# Compilação, modernização e anotação morfossintática de um *corpus* histórico do nheengatu segundo o modelo das Dependências Universais

**Antônio Levy Melo Nogueira**[1], **Letícia Farias Nunes**[1], **Dominick Maia Alexandre**[1], **Leonel Figueiredo de Alencar**[1]

[1]Universidade Federal do Ceará (UFC), Brazil
Av. da Universidade 2683 – 60.020-181 – Fortaleza – CE – Brazil

levy.melo.lm@gmail.com, leticiafariasnunes@alu.ufc.br,
dominick@letras.ufc.br, leonel.de.alencar@ufc.br

***Abstract.*** *This work presents the compilation, orthographic adaptation, and morphosyntactic annotation of the Nheengatu variant spoken in the Solimões River region in the 19th century. Nheengatu, the only living language descended from Old Tupi, like many minority languages, lacked syntactically annotated corpora until 2022, the year the UD_Nheengatu-CompLin treebank was released as part of the Universal Dependencies (UD) collection. The findings described here indicate the expansion of this treebank, contributing to the strengthening of resources available for the linguistic description and computational processing of Nheengatu.*

***Resumo.*** *Este trabalho apresenta a compilação, a adaptação ortográfica e a anotação morfossintática da variante do nheengatu falada na região do rio Solimões no século XIX. O nheengatu, única língua viva descendente do tupi antigo, assim como muitas línguas minoritárias, não dispunha de corpora anotados sintaticamente até 2022, ano em que foi lançado o treebank UD_Nheengatu-CompLin na coleção Universal Dependencies (UD). As etapas aqui descritas indicam a expansão desse treebank, contribuindo para o fortalecimento dos recursos disponíveis para a descrição linguística e o processamento computacional do nheengatu.*

## 1. Introdução

Esta pesquisa é vinculada a um projeto guarda-chuva cujo objetivo é desenvolver ferramentas e recursos para a anotação automática de *corpora* linguísticos [de Alencar 2024b]. A presente etapa concentra-se na expansão do *treebank* UD_Nheengatu-CompLin, que contém 2.120 árvores[1], por meio da inclusão de sentenças extraídas da obra *Christu Muhençáua* [Aguiar 1898], representativa de uma variante histórica do nheengatu falada na região do rio Solimões no final do século XIX.

Este estudo insere-se no contexto de iniciativas voltadas à criação de *treebanks* para línguas de baixo recurso – isto é, línguas com escassez de dados linguísticos disponíveis para tarefas de processamento de linguagem natural (PLN) – e à aplicação

[1]Informação baseada na versão de 15 de maio de 2025. Disponível em: https://github.com/UniversalDependencies/UD_Nheengatu-CompLin

do modelo *Universal Dependencies* (UD) a línguas indígenas, áreas em que ainda há lacunas significativas. Trabalhos como [Galves et al. 2017, Tyers and Henderson 2021, Martín Rodríguez et al. 2022, Sandalo and Galves 2023, Santos et al. 2024] mostram avanços nesse sentido e contribuem para o desenvolvimento de recursos computacionais para línguas minoritárias.

As atividades desta etapa incluem: a transcrição e a adaptação ortográfica do texto em nheengatu; a atualização ortográfica da versão em português desses textos; e a anotação morfossintática das sentenças em nheengatu com base no modelo UD [de Marneffe et al. 2021]. Espera-se que a ampliação do *corpus* anotado e do inventário lexical, por meio da incorporação de lemas oriundos de variantes históricas e de vocabulário atestado em [Aguiar 1898], contribuam para a melhoria da cobertura linguística da ferramenta de anotação automática Yauti [de Alencar 2023], e, por consequência, para o aumento da acurácia no processo de anotação morfossintática do nheengatu.

## 2. Metodologia

### 2.1. Transcrição

A primeira tarefa consistiu na transcrição integral do texto *Upãin mahã munhançáua* ("Criação do mundo"), de [Aguiar 1898], sentença por sentença. Diferentemente de etapas anteriores, desta vez a transcrição foi feita no formato *CoNLL-U* [de Marneffe et al. 2024], padrão usado nas anotações conforme o *framework* UD.

Seguindo as convenções adotadas pelos anotadores do *treebank* UD_Nheengatu-CompLin, a estrutura de metadados que antecede cada sentença (e na qual são registrados os textos transcritos) foi organizada conforme ilustrado na Listing 1. Cada entrada é composta pelos seguintes campos: `sent_id` (identificador da sentença); `text_sec` (adaptação secundária); `text_por_sec` (tradução secundária); `text_sec_source` (fonte da adaptação secundária); `text_por_sec_source` (fonte da tradução secundária); `text` (adaptação nossa da sentença original em nheengatu); `text_orig` (sentença conforme a grafia original); `text_source` (fonte original da sentença); `text_por` (versão em português de acordo com a norma-padrão); `text_orig_transcriber` (pessoa responsável pela transcrição do texto original); `text_por_modernizer` (pessoa responsável pela modernização da tradução em português); e `text_modernizer` (pessoa responsável pela modernização da sentença em nheengatu).

**Listing 1. Metadados informados no *treebank* UD_Nheengatu-CompLin**

```
# sent_id = Aguiar1898:21-8:19:219
# text_sec = Aé umukuruí ne akanga.
# text_por_sec = Ela esmagará a tua cabeça.
# text_sec_source = Avila (2021)
# text_por_sec_source = Avila (2021)
# text = -- Aé umukuruí kurí ne akã.
# text_orig = -- Aé u-mucururi curi ne acan.
# text_source = p. 85
# text_por = -- Ela esmagará a tua cabeça.
# text_orig_transcriber = Antônio Levy Melo Nogueira
# text_por_modernizer = Antônio Levy Melo Nogueira
# text_modernizer = Letícia Farias Nunes
```

### 2.2. Adaptação ortográfica

[Aguiar 1898] constitui um dos principais registros escritos do nheengatu do século XIX e integra o *corpus* principal do dicionário proposto por [Avila 2021], sendo citado em apro-

ximadamente 80 verbetes [de Alencar 2024b]. Embora tenhamos considerado as adaptações previamente realizadas por [Avila 2021], fizemos nossa própria adaptação, tomando como referência o seu material, mas priorizando formas históricas fonológica e morfologicamente mais próximas do original.

Um exemplo ilustrativo encontra-se na Listing 1, em que a forma *acan* ('cabeça'), registrada no texto original, foi adaptada por [Avila 2021] como *akanga*, conforme observado no metadado `text_sec`. Em nossa versão, no entanto, optamos por *akã*, forma histórica também atestada no dicionário de [Avila 2021].

Essa escolha visou preservar a integridade do texto original e incluir de forma consistente a variante do nheengatu do rio Solimões do século XIX ao *treebank* UD_Nheengatu-CompLin, respeitando suas especificidades históricas e dialetais.

### 2.3. Anotação morfossintática

Após a transcrição e a adaptação ortográfica do *corpus*, a etapa de anotação morfossintática foi iniciada com o uso do Yauti, analisador desenvolvido em Python para o nheengatu [de Alencar 2023, de Alencar 2024a]. Para permitir a análise automatizada das sentenças oriundas da obra de [Aguiar 1898], o Yauti oferece a função `parseExampleAguiar`, que recebe como entrada os metadados completos de cada sentença (conforme ilustrado na Listing 1), processando o conteúdo do campo `text`, que corresponde à sentença pós-normalização ortográfica.

A função realiza, de forma automatizada, a segmentação em *tokens* e a atribuição de etiquetas morfossintáticas (`UPOS` e `XPOS`), traços morfológicos (`FEATS`), bem como a identificação do `head` e das relações de dependência sintática (`DEPREL`). Contudo, por se tratar de um analisador em desenvolvimento e dependente de dados lexicais previamente inseridos, a saída gerada pelo Yauti requer revisão e correção manuais. A curadoria humana é indispensável não apenas para ajustar erros de análise, mas também para a desambiguação de formas ambíguas.

A Figura 1 apresenta um exemplo de anotação automática contendo dois casos de ambiguidade morfossintática. O primeiro envolve a forma *awá*, que pode desempenhar as funções de pronome indefinido (e.g., 'alguém'), interrogativo (e.g., 'quem', 'qual') ou relativo livre ('quem'). O segundo caso refere-se à forma *maã*, que pode corresponder à partícula de condicional, a um pronome interrogativo (e.g., 'o que', 'qual'), indefinido ou relativo livre ('o que'), ao substantivo 'coisa', ou, ainda, à forma verbal não finita 'ver'.

Tais casos de ambiguidade são tratados diretamente no texto de entrada da função `parseExampleAguiar`, por meio da atribuição manual da etiqueta morfossintática apropriada. As correções das colunas `HEAD` e `DEPREL`, por sua vez, são realizadas posteriormente no arquivo em formato `.conllu`. A Figura 2 exibe as relações de dependência da referida sentença após a etapa de correção manual das anotações morfossintáticas.

Além disso, a inclusão de novas entradas no arquivo `lexicon.json` do Yauti também é feita de forma manual e contínua pelos anotadores, visto que formas ausentes não são automaticamente reconhecidas e anotadas.

```
Yauti.parseExampleAguiar(s,'lev')
# sent_id = Aguiar1898:1:1:1
# text = – Awá taá umunhã upaĩ maã?
# text_source = p. 23
# text_orig = – Auá tahá u-munhã upäin mahã?
# text_por = – Quem criou todas as cousas?
# title = CHRISTU MUÉSAWA TUPANA-MUNHANGARA
# title_orig = CHRISTU MUHENÇÁUA TUPANA-MUNHANGÁRA
# title_por = DOUTRINA CRISTÃ DE DEUS CRIADOR
# text_orig_transcriber = Davi Sampaio Santos
# text_por_modernizer = Davi Sampaio Santos
# text_modernizer = Antônio Levy Melo Nogueira
# text_eng = TODO
# inputline = – Awá taá umunhã upaĩ maã?
# text_annotator = Antônio Levy Melo Nogueira
1       –       –       PUNCT   PUNCT   _       4       punct   _       TokenRange=-1:0
2       Awá     awá     PRON    IND     PronType=Ind    2       det     _       TokenRange=1:4
2       Awá     awá     PRON    INT     PronType=Int    2       det     _       TokenRange=1:4
2       Awá     awá     PRON    RELF    PronType=Rel    4       nsubj   _       TokenRange=1:4
3       taá     taá     PART    CQ      PartType=Int    4       advmod  _       TokenRange=5:8
4       umunhã  munhã   VERB    V       Mood=Ind|Person=3|VerbForm=Fin  0       root    _       TokenRange=9:15
5       upaĩ    upaĩ    DET     TOT     PronType=Tot    6       det     _       TokenRange=16:20
6       maã     maã     PART    COND    Modality=Cond|PartType=Mod      4       advmod  _       SpaceAfter=No|TokenRange=21:24
6       maã     maã     PRON    IND     PronType=Ind    6       det     _       SpaceAfter=No|TokenRange=21:24
6       maã     maã     PRON    INT     PronType=Int    4       obj     _       SpaceAfter=No|TokenRange=21:24
6       maã     maã     NOUN    N       Number=Sing     4       obj     _       SpaceAfter=No|TokenRange=21:24
6       maã     maã     PRON    RELF    PronType=Rel    4       obj     _       SpaceAfter=No|TokenRange=21:24
6       maã     maã     VERB    V       VerbForm=Inf    4       parataxis       _       SpaceAfter=No|TokenRange=21:24
7       ?       ?       PUNCT   PUNCT   _       4       punct   _       SpaceAfter=No|TokenRange=24:25
```

**Figura 1. Exemplo de uso do Yauti para anotação de sentença de [Aguiar 1898].**

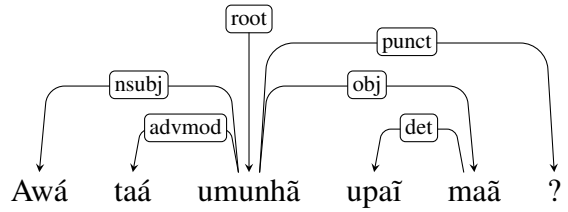


**Figura 2. Relações de dependência da sentença *Awá taá umunhã upaĩ maã?* 'Quem criou todas as cousas?' [Aguiar 1898, p. 23]**

## 3. Resultados

Foram extraídas 226 sentenças de [Aguiar 1898], provenientes de 21 textos catequéticos. Todas as sentenças foram transcritas e adaptadas, em conformidade com as convenções supracitadas. Até o momento, 86 sentenças foram anotadas, das quais 75 passaram por revisão de um segundo anotador. As anotações também contribuíram para ampliar a cobertura lexical do Yauti, com a incorporação de 24 novas entradas, totalizando 2.113 lemas.

## 4. Conclusões

A variedade do nheengatu presente na obra de [Aguiar 1898] apresenta estruturas sintáticas mais complexas, como inversões na ordem dos constituintes em relação ao padrão observado em sentenças previamente integradas ao UD_Nheengatu-CompLin, além de arcaísmos lexicais e gramaticais. Essas particularidades constituem desafios que serão aprofundados nas próximas etapas. A pesquisa segue em desenvolvimento por meio da atualização do repositório público[2] do projeto, mantido ativamente com a inclusão de dados e discussões via *issues*. A inclusão definitiva de novas sentenças no *treebank* ocorrerá após revisão humana e validação pelo *script* oficial de UD (`validate.py`), que lê arquivos CoNLL-U e verifica sua conformidade com a especificação UD, incluindo a checagem de traços e relações de dependência do *treebank*.

[2] `https://github.com/CompLin/nheengatu`

## Referências


[Aguiar 1898] Aguiar, C. (1898). *Doutrina christã destinada aos naturaes do amazonas em nhihingatú com traducção portugueza em face*. Pap. e Tip. Pacheco, Silva & C., Petrópolis.

[Avila 2021] Avila, M. T. (2021). *Proposta de dicionário nheengatu-português*. PhD thesis, Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo.

[de Alencar 2023] de Alencar, L. F. (2023). 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*, pages 135–145, Porto Alegre, RS, Brasil. SBC.

[de Alencar 2024a] de Alencar, L. F. (2024a). Aspectos da construção de um corpus sintaticamente anotado do nheengatu no modelo dependências universais. *Texto Livre*, 17:e52653.

[de Alencar 2024b] de Alencar, L. F. (2024b). Aspectos léxico-gramaticais do nheengatu na obra christu muhençáua, de dom josé lourenço. Projeto de pesquisa de Iniciação Científica (PIBIC), não publicado.

[de Marneffe et al. 2024] de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., McDonald, R., Nivre, J., Petrov, S., Pyysalo, S., Schuster, S., Silveira, N., Tsarfaty, R., Tyers, F., and Zeman, D. (2024). Conll-u format. Accessed: 2024-01-09.

[de Marneffe et al. 2021] de Marneffe, M.-C., Manning, C. D., Nivre, J., and Zeman, D. (2021). Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

[Galves et al. 2017] Galves, C., Sandalo, F., de Sena, T. A., and Veronesi, L. (2017). Annotating a polysynthetic language: From Portuguese to Kadiwéu. *Cadernos de Estudos Linguísticos*, 59(3):631–648.

[Martín Rodríguez et al. 2022] Martín Rodríguez, L. et al. (2022). Tupían language ressources: Data, tools, analyses. In Melero, M., Sakti, S., and Soria, C., editors, *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 48–58, Marseille, France. European Language Resources Association.

[Sandalo and Galves 2023] Sandalo, M. F. S. and Galves, C. M. C. (2023). Anotando sintaticamente uma língua originária do brasil: O problema de anchieta. *Cadernos de Estudos Linguísticos*, 65(00).

[Santos et al. 2024] Santos, L. L., Aragon, C. C., and Gerardi, F. (2024). Línguas minoritárias e anotações sintáticas de corpora: experiências de pesquisa na iniciação científica. *Letras de hoje*, 59(1):1–9.

[Tyers and Henderson 2021] Tyers, F. M. and Henderson, R. (2021). A corpus of k'iche' annotated for morphosyntactic structure. In *Proceedings of the First Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*.