v. 34, n. 1, p. 8–50, 2026
ISSN: 2237-2083 · DOI: 10.17851/2237-2083.34.1.8–50
Submetido em: 08/04/2025 · Aprovado em: 20/01/2026

REVISTA DE ESTUDOS DA LINGUAGEM

Per aspera ad astra: Improving the Automatic Evaluation of a Universal Dependencies Treebank for a Low-Resource Language

Per aspera ad astra: *melhorando a avaliação automática de um treebank no modelo Dependências Universais para uma língua com poucos recursos*

Leonel Figueiredo de Alencar
Universidade Federal do Ceará (UFC)
Fortaleza | CE | BR
leonel.de.alencar@ufc.br
https://orcid.org/0000-0001-8148-6994

Hélio Leonam Barroso Silva
Universidade Federal do Ceará (UFC)
Fortaleza | CE | BR
heliolbs@alu.ufc.br
https://orcid.org/0000-0003-2752-9871

Juliana Lopes Gurgel
Universidade Federal do Ceará (UFC)
Fortaleza | CE | BR
julianagurgel@letras.ufc.br
https://orcid.org/0009-0005-4640-454X

Dominick Maia Alexandre
Universidade Federal do Ceará (UFC)
Fortaleza | CE | BR
dominick@letras.ufc.br
https://orcid.org/0009-0000-7749-7762

Abstract: Until recently, advancements in digital humanities have favored majority languages. The construction of treebanks for 14 Brazilian Indigenous languages within the Universal Dependencies (UD) framework marks a significant step toward bridging the digital divide affecting these minority languages. However, aside from the Nheengatu treebank, these corpora are small and/or receive low quality ratings. This paper details our recent efforts to enhance the Nheengatu treebank. We examined UD's automatic evaluation methodology to pinpoint the areas we should focus on. Among other improvements, we corrected nearly all 2,726 errors flagged by the Udapi framework, a core component of the UD rating system. As a result, the treebank's rating advanced from 2.0 stars in UD v2.14 to 3.5 stars in v2.15 and 4.0 stars in the upcoming v2.16. It is now the largest and best evaluated among the 21 UD treebanks for Amerindian languages. A corollary contribution of the annotation revision was the identification of discrepancies between the UD documentation and the Udapi algorithm. Specifically, while the documentation permits (i) assigning the Degree feature to nouns and (ii) using the ExtPos feature to mark an unexpected POS tag on the head of an exocentric MWE, Udapi systematically treats both configurations as annotation errors.

Keywords: Nheengatu; Universal Dependencies; treebank; morphosyntactic annotation; computational linguistics.

Resumo: Até recentemente, os avanços nas humanidades digitais privilegiaram as línguas hegemônicas. A construção de *treebanks* para 14 línguas indígenas brasileiras no modelo Dependências Universais (UD) representa uma iniciativa importante para reduzir a exclusão digital que afeta essas línguas minoritárias. No entanto, com exceção do *treebank* de nheengatu, esses *corpora* são pequenos e/ou recebem notas baixas de avaliação. Neste artigo, detalhamos nossos esforços recentes para aprimorar o *treebank* de nheengatu. Analisamos a metodologia de avaliação automática de UD a fim de identificar os pontos que exigiam maior atenção. Entre outras melhorias, corrigimos quase todos os 2.726 erros apontados pelo *framework* Udapi, um componente central do sistema de pontuação de UD. Como resultado, a avaliação do *treebank* avançou de 2,0 estrelas na UD v2.14 para 3,5 estrelas na v2.15 e 4,0 estrelas na próxima v2.16. Atualmente, esse é o maior e mais bem avaliado entre os 21 *treebanks* de línguas ameríndias da coleção UD. Uma contribuição corolária da revisão das anotações foi revelar discrepâncias entre a documentação de UD e o algoritmo do Udapi. Especificamente, enquanto a documentação permite (i) a atribuição do traço *Degree* a substantivos e (ii) o uso do traço ExtPos para marcar uma classe de palavra inesperada no núcleo de uma MWE exocêntrica, o Udapi trata sistematicamente ambas as configurações como erros de anotação.

Palavras-chave: Nheengatu; Dependências Universais; treebank; anotação morfossintática; linguística computacional.

1 Introduction

Rodrigues (1966, p. 5) defined the documentation, description, and comparison of Indigenous languages as "the greatest task of linguistics in Brazil." More than half a century later, it is evident that Indigenous linguistics never reached such magnitude in the country, despite the considerable accumulation of publications over this period (Oliveira; Camacho, 2013; Rodrigues, 1986; Seki, 1999; Storto, 2019). On the other hand, natural language processing, computational linguistics, and corpus linguistics have expanded in Brazil at an increasing pace since the 1990s, completely detached from the Indigenous languages. Thus, Seki's (1999, p. 285) proposal to create a unified database compiling data from various studies on these

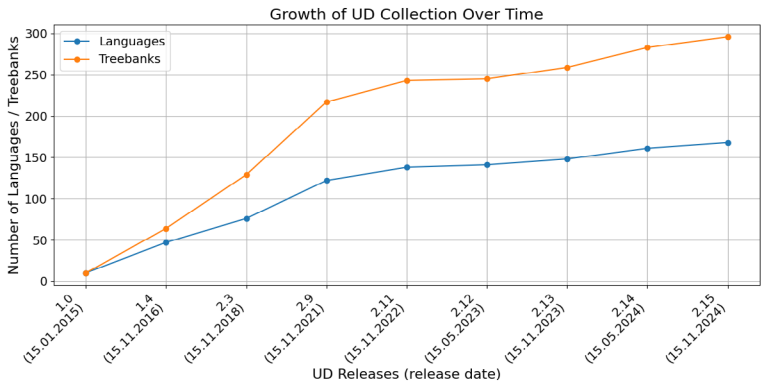
languages, especially the "corpora that mediate between spoken languages and the produced works", was ignored.

Fortunately, the marginalization of Indigenous languages from digital humanities begins to be reversed. We highlight some recent initiatives. Galves *et al.* (2017) and Sandalo and Galves (2023) propose an annotation scheme for a constituency treebank of Kadiwéu. Rueter *et al.* (2021) and Santos, Aragon, and Gerardi (2024) address the development of treebanks in the Universal Dependencies framework (Marneffe *et al.*, 2021) for the Apurinã language and various languages of the Tupian language family, respectively. D'Angelis, Oliveira, and Schwade (2021) deals with the translation of a smartphone operating system into Nheengatu and Kaingang. Cavalin *et al.* (2023), Pinhanez, Cavalin, and Nogima (2024), and Silva and Pardo (2024) report on experiments in the field of artificial intelligence involving, respectively, language identification, machine translation, and grammar induction for various Brazilian Indigenous languages.

The Universal Dependencies (UD) model underpins the largest collection of syntactically annotated corpora (Marneffe *et al.*, 2021; Zeman, 2023). Launched in 2015 with 10 treebanks for 10 languages, this collection has grown exponentially, reaching, in less than a decade, 296 treebanks for 168 languages (Figure 1). This represents, according to Equation 1, a compound annual growth rate (CAGR) of 41.15%, reflecting the increasingly widespread adoption of the UD model and the growing interest in multilingual morphosyntactic annotation and parsing.

$$CAGR = \left(\left(\frac{final_{value}}{initial_{value}}\right)^{\left(\frac{1}{years}\right)} - 1\right) \times 100\% = \left(\left(\frac{296}{10}\right)^{\left(\frac{1}{9.88}\right)} - 1\right) \times 100\% \approx 41.15\% \quad (1)$$

Figure 1 – Growth of the UD collection in terms of treebanks and languages since 2015



Treebanks have been used for language documentation, morphosyntactic research—particularly in linguistic typology—and the development of computational tools such as syntactic parsers. We have identified, in release 2.15 of this collection, a total of 21 treebanks for 20 languages of the Americas, 14 of which are from Brazil, including Nheengatu. First introduced in the November 15, 2022 release as the seventh-largest Amerindian language treebank in terms of token count, UD_Nheengatu-CompLin became the largest in the UD collection by this criterion in the November 15, 2023 release (Alencar 2024a, 2024b).

Once widespread in Amazonia due to its adoption as *lingua franca* during the colonial period, Nheengatu—also known, *inter alia*, as Modern Tupi and Amazonian General Language (*Língua Geral Amazônica* in Portuguese)—has rapidly shrunk with the advancement of Western civilization in the region since the second half of the 19th century (Borges, 1996; Freire, 2011; Navarro; Ávila; Trevisan, 2017; Rodrigues, 1986; Rodrigues, 1993; Rodrigues, 1996; Rodrigues; Cabral, 2011). Reduced to an estimated 6,000 speakers in Brazil and 8,000 in Colombia, Nheengatu faces extinction, ranking at level 6b on the Expanded Graded Intergenerational Disruption Scale (EGIDS) (Eberhard; Simons; Fennig, 2025). The recent digitalization efforts have increased its Digital Language Support (DLS) from 0.07 (Eberhard; Simons; Fennig, 2023) to 0.21 (Eberhard; Simons; Fennig, 2025). However, Nheengatu remains threatened by the digital divide that marginalizes Indigenous languages. In contrast, majority languages such as Portuguese, Spanish, and English reach the maximum DLS score of 1.00.

Over five releases of the collection, UD_Nheengatu-CompLin has grown not only quantitatively but also qualitatively. In the UD project's automatic rating system, which ranges from zero to five stars, UD_Nheengatu-CompLin was, as of the May 15, 2024 version, one of only four Amerindian language treebanks rated at 2.0 stars, whereas the ratings of the other 17 treebanks of this language group ranged between 0 and 1.5 stars. The 2.0-star rating, however, is only half the number of stars assigned to treebanks of majority languages such as UD_Portuguese-Porttinari (Duran *et al.*, 2023) and UD_Portuguese-CINTIL (Branco *et al.*, 2022).

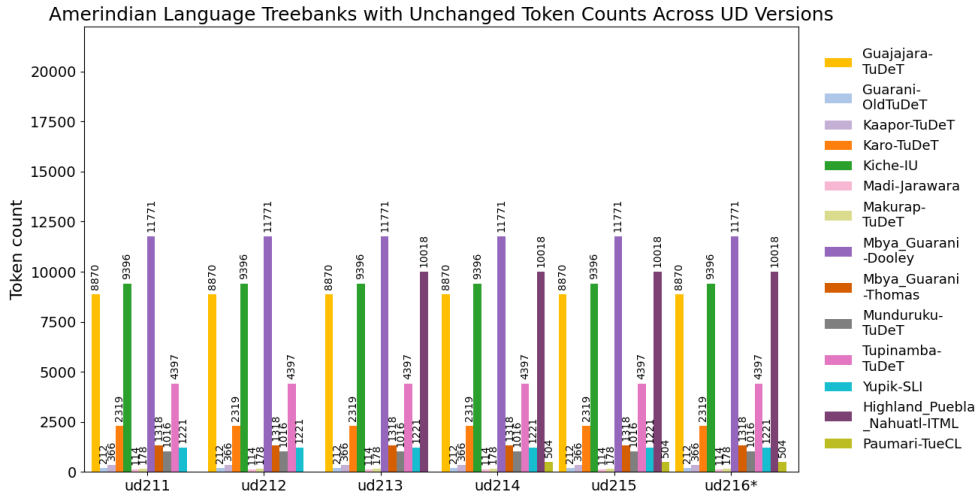
This paper describes our latest efforts to enhance UD_Nheengatu-CompLin, which led to a 1.5-star increase in the UD v2.15 version, released on November 15, 2024, making it the only Amerindian language treebank with more than two stars. The primary factor behind this rating boost was the systematic correction of 98.42% of the errors identified by the Udapi framework (Popel; Žabokrtský; Vojtek, 2017). This revision involved a reworking of our approach to Nheengatu's verb valency and verbal inflectional morphology. It also revealed inconsistencies between Udapi's constraints and the UD documentation. For the upcoming May 15, 2025 release of the UD collection, we have continued improving UD_Nheengatu-CompLin. By dividing the treebank into training and test sets that meet the minimum word count thresholds established in the scoring algorithm, we increased its rating to 4.0 stars in the development version of February 12, 2025. Meanwhile, other Amerindian treebanks have either maintained or lost stars.

In the next section, we compare UD_Nheengatu-CompLin with other Amerindian language treebanks based on statistical parameters such as token count, sentence count, and annotation richness across different UD versions. In Section 3, we present the two evaluation methodologies applied to all UD treebanks: validation and rating. We also highlight UD_Nheengatu-CompLin's recent achievements in this domain. Since joining the UD collection in the November 15, 2022 v2.11 release as a fully validated treebank, UD_Nheengatu-CompLin has maintained this status across all subsequent releases. In UD v2.15, it became the only Amerindian treebank ranked within the top 33.5% according to the star rating system. In Section 4, we discuss the errors detected in UD_Nheengatu-CompLin by the Udapi framework, one of the key components of the rating system, as well as the annotation scheme revisions that successfully resolved most of these errors, thereby increasing the treebank's star rating. In the final section, we summarize our main contributions and outline strategies to reach the top of the rating system, which currently consists of 4.5 stars.

2 Overview of UD_Nheengatu-CompLin

The UD_Nheengatu-CompLin treebank joined the UD collection on November 15, 2022, in release 2.11, ranking seventh and tenth among the then 18 treebanks of Amerindian languages in terms of token and sentence count, respectively[1]. A year later, in version 2.13, it became the largest Amerindian language treebank in terms of tokens, and in version 2.14, it also took first place in sentence count. In the 2.15 release of the UD collection, it remains the leader in both criteria, with 61.7% more tokens than UD_Mbya_Guarani-Dooley (the second-largest treebank in token count) and 27.7% more sentences than UD_Kiche-IU (the second-largest in sentence count). In the development snapshot of January 20, 2025, these differences have increased to 75% and 40.6%, respectively. Figure 2 displays the treebanks whose token counts remained unchanged between UD 2.11 and this snapshot, whereas Figure 3 shows those whose token counts changed over this period. Figure 4, by contrast, tracks across these UD versions only the treebanks that rank among the ten largest by sentence count in the January 20, 2025 snapshot.

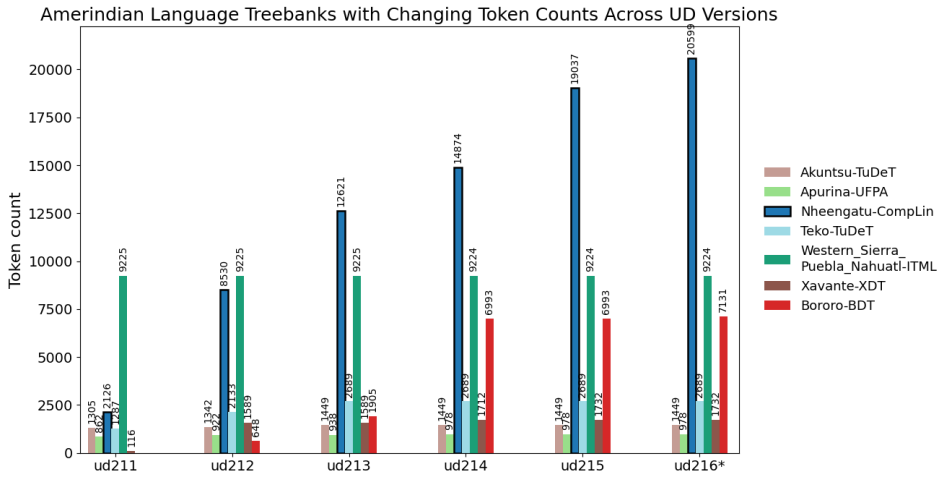
Figure 2 – Token counts of Amerindian language treebanks whose token counts remained unchanged between UD release 2.11 and the pre-release UD 2.16* development snapshot of January 20, 2025. Treebank identifiers are abbreviated in the figure for readability.



Fonte: elaboração própria.

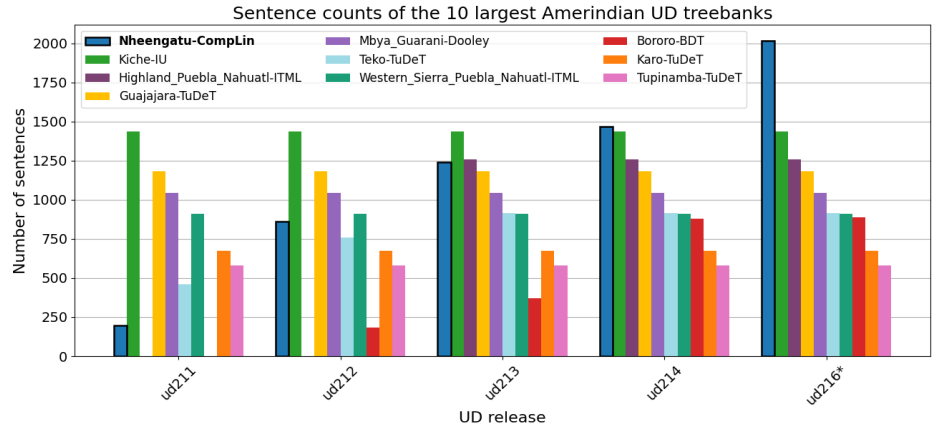
[1] All UD versions are stored in the digital library LINDAT/CLARIAH-CZ, available at <https://lindat.mff.cuni.cz/repository/xmlui/>.

Figure 3 – Token counts of Amerindian language treebanks whose token counts changed between UD release 2.11 and the pre-release UD 2.16* development snapshot of January 20, 2025. Treebank identifiers are abbreviated in the figure for readability.



Fonte: elaboração própria.

Figure 4 – Sentence counts across different UD versions for the ten largest Amerindian language treebanks by sentence count, as determined in the pre-release **UD 2.16*** development snapshot of January 20, 2025. Treebank identifiers are abbreviated in the figure for readability.



Fonte: elaboração própria.

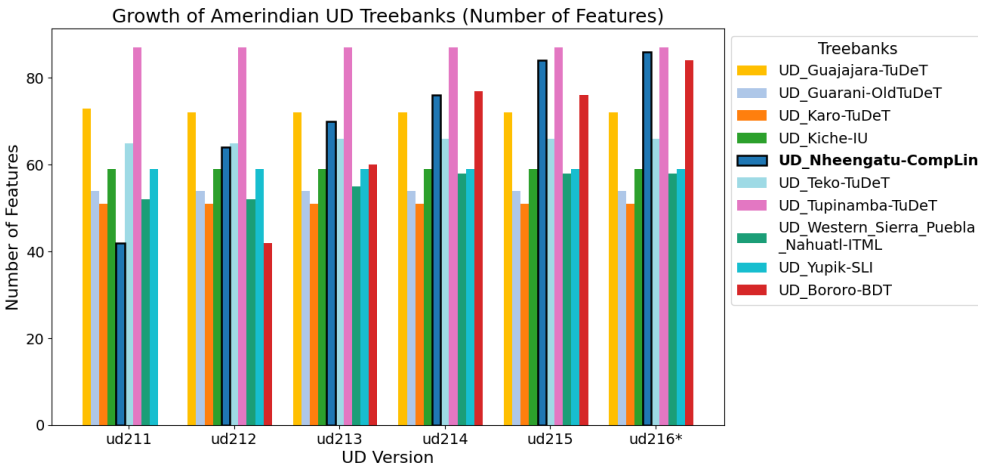
Given the disparity in growth rates between UD_Nheengatu-CompLin and the other treebanks of Amerindian languages, it is likely that the gap between them will widen even further in the next UD version. As shown in Figures 2 and 3, UD_Nheengatu-CompLin is the only corpus in this group of languages that has substantially grown with each UD release since its launch in UD v2.11. We observe that 15 treebanks have seen no increase in size during this period[2]. Only the Bororo and Xavante language treebanks have had token increments since UD v2.13. The UD_Teko-TuDeT treebank experienced significant growth only in the two versions following its release. In UD v2.12, its token count increased by 65.7%, and in the next

[2] Included in Figure 3 among the treebanks with variation in token counts, UD_Western_Sierra_Puebla_Nahuatl-ITML did not increase in size over the period considered; instead, its token count decreased by one token between UD v2.13 and v2.14.

release, it grew by another 26.1%. However, since v2.14, no further increases have been recorded, indicating a stabilization in treebank size. The Apurinã, Akuntsu, and Xavante treebanks also seem to have reached their final size plateau.

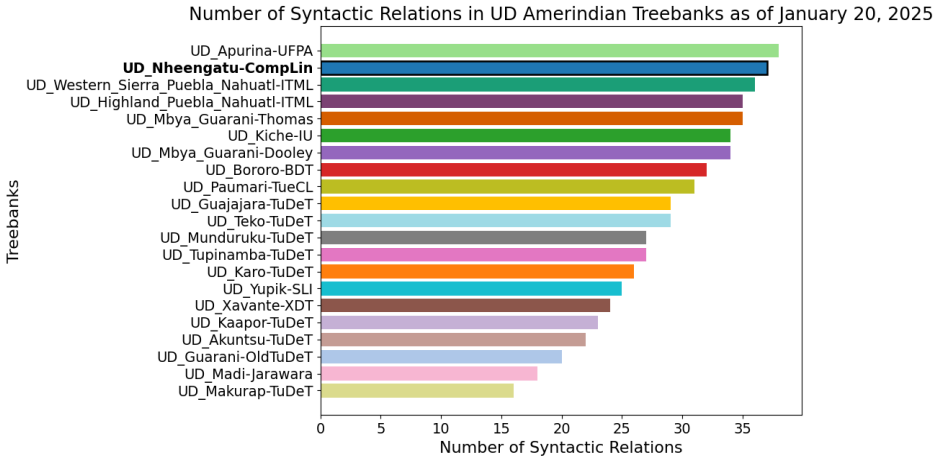
Regarding the different layers of UD annotation, UD_Nheengatu-CompLin ranks among the Amerindian language treebanks with the highest counts of part-of-speech tags, morphological features, and syntactic relations. As shown in Figure 5, the Nheengatu treebank has more than doubled its inventory of morphological features between its first version and the present, now ranking second in this category, the same position it holds for syntactic relations (Figure 6). On the other hand, UD_Nheengatu-CompLin is among the five Amerindian language treebanks with the highest number of part-of-speech tags, totaling 16 out of the maximum of 17 (Figure 7).

Figure 5 – Evolution of the number of features in the 10 treebanks with the highest feature counts in the pre-release **UD 2.16*** development snapshot of January 20, 2025, tracked across the successive versions since 2.11



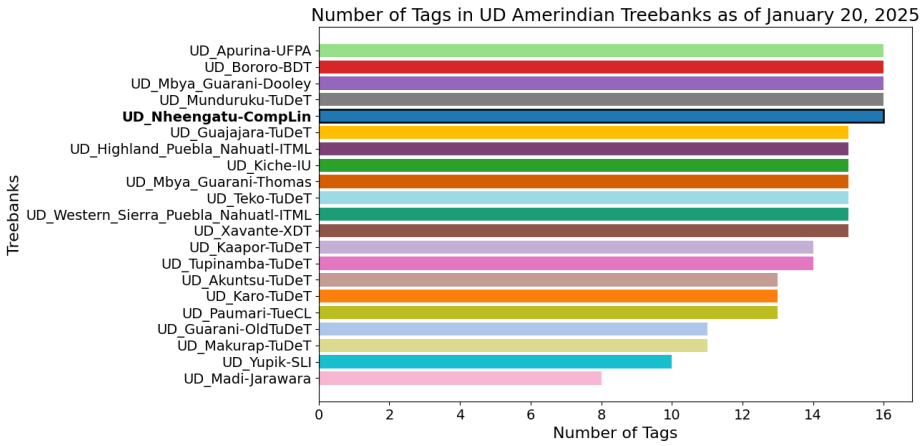
Fonte: elaboração própria.

Figure 6 – Number of dependency relations per Amerindian language treebank in the pre-release **UD 2.16*** development snapshot of January 20, 2025. The data was computed by UD project's conllu-stats.pl script, including subtyped relations.



Fonte: elaboração própria.

Figure 7 – Number of part-of-speech tags per Amerindian language treebank in the pre-release **UD 2.16*** development snapshot of January 20, 2025



Fonte: elaboração própria.

In summary, UD_Nheengatu-CompLin stands out in release 2.15 of the UD collection as the largest of the 21 treebanks of Amerindian languages. It surpasses UD_Mbya_Guarani-Dooley by 61.7% in the number of tokens and UD_Kiche-IU by 40.6% in the number of sentences, which rank second in these two criteria, respectively. In the next section, we will see how UD_Nheengatu-CompLin, which in UD 2.14 was among the group of only four treebanks of these languages with 2 stars, advanced by 1.5 in this ranking in UD v2.15, reaching 4 stars in the development version of February 12, 2025.

3 The automatic evaluation of UD treebanks

The exponential growth of the UD collection in just under a decade would have been impractical if manual validation were required for all 296 treebanks. Without this validation, the usability of these resources would be drastically reduced, if not entirely infeasible, for the development of natural language processing applications and linguistic investigations. On the other hand, the increasing coverage of the collection, with a growing number of languages represented by more than one treebank, has made it necessary to compare the treebanks in qualitative terms to guide the selection of the most appropriate resources for specific scenarios.

To address these requirements, every treebank is subject to an automatic evaluation that assesses both quantitative and qualitative aspects across various levels of annotation, in addition to analyzing textual genre diversity and the structure of its GitHub repository. This evaluation utilizes two distinct methodologies: one focused on validating the treebank and the other on ranking it on a scale from 0 to 5 stars. In the following two subsections, we will explore each process in detail.

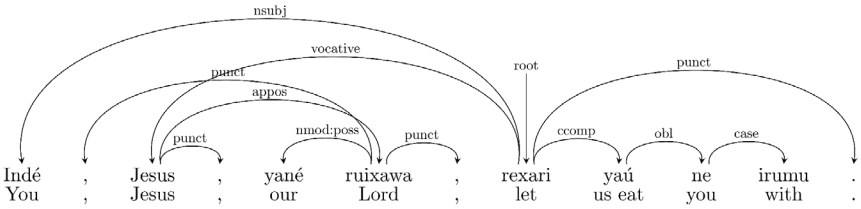
3.1 The automatic validation

The validation process employs the programs check_files.pl and validate.py. While the former verifies whether the treebank repository's file structure meets a series of requirements, the latter examines, for each sentence, whether it constitutes a tree-like graph, whether its annotation conforms to the CoNLL-U format, and whether it adheres to certain constraints of the UD theory. The CoNLL-U format consists of a table with ten columns, each specifying a particular piece of information about a word in a sentence: (i) ID, (ii) form, (iii) lemma, (iv) universal part-of-speech (UPOS), (v) treebank-specific part-of-speech, (vi) morphological features (FEATS), (vii) governing word (HEAD), (viii) universal dependency relation (DEPREL), (ix) enhanced dependency, and (x) other annotation (Marneffe *et al.*, 2024a). At a minimum, the fields UPOS, HEAD, and DEPREL must be filled with categories specified by the model (Marneffe *et al.*, 2024b). The FEATS column must be either underspecified with an underscore or consist of attribute-value pairs in the prescribed format.

Additionally, the validity of specific combinations of word classes with syntactic relations is checked. For example, advmod requires one of the labels from the set {ADV, ADJ, CCONJ, DET, PART, SYM}. Similarly, only a pronoun (PRON) or a determiner (DET) can serve as a syntactic determiner (det). Certain tree configurations are identified as invalid, such as a single node governing two subjects or a punctuation mark being attached to a governor in a non-projective manner, as shown in Figure 8, an error corrected in Figure 9. Violations of these and other constraints are classified as errors, preventing the treebank from being deemed valid.

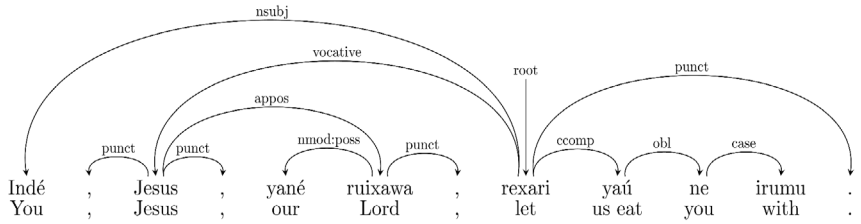
1) Indé, Jesus, yané r-uixawa, re-xari ya-ú ne irumu.
you Jesus our CONT-Lord 2SG.ACT-let 1PL.ACT-eat you with
'You, Jesus, our Lord, let us eat with you.' (Avila, 2021, p. 777)

Figure 8 – Dependency graph of (1) with punctuation attached non-projectively



Fonte: elaboração própria.

Figure 9 – Dependency graph of sentence Avila2021:0:0:2 from UD_Nheengatu-CompLin with all punctuation attached projectively

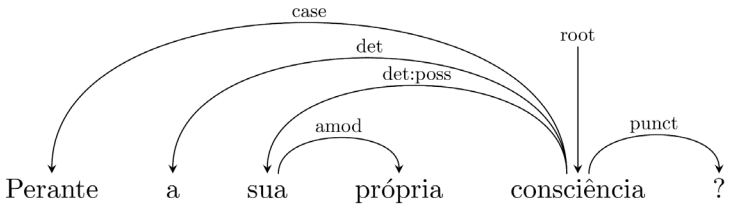


Fonte: elaboração própria.

Other requirements are less strict, and when violated, they generate messages classified as warnings, which do not affect the validity of the treebank. One example of a configuration that triggers such a message is when a determiner governs another node, as shown in Figure 10, which contradicts the UD principle that only content words can function as governors (Marneffe *et al.*, 2021).

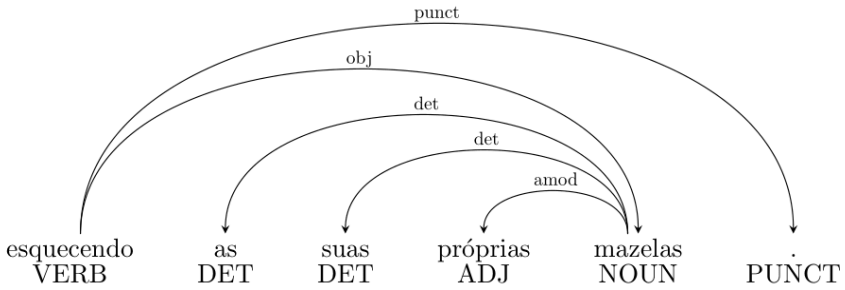
Apparently, classifying a deviation from annotation guidelines as undesirable rather than an error stems from the lack of sufficient grounds to consider a given phenomenon as universal, even though it is hypothesized to be so. The warning against configurations like the one in Figure 10, which is currently ignored by the maintainers of UD_Portuguese-CINTIL, seems to point to an analysis that is indeed unlikely—so much so that it has not been adopted by UD_Portuguese-Porttinari, as evidenced by Figure 11.

FIGURE 10 – Example of a determiner governing an adjective in the UD_Portuguese-CINTIL treebank



Fonte: elaboração própria.

FIGURE 11 – Excerpt from sentence FOLHA_DOC000115_SENT004 of UD_Portuguese-Porttinari with an alternative analysis of the construction of Figure 10



Fonte: elaboração própria.

As a free and open-source software, the program validate.py is easy to run from the command line of operating systems like Linux. By detecting errors that often go unnoticed by annotators and human reviewers, it plays a key role in the quality control of annotations. Thus, it is integrated into the development pipeline of UD_Nheengatu-CompLin, being executed periodically after changes to the treebank. This has ensured that successive development versions of UD_Nheengatu-CompLin consistently rank among the treebanks with the highest validation status.

The validity status, as determined by check_files.pl and validate.py, is a necessary condition for the admission of a new treebank into a release of the UD collection. On the other hand, veteran treebanks that become invalid between one release and the next must address the issues detected within four years, after which they are removed from the collection.

The results of the check_files.pl and validate.py programs are consolidated into an online report that classifies the treebanks in the UD collection into various categories based on their validation status, depending on the types of problems detected. This report is always updated every time the development version of a treebank is uploaded to the UD repository. Table 1 shows the number of treebanks and languages in each of these categories based on data collected from the report on January 18, 2025[3].

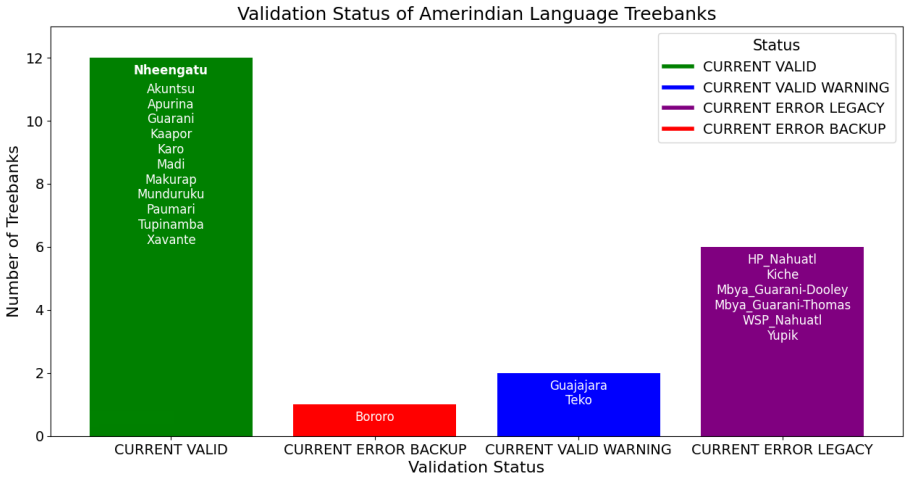
Table 1 – Frequencies of the different validation categories of the development versions of the UD treebanks as of January 18, 2025

Category	Treebanks	Languages
CURRENT VALID WARNING	48	40
CURRENT VALID	51	44
SAPLING EMPTY	71	61
CURRENT ERROR LEGACY	181	96
SAPLING ERROR	24	24
CURRENT ERROR BACKUP	11	9
RETIRED ERROR	6	5
CURRENT ERROR NEGLECTED	5	5
TOTAL VALID/BACKUP/LEGACY/NEGLECTED	296	168

Fonte: elaboração própria.

Version 2.15, the most recent, includes treebanks in the categories CURRENT VALID, CURRENT VALID WARNING, CURRENT ERROR BACKUP, CURRENT ERROR LEGACY, and CURRENT ERROR NEGLECTED, totaling 296 corpora of 168 languages. The UD_Nheengatu-CompLin is included in the first category, along with eleven other treebanks of Amerindian languages, representing the highest adherence to the principles of UD (Figure 12).

Figure 12 – Validation status of the treebanks for Indigenous languages of the Americas as of January 18, 2025. Treebank identifiers are abbreviated in the figure for readability.



Fonte: elaboração própria.

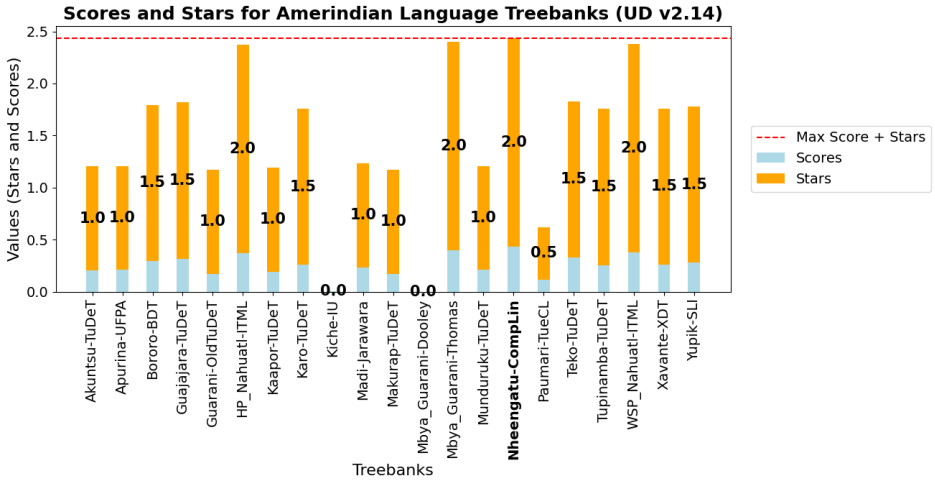
[3] https://quest.ms.mff.cuni.cz/udvalidator/cgi-bin/unidep/validation-report.pl

The remaining categories reflect decreasing levels of validation. The categories with the keywords WARNING and ERROR characterize the treebanks whose validation generated warning and error messages, respectively. The keywords BACKUP, LEGACY, and NEGLECTED reflect an increasing period, counted from the last UD release, during which errors persist in the treebanks. BACKUP indicates errors produced after the most recent release, while LEGACY and NEGLECTED refer to errors persisting up to and after three years, respectively. RETIRED designates treebanks that were excluded from the latest release of the collection for not correcting the detected errors in the prescribed time frame. The categories starting with the term SAPLING identify treebanks that have never been included in any UD release, among which those marked additionally with VALID and VALID WARNING are eligible for admission in the next release.

3.2 The stars classification system

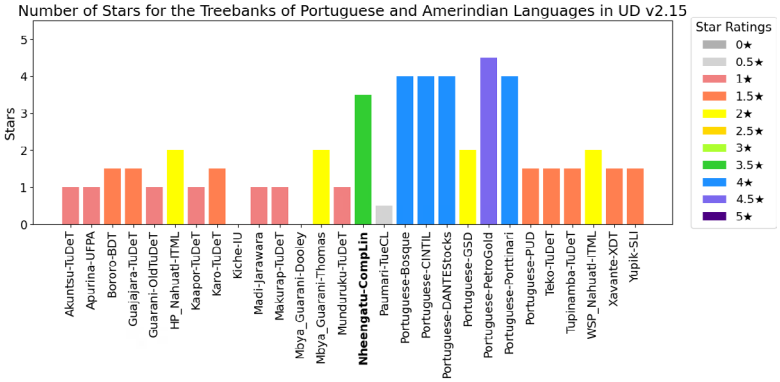
The second evaluation system uses the program evaluate_treebank.pl to compute a score between 0 and 1 for each treebank in the UD collection, considering a series of criteria, which will be explained later. This score serves as the basis for ranking the treebanks on a discrete scale of 0 to 5 stars, with increments of 0.5 (i.e., the possible values are 0, 0.5, 1.0, 1.5, ..., 5.0). Figure 13 shows the scores and star quantities of the treebanks of Amerindian languages in version 2.14 of the UD collection. Only four of these corpora achieve a two-star rating, among which is UD_Nheengatu-CompLin, which outperforms all others in the sum of score and star quantity. In UD v2.15, the number of stars for UD_Nheengatu-CompLin increased by 75% to 3.5 (Figure 14), making it the only Amerindian language treebank to enter the third-highest rating range (Figure 15), close to the major language treebanks such as Portuguese.

Figure 13 – Total scores and number of stars of the treebanks for Amerindian languages in UD v2.14. Treebank identifiers are abbreviated in the figure for readability.



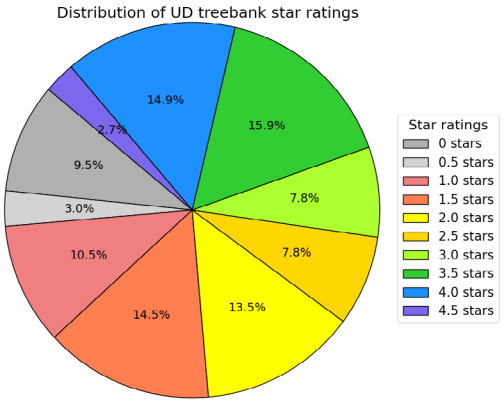
Fonte: elaboração própria.

Figure 14 – Number of stars of Portuguese and Amerindian language treebanks in UD v2.15. Treebank identifiers are abbreviated in the figure for readability.



Fonte: elaboração própria.

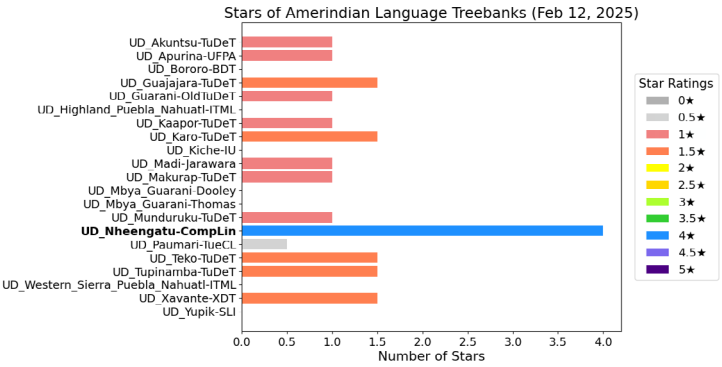
Figure 15 – Frequency distribution of star ratings in UD v2.15. Treebank identifiers are abbreviated in the figure for readability.



Fonte: elaboração própria.

In the development version of January, 20, 2025, UD_Nheengatu-CompLin has reached the 4-star rating of a treebank like UD_Portuguese-Porttinari in UD 2.15, while the number of stars for four Indigenous language treebanks has dropped to zero, with the others maintaining the rating from UD v2.15 (Figure 16).

Figure 16 – Number of stars of the development versions of the treebanks for Amerindian languages as of February 12, 2025



Fonte: elaboração própria.

To assign a score from 0 to 1 for a treebank, eight dimensions are evaluated, namely: features, genres, lemmas, size, split, tags, udapi, udeprels. Each dimension has a weight and a score. The same weights, as shown in Table 2, apply to all treebanks, while the scores are calculated for each treebank.

Table 2 – Weights assigned to different dimensions in the scoring system

Dimension	Weight
features	3
genres	3
lemmas	3
size	10
split	2
tags	3
udapi	12
udeprels	3

Fonte: elaboração própria.

The scores for genres, size, split, and udapi are calculated according to Equation (2), while the scores for features, lemmas, tags, and udeprels are calculated according to Equation (3), where FS_{dim} represents the final score of the dimension, w_{dim} is the weight assigned to the dimension, f_{dim} is the additional multiplicative factor (when applicable), and s_{dim} is the raw score of the dimension.

$$FS_{dim} = w_{dim} \cdot s_{dim} \quad (2)$$

$$FS_{dim} = w_{dim} \cdot f_{dim} \cdot s_{dim} \quad (3)$$

The multiplicative factor depends on information contained in the <pre> element, which constitutes the last part of the README.md file of the treebank repositories on GitHub[4]. Table 3 reproduces the specifications for these four dimensions in the README.md file of four treebanks in the UD collection. For each dimension **considered independently**, the multiplicative factor reaches the maximum value of 1.0 when its annotation is natively manual, whereas the automatic conversion of a manual annotation is assigned a lower factor of 0.8. In these four dimensions, UD_Nheengatu-CompLin achieves the maximum multiplicative factor.

Table 3 – Comparison of Four UD Treebanks on Key Dimensions

Treebank	Relations	Lemmas	UPOS	Features
Nheengatu	manual native	manual native	manual native	manual native
Paumari	manual native	not available	manual native	not available
CINTIL	converted from manual	converted from manual	converted from manual	converted from manual
Porttinari	manual native	manual native	manual native	manual native

Fonte: elaboração própria.

The features score reflects the richness and reliability of the morphological information in the FEATS column, taking into account the number of features and the nature of the annotation, as shown in Table 3. Based on the percentage of words with at least one feature, a value between 0.01 and 1.0 is computed. The higher the percentage, the higher the value, which is then multiplied by the corresponding factor for the nature of the annotation. The UD_Paumari-TueCL, for example, has 504 words, none of which have morphological features, resulting in the minimum score value of 0.01. Just under half of the words in UD_Apurina-UFPA have some feature, resulting in the value 0.5, which is multiplied by the factor 0.8, giving a score of 0.4. UD_Nheengatu-CompLin has more than two-thirds of its words annotated morphologically, all of which are done manually, achieving the maximum score of 1.0.

The genres score takes into account the diversity of textual genres in the treebank. Currently, the program evaluate_treebank.pl lists 17 different genres, although the UD documentation specifies 18, including an additional government genre, used, for example, by UD_English-GUM (Müller-Eberstein; Goot; Plank, 2021). To calculate this score, one simply takes the number of textual genres in the treebank, which is provided in the aforementioned block delimited by the <pre> and </pre> tags, and divides it by the maximum number accepted by UD. With five different genres, UD_Nheengatu-CompLin surpasses the seven Portuguese treebanks and only loses in this regard, among the Amerindian language treebanks, to UD_Kiche-IU, which identifies 6 genres.

The lemmas score, like the previous ones, ranges from 0.01 to 1. At the lower end are treebanks that are predominantly or entirely without lemmatization, meaning that the LEMMAS column is mostly or entirely filled with underscores ("_"). At the other end, we have treebanks like UD_Nheengatu-CompLin, with natively manual lemmatization. With a score of 0.8, UD_Portuguese-CINTTIL occupies an intermediate position due to the automatic conversion of manual lemmatization, as shown in Table 3.

[4] https://github.com/UniversalDependencies

The size score takes into account the number of words in the treebank. To compute it, minimum and maximum reference values were defined: 1,000 and 1,000,000 words, respectively. Any treebank with fewer than 1,000 words receives a score of 0. This applies to treebanks of the following Indigenous languages: Apurinã, Old Guarani, Kaapor, Jarawara, Makurap, and Paumari. Conversely, any treebank with 1,000,000 words or more is assigned the maximum score value.

Since treebank sizes vary across several orders of magnitude, normalization on a logarithmic scale was necessary. To this end, a size coefficient (SC) was defined as the natural logarithm of the square of one-thousandth of the number of words N in the treebank, as shown in Equation 4:

$$SC = \ln\left[\left(\frac{N}{1{,}000}\right)^2\right] \quad (4)$$

Substituting the maximum word count into the equation yields:

$$SC = \ln\left[\left(\frac{1{,}000{,}000}{1{,}000}\right)^2\right] = ln[(1{,}000)^2] = ln[1{,}000{,}000] = 13.815511 \quad (5)$$

These maximum and minimum logarithmic limits allow for a linear comparison of treebank sizes. The final size score is calculated by dividing the result of Equation 4 by the maximum value from Equation 5. For example, UD_Portuguese-Porttinari contains N = 168,080 words. Substituting this value into Equation 4, we obtain $\ln[(168{,}080/1{,}000)^2] = 10.248880$. The size score is then 10.248880/13.815511 = 0.741838. Similarly, version 2.14 of UD_Nheengatu-CompLin, with 15,036 words, received a size score of 0.392377. In version 2.15, with 19,278 words, the score increased to 0.428353.

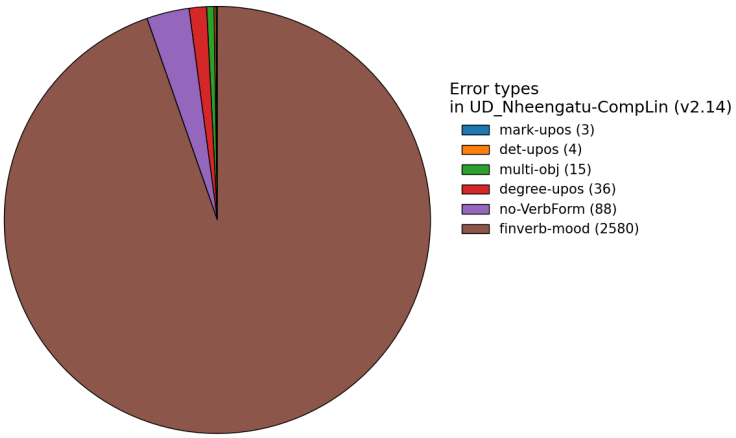
The split score takes into account the suitability of the treebank for training a parser using machine learning algorithms, which typically require a division into three sets: training, development, and testing (Jurafsky; Martin, 2009). For a treebank to obtain the maximum score of 1, the first set must contain more than 10,000 words, while the others must consist of at least 10,000 words. Each of these conditions is worth 0.33. To the sum of these values, 0.01 is added, which is the value assigned to treebanks that do not meet any of the conditions. This situation applies to most of the 21 Indigenous language treebanks in UD v2.15, with only five achieving a split score of 0.34, including UD_Nheengatu-CompLin in versions 2.14 and 2.15. In the development version of January 20, 2025, UD_Nheengatu-CompLin, as the only one among the 21 Amerindian language treebanks, achieved a split score of 0.64, while six of the seven Portuguese treebanks reached the maximum value.

The tags score takes into account the use of the part-of-speech tags provided by UD in the treebank. The approach is similar to that of textual genres. Currently, UD accepts 17 different tags. To calculate this score, simply take the number of tags that occur in the treebank and divide by the 17 provided by UD. UD_Nheengatu-CompLin, for example, uses 15 tags in version 2.14, resulting in a score of 15 / 17 = 0.882352. With 16 tags, version 2.15 attains a score of 16/17 = 0.941176. UD_Portuguese-Porttinari, on the other hand, uses 16 labels, reaching a score of 16 / 17 = 0.941176.

The udapi score takes into account the results of the ud.MarkBugs block of the Udapi framework (Popel; Žabokrtský; Vojtek, 2017). This program checks whether the annotations violate a set of restrictions, generating a report with the types and frequencies of detected

errors (referred to as “bugs”). Figure 17 displays the number of errors for UD_Nheengatu-CompLin in version 2.14 of the UD collection, which we will address in detail in Section 4.

Figure 17 – Frequencies of types of the 2726 errors detected by Udapi



Fonte: elaboração própria.

The udapi score is calculated similarly to the lemma score, comparing the number of errors B with the number N of words in the treebank. The worst case is one error for every 10 words. If the ratio of errors to words, i.e., N/B, for B > 0, is equal to or smaller than this threshold, the score will be the minimum value of 0.01. Otherwise, Equation 6 is applied, by which a treebank with no errors achieves the maximum score of 1.

$$US = 1 - \frac{B \times 10}{N} \quad (6)$$

Version 2.14 of UD_Nheengatu-CompLin has 2,726 errors for 15,036 words, which equates to one error every 5.5 words, resulting in the minimum score of 0.01. In version 2.15 of this treebank, we managed to reduce the number of errors to 58, while the total number of words increased to 19,278, catapulting the score to 0.969913. The corresponding version of UD_Portuguese-Porttinari, in turn, has 67 errors for 168,080 words, which equates to one error every 2,626 words, yielding a score of 0.996013. Among the 21 Amerindian language treebanks, only the two Mbyá Guaraní treebanks, being error-free, achieve the maximum score of 1. On the other hand, in version 2.15 of the UD collection, none of the seven Portuguese treebanks reach the score of 1. Two of these treebanks achieve the minimum value of 0.01, with the scores of the others varying between 0.996013 and 0.988368, values obtained by UD_Portuguese-Porttinari and UD_Portuguese-Bosque, respectively.

Finally, the udeprels score reflects the use of the inventory of universal dependency relations by the treebank. This metric only considers the set of 37 main relations, such as subject (nsubj), direct object (obj), and nominal modifier (nmod), excluding subtypes like passive subject (nsubj:pass) and possessive modifier (nmod:poss). The calculation is analogous to the textual genre and part-of-speech tag scores, dividing the number of universal relations in the treebank by the maximum number of 37 universal relations. Version 2.15 of UD_Nheengatu-CompLin uses 33 of these 37 relations, as does UD_Portuguese-Porttinari, resulting in a score of 33 / 37 = 0.891891. Among the 21 Amerindian language treebanks in version 2.15 of the UD collection, only the two Nahuatl treebanks surpass Nheengatu in this

respect. With 35 items each, only UD_Portuguese-Bosque and UD_Portuguese-DANTEStocks, among the Portuguese treebanks, have a larger set of dependency relations.

To calculate the total score of a treebank, the sum of the partial scores in the eight dimensions is multiplied by the factors availability ∈ {0.01, 1} and validity ∈ {0.1, 1}. The first factor is applied at its minimum value to a treebank lacking sentence text; otherwise, the value is the maximum. In UD v2.15, among the 21 Amerindian language treebanks, only UD_Mbya_Guarani-Dooley falls into the first condition. The second factor takes the maximum value if the treebank is valid according to the validate.py program; otherwise, it takes the minimum value, a condition that applies to both UD_Mbya_Guarani-Dooley and UD_Kiche-IU. The number of stars of a treebank is computed using Equation 7.

$$stars = \frac{\lfloor score \times 10 + 0.5 \rfloor}{2} \quad (7)$$

Summing up, version 2.15 of UD_Nheengatu-CompLin attains the individual dimension scores reported in Table 4. Each dimension score is multiplied by its corresponding weight, yielding the weighted scores shown in the last column. The weights shown in Table 4 are the normalized versions of the raw dimension weights listed in Table 2, obtained by dividing each raw weight by their total sum 39, so that they sum to 1 and can be used to compute a weighted average score between 0 and 1. The resulting total score is 0.743181144774127, which—since both availability and validity factors equal 1—is converted into stars using Equation 7, yielding an overall rating of 3.5 stars.

Table 4 – Scores computed by evaluate-treebank.pl for version 2.15 of UD_Nheengatu-CompLin the eight evaluation dimensions

Dimension	Weight	Score	Weighted score
features	0.0769230769230769	1	0.0769230769230769
genres	0.0769230769230769	0.294117647058824	0.0226244343891403
lemmas	0.0769230769230769	1	0.0769230769230769
size	0.256410256410256	0.428353991978387	0.109834356917535
split	0.0512820512820513	0.34	0.0174358974358974
tags	0.0769230769230769	0.941176470588235	0.0723981900452489
udapi	0.307692307692308	0.969913891482519	0.298435043533083
udeprels	0.0769230769230769	0.891891891891892	0.0686070686070686
TOTAL			**0.743181144774127**

Fonte: elaboração própria.

4 Analysis and correction of the Udapi errors

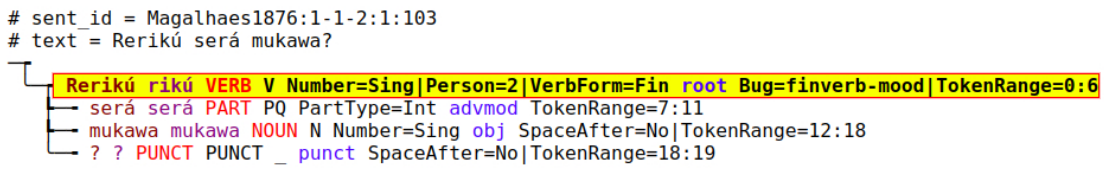
In this section, we detail each error type shown in Figure 17 and explain how we resolved the majority of them. Subsections 4.1, 4.2, and 4.3 address the corrected errors, which account for 98.4% of the total, while subsection 4.4 discusses the small remaining percentage that has yet to be corrected.

4.1 Underspecified verbal mood

The finverb-mood error occurs when verb forms annotated as finite lack the Mood feature, as illustrated in Figure 18. This type makes up 94.65% of the 2,726 errors flagged by Udapi in version 2.14 of UD_Nheengatu-CompLin. In the following, we summarize different accounts of Nheengatu's verb inflection. These generally converge on the observation that Nheengatu verb forms, abstracting away from historical or irregular imperative forms, only encode person and number. Tense and mood are conveyed by other means, typically particles and auxiliaries. This motivated our previous decision to specify mood as a morphological feature in the treebank only if the verb form is, regardless of context, unambiguously imperative. To address these errors, we have revised our annotation policy to always specify the Mood feature for finite verbs. In version 2.15 of the treebank, the number of such errors was reduced to zero.

2) Re-rikú será mukawa?
2SG.ACT-have PQ shotgun
'Do you have the shotgun?' (Magalhães, 1876, p. 14)

FIGURE 18 – Analysis of example (2) with a finverb-mood error detected by Udapi



Fonte: elaboração própria.

Following Hartt (1872), Magalhães (1876, p. 10) observes that, in Nheengatu, "tense roots are not yet incorporated into the verb or the attributive root, as occurs in inflectional languages"[5]. For both authors, the "undefined present", exemplified in (2), is decomposed into the personal prefix *re-* and the root *rikú* ("have"), forming other verb tenses through the auxiliary *ikú* ("be") and certain particles, such as *kurí* and *ana*[6]. Similarly, Sympson (1877, p. 11) states that "[t]he verbs of the *Brasilica* language never change their endings."[7] Except for the present indicative and the imperative, other moods, tenses, and verb forms are expressed through auxiliaries and "signs", which roughly correspond to the particles described by Magalhães (1876). Compare (2) in the "undefined present" with sentences (3)–(5) in the "defined present", future, and past tenses, respectively, according to Magalhães (1876).

3) Maã taá re-munhã re-ikú?
what CQ 2SG.ACT-do 2SG.ACT-be
'What are you doing?' (Magalhães, 1876, p. 28)

[5] Our translation from Portuguese: "as raízes de tempo ainda não estão incorporadas ao verbo, ou à raiz atributiva, como sucede nas línguas de flexão."

[6] Magalhães (1876, p. 10) includes the subordinator *ramé* ("when") among the particles forming verb tenses, whereas Hartt (1872) classifies this word as an adverb.

[7] Our translation from Portuguese: "[o]s verbos da língua brasílica nunca mudam de terminação."

4) Re-maã kurí ixé wirandé.
2SG.ACT-see FUT me tomorrow
'You will see me tomorrow.' (Magalhães, 1876, p. 73)

5) Mairamé taá re-maã ana ixé?
when PQ 2SG.ACT-see PFV me
'When did you see me?' (Magalhães, 1876, p. 73)

From the perspective of the grammatical exposition by Magalhães (1876), the notion of verbal mood, characteristic of languages like Portuguese, does not participate in the inflectional mechanism of Nheengatu. The same second-person singular personal prefix constitutes the only inflection of the main verb, regardless of whether the sentence is a question, statement, or command, as demonstrated in examples (2), (6), and (7). The illocutionary force of these sentences is determined by the syntactic structure, prosody, or discourse context rather than by verb inflection.

6) re-puká sé nhaã kunhã-etá r-enundé
2SG.ACT-laugh happy those woman-PL CONT-in_front_of
'you laughed happily in front of those women'

7) Re-tirika, yautí, kurumú xa-pirú indé.
2SG.ACT-get_out, tortoise, or_else 1SG.ACT-step_on you
'Get out, tortoise, or else I'll step on you.'

The term imperative occurs only once in Magalhães (1876, p. 94): "The verb *ço*, to go, forms the imperative *cóĩ*, which is read as: *cóin*."[8]. Sentences with this form, such as (8), are unequivocally imperative.

8) Re-kũi apigawa úkupi.
2SG.ACT-go[IMP] man house:LOC
'Go to the man's house.'

Unlike Magalhães (1876), imperative forms are part of the conjugation paradigms of both Hartt (1872) and Sympson (1877), as in examples (9)–(11).

9) saisú iné
[2SG.IMP]love 2SG
'You love!' (Sympson, 1877, p. 43)

10) pe-saisú penhẽ
2SG.ACT-love 2PL
'You guys love!' (Sympson, 1877, p. 43)

11) E-kũi e-sikináu nhaã kawasú yurú.
2SG.IMP-go 2SG.IMP-cover that bucket mouth
'Go cover the mouth of the bucket.' (Hartt, 1938, p. 320)

[8] Our translation from Portuguese: "*O verbo* ço, *ir, faz no imperativo* cóĩ, *que se lê*: cóin."

Hartt (1872) contrasts the prefix *e-* of the second-person singular imperative, represented according to Sympson (1877) by the absence of a personal prefix, with the corresponding prefix *pe-* of the second-person plural. This prefix, according to Sympson (1877, p. 43), is distinguished from the homonymous form of the present indicative "by usually adding the particle *penhê*"[9]. Hartt (1872) and Sympson (1877) discuss the irregular formation of the imperative of *sú* ("go"), exemplified in the second-person singular by *ekóiŋ* and *icúen*, respectively. Hartt (1872) also highlights the prohibitive particle *teñé*, which Sympson (1877, p. 11) spells as *ten* and calls an imperative "sign", used "when the verb is conjugated negatively"[10].

Following Magalhães (1876), Stradelli (1929, p. 40) states that "the conjugation paradigm of Nheengatu is reduced to the paradigm of the present tense"[11]. The same personal prefixes are used in all tenses and moods, rarely being omitted. The different tenses are formed either with the suffix *ana* or with adverbs such as *kwera*, *kurí*, etc., which are not integral parts of the verb. In his grammatical synopsis, Stradelli (1929, p. 41), differing from Magalhães (1876) but aligning with Sympson (1877), specifically addresses the formation of the imperative, which "is obtained with the present indicative used without the personal pronoun and lacking the first person"[12]. He exemplifies this with *rikú* (*recô*, "you have") and *urikú* (*orecô*, "he has")[13]. He highlights both the regional variation between the irregular imperative forms *rekũi* (*recoin*) and *kũi* (*coin*), on one hand, and the "imperfect" form *resú ana* (*resoana*), on the other, as well as the *i-* allomorph of the second-person singular prefix, which he observed in localities along the Solimões River.

Corroborating the observations of Sympson (1877) and Stradelli (1929), Avila (2021) includes *i-*, *u-*, and *ø-* as variants of ***e-***, the historical lemma corresponding to ***re-***, used in contemporary Rio Negro Nheengatu in both the indicative and imperative moods. Avila (2021) lists a large number of historical records of these variants, attested, for example, in the forms *eruri*, *ururi*, and *munhã* from examples (12)–(15). The second form coincides with the third-person indicative form, while the third corresponds to the basic verb form, devoid of inflection. Avila (2021) describes in specific entries different irregular imperative forms, such as *rekũi* or *ekũi*, covered by Hartt (1872), Magalhães (1876), Sympson (1877), and Stradelli (1929), also dedicating a separate entry to *yuri*, the irregular second-person imperative form of the verb *yuri* ("come"). These entries list diverse citations and historical spelling variants, most of which we incorporated into UD_Nheengatu-CompLin.

---

[9] The original Portuguese passage reads as follows: "A segunda pessoa do plural do imperativo diferença-se da segunda pessoa do presente do indicativo, em todos os verbos, por se lhe acrescentar usualmente a partícula — *penhê*."

[10] Our translation from Portuguese: "quando o verbo é conjugado negativamente."

[11] Our translation from Portuguese: "o paradigma da conjugação nheengatu se resume no paradigma do tempo presente."

[12] The original passage reads as follows: "O imperativo se obtém com o presente indicativo usado sem o pronome pessoal e com a falta da primeira pessoa."

[13] The original examples read as follows: "*Recô* — tenha; *Orecô* — tenha." There is probably a typo in the first example. The second-person singular prefix *re-* is missing. The form should be *rerikú* (*rerecô*), which appears on the same page in a past tense example. An Italian naturalized as a Brazilian, Stradelli (1929) translates the imperative second-person singular form of *rikú* ("have") into Portuguese *tenhas* instead of *tem*. The former is restricted to the negative imperative (Cunha; Cintra, 2017).

12) Mitú e-ruri se igara a-sú arama
curassow 2SG.IMP-bring my canoe 1SG.ACT-go for
a-ú tuyuka.
1SG.ACT-eat clay
'Curassow, bring my canoe so I can go eat clay.' (Rodrigues, 1890, p. 159)

13) I-ruri xa-ú ne piá.
2SG.IMP-bring 1SG.ACT-eat your heart
'Give me your heart to eat.' (Rodrigues, 1890, p. 36)

14) U-ruri Karã se muirakitã!
2SG.IMP-bring limpkin my talisman
'Limpkin, bring my talisman!' (Rodrigues, 1890, p. 132)

15) Munhã tatá, siusí-etá u-sema ikú.
[2SG.IMP]make fire, Pleiades 3SG.ACT-go.out[INF]be
'Make the fire, the Pleiades are rising.' (Rodrigues, 1890, p. 174)

The specific morphological markings of the affirmative imperative have not survived in contemporary Rio Negro Nheengatu, as described by Casasnovas (2006), Cruz (2011), Navarro (2016), and Avila (2021). Only the marking of the negative imperative has been preserved, through *tenhẽ* or *te* [te], the latter form classified by Cruz (2011) as a clitic and by Avila (2021) as a particle, as in example (16).

16) Te re-kiri, Aline!
PROH 2SG.ACT-sleep, Aline
'Don't sleep, Aline!' (Cruz, 2011, p. 407)

Since the Nheengatu verb, as a rule, does not express verbal mood by itself, the morphological analysis of examples such as (2)–(7) in UD_Nheengatu-CompLin as of UD v2.14 does not include this feature. It limits itself to the Number, Person, and VerbForm features. The latter allows distinguishing between finite and non-finite forms, such as infinitive, participle, and gerund (see Section 4.2). Only specific imperative forms, such as in (8) or (11), are annotated with the Mood feature. In the case of the negative imperative with the particle *te* or the historical variant *tenhẽ*, the Mood feature applies to the particle, while the verb remains underspecified regarding mood. Table 5 displays the frequencies of mood specifications by word classes. Only a tiny fraction of auxiliaries and full verbs, totaling 14 occurrences, have the Mood feature, in this case with the imperative value, while 18 particles are specified for the imperative mood and 14 for the conditional mood, expressed in Nheengatu by the *amú* and *maã* particles. The latter is exemplified in (17).

Table 5 – Parts of speech and mood frequencies in UD_ Nheengatu-CompLin v2.14. The dash indicates that the feature is absent.

Part of Speech	Mood	Frequency
AUX	Imp	2
PART	Cnd	14
PART	Imp	18
VERB	Imp	12
AUX	—	356
VERB	—	2403

Fonte: elaboração própria.

In the analysis of (17), we follow Navarro (2016) and Avila (2021). The former distinguishes, in Nheengatu, three types of conditional mood, namely, (i) real hypothesis, (ii) possible hypothesis, as in the example at hand, and (iii) unreal hypothesis. According to Avila (2021), *maã* in this example functions as a conditional particle, applying to both the protasis and the apodosis. Cruz (2011, p. 502, 385), on the other hand, analyzes this word as a subordinator, assigning it the function of a "marker of hypothetical modality"[14], a classification that does not seem to align with her definition of subordinators as "specialized marks for indicating the syntactic dependency between two clauses."[15]

17) mira ramé maã indé, indé intí maã re-xari ixé xa-manú
human if COND you, you NEG COND 2SG.ACT-let me 1SG.ACT-die
'if you were human, you wouldn't let me starve.' (Amorim, 1928, p. 28)

To fix the finverb-mood errors flagged by Udapi, we automatically included the Mood feature in all verb forms. In structurally unambiguous cases, such as in interrogative sentences, adverbial subordinate clauses, or with the negative imperative particle *te* (or *tenhẽ*), we used the values Ind and Imp for indicative and imperative, respectively. Otherwise, we used the value Imp,Ind to indicate that the verbal form may be either imperative or indicative, requiring manual disambiguation, something to be done in a next version of the treebank.

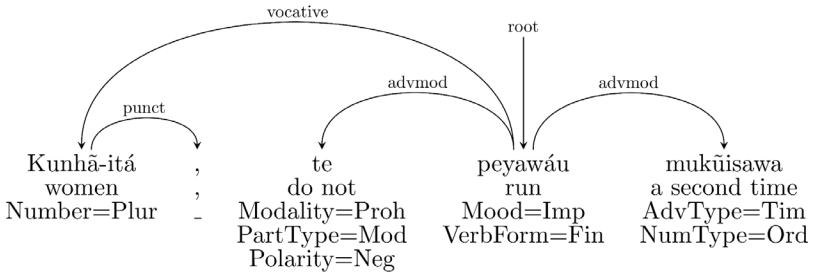
On the other hand, we removed the Mood feature in the FEATS field of the negative imperative and conditional particles, while maintaining the other features. The former is additionally annotated with features specifying it as a negative polarity negation particle, while the latter is characterized as a modal particle. Instead of Mood=Imp and Mood=Cnd, we now annotate these two particles with Modality=Proh, as exemplified in Figure 19, and Modality=Cond, respectively. The abbreviation proh is widely used in linguistic typology to designate the prohibitive modality (Bybee; Perkins; Pagliuca, 1994; Palmer, 2001), a practice that seems to underlie hartt (1872)'s classification of *teñé* as a prohibitive particle. The substitution of Mood by Modality aligns with the UD documentation, which restricts the application of the Mood feature to verbs, defining Mood as "a feature that expresses modality and

[14] Our translation from Portuguese: "marcador de modalidade 'hypotética' [*sic*]."

[15] Our translation from Portuguese: "marcas especializadas na função de marcar a dependência sintática entre duas orações."

subclassifies finite verb forms" (Marneffe *et al.*, 2024a), although the validator does not prohibit its application to particles.

Figure 19 – Annotation of the prohibitive particle *te* and an imperative verb in version 2.15 of the UD_Nheengatu-CompLin treebank. The example is a fragment from a sentence of Amorim (1928, p. 23–24).



Fonte: elaboração própria.

The Modality feature is an innovation of UD_Nheengatu-CompLin version 2.15. The treebanks of Egyptian and Polish limit themselves to annotating modal particles with PartType=Mod, while the Bulgarian and Czech treebanks leave the FEATS field empty for the epistemic modality particle and the deontic modality particle, respectively, that, according to UD documentation, express sentential modality.

4.2 Missing verb form specification

The no-VerbForm error consists of a verb without the VerbForm feature, which characterizes finite forms, infinitives, participles, etc. This type makes up 3.23% of the bugs flagged by Udapi in UD_Nheengatu-CompLin version 2.14. The rationale underpinning our annotation practice in this regard was similar to that behind the finverb-mood errors. Analogously, we only annotated a verb with the VerbForm feature when it formally encodes this property, as is the case in (18) with *yamburi*, but not *membeka*.

18) Ya-mburi maniáka paranã upé i
1PL.ACT-put manioc river in 3SG.INACT
membeka arama.
become_soft to
'We put the manioc in the river to soften it.' (Moore; Facundes; Pires, 1994, p. 105)

Hartt (1872), Magalhães (1876), and Stradelli (1929) highlight a restricted group of auxiliary verbs that do not take personal prefixes when following another verb in the same sentence. An example of this particularity is *putari* ("want") in (19). Magalhães (1876, p. 11) explains: "The verb *putári* 'want' has a very peculiar way of appearing in the sentence; whenever it comes together with another verb, it is the other verb that receives the pronominal prefix, while *putári* remains unchanged [...]."[16]

[16] Our translation from Portuguese: "O verbo *putári* querer, [*sic*] tem um mui singular modo de figurar na oração; sempre que ele vem junto com outro verbo, é esse outro verbo que recebe o prefixo pronominal, ao passo que

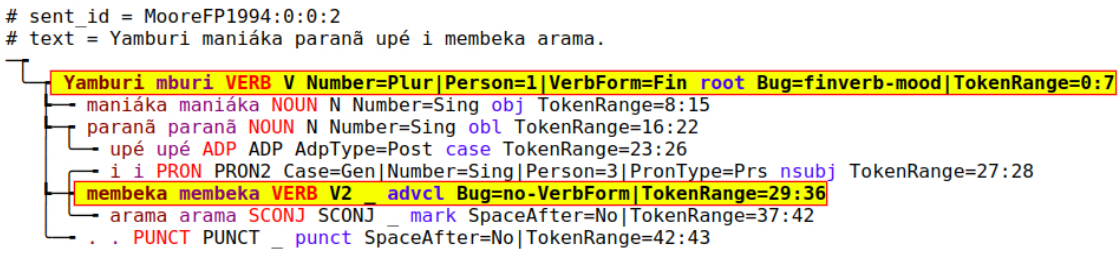
19) Re-mixiri-putari será pirá?
2SG.ACT-grill-want[NFIN] PQ fish
'Do you want to grill fish?' (Magalhães, 1876, p. 96)

The verb *putari* ("want") is conjugated when functioning as the main verb, as in (20). Based on the contrast in pairs like *putari* and *xaputari* in (19) and (20), the morphological analysis of active verbs in UD_Nheengatu-CompLin version 2.14 included the VerbForm feature, with the values Inf and Fin for non-finite and finite forms, respectively.

20) Intí maã; xa-putari yepé panera xa-memúi arama.
no thing 1SG.ACT-want one pot 1SG.ACT-cook to
'No; I want a pot to cook in.' (Magalhães, 1876, p. 96–97)

The spelling of inactive prefixes is one of the many aspects in which Nheengatu orthographies differ. In this regard, we follow the approach of Navarro (2016) and Avila (2021), adopted by Yamã *et al*. (2021), treating these elements as second-class pronouns, spelled separately from the verbal root, as in (18)[17]. As a consequence of this decision, the FEATS field of inactive verbs was left without any morphological features since these verbs are invariable. This led to 88 instances of the no-VerbForm error detected by Udapi (Figure 20).

Figure 20 – Analysis of example (20) with no-VerbForm errors identified and highlighted by the Udapi tool



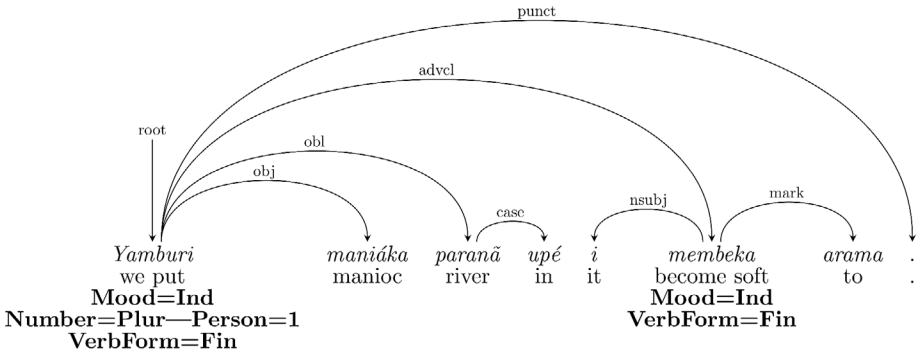
Fonte: elaboração própria.

In the UD_Nheengatu-CompLin version 2.15, all these errors were corrected, providing the missing Mood and VerbForm features, as exemplified in Figure 21. We revised our annotation policy to ensure that the VerbForm feature is consistently included in the FEATS field of all verbs.

ele fica invariável [...]."

[17] Cruz (2011), on the other hand, treats these elements as agreement morphemes prefixed directly to the verbal roots. This spelling appears to be more widespread in the Rio Negro region, e.g., Casasnovas (2006).

Figure 21 – Corrected analysis of the example in Figure 20



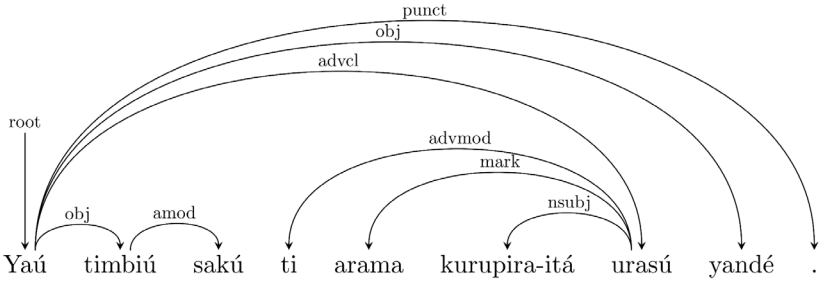
Fonte: elaboração própria.

4.3 Multiple objects

Udapi detected 15 multi-obj errors in version 2.14 of UD_Nheengatu-CompLin. These errors fall into two subtypes. In the first subtype, a verb governs two direct objects within the same sentence, as shown in Figure 22. The error in this tree resulted from the oversight of the human annotator, who incorrectly identified the head verb of the sentence as the governor of the embedded verb's object. The remaining errors of this subtype arose from analogous mistakes in identifying the governor or the syntactic relation of a nominal, which, for example, was analyzed as a direct object instead of a nominal complement when preceding another in a possessive genitive relation. We eliminated all errors involving double direct objects by modifying the governor or the syntactic relation of the spurious direct object.

21) Ya-ú t-imbiúsakú ti arama kurupira-itá
1PL.ACT-eat ABS-food hot not to cupurira-PL
u-rasú yandé.
3.ACT-take_away us
'We would eat hot food for the curupiras not to take us away.' (Moore; Facundes; Pires, 1994, p. 108)

Figure 22 – Dependency tree of (21) with a verb incorrectly governing two direct objects

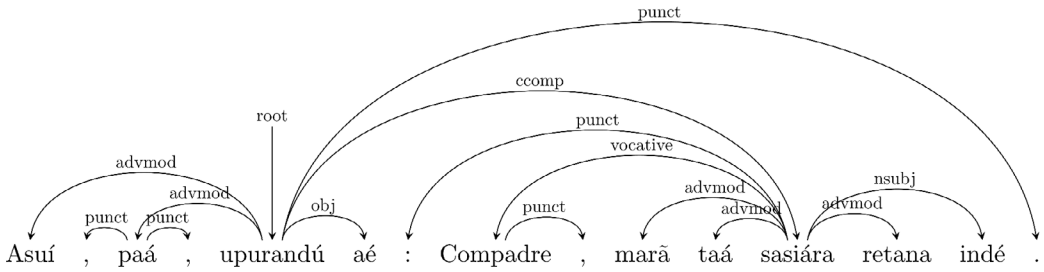


Fonte: elaboração própria.

In the second subtype, illustrated in Figure 23, the same verb governs both a direct object and a ccomp clausal complement. This subtype accounts for about two-thirds of the multi-obj instances in the Udapi report. It involves sentences with communication verbs such as *purandú* (“ask”) and *suaxara* (“answer”), where the recipient semantic role is realized by a bare nominal, that is, without an adposition, and the patient by the ccomp clausal complement.

22) Asuí, paá, u-purandú aé: Compadre, marã taá sasiára retana indé.
and RPRT 3SG.ACT-asked him: Compadre, why CQ sad very you
‘And he asked him: Compadre, why are you so sad?” (Casasnovas, 2006, p. 66)

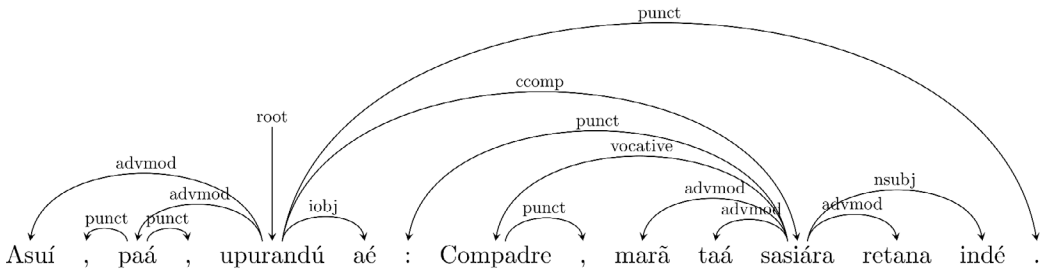
Figure 23 – Dependency tree of (22) with a verb incorrectly governing a direct object and a complement clause



Fonte: elaboração própria.

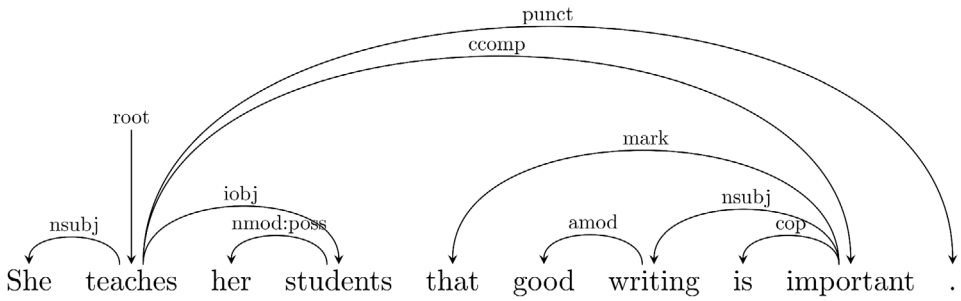
The solution in this case was to analyze the nominal that realizes the recipient not as a direct object, but as an indirect object, as shown in Figure 24. This solution coheres with the analysis in the UD documentation for an analogous example in English (Figure 25).

Figure 24 – Corrected dependency tree of (22) with the main verb governing an indirect object and a complement clause



Fonte: elaboração própria.

Figure 25 – Dependency tree for an example from the UD documentation with a verb governing an iobj and a ccomp generated by UDPipe 2 using the english-ewt-ud-2.15-241121 model



Fonte: elaboração própria.

The type of indirect object exemplified in Figure 24 was not part of the annotation scheme of UD_Nheengatu-CompLin before the revision triggered by the errors discussed in this section. The assignment of the iobj syntactic relation was limited to dative case pronouns, as in (23), or nominals governing the postpositions *arama* and *supé*, which Cruz (2011) classifies as denoting the "prospective" (or "intralocutive") dative and the "extralocutive", respectively, as illustrated in (24) – (26).

23) E-purú ne kiwawa ixé-u.
2SG.IMP-purú your comb 1SG-DAT
'Lend me your comb.' (Hartt, 1938, p. 319)

24) E-purú xinga ixé arama.
2SG.IMP-purú some me to
'Lend me some.' (Magalhães, 1876, p. 201)

25) "Mamãe", u-nheẽ-wera ixé arã nhaã se kurumĩ.
mom 3SG.ACT-say-FREQ me to that my boy
"Mom," that little boy of mine always used to say to me.' (Cruz, 2011, p. 217)

26) "Kuíri-ta?" u-nheẽ i mú supé.
now=PQ 3SG.ACT-say his brother to
'"What now?" he said to his brother.' (Cruz, 2011, p. 217)

The UD documentation on the obj and ccomp syntactic relations does not explicitly forbid a verb to govern an obj and a ccomp simultaneously. Similarly, the validate.py program does not flag an annotation error in this case, which only Udapi does[18]. The reason for excluding this valency frame lies in the parallelism between obj and ccomp, which is manifested in alternations such as in (27)–(29) (Levin, 1993, p. 205). From the perspective of Frame Semantics, in all of

[18] Udapi's ud.MarkBugs block was based on constraints listed on a legacy page at https://universaldependencies.org/svalidation.html (Popel; Žabokrtský; Vojtek, 2017). We are not aware of any direct reference to this page in the main UD annotation guidelines. However, the main UD documentation includes Nivre *et al.* (2017) among the UD project publications, and this tutorial mentions the URL. According to the tutorial, a previous version of the validation system used these constraints. One of them states: "No predicate can have more than one direct object. Ccomp counts as [a] direct object." It is unclear to us why subsequent versions of the validator discarded this constraint.

these sentences, the verb expresses the same elements of the Communication_manner frame: SPEAKER, ADDRESSEE, and MESSAGE (Ruppenhofer *et al.*, 2016), differing only in the syntactic expression of the latter. These elements correspond to the more general semantic roles of Agent, Recipient (or Goal), and Theme (Levin, 1993; Perini, 2015, 2019). In (27), the MESSAGE element is realized by an NP, whereas in the other two examples, it is expressed through a subordinate clause introduced by the complementizer *that* and through direct speech, respectively. Zwicky (1971) classifies all three of these syntactic realizations as direct objects of the verb, constituting mutually exclusive forms of expression of the same semantic element.

27) Susan whispered the news to Rachel.

28) Susan whispered to Rachel that the party would be tonight.

29) Susan whispered to Rachel, "Leave the room."

Cruz (2011, p. 424) denies the existence of ditransitive verbs in Nheengatu, arguing that the Beneficiary semantic role is compatible with any predicate. However, this claim lacks sufficient justification when examined through the lens of semantic role theory and valency theory. First, it is essential to distinguish between the semantic roles of Recipient or Goal, as in (27), and Beneficiary, as in the following example (Levin, 1993, p. 49):

30) Martha carved a toy for the baby.

Second, in communication and possession-transfer verbs, the syntactic relation that realizes the recipient of the message or the transferred entity functions as an argument of the verb's logical-semantic structure, alongside the arguments corresponding to the Agent and the Theme (Helbig, 1992). Similarly, in Frame Semantics, the three elements SPEAKER, ADDRESSEE, and MESSAGE are classified as core arguments.

4.4 Unsolved errors

Of the 2726 bugs flagged by Udapi in version 2.14 of UD_Nheengatu-CompLin, we failed to correct only 43. In 4.4.1, we address the 36 errors triggered by nouns with the diminutive or augmentative degree. In 4.4.2, we discuss the det-upos and mark-upos errors, identified in a total of seven sentences. These cases involve the incompatibility between the part-of-speech tag of the head of an exocentric multiword expression and the syntactic relation assigned to that expression.

4.4.1 Errors involving degree features

With 36 occurrences, the degree-upos error results from assigning the degree feature to nouns. The program only accepts this feature for adjectives and adverbs, as shown in Listing 1:

Listing 1 – Validation of the Degree feature against UPOS values

```
if feats['Degree'] and upos not in ('ADJ', 'ADV'):
    self.log(node, 'degree-upos',
             'Degree=%s␣upos!=ADJ|ADV␣(but␣%s)' % (feats['Degree'], upos))
```

Fonte: elaboração própria.

This rule appears to contradict the UD guidelines, which state:

> Degree of comparison is typically an inflectional feature of some adjectives and adverbs. A different flavor of degree is diminutives and augmentatives, which often apply to nouns but are not restricted to them (Marneffe *et al.*, 2024a).

In compliance with this explanation, the Degree feature includes, among others, the values Dim and Aug, which apply to morphologically derived forms of nouns that express, respectively, "small size, or, metaphorically, affection towards the entity described by the noun" and "large size or force", such as *appeltje* ("little apple" in Dutch) and *apartamentão* ("big apartment" in Portuguese) (Marneffe *et al.*, 2024a). The UD documentation emphasizes that this feature also applies to verbs and adjectives in some languages. According to Rio-Torto (2015), diminutives and augmentatives are part of the formation of evaluatives, which in Portuguese is both isocategorial and pluricategorial, meaning it preserves the category of the base, which may belong to different categories, namely nouns, adjectives, adverbs, verbs, and pronouns.

According to Cruz (2011, p. 242), the derivational suffixes "*miri* 'diminutive' and *wasu* 'augmentative' allow the creation of new nouns from nominal bases" to refer to "entities with larger or smaller dimensions than the prototypical entity designated by the simple base noun"[19], as in (31) and (32).

31) U-munhã kurusa-mirĩ-etá Kurupira r-apé upé.
3.ACT-make cross-DIM-PL Curupira CONT-path on
'They made little crosses on Curupira's path.' (RODRIGUES, 1890, p. 78)

32) Ape ana tẽ paá tipusi-wasú u-mungiri aé.
then PFV FOC RPRT sleep-AUG 3.ACT-make_sleep her
'At the same moment, they say, a great sleep put her to sleep.' (Amorim, 1928, p. 175)

Following Cruz (2011), Avila (2021) classifies *wasú* as an augmentative suffix but diverges from her regarding the status of *mirĩ*. Avila (2021) classifies it as an adjective or adverb, writing it separately from the word it modifies, although acknowledging that it is in the process of grammaticalization as a suffix, given its combination with the plural particle *itá* (or *etá*), as in (31).

Cruz (2011) documents the use, albeit limited, of *miri* as an independent word, as does Avila (2021). The latter observes the incipient nominalization of *wasú* as an adjective. Previously, Stradelli (1929, p. 33) noted the ambivalent nature of these two morphemes, func-

[19] The original Portuguese text reads as follows: "Os morfemas derivativos *miri* 'diminutivo' e *wasu* 'aumentativo' permitem criar novos nomes a partir de bases nominais [...] esses novos nomes designam entidades com dimensões maiores ou menores do que a entidade prototípica, designada pelo nome simples de base."

tioning either as a "qualifying adjective" or as an adverbial modifier of another adjective[20], or as diminutive and augmentative suffixes, as in (33)–(36).

33) Se s-etama ne r-etama mirĩ piri.
my NCONT-country your CONT-country small CMPR
'My homeland is smaller than yours.' (Stradelli, 1929, p. 28)

34) Se t-uixawa amú t-uixawa-etá wasú piri.
my NCONT-chief other NCONT-chief-PL big CMPR
'My chief is bigger than the other chiefs.' (Stradelli, 1929, p. 28)

35) parana-miri-etá
river-DIM-PL
'tributaries' (Stradelli, 1929, p. 31)

36) pira-usu-etá
fish-AUG-PL
'wales' (Stradelli, 1929, p. 31)

Stradelli (1929) also highlights the reduction of *wasú* to *asú* in certain contexts, such as in *cunhãmucuasú* (Port. "mocetona", Engl. "big girl") and *cáuasú* (Port. "vespa grande", Engl. "large wasp"), the augmentative forms of *cunhãmucú* (Port. "moça", Engl. "girl") and *cáua* (Port. "vespa", Engl. "wasp")[21]. In our view, this allomorphy further supports the affixal nature of the element.

Regardless of the status of the morphemes *mirĩ* and *wasú*, the bound morpheme *-í* productively suffixes to nouns in 19th-century Nheengatu (Stradelli, 1929), as in (37), corroborating the need to include the degree feature in the morphological analysis of the respective derived words. For this reason, we prefer to maintain this aspect of our annotation scheme, which, as we have shown, aligns with UD guidelines, despite the errors flagged by Udapi.

37) aresé pe-yeréu makaka-í aramapanhẽ ara upé
therefore 2PL.ACT-turn monkey-DIM to all life LOC
'therefore, you will become little monkeys forever' (Magalhães, 1876, p. 171)

4.4.2 Errors involving exocentric multiword expressions

Among the six error types identified by the Udapi tool, there were only four occurrences of det--upos and three of mark-upos. These two types of errors involve the multiword expressions *mayé waá* and *waá upé*, which function as det and mark, respectively. These syntactic relations are deemed incompatible with the part-of-speech tags of the heads of the corresponding multiword expressions (MWEs). Below, we first address the mark-upos errors, triggered by the classification of *waá*, the head of the MWE *waá upé*, as a relative pronoun, an analysis that contradicts the approaches of Cruz (2011) and Avila (2021), requiring a more in-depth discussion.

[20] The original Portuguese text reads as follows: "O adjetivo qualificativo admite, tal como o substantivo, três graus de qualificação: diminutiva, aumentativa e superlativa. Aumentativo, superlativo, diminutivo e adjetivo que modifica outro adjetivo é sempre posposto a este."

[21] In the original text by Stradelli (1929, p. 30), the first augmentative lacks the acute accent that marks the final syllable of the other examples of these formations.

The mark-upos errors occur in sentences such as (38) and (39), where the relative pronoun *waá* is governed by a verb under the syntactic relation mark, which is typically assigned to conjunctions.

38) Sampaio u-sendú nheengatú puapuãmu waá upé
Sampaio 3.ACT-hear Nheengatu wander REL LOC
Barcelos upé.
Barcelos LOC
'Sampaio heard Nheengatu while wandering in Barcelos.' (Avila, 2021, p. 802)

39) Ti waá upé yawaraté u-sarú, mairamé u-saã
not REL LOC jaguar 3.ACT-expect, when 3.ACT-feel
u-suú ana [...].
3.ACT-bite PRF
'Unexpectedly, the jaguar felt itself being bitten [...].' (Casasnovas, 2006, p. 70)

According to the code in Udapi's ud.MarkBugs block in Listing 2, mark-upos errors are flagged exactly when PRON and mark occupy the UPOS and DEPREL columns of the annotation for a word, i.e., when a pronoun functions as a conjunction.

Listing 2 – Validation of the mark relation against UPOS value

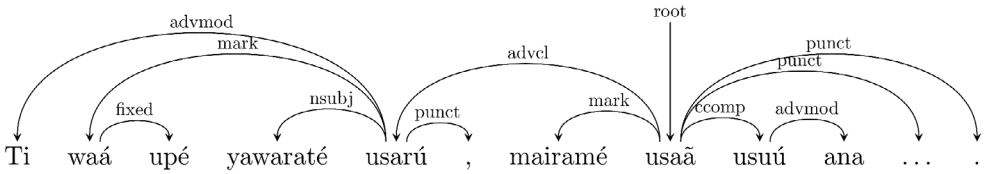
```
1 if udeprel == 'mark' and upos == 'PRON':
2     self.log(node, 'mark-upos', 'deprel=mark␣upos=PRON')
```

Fonte: elaboração própria.

Following Avila (2021), the combination of the relative *waá* with the locative postposition *upé* in (38) and (39) makes up a MWE equivalent to the Portuguese conjunction *quando* ("when"). This expression exhibits the typical behavior of postpositional subordinating conjunctions in Nheengatu, which precede the predicate of the clause when it falls under the scope of negation.

In UD, the first member of an MWE functions as the head, to which the following members are attached via the fixed syntactic relation (Figure 26). It is in the very nature of these constructions to engage in syntactic relations that are incompatible with the grammatical class of the expression's head (Constant *et al.*, 2017). To ensure compatibility between determined word classes and specific syntactic relations, UD introduced, after the release of version 2.14 on 15/05/2024, the ExtPos attribute to encode, in the FEATS field, the "external" tag of an exocentric MWE. Following this change, we began annotating *waá upé* with the feature ExtPos=SCONJ, without which the annotation of sentences like (38) and (39) triggers a validation error in the current validator version. However, Udapi continues to flag mark-upos errors for occurrences of *waá upé* in UD_Nheengatu-CompLin.

Figure 26 – Dependency tree of example (39)

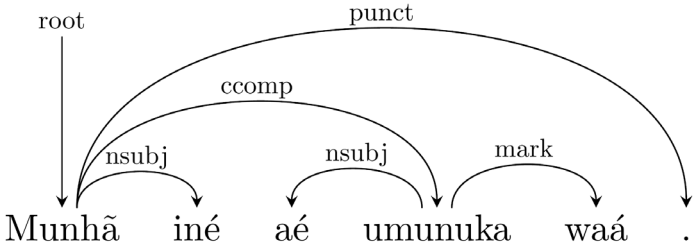


Fonte: elaboração própria.

This mismatch between the validation tool and Udapi suggests the need for an update to the latter's code, something we hope will be addressed soon. Otherwise, one way to resolve the mark-upos errors would be to annotate the relative not as a pronoun but as a subordinating conjunction. Indeed, instances of *waá* as a subordinator occur in UD_Nheengatu-CompLin, as shown in Figure 27. However, this use of *waá* as a complementizer is not documented by Avila (2021). It appears in UD_Nheengatu-CompLin only in this example and another from Casasnovas (2006, p. 40), seemingly constituting a calque from Portuguese.

40) Munhã iné aé u-munuka waá.
[2SG.IMP]make you he 3.ACT-cut that
'You make him cut it.' (Seixas, 1853, p. XII)

Figure 27 – Dependency tree of example (40)



Fonte: elaboração própria.

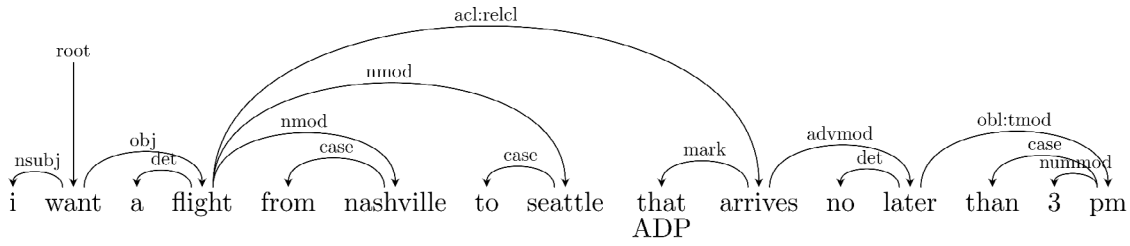
According to the UD documentation (Marneffe *et al.*, 2024a), "one needs to distinguish between relative clause markers, which are mark, from relative pronouns such as [en] *who* or *that*, which fill a regular verbal argument or modifier grammatical relation." In UD, relative pronouns are generally annotated with the PRON tag and the feature PronType=Rel, as in UD_Portuguese-Porttinari (Duran, 2021).

However, the classification of *that* as a relative pronoun is not unanimous in examples like (41), extracted from UD_English-Atis. As the only one among the six largest English treebanks in version 2.15 of the UD collection, UD_English-Atis classifies *that* in this type of example as an adposition linked via mark to the verb of the relative clause. According to Andrews (2007, p. 218), "[i]t is generally assumed *that* in relative clauses is not a relative pronoun."

41) i want a flight from nashville to seattle that arrives no later than 3 pm

In UD_English-GUM, we found 690 occurrences of *that* as a relative pronoun, including examples analogous to (41). However, there are 6 occurrences of *that* as SCONJ in relative clauses, such as (42).

Figure 28 – Dependency tree of (41)



Fonte: elaboração própria.

42) We've seen mass strikes all around the world, in countries that we wouldn't expect it.

The treatment of the relative *waá* in Nheengatu descriptions, regarding the dichotomy established in UD between relative clause markers and pronouns, is not uniform. While Hartt (1872) and Sympson (1877) classify this element as a relative pronoun and a "relative adjective or pronoun", respectively, Stradelli (1929) refers to it as a "conjunctive adjective."

According to Moore, Facundes e Pires (1994), *waá* belongs to the set of subordinating particles, alongside elements such as *ramé* and *arama*, which mark temporal and purposive subordinate clauses, respectively. Relative clauses with *waá* conform to the structural model of subordinate clauses that Moore, Facundes e Pires (1994) designate as "indigenous", in which the subordinating particle immediately follows the head of the VP, which, in example (43), is the adjective *nharú* ("wild").

Similarly, Cruz (2011) classifies *waá* as a relativizing particle, used to construct clauses with the same degree of subordination as subordinators. Avila (2021) refers to *waá* both as a "relativizing particle" and a "nominalizing particle." Both highlight that it admits the plural suffix *-itá* and precedes the predicate of the clause when under the scope of a negation particle. Navarro (2016), in turn, adopts the traditional label of relative pronoun. Finally, Melgueiro, Cabral e Martins (2019), despite calling *waá* a relativizing particle, acknowledge that it is responsible for the "retrieval of the relativized nominal", implementing the "relative pronoun" strategy described by Givón (1990, p. 187), in which the relative pronoun also functions as a subordinating morpheme.

43) Asuí aikwé yuíri makú-itá nharú waá-itá kaá rupí.
and EXST also Indian-PL fierce REL-PL forest PERL
'And there were also fierce Indians in the woods.' (Magalhães, 1876, p. 171)

According to Cruz (2011, p. 514), in headless or free relative clauses, "the head of the modified noun phrase is not explicit, or, depending on the adopted theory, the noun phrase is expressed by a zero"[22], as in (44). In the analysis she proposes for this example, the relativi-

[22] Our translation from Portuguese: "o núcleo do sintagma nominal modificado não é explicito, ou, dependendo da teoria adotada, o sintagma nominal é expresso por um zero."

zing particle is reduced to the clitic *wa*, which adjoins to *ita*[23]. Cruz (2011) analyzes this host as a plural particle with scope over the noun phrase. The relative clause with the plural particle functions as a genitive complement of the noun *ara* ("day").

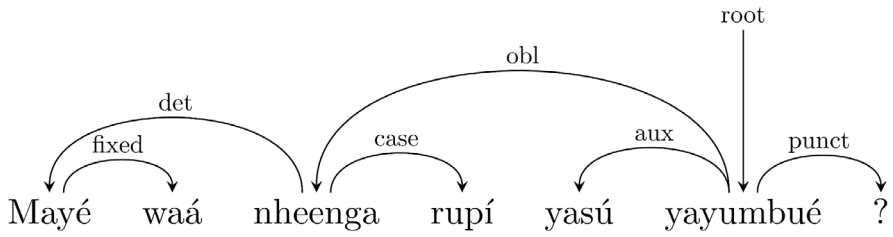
44) uwiara [u-manu wa]=ita ara
today 3.ACT-die REL=PL day
'Today is the day of those who died.' (Cruz, 2011, p. 515)

In UD_Nheengatu-CompLin, we prefer to maintain the analysis of the relativizer *waá* as a relative pronoun rather than a relativizing particle, despite the errors pointed out by Udapi in sentences with *waá upé*. Since UD is a lexicalist theory that tends to reject empty categories, this analysis seems to align better with the role of *itá* as a pluralizer of the noun phrase. Otherwise, one must postulate an empty nominal head in (44) to assign a noun phrase status to the expression in brackets. Instead, we argue that the head of this phrase is the relativizer, which, like the proximal and distal demonstrative pronouns *kwá* and *nhaã* and nouns, can be pluralized by *itá*, which, following Moore, Facundes e Pires (1994), among others, we treat as a plural suffix.

The det-upos errors result from assigning the syntactic relation det to *mayé* as head of the MWE *mayé waá*, as in (45), whose analysis is depicted in Figure 29. This MWE functions as a determiner or pronoun (Avila, 2021). Since *mayé* as an independent word is an interrogative adverb, a conflict arises with a specification in Udapi's code constraining det to pronouns and determiners. Similar to the ExtPos=SCONJ feature that specifies the external part-of-speech tag of *waá upé*, Udapi ignores the ExtPos=DET feature that we inserted in the FEATS field for examples with *mayé waá* to comply with the current version of the validator.

45) Mayé waá nheenga rupí ya-sú ya-yu-mbué?
how REL language PERL1PL.ACT-go 1PL.ACT-MID-study
'Through which language are we going to study?' (Avila, 2021, p. 465)[24]

Figure 29 – Dependency tree of example (45)



Fonte: elaboração própria.

[23] This morpheme is spelled *itá* in the orthography of Avila (2021).

24 *Apud* Oliveira and Schwade (2012, p. 82).

5 Final remarks

In just a decade, the UD collection has grown to 296 treebanks covering 168 languages, as of the 2.15 release. This unprecedented milestone for such a resource highlights the growing interest in multilingual parsing and the increasing adoption of UD as a standard for morphosyntactic annotation. Brazilian Indigenous languages, which were initially neglected in natural language processing, corpus linguistics, and computational linguistics— fields traditionally focused on major languages like Portuguese—now represent 4.7% of the treebanks and 8.3% of the languages in UD. Additionally, the collection includes seven more Amerindian language treebanks.

Despite these percentages, the vast linguistic diversity of Brazil consisting of circa 150 living languages according to Storto (2019) remains underrepresented in UD, as is the case with other natural language processing domains. Another challenge faced by these minority languages, most of which are endangered, is data scarcity. Most of the 14 treebanks for Brazilian Indigenous languages are characterized by their small size: five contain fewer than 1,000 tokens, another five range between 1,000 and 5,000, and two between 5,000 and 10,000. Only UD_Mbya_Guarani-Dooley and UD_Nheengatu-CompLin exceed 10,000 tokens, the latter surpassing the former by 61.7% in size.

Another major discrepancy between Brazilian Indigenous language treebanks and those for majority languages like Portuguese results from their evaluation. Except for UD_Nheengatu-CompLin, these treebanks fall within the lower end of the star ranking system, receiving ratings between 0 and 1.5. Ironically, UD_Mbya_Guarani-Dooley is the only treebank in this group with a 0-star rating. In UD's most recent release, ratings range from 0 to 4.5 stars. A treebank's star rating is based on a score derived from multiple factors, including its size, annotation richness, textual genre diversity, and validation status.

This paper has detailed our efforts to improve UD_Nheengatu-CompLin's rating. In UD v2.14, the treebank received the highest score among all 21 Amerindian language treebanks, followed by UD_Mbya_Guarani-Thomas. Along with two modern Nahuatl treebanks, these two formed a group of four treebanks rated at 2.0 stars. Between UD v2.14 and v2.15, we raised UD_Nheengatu-CompLin's rating to 3.5 stars. Since UD v2.13, it has been the largest of all 21 treebanks for Amerindian languages. With the v2.15 update, it also became the only one in this group to surpass 2 stars.

To achieve this result, we focused on correcting errors detected by the Udapi framework. The number of words per error is the most heavily weighted factor in computing a treebank's score. We managed to reduce these errors drastically. While in UD v2.14, there was one error every 5.5 words, in UD v2.15, the ratio improved to one error per 332.38 words. As a result, the Udapi score increased from a minimum of 0.01 to 0.969913. Between versions 2.14 and 2.15, not only did we reduce errors, but we also continued expanding UD_Nheengatu-CompLin, as in previous update cycles. The word count grew by 78%, resulting in a 9.17% increase in the size score and a 4.87% increase in the part-of-speech tag score. Together, these improvements boosted UD_Nheengatu-CompLin's rating by 71.2%, elevating it to 3.5 stars.

As noted by 19th-century grammarians, including Hartt (1872), Magalhães (1876), and Sympson (1877), the Nheengatu verb in general does not express in itself tense or mood distinctions, only inflecting for person and number. Except for certain irregular forms and the

now-obsolete second-person singular imperative prefixes, interpreting a verb form as indicative or imperative depends on immediate syntactic context, prosody, or pragmatic factors. In addition, finite verbs occur in constructions where languages like Portuguese employ participles and infinitives. Only a few auxiliaries incorporate uninflected into the main verb, establishing a formal contrast with the corresponding inflected forms as full verbs.

A total of 97.87% of the errors flagged by Udapi resulted from our initial treatment of Nheengatu's verbal inflectional morphology. These errors involve the absence of Mood and VerbForm features in verb forms without a corresponding formal marking. These errors were systematically corrected by adding the missing features. Our annotation policy was revised accordingly. In addition to these corrections, we also eliminated the multi-obj errors, which made up 0.55% of the total error instances. Some of these cases resulted from human annotation oversights. Another part led to a revision of the annotation policy, whereby bare nominals expressing recipients of communication verbs with complement clauses are now treated as indirect objects. The remaining errors represent a mismatch between the constraints implemented in Udapi and UD's annotation guidelines. Therefore, we have postponed handling them until further discussion clarifies this matter.

In preparation for UD v2.16, we have raised the rating of the February 12, 2025 development version of UD_Nheengatu-CompLin to 4 stars, primarily due to a 97.06% increase in the split score. In UD v2.15, only 44 treebanks, or 14.9% of the UD collection, achieved a 4-star rating, covering 23 languages[25]. All these languages are European. Except for Catalan and Scottish Gaelic, they are the majority languages of sovereign states, with the highest vitality index, i.e. 1, on the EGIDS scale. Catalan and Scottish Gaelic are statutory provincial languages in Spain and the United Kingdom, respectively, ranking at level 2 on the EGIDS scale (Eberhard; Simons; Fennig, 2025).

An even greater challenge will be raising UD_Nheengatu-CompLin's rating to 4.5 stars. Only eight treebanks for eight languages achieved this rating in UD v2.15, including UD_Portuguese-PetroGold. While five languages enjoy the highest vitality status, two—Western Armenian and Belarusian—are endangered, ranked at levels 6b and 7 on the EGIDS scale, respectively. However, their respective treebanks are among the largest in the UD collection, with approximately 122,000 and 305,000 tokens each, underscoring the scale of our challenge. Indeed, size is the second most heavily weighted factor in a treebank's rating. To achieve 4.5 stars, UD_Nheengatu-CompLin will need to reach approximately 59,000 words, assuming improvements in split, gender, POS tagging, syntactic relations, and Udapi scores, as outlined below. This growth represents nearly tripling its current size, making UD_Nheengatu-CompLin, *ceteris paribus*, the only treebank of a minority and non-European language at the top of the UD ranking system.

In the short to medium term, we will gradually increase the size score by adding new examples to UD_Nheengatu-CompLin. In this expansion, we will enrich UD_Nheengatu-CompLin in various dimensions. First, we will broaden the range of textual genres, increasing the respective score by 60% by incorporating excerpts from the Nheengatu translation of the Brazilian Constitution (Lucchesi *et al.*, 2023), songs, and abstracts, representing the legal, poetic, and academic genres. Once the corpus surpasses 30,000 words, we will establish a 10,000-

[25] This group includes two extinct languages: Latin and Old East Slavic, the latter being the ancestor of Russian, Ukrainian, and Belarusian (Simone, 2018).

word test set, thereby achieving the maximum split score. Second, we will seek to increase POS tagging and syntactic relation scores. Currently, UD_Nheengatu-CompLin does not use one of UD's 17 POS tags, namely SYM, which is applied to mathematical operators and currency symbols such as $ (dollar), € (euro), and ¥ (yen). To incorporate this tag and achieve the maximum tag score, we could ask Nheengatu speakers to translate examples from other treebanks or create analogous ones. Of UD's 37 syntactic relations, three are not used in the current development version of UD_Nheengatu-CompLin: clf, list, and orphan. The first applies to nominal classifiers, a phenomenon common in Asian languages such as Chinese but absent in Nheengatu. The list relation applies to list elements, while orphan links a dependent word to an alternative head when the original head is elided, as in (46). Both relations seem possible in Nheengatu, potentially increasing the syntactic relation score from 0.94 to 0.97.

46) Marie won gold and Peter bronze.

At the same time, we will strive to eliminate all errors detected by Udapi to maximize this score. To this end, we will explore two strategies. On the one hand, we will engage with the Udapi maintenance team to address cases where sanctioned UD annotation configurations are flagged as errors. The first of these configurations is the marking of noun degree. The second is assigning a POS tag to the head of an exocentric MWE in violation of the set of syntactic relations licensed for that tag. Udapi ignores the UD v2.15 ExtPos feature, which encodes the POS tag of these MWEs. On the other hand, we will assess the feasibility of tagging these MWE heads based on the ExtPos value, though this does not seem linguistically satisfactory at present.

By raising these questions, we hope this paper, in addition to contributing to greater inclusion of Nheengatu in digital humanities, also helps refine UD's rating system and inspires other treebank developers, particularly those working with minority and endangered languages, to improve the evaluation of their resources.

Authorship contribution statement

Leonel Figueiredo de Alencar: Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Project administration, Resources, Software, Validation, Visualization, Writing – original draft, Writing – review & editing. Hélio Leonam Barroso Silva: Data curation, Formal analysis, Investigation, Methodology, Resources, Software, Writing – original draft, Writing – review & editing. Juliana Lopes Gurgel: Data curation, Formal analysis, Visualization, Writing – review & editing. Dominick Maia Alexandre: Data curation, Formal analysis, Writing – review & editing.

Acknowledgments

We are grateful to the following institutions for providing scholarships and other financial assistance to transcribers and annotators involved in the construction of UD_Nheengatu-CompLin: The Ceará State Foundation to Support Scientific and Technological Development

(Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico – Funcap), the Coordination for the Improvement of Higher Education Personnel (Coordenação de Aperfeiçoamento de Pessoal de Nível Superior – CAPES), and, last but not least, the São Paulo State Research Support Foundation (Fundação de Amparo à Pesquisa do Estado de São Paulo – FAPESP), under Grant No. 22/09158-5, awarded to the DACILAT project at the Department of Linguistics of the University of Campinas (UNICAMP). We are also very grateful to the two anonymous reviewers for their numerous and extremely valuable comments and suggestions. We acknowledge the use of AI-based tools, including Grammarly, Quillbot, and ChatGPT, to assist with spelling, grammar, vocabulary, and style, as well as to facilitate the drafting of Python and Bash scripts used in data processing. We carefully examined, tested, and, when necessary, corrected the AI suggestions, taking full responsibility for the form and content of the paper.

References

ALENCAR, L. F. de. Aspectos da construção de um corpus sintaticamente anotado do nheengatu no modelo Dependências Universais. *Texto Livre*, Belo Horizonte, v. 17, p. e52653, 2024a. DOI: https://doi.org/10.1590/1983-3652.2024.52653.

ALENCAR, L. F. de. A Universal Dependencies treebank for Nheengatu. *In*: GAMALLO, P. *et al*. (Eds.). *Proceedings of the 16th International Conference on Computational Processing of Portuguese*. Santiago de Compostela, Galicia, Spain: Association for Computational Linguistics, 2024b. v. 2, p. 37–54. Available at: https://aclanthology.org/2024.propor-2.8.pdf. Accessed on: Apr. 2, 2025.

AMORIM, A. B. de. Lendas em Nheêngatu e em Portuguez. *Revista do Instituto Historico e Geographico Brasileiro*, Rio de Janeiro, v. 154, t. 100, p. 9–475, 1928.

ANDREWS, A. D. Relative clauses. *In*: SHOPEN, T. (ed.). *Language typology and syntactic description*. v. 2 — *Complex constructions*. Cambridge, UK: Cambridge University Press, 2007. p. 206–236.

AVILA, Marcel Twardowsky. *Proposta de dicionário nheengatu-português*. 2021. Tese (Doutorado em Estudos da Tradução) - Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo, 2021. doi:10.11606/T.8.2021.tde-10012022-201925. Acesso em: 2026-04-12.

BRANCO, A. *et al*. Universal grammatical dependencies for Portuguese with CINTIL data, LX processing and CLARIN support. *In*: CALZOLARI, N. *et al*. (Eds.). *Proceedings of the Thirteenth Language Resources and Evaluation Conference*. Marseille, France: European Language Resources Association, 2022. p. 5617–5626. Available at: https://aclanthology.org/2022.lrec-1.603/. Accessed on: Apr. 2, 2025.

BORGES, L. C. O nheengatú: uma língua amazônica. *Papia*, Brasília, v. 4, n. 2, p. 44–55, 1996. Available at: https://etnolinguistica.wdfiles.com/local--files/artigo:borges-1996/borges_1996_nheengatu.pdf. Accessed on: Apr. 2, 2025.

BYBEE, J. L.; PERKINS, R.; PAGLIUCA, W. *The evolution of grammar*: Tense, aspect, and modality in the languages of the world. Chicago: The University of Chicago Press, 1994.

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

CAVALIN, P. *et al*. Understanding native language identification for Brazilian indigenous languages. *In*: *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*. Toronto, Canada: Association for Computational Linguistics, 2023. p. 12–18. Available at: https://aclanthology.org/2023.americasnlp-1.3. Accessed on: Apr. 2, 2025.

CONSTANT, M. *et al*. Multiword expression processing: A survey. *Computational Linguistics*, Cambridge, v. 43, n. 4, p. 837–892, dez. 2017. Available at: <https://aclanthology.org/J17-4005/>. Accessed on: Apr. 2, 2025.

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

CUNHA, C.; CINTRA, L. *Nova gramática do português contemporâneo*. 7. ed. Rio de Janeiro: Lexicon, 2017.

D'ANGELIS, W. da R.; OLIVEIRA, M. C. de; SCHWADE, M. C. de D. L. Acesso ao mundo digital ou acesso digital ao mundo? *Revista Digital de Políticas Linguísticas*, Córdoba, v. 15, p. 134–158, 2021.

DURAN, M. *et al*. The dawn of the Porttinari multigenre treebank: Introducing its journalistic portion. *In*: *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*. Porto Alegre: SBC, 2023. p. 115–124. Available at: <https://sol.sbc.org.br/index.php/stil/article/view/25443>. Accessed on: Apr. 2, 2025.

DURAN, M. S. *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: Instituto de Ciências Matemáticas e de Computação da Universidade de São Paulo, 2021. Available at: <https://sites.google.com/icmc.usp.br/poetisa/publications> Accessed on: Apr. 2, 2025.

EBERHARD, D. M.; SIMONS, G. F.; FENNIG, C. D. (Eds.). *Ethnologue:* Languages of the World. 26. ed. Dallas: SIL International, 2023. Available at: http://www.ethnologue.com. Accessed on: Jul. 30, 2023.

EBERHARD, D. M.; SIMONS, G. F.; FENNIG, C. D. (Eds.). *Ethnologue:* Languages of the World. 28. ed. Dallas: SIL International, 2025. Available at: http://www.ethnologue.com. Accessed on: Apr. 2, 2025.

FREIRE, J. R. B. *Rio Babel*: A história das línguas na Amazônia. 2nd ed. Rio de Janeiro: EdUERJ, 2011.

GALVES, C. *et al*. Annotating a polysynthetic language: From Portuguese to Kadiwéu. *Cadernos de Estudos Linguísticos*, Campinas, v. 59, n. 3, p. 631–648, 2017. DOI: https://doi.org/10.20396/cel.v59i3.8651003

GIVÓN, T. *Syntax: A functional-typological introduction*. Amsterdam: John Benjamins, 1990. v. 2.

HARTT, C. F. Notes on the Lingoa Geral or Modern Tupi of the Amazonas. *Transactions of the American Philological Association*, Baltimore, v. 3, p. 58–76, 1872. Available at: https://www.jstor.org/stable/310258. Accessed on: Apr. 2, 2025.

HARTT, C. F. Notas sobre a língua geral, ou tupí moderno do Amazonas. *Anais da Biblioteca Nacional do Rio de Janeiro*, Rio de Janeiro, vol. 51, p. 305–390, 1938.

HELBIG, G. *Probleme der Valenz- und Kasustheorie*. Tübingen: Max Niemeyer Verlag, 1992.

JURAFSKY, D.; MARTIN, J. H. *Speech and language processing: An introduction to natural language processing, computational linguistics, and speech recognition*. 2nd ed. London: Pearson, 2009.

LEVIN, B. *English verb classes and alternations: A preliminary investigation*. Chicago: University of Chicago Press, 1993.

LUCCHESI, M. *et al.* (Eds.). *Mundu Sa Turusu Waá:* Ubêuwa Mayé Míra Itá Uikú Arãma Purãga Iké Braziu Upé. Brasília: Supremo Tribunal Federal, Conselho Nacional de Justiça, 2023. Available at: https://bibliotecadigital.cnj.jus.br. Accessed on: Apr. 2, 2025.

MAGALHÃES, J. V. C. de. *O selvagem*. Rio de Janeiro: Typographia da Reforma, 1876.

MARNEFFE, M.-C. de *et al*. Universal Dependencies. *Computational Linguistics*, Cambridge, v. 47, n. 2, p. 255–308, 2021. Available at: https://aclanthology.org/2021.cl-2.11. Accessed on: Apr. 2, 2025.

MARNEFFE, M.-C. de *et al*. *Universal Dependencies Guidelines.* [S. n.]: 2024a. Available at: <https://universaldependencies.org/guidelines.html>. Accessed on: Apr. 2, 2025.

MARNEFFE, M.-C. de *et al*. *UD Validation since release 2.5*. [S. n.]: 2024b. Available at: https: //universaldependencies.org/validation-rules.html. Accessed on: Apr. 2, 2025.

MELGUEIRO, E. M.; CABRAL, A. S. A. C.; MARTINS, M. F. Orações relativas em nheengatú ou ingatú. *Revista Brasileira de Linguística Antropológica*, Brasília, v. 11, n. 2, p. 151–166, 2019. DOI: https://doi.org/10.26512/rbla.v11i02.28115

MOORE, D.; FACUNDES, S.; PIRES, N. 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. Available at: https://escholarship.org/uc/item/7tb981s1. Accessed on: Jul. 26, 2024.

MÜLLER-EBERSTEIN, M.; GOOT, R. van der; PLANK, B. How universal is genre in Universal Dependencies? *In*: DAKOTA, D.; EVANG, K.; KÜBLER, S. (Eds.). *Proceedings of the 20th International Workshop on Treebanks and Linguistic Theories (TLT, SyntaxFest 2021)*. Sofia, Bulgaria: Association for Computational Linguistics, 2021. p. 69–85. Available at: https://aclanthology.org/2021.tlt-1.7/. Accessed on: Apr. 2, 2025.

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

NAVARRO, E. d. A.; ÁVILA, M. T.; TREVISAN, R. G. O Nheengatu, entre a vida e a morte: a tradução literária como possível instrumento de sua revitalização lexical. *Revista Letras Raras*, Campina Grande, v. 6, n. 2, p. 9–29, 2017. DOI: https://dx.doi.org/10.35572/rlr.v6i2.768

NIVRE, J.; ZEMAN, D.; GINTER, F.; TYERS, F. M. Tutorial on Universal Dependencies: Infrastructure, resources and tools for UD. [S. l.: s. n.], 2017. Tutorial presented at the *15th Conference of the European Chapter of the Association for Computational Linguistics* (EACL 2017). Available at: https://universaldependencies.org/eacl17tutorial/infrastructure.pdf. Accessed on: Apr. 3, 2025.

OLIVEIRA, G. M. de; CAMACHO, R. G. *Estratégias de relativização e construções alternativas nas línguas indígenas do Brasil*. São Paulo: Cultura Acadêmica, 2013. Available at: http://hdl.handle.net/11449/109292. Accessed on: Apr. 2, 2025.

PALMER, F. R. *Mood and modality*. 2nd ed. Cambridge, United Kingdom: Cambridge University Press, 2001.

PERINI, M. A. *Describing verb valency*: Practical and theoretical issues. Cham, Switzerland: Springer, 2015. DOI: https://doi.org/10.1007/978-3-319-20985-2

PERINI, M. A. *Thematic relations:* A study in the grammar-cognition interface. Cham, Switzerland: Springer, 2019. DOI: https://doi.org/10.1007/978-3-030-28538-8

PINHANEZ, C.; CAVALIN, P.; NOGIMA, J. Human evaluation of the usefulness of fine-tuned English translators for the Guarani Mbya and Nheengatu indigenous languages. *In*: GAMALLO, P. *et al*. (eds.). *Proceedings of the 16th International Conference on Computational Processing of Portuguese*. Santiago de Compostela, Galicia/Spain: Association for Computational Linguistics, 2024. v. 2, p. 32–36. Available at: https://aclanthology.org/2024.propor-2.7/. Accessed on: Apr. 2, 2025.

POPEL, M.; ŽABOKRTSKÝ, Z.; VOJTEK, M. Udapi: Universal API for Universal Dependencies. *In*: *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*. Gothenburg, Sweden: Association for Computational Linguistics, 2017. p. 96–101. Available at: <https://aclanthology.org/W17-0412>. Accessed on: Apr. 2, 2025.

RIO-TORTO, G. Formação de avaliativos. *In*: RIO-TORTO, G. *et al*. (Eds.). *Gramática derivacional do Português*. 2nd ed. Coimbra: Imprensa da Universidade de Coimbra, 2015. p. 357–389. Available at: http://dx.doi.org/10.14195/978-989-26-0864-8. Accessed on: Apr. 2, 2025.

RODRIGUES, A. D. Tarefas da lingüística no Brasil. *Estudos Lingüísticos (Revista Brasileira de Lingüística Teórica e Aplicada)*, Rio de Janeiro, v. 1, n. 1, p. 4–15, 1966. Available at: http://www.etnolinguistica.org/biblio:rodrigues-1966-tarefas. Accessed on: Apr. 2, 2025.

RODRIGUES, A. D. *Línguas brasileiras: para o conhecimento das línguas indígenas*. São Paulo: Loyola, 1986.

RODRIGUES, A. D. Línguas indígenas: 500 anos de descobertas e perdas. *DELTA: Documentação e Estudos em Linguística Teórica e Aplicada*, São Paulo, v. 9, n. 1, p. 83–103, 1993. Available at: https://revistas.pucsp.br/index.php/delta/article/view/45596. Accessed on: Apr. 2, 2025.

RODRIGUES, A. D. As línguas gerais sul-americanas. *Papia*, São Paulo, v. 4, n. 2, p. 6–18, 1996. Available at: https://etnolinguistica.wdfiles.com/local--files/artigo%3Arodrigues-1996/rodrigues_1996_linguas_gerais.pdf. Accessed on: Apr. 2, 2025.

RODRIGUES, A. D.; CABRAL, A. S. A. C. A contribution to the linguistic history of the Língua Geral Amazônica. *ALFA: Revista de Linguística*, São José do Rio Preto, v. 55, n. 2, 12 2011. DOI: https://doi.org/10.1590/S1981-57942011000200012

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

RUETER, J. *et al*. Apurinã Universal Dependencies treebank. *In*: MAGER, M. *et al*. (eds.). *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*. Online: Association for Computational Linguistics, 2021. p. 28–33. Available at: https://aclanthology.org/2021.americasnlp-1.4. Accessed on: Apr. 2, 2025.

RUPPENHOFER, J. *et al*. *FrameNet II: Extended theory and practice*. [Berkeley: International Computer Science Institute]: 2016. Revised version. Available at: https://akb89.github.io/myValencer/framenet_book.pdf. Accessed on: Apr. 2, 2025.

SANDALO, M. F. S.; GALVES, C. M. C. Anotando sintaticamente uma língua originária do brasil: O problema de anchieta. *Cadernos de Estudos Linguísticos*, Campinas, v. 65, n. 00, p. e023007, 2023. DOI: https://doi.org/10.20396/cel.v65i00.8673592

SANTOS, L. L.; ARAGON, C. C.; GERARDI, F. Línguas minoritárias e anotações sintáticas de corpora: experiências de pesquisa na iniciação científica. *Letras de hoje*, Porto Alegre, v. 59, n. 1, p. 1–9, 2024. DOI: https://doi.org/10.15448/1984-7726.2024.1.44734

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

SEKI, L. A lingüística indígena no Brasil. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, São Paulo, v. 15, n. especial, p. 257–290, 1999.

SILVA, D. P. G. da; PARDO, T. A. S. Grammar induction for Brazilian indigenous languages. *In*: GAMALLO, P. *et al*. (Eds.). *Proceedings of the 16th International Conference on Computational Processing of Portuguese*. Santiago de Compostela, Galicia/Spain: Association for Computational Linguistics, 2024. v. 2, p. 64–72. Available at: https://aclanthology.org/2024.propor-2.10/. Accessed on: Apr. 2, 2025.

SIMONE, L. R. Uma breve introdução ao idioma eslavo oriental antigo. *Slovo – Revista de Estudos em Eslavística*, Rio de Janeiro, v. 1, n. 1, p. 16–17, 2018. Available at: https://revistas.ufrj.br/index.php/slovo/article/view/17473/11271. Accessed on: Apr. 2, 2025.

STORTO, L. R. *Línguas indígenas:* tradição, universais e diversidade. Campinas: Mercado de Letras, 2019.

STRADELLI, E. 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*, Rio de Janeiro, v. 158, n. 104, p. 9–768, 1929.

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

YAMÃ, Y. *et al*. *Dicionário e estudo de nheengatu tradicional*. 2nd ed. São Paulo: Cintra, 2021.

ZEMAN, D. *Cross-Language Harmonization of Linguistic Resources*. Prague: Institute of Formal and Applied Linguistics (ÚFAL), 2023. Habilitation thesis. Available at: https://chres.is.cuni.cz/media/documents/2024/02/25/thesis-without-papers.pdf. Accessed on: Apr. 2, 2025.

ZWICKY, A. M. In a manner of speaking. *Linguistic Inquiry*, Cambridge, v. 2, n. 2, p. 223–233, 1971.