. *Proposta de dicionário nheengatu-português*. 2021. Tese (Doutorado) – Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo. Disponível em: https://doi.org/10.11606/T.8.2021.tde-10012022-201925.

BIRD, Steven; GELBART, Katie; MCALISTER, Isaac (ed.). *Fábulas de Terra Preta: Uma coletânea bilíngue*. Manaus: sine nomine, 2013.

BRASIL, Missão Novas Tribos do (ed.). *Novo Testamento na língua Nyengatu*. 2nd. Barueri, SP: Sociedade Bíblica do Brasil, 2019. Primeira edição publicada em 1973.

CASASNOVAS, Afonso. *Noções de língua geral ou nheengatú: gramática, lendas e vocabulário*. 2. ed. Manaus: Editora da Universidade Federal do Amazonas; Faculdade Salesiana Dom Bosco, 2006.

COSTA, D. Frederico. *Carta pastoral de D. Frederico Costa bispo do Amazonas a seus amados diocesanos*. Fortaleza: Typ. Minerva, 1909.

CRUZ, Alina da. *Fonologia e gramática do nheengatú: A língua falada pelos povos Baré, Warekena e Baniwa*. Utrecht: Lot, 2011.

CUNHA, Celso; CINTRA, Lindley. *Nova gramática do português contemporâneo*. 2. ed. Rio de Janeiro: Lexicon, 2017.

D'ANGELIS, Wilmar da Rocha. A língua Nheengatu e suas ortografias: questões técnicas e de política linguística. *LIAMES: Línguas Indígenas Americanas*, v. 23, n. 00, p. 1–22, fev. 2023.

D'ANGELIS, Wilmar da Rocha; OLIVEIRA, Mateus Coimbra de; SCHWADE, Michéli Carolíni de Deus Lima. Acesso ao mundo digital ou acesso digital ao mundo? *Revista Digital de Políticas Linguísticas*, v. 15, p. 134–158, 2021.

DURAN, Magali Sanches. *Manual de Anotação de POS Tags: Orientações para Anotação de Etiquetas Morfossintáticas em Língua Portuguesa, Seguindo as Diretrizes da Abordagem Universal Dependencies (UD)*. São Carlos, SP, set. 2021. (Relatórios Técnicos, 434).

EBERHARD, David M.; SIMONS, Gary F.; FENNIG, Charles D. (ed.). *Ethnologue: Languages of the World*. twenty-sixth. Dallas: SIL International, 2023. Disponível em: http://www.ethnologue.com.

EVANS, Nicholas. Word classes in the world's languages. *In:* BOOIJ, Geert; LEHMANN, Christian; MUGDAN, Joachim; KESSELHEIM, Wolfgang; SKOPETEAS, Stavros (ed.). *Morphology: An International Handbook on Inflection and Word-Formation*. Berlin, New York: Walter de Gruyter, 2000. v. 1. p. 708–732.

FARACO, Carlos Alberto. Por que as línguas mudam? *In:* ÁVILA OTHERO, Gabriel de; NASCIMENTO FLORES, Valdir do (ed.). *O que sabemos sobre a linguagem*. São Paulo: Parábola, 2022.

FRANCIS, W. Nelson; KUČERA, Henry. *Manual of Information to Accompany A Standard Corpus of Present-Day Edited American English, for Use with Digital Computers*. 3. ed. Providence, RI: Brown University, 1979. Primeira edição publicada em 1964. Disponível em: http://clu.uni.no/icame/manuals/BROWN/INDEX.HTM.

FREIRE, José Ribamar Bessa. *Rio Babel: A história das línguas na Amazônia*. 2. ed. Rio de Janeiro: EdUERJ, 2011.

GALVES, Charlotte; SANDALO, Filomena; SENA, Ticiana A. de; VERONESI, Luiz. Annotating a polysynthetic language: From Portuguese to Kadiwéu. *Cadernos de Estudos Linguísticos*, v. 59, n. 3, p. 631–648, dez. 2017.

GERARDI, Fabrício Ferraz; REICHERT, Stanislav; ARAGON, Carolina Coelho. TuLeD (Tupían lexical database): introducing a database of a South American language family. *Language Resources and Evaluation*, v. 55, n. 4, p. 997–1015, dez. 2021.

GÓES NETO, Antônio Fernandes. *O Novo Testamento em nyengatu (1973): um capítulo na história das traduções bíblicas para línguas indígenas*. Jun. 2015. Dissertação de Mestrado – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo. Disponível em: https://doi.org/10.11606/D.8.2015.tde-15102015-141005.

GREENBERG, Joseph H. Numeral. *In:* BOOIJ, Geert; LEHMANN, Christian; MUGDAN, Joachim; KESSELHEIM, Wolfgang; SKOPETEAS, Stavros (ed.). *Morphology: An International Handbook on Inflection and Word-Formation*. Berlin, New York: Walter de Gruyter, 2000. v. 1. p. 770–783.

HIRSCHMANN, Hagen. *Korpuslinguistik: Eine Einführung*. Stuttgart: J.B. Metzler, 2019.

IONIN, Tania; MATUSHANSKY, Ora. *Cardinals: The Syntax and Semantics of Cardinal-Containing Expressions*. Cambridge, Massachusetts: The MIT Press, 2018.

JURAFSKY, Daniel; MARTIN, James H. *Speech and language processing: an introduction to natural language processing, computational linguistics, and speech recognition*. 2. ed. London: Pearson, 2009.

LEHMANN, Christian. The nature of parts of speech. *Sprachtypologie und Universalienforschung*, v. 66, n. 2, p. 141–177, 2013.

LEHMANN, Christian. *Theoretical foundation for word classes*. [*S. l.: s. n.*], 2023. Disponível em: https://christianlehmann.eu/publ/lehmann_word_classes.pdf.

LIMA, Rocha. *Gramática normativa da língua portuguesa*. 49. ed. Rio de Janeiro: José Olympio, 2011.

LOPES, Lucelene; DURAN, Magali Sanches; GRAÇAS VOLPE NUNES, Maria das; PARDO, Thiago Alexandre Salgueiro. *Corpora Building Process According to the Universal Dependencies Model: An Experiment for Portuguese*. São Carlos, SP, mar. 2022.

LOPES, Lucelene; PARDO, Thiago. Towards Portparser - a highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework. *In:* GAMALLO, Pablo *et al.* (ed.). *Proceedings of the 16th International Conference on Computational Processing of Portuguese*. Santiago de Compostela, Galicia/Spain: Association for Computational Lingustics, mar. 2024. p. 401–410. Disponível em: https://aclanthology.org/2024.propor-1.41.

MACAMBIRA, José Rebouças. *Estrutura Morfo-sintática do Português*. 9. ed. São Paulo: Pioneira, 1999.

MAGALHÃES, José Vieira Couto de. *O selvagem*. Rio de Janeiro: Typographia da Reforma, 1876.

MARNEFFE, Marie-Catherine de *et al. Syntax: General Principles*. [*S. l.: s. n.*], 2024. https://universaldependencies.org. Acesso em: 18. jul. 2024.

MARNEFFE, Marie-Catherine de *et al. UD Validation since release 2.5*. [*S. l.: s. n.*], 2024. https://universaldependencies.org/validation-rules.html. Acesso em: 18. jul. 2024.

MARNEFFE, Marie-Catherine de *et al. Universal POS tags*. [*S. l.: s. n.*], 2024. https://universaldependencies.org/u/pos/all.html. Acesso em: 18. jul. 2024.

MARNEFFE, Marie-Catherine de; MANNING, Christopher D.; NIVRE, Joakim; ZEMAN, Daniel. Universal Dependencies. *Computational Linguistics*, MIT Press, Cambridge, MA, v. 47, n. 2, jun. 2021.

MARTÍN RODRÍGUEZ, Lorena *et al.* Tupían Language Ressources: Data, Tools, Analyses. *In:* MELERO, Maite; SAKTI, Sakriani; SORIA, Claudia (ed.). *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*. Marseille, France: European Language Resources Association, jun. 2022. p. 48–58. Disponível em: https://aclanthology.org/2022.sigul-1.7.

MOORE, Denny. Historical Development of Nheengatu (Língua Geral Amazônica). *In:* MUFWENE, Salikoko S. (ed.). *Iberian Imperialism and Language Evolution in Latin America*. Chicago: University of Chicago Press, 2014. p. 108–142.

MOORE, Denny; FACUNDES, Sidney; PIRES, Nádia. Nheengatu (Língua Geral Amazônica), its History, and the Effects of Language Contact. *In:* PROCEEDINGS of the Meeting of the Society for the Study of the Indigenous languages of the Americas, July 2-4, 1993 and the Hokan-Penutian Workshop, July 3, 1993. Berkeley, CA: [University of California], 1994. p. 93–118. Disponível em: https://escholarship.org/uc/item/7tb981s1.

NAVARRO, Eduardo de Almeida. *Curso de Língua Geral (nheengatu ou tupi moderno): A língua das origens da civilização amazônica*. 2. ed. São Paulo: Centro Angel Rama da Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo, 2016.

NAVARRO, Eduardo de Almeida; ÁVILA, Marcel Twardowsky; TREVISAN, Rodrigo Godinho. O Nheengatu, entre a vida e a morte: A tradução literária como possível instrumento de sua revitalização lexical. *Revista Letras Raras*, v. 6, n. 2, p. 9–29, 2017.

NIVRE, Joakim *et al.* Universal Dependencies v1: A Multilingual Treebank Collection. *In:* PROCEEDINGS of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Portorož, Slovenia: European Language Resources Association (ELRA), maio 2016. p. 1659–1666. Disponível em: https://aclanthology.org/L16-1262.

POPEL, Martin; ŽABOKRTSKÝ, Zdeněk; VOJTEK, Martin. Udapi: Universal API for Universal Dependencies. *In:* PROCEEDINGS of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017). Gothenburg, Sweden: Association for Computational Linguistics, maio 2017. p. 96–101. Disponível em: https://aclanthology.org/W17-0412.

PTASZYNSKI, Michal; MUKAICHI, Kazuki; MOMOUCHI, Yoshio. NLP for Endangered Languages: Morphology Analysis, Translation Support and Shallow Parsing of Ainu Language. *In:* PROCEEDINGS of the 19th Annual Meeting of the Association for Natural Language Processing. Nagoya, Japan: [*s. n.*], mar. 2013. p. 418–421. Disponível em: https://www.anlp.jp/proceedings/annual_meeting/2013/pdf_dir/P2-5.pdf.

ROBINS, Robert Henry. The Development of the Word Class System of the European Grammatical Tradition. *Foundations of Language*, Springer, v. 2, n. 1, p. 3–19, 1966.

RODRIGUES, Aryon Dall'Igna. *Línguas brasileiras: Para o conhecimento das línguas indígenas*. São Paulo: Loyola, 1986.

RODRIGUES, Aryon Dall'Igna. Línguas indígenas: 500 anos de descobertas e perdas. *DELTA: Documentação e Estudos em Linguística Teórica e Aplicada*, v. 9, n. 1, p. 83–103, 1993. Disponível em: https://revistas.pucsp.br/index.php/delta/article/view/45596.

RODRIGUES, Aryon Dall'Igna. As línguas gerais sul-americanas. *Papia*, São Paulo, v. 4, n. 2, p. 6–18, 1996.

RODRIGUES, Ayron Dall'Igna; CABRAL, Ana Suelly Arruda Câmara. A contribution to the linguistic history of the Língua Geral Amazônica. *ALFA: Revista de Linguística*, v. 55, n. 2, dez. 2011.

RODRIGUES, João Barbosa. *Poranduba amazonense ou kochiyma-uara porandub, 1872-1887*. Rio de Janeiro: Typ. de G. Leuzinger & Filhos, 1890.

RUETER, Jack *et al.* Apurinã Universal Dependencies Treebank. *In:* MAGER, Manuel *et al.* (ed.). *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*. Online: Association for Computational Linguistics, jun. 2021. p. 28–33. Disponível em: https://aclanthology.org/2021.americasnlp-1.4.

SANDALO, Maria Filomena Spatti; GALVES, Charlotte Marie Chambelland. Anotando sintaticamente Uma língua originária do Brasil: O problema de Anchieta. *Cadernos de Estudos Linguísticos*, v. 65, n. 00, 2023.

SANTORINI, Beatrice. *Part of Speech Tagging Guidelines for the Penn Treebank Project*. 3. ed. [*S. l.: s. n.*], 1990. Disponível em: https://catalog.ldc.upenn.edu/docs/LDC99T42/tagguid1.pdf.

SANTOS, Luana Luiza; ARAGON, Carolina Coelho; GERARDI, Fabrício. Línguas minoritárias e anotações sintáticas de corpora: experiências de pesquisa na iniciação científica. *Letras de hoje*, v. 59, n. 1, p. 1–9, 2024.

SCHUSTER, Sebastian; MANNING, Christopher D. Enhanced English Universal Dependencies: An Improved Representation for Natural Language Understanding Tasks. *In:* PROCEEDINGS of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Portorož, Slovenia: European Language Resources Association (ELRA), maio 2016. p. 2371–2378. Disponível em: https://aclanthology.org/L16-1376.

SEIXAS, Manoel Justiniano de. *Vocabulario da lingua indigena geral para o uso do Seminario Episcopal do Pará*. Pará: Typ. de Mattos e Compª., 1853.

SIMONS, Gary F.; THOMAS, Abbey L. L.; WHITE, Chad K. K. Assessing Digital Language Support on a Global Scale. *In:* CALZOLARI, Nicoletta *et al.* (ed.). *Proceedings of the 29th International Conference on Computational Linguistics*. Gyeongju, Republic of Korea: International Committee on Computational Linguistics, out. 2022. p. 4299–4305. Disponível em: https://aclanthology.org/2022.coling-1.379.

STORTO, Luciana Raccanello. *Línguas indígenas: tradição, universais e diversidade*. Campinas, SP: Mercado de Letras, 2019.

STRADELLI, Ermano. Vocabularios da lingua geral portuguez-nheêngatú e nheêngatú-portuguez, precedidos de um esboço de Grammatica nheênga-umbuê-sáua mirî e seguidos de contos em lingua geral nheêngatú poranduua. *Revista do Instituto Historico e Geographico Brasileiro*, v. 158, n. 104, p. 9–768, 1929.

STRAKA, Milan; HAJIČ, Jan; STRAKOVÁ, Jana. UDPipe: Trainable Pipeline for Processing CoNLL-U Files Performing Tokenization, Morphological Analysis, POS Tagging and Parsing. *In:* PROCEEDINGS of the Tenth International Conference on Language Resources and Evaluation (LREC'16). Portorož, Slovenia: European Language Resources Association (ELRA), maio 2016. p. 4290–4297. Disponível em: https://aclanthology.org/L16-1680.

SYMPSON, Pedro Luiz. *Grammatica da lingua brazilica geral, fallada pelos aborigines das provincias do Pará e Amazonas*. Manaus: Typographia do Commercio do Amazonas, 1877.

TESNIÈRE, Lucien. *Éléments de syntaxe structurale*. Paris: Librairie C. Klincksieck, 1959.

TREVISAN, Rodrigo Godinho. *Tradução comentada da obra Le Petit Prince, de Antoine de Saint-Exupéry, do francês ao nheengatu*. Mar. 2017. Diss. (Mestrado) – Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo. DOI: 10.11606/D.8.2017.tde-07082017-124328. Disponível em: https://doi.org/10.11606/D.8.2017.tde-07082017-124328.

VASQUEZ, Alonso *et al.* Toward Universal Dependencies for Shipibo-Konibo. *In:* PROCEEDINGS of the Second Workshop on Universal Dependencies (UDW 2018). Brussels, Belgium: Association for Computational Linguistics, nov. 2018. p. 151–161. Disponível em: https://aclanthology.org/W18-6018.

WILKINS, David P. Interjections as Deictics. *Journal of Pragmatics*, v. 18, p. 119–158, 1992.

ZEMAN, Daniel *et al. Universal Dependencies 2.14*. [*S. l.: s. n.*], 2024. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University. Disponível em: http://hdl.handle.net/11234/1-5502.