Parsing Nheengatu: Performance Gains for a Brazilian Indigenous Universal Dependencies Treebank

Dominick Maia Alexandre and Leonel Figueiredo de Alencar
Universidade Federal do Ceará (UFC), Brazil
Av. da Universidade 2683 – 60.020-181 – Fortaleza – CE – Brazil
dominick@letras.ufc.br, leonel.de.alencar@ufc.br

Abstract

This paper evaluates the impact of expanding the UD_Nheengatu-CompLin treebank on parsing performance for Nheengatu, a Brazilian endangered Indigenous language. We hypothesized that the inclusion of annotated data would result in a 10% improvement in the Labeled Attachment Score (LAS). To test this hypothesis, we conducted a 10-fold cross-validation experiment using UDPipe 1.4.0 under two conditions: parsing with gold tokenization and gold tags, and automatic parsing from raw text. Statistical significance was determined using the Mann–Whitney U test. Although the expected gain was not achieved, the results show improvements in parsing accuracy and reduced variance across folds. The findings highlight the importance of corpus expansion and standardized annotation workflows for improving parsing performance in low-resource language scenarios and for supporting reproducible evaluation methods in the computational modeling of minority languages.

1 Introduction

Despite significant advances in Natural Language Processing (NLP) over the last decade, a substantial gap persists between high-resource and low-resource languages. This disparity is driven by the scarcity of annotated data for the latter and the concentration of technological development on the former (Joshi et al., 2020; Bird, 2020). In addition, many Indigenous languages present challenges for computational modeling, such as rich morphology, orthographic variation, and limited standardization, which hinder the direct application of existing NLP methods (Mager et al., 2018).

In this context, the Universal Dependencies (UD) project provides a shared cross-linguistic annotation framework that has been widely adopted for syntactic modeling across many well-resourced languages (Nivre et al., 2017; Church and Liberman, 2021), while also offering opportunities for extending NLP methods to low-resource ones (Thomas, 2019; Tyers and Henderson, 2021; Martín Rodríguez et al., 2022; da Silva and Pardo, 2024).

The Nheengatu treebank in the UD collection is a useful case study for examining the morphosyntactic annotation and parsing for low-resource languages. Once used as a *lingua franca* across the Amazon basin during the 18th and 19th centuries, Nheengatu is today an endangered language spoken primarily in Brazil. The UD_Nheengatu-CompLin treebank constitutes the first effort to provide a syntactically annotated corpus for the language following the UD framework (de Alencar, 2023; Alencar, 2024; de Alencar, 2024b,a, 2025).

Since its release in 2022, the treebank has been expanded and revised to improve coverage and consistency. In UD version 2.17, it received a 3.5-star rating, exceeding all other 24 Indigenous-language treebanks of the Americas ($\leq$ 2 stars) and approaching much larger treebanks of high-resource languages, such as UD_Portuguese-Porttinari (Duran et al., 2023) and UD_Portuguese-PetroGold (Souza et al., 2021) (4 stars).

Earlier parsing experiments on the UD_Nheengatu-CompLin were conducted by de Alencar (2024a), but they relied on a smaller and imbalanced dataset. Recent annotation efforts have incorporated historical data from Hartt's (1938), introducing a nineteenth-century Lower Amazon variety and enabling a reassessment of parsing performance under a more linguistically diverse dataset.

In this work, we evaluate parsing accuracy with and without the inclusion of Hartt's (1938), using an updated parsing pipeline and reproducible evaluation scripts. Although we hypothesized that this expansion would lead to an improvement of at least 10% in Labeled Attachment Score (LAS), the experiments revealed a more modest but still significant enhancement in parsing quality.

Proceedings of the 17th International Conference on Computational Processing of Portuguese (PROPOR 2026)
Vol. 2, pages 210–219

We conduct our experiments using UDPipe 1.4.0 (Straka et al., 2016), whose sensitivity to annotation quality makes it well suited for assessing the impact of treebank expansion in a low-resource setting.

The paper is organized as follows: Section 2 reviews related work on parsing for Brazilian and Amerindian languages within the UD framework; Section 3 presents Nheengatu and the linguistic phenomena recently included in the treebank; Section 4 describes the methodology; Section 5 reports the parsing results; Section 6 discusses the most frequent parser errors; and Section 7 concludes and outlines directions for future work.

2 Related work

Further improvements in the Labeled Attachment Score (LAS), the main metric for dependency parsing, are now limited less by model architecture than by the quality and size of the data (Lopes et al., 2024). This reliance on data contributes to disparities between high- and low-resource languages, with LAS values above 90% for languages such as English, Portuguese, and Russian, compared to below 40% for many minority languages.

Lopes and Pardo (2024) demonstrate that parsing accuracy degrades systematically as training data is reduced, with LAS dropping from 91.74% when trained on 5,893 sentences to 88.78% with only 1,473 sentences. These results show that even for well-resourced languages and gold-standard annotations, corpus size remains a decisive factor for parsing performance. Importantly, this effect becomes substantially more pronounced in low-resource settings, helping to explain the markedly lower LAS values reported by Vasquez et al. (2018) for Shipibo-Konibo and by Pugh et al. (2022) for Nahuatl. In such cases, limited training data reduces lexical and grammatical coverage, leading to severe constraints on parser generalization.

In the context of Indigenous languages and the Tupian language family, Blum's (2022) study shows that large multilingual models such as mBERT and RoBERTa perform poorly on Brazilian Indigenous languages. In zero-shot settings, accuracy rarely exceeds 40%. By contrast, models trained on closely related languages consistently achieve better results than those relying on typologically distant sources. These findings indicate that large multilingual models often fail to capture the grammatical structure of languages that are underrepresented in their training data (Blum, 2022).

Blum (2022) further shows that combining data from multiple related languages can outperform single-source models, even with small training sets. Their experiments indicate that as few as 50–60 annotated sentences can already yield measurable gains, particularly for PoS tagging. In contrast, dependency parsing remains more challenging, with lower transfer performance due to syntactic complexity (Blum, 2022).

3 The Nheengatu language

Once the most widely spoken language in the Brazilian Amazon, reaching parts of present-day Venezuela and Colombia, Nheengatu is now spoken by about 6,000 speakers in Brazil and faces challenges in intergenerational transmission (Navarro, 2012; Navarro et al., 2017; Eberhard et al., 2025).

Originating from Tupinambá and widely adopted by Jesuits, settlers, and different Indigenous groups, Nheengatu served social, political, and religious functions in Brazil during the seventeenth and eighteenth centuries. Despite a royal ban in the eighteenth century, it continued to be spoken into the early twentieth century (Borges, 1996; Rodrigues, 1986; Moore, 2014; Stradelli, 2014).

Today, Nheengatu is mainly spoken in the Upper Rio Negro region and remains an important marker of Amazonian identity. The language has been documented since the nineteenth century through grammars, religious texts, and literary works, which form a key part of the textual sources used in the UD_Nheengatu-CompLin treebank.

3.1 The UD_Nheengatu-CompLin treebank

Since its release in 2022, the UD_Nheengatu-CompLin treebank has been under continuous development. In the UD version 2.17 release (November 15, 2025), it underwent a substantial expansion in size and linguistic coverage, reaching 2,742 trees and 26,033 words. More than 600 sentences were added from nineteenth-century Nheengatu documented during Charles Frederick Hartt's Morgan Expedition (1870–1871) in the Lower Amazon region (Hartt, 1872, 1938). All new sentences were annotated according to the project guidelines and reviewed by a second annotator. This addition made Hartt's (1938) the second largest source in the treebank, after Avila's (2021), and the largest source of spoken-genre data. The updated distribution of sources is shown in Figure 1.

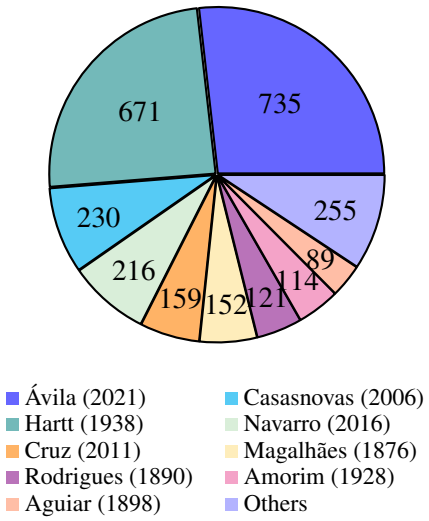

Figure 1: Sentence sources in UD_Nheengatu-CompLin ($n = 2742$). Sources with fewer than 89 sentences are grouped as *Others*.

(1) *E-kũi* *ne* *kiwira* *senúi*
2SG.IMP-go your brother [INF]call
u-ruri *arama* *ií* *ixéu.*
3SG.ACT-bring to water 1SG.DAT
'Go get your brother to bring me water.' (Hartt, 1938, p. 335)

As a result of this expansion, the treebank now includes a broader range of linguistic phenomena, such as archaisms, additional lexical forms, and morphosyntactic patterns not previously represented in the treebank (Alexandre and de Alencar, 2025). Example (1) illustrates morphosyntactic patterns no longer attested in contemporary Nheengatu. In present-day usage, imperative meanings are expressed using forms identical to the indicative. In contrast, the historical construction employs the auxiliary verb *ekũi* 'go' with archaic imperative morphology, including the second-person singular prefix *e-* and an irregular imperative form derived from the verb *sú* 'to go'.

The example also exhibits an SOV constituent order, with the full nominal object *ne kiwira* 'your brother' preceding the bare verb root. This indicates that historical Nheengatu allowed SOV order with full noun phrases, unlike contemporary Nheengatu, which displays a stable SVO pattern, likely influenced by long-term contact with Portuguese (da Cruz, 2011). From a typological perspective, however, the SOV order is more compatible with the language's postpositional system (Aikhenvald and Dixon, 2001).

Another archaic feature illustrated by the example is the absence of the double subject agreement that is typical of auxiliary constructions with verbs like *sú* 'to go' in contemporary Nheengatu. The sentence also contains the first-person dative pronoun *ixéu* 'to me', an inherited form from Old Tupi that had largely been replaced by the postposition *arama* by the time of Hartt's documentation (Avila, 2021), yet was still attested in nineteenth-century Lower Amazon Nheengatu.

Figure 2 displays the dependency graph for (1), illustrating how these historical morphosyntactic properties are encoded in UD_Nheengatu-CompLin. The annotation of Hartt's (1938) expanded the treebank with morphosyntactic patterns that are absent from contemporary data. These phenomena can now be analyzed using Yauti (de Alencar, 2023, 2025), an automatic annotation tool for Nheengatu, extending the treebank's linguistic coverage and supporting further computational analysis.

4 Methods

4.1 Parsing Experiments

Two parsing experiments were conducted using a 10-fold cross-validation procedure, following the methodology adopted in previous work on the UD_Nheengatu-CompLin treebank (de Alencar, 2024a). Parsing performance was evaluated using the development version of the treebank before and after the inclusion of historical material from Hartt's (1938).

The experiments were conducted using UDPipe 1.4.0, a trainable, language-agnostic pipeline for tokenization, morphosyntactic tagging, lemmatization, and dependency parsing of CoNLL-U data, which provides pre-trained models for Universal Dependencies treebanks and supports both gold-standard and fully automatic processing pipelines (Straka et al., 2016).

In **Experiment 1**, the parser was trained and evaluated on a version of the treebank excluding all 19th-century Lower Amazon data from Hartt (1938) (`nohartt.conllu`, 2,091 sentences), whereas **Experiment 2** used an expanded version including 743 sentences from Hartt (1938) (`all.conllu`, 2,834 sentences).[1]

[1]All data are available at: https://github.com/CompLin/nheengatu.

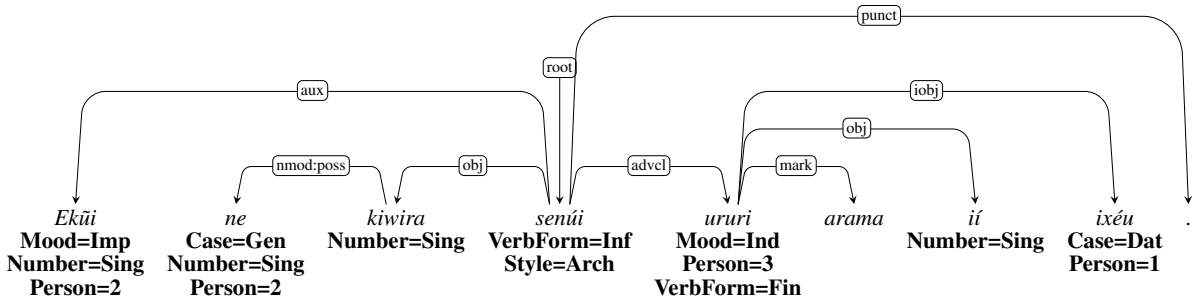

Figure 2: Dependency tree for (1).

For both experiments, the data were partitioned into 10 non-overlapping subsets of sentences of approximately equal size. In each fold, nine subsets were used for training and one for testing, such that every file and its sentences were used exactly once, as illustrated in Figure 3. This setup reduces dependence on a single train–test split and is well-suited for low-resource treebanks.

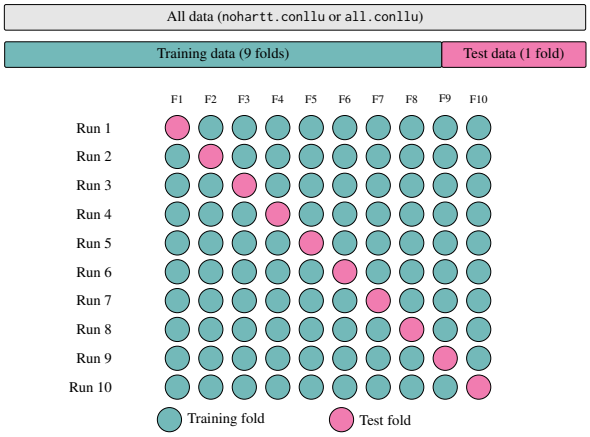

Figure 3: Illustration of 10-fold cross-validation. Each run uses one fold as test data (pink) and the remaining nine folds as training data (green).

Parsing performance was evaluated using the metrics **Labeled Attachment Score (LAS)** and **Unlabeled Attachment Score (UAS)**. Two evaluation conditions were considered: (i) parsing with gold tokenization and gold morphosyntactic annotations, and (ii) fully automatic parsing from raw text, in which UDPipe performed tokenization, tagging, and dependency parsing.

4.2 Parsing Performance Comparison

To assess the statistical significance of parsing performance differences between experimental settings, we used a Python script that reads per-fold LAS values directly from the parser output files and compares the distributions of LAS scores obtained under different experiments.

Statistical significance was evaluated using the **Mann–Whitney U test**, a nonparametric test for comparing two independent samples based on rank ordering (Mann and Whitney, 1947). For each comparison, a two-sided Mann–Whitney U test was performed on the two sets of LAS values.

The test was implemented using the `mannwhitneyu` function from the SciPy scientific computing library (Virtanen et al., 2020). In addition to reporting the U statistic and the corresponding two-tailed p-value, the script generates a fold-wise plot of LAS scores for visual inspection of performance differences across experiments.

The `CompareParsingResults.py` script was used to compare the ten fold-level LAS scores from Experiments 1 and 2, as well as LAS scores reported in prior work on the same treebank (de Alencar, 2024a). Because the earlier study used UDPipe 1.2, while our experiments used UDPipe 1.4.0, we performed two comparisons: one against the original UDPipe 1.2 results and one against a re-run of the experiment reported by (de Alencar, 2024a) using UDPipe 1.4.0.[2]

4.3 Error analysis

To identify the most frequent parsing errors, we analyze the outputs of 10 test splits under gold tokenization and tagging. For each split, gold and predicted CoNLL-U files were aligned and compared. We then compute a confusion matrix over UD dependency relations using a custom Python script (`deprel_confusion.py`), capturing label confusions independently of head assignment.

We further perform a diagnostic analysis that reports UAS and label accuracy and classifies errors into attachment-only, label-only, and combined

[2]Due to time constraints, we used the same models trained by de Alencar (2024a) for parsing with UDPipe 1.4.0.

	Exp. 1 (%)		Exp. 2 (%)	
Metric	**UAS**	**LAS**	**UAS**	**LAS**
Mean	87.06	82.70	88.12	84.29
SD	1.14	1.39	0.90	1.13

Table 1: Mean and standard deviation (SD) of parsing performance with **gold input** for Experiments 1 and 2.

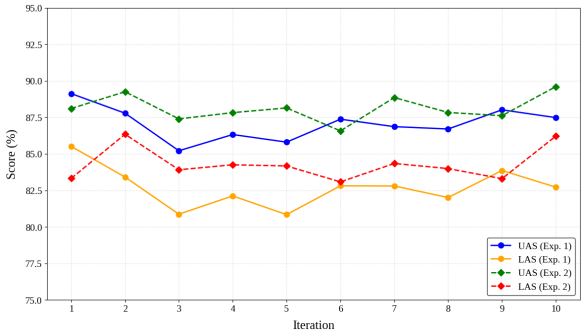

Figure 4: UAS and LAS with **gold input** across 10 folds for Experiments 1 and 2.

types. The most frequent errors are discussed in the following section.

5 Parsing Experiment Results

This section reports the results of the parsing experiments comparing two versions of the UD_Nheengatu-CompLin treebank. Although the initial hypothesis of a 10% improvement in parsing performance was not supported, the results indicate a positive effect of increased corpus size and coverage on parsing accuracy.

5.1 Parsing with Gold Input

Table 1 reports the mean and standard deviation of UAS and LAS for the two experiments with gold tokenization and gold tags. Across both metrics, Experiment 2 achieves higher mean scores than Experiment 1, with gains of 1.06 points in UAS and 1.59 points in LAS.

Experiment 2 also shows lower variance across folds for both metrics, indicating more stable performance under cross-validation (UAS SD = 0.90, LAS SD = 1.13) compared to Experiment 1 (UAS SD = 1.14, LAS SD = 1.39). Figure 4 shows the distribution of UAS and LAS scores across the ten iterations for both experiments.

5.2 Tokenization from Raw Text

We next evaluate tokenization performance under a fully automated parsing setting in which test sentences are provided as raw text, allowing the parser to perform tokenization. This setting allows us to assess the effects of corpus expansion when errors from earlier processing stages are not controlled.

	Exp. 1 (%)		Exp. 2 (%)	
Metric	**F1**	**SD**	**F1**	**SD**
Tokens	94.64	0.83	94.24	0.72
Multiword tokens	85.44	5.35	87.70	3.27
Words	94.42	0.80	94.09	0.69
Sentences	59.35	6.15	62.55	4.02

Table 2: Mean F1-score and standard deviation (SD) of tokenizer performance from **raw text** for Experiments 1 and 2.

	Exp. 1 (%)		Exp. 2 (%)	
Metric	**F1**	**SD**	**F1**	**SD**
UPOS tags	90.01	0.84	89.52	0.87
XPOS tags	89.18	0.87	88.73	0.91
Features	86.77	1.03	85.97	1.04
Lemmas	91.47	0.98	90.89	1.01

Table 3: Mean F1-score and standard deviation (SD) of tagging performance from **raw text** for Experiments 1 and 2.

Table 2 reports mean F1-scores and standard deviations for tokenization across Experiments 1 and 2, evaluated on tokens, multiword tokens, words, and sentences.

Token- and word-level segmentation remains high and stable across both experiments, with F1-scores above 94%. For multiword tokens, Experiment 2 achieves a higher mean F1-score than Experiment 1 (+2.26), along with a lower standard deviation across folds. Sentence segmentation remains the most challenging subtask in both experiments; however, Experiment 2 again shows higher mean performance (+3.20) and reduced variability.

5.3 Tagging from Raw Text

Table 3 reports tagging performance for UPOS, XPOS, morphological features, and lemmatization. Across all tagging components, Experiment 2 shows slightly lower mean scores than Experiment 1.

Within the UD framework, UPOS tagging relies on a smaller and more abstract label set than XPOS tagging and morphological feature prediction. In contrast, morphological features encode language-specific distinctions and combinations of features, which introduce additional complexity.

The larger differences observed for FEATS are

	Exp. 1 (%)		Exp. 2 (%)	
Metric	**UAS**	**LAS**	**UAS**	**LAS**
Mean	74.46	68.73	74.80	69.56
SD	1.96	2.17	0.95	0.91

Table 4: Mean and standard deviation (SD) of parsing performance from **raw text** for Experiments 1 and 2.

consistent with the increased linguistic and orthographic variability introduced in Experiment 2, which incorporates historical data with a wider range of morphosyntactic patterns, part-of-speech and feature combinations, and preserved orthographic variation from the source material.

5.4 Parsing from Raw Text

Table 4 reports UAS and LAS results for Experiments 1 and 2. As expected, performance is substantially lower than in the gold-input setting, reflecting the accumulation of errors from tokenization and morphosyntactic tagging. Experiment 2 shows small gains in mean UAS (+0.34) and LAS (+0.83), as well as reduced variance across folds, with standard deviations below 1% for both metrics compared to Experiment 1, which shows standard deviations close to 2%.

This pattern indicates a more stable parsing behavior under fully automatic conditions, even when errors from earlier processing stages may propagate to dependency parsing.

5.5 Statistical Significance

To assess whether the differences are statistically reliable, we apply a two-sided Mann–Whitney U test to the fold-level LAS scores. Under gold-input conditions, the test yields $U = 14.0$ and $p = 0.0073$, indicating a statistically significant difference between Experiments 1 and 2 (Figure 5).

We further compare Experiment 2 with the results reported by de Alencar (2024a), obtained using UDPipe version 1.2. As shown in the left panel of Figure 6, Experiment 2 consistently achieves higher LAS values across all folds when compared to the 2024 experiment. The observed differences are modest but systematic, generally ranging between approximately 1 and 3 percentage points, and are statistically significant ($p = 0.00032$).

Together with the fold-level patterns shown in Figure 5, this result suggests that the higher LAS scores observed for Experiment 2 are unlikely to be due to random variation across folds.

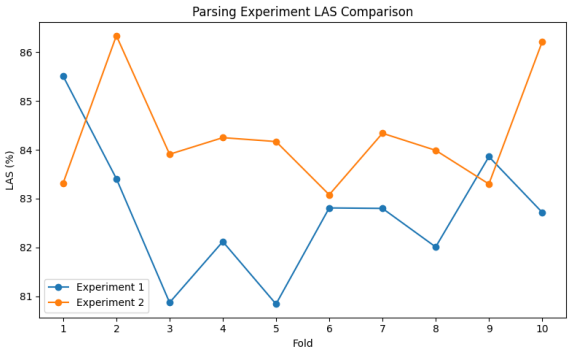

Figure 5: LAS comparison between Experiments 1 and 2.

However, this improvement primarily reflects the effects of corpus expansion, as the 2024 experiment was based on a smaller dataset (1,336 sentences). Contrary to initial expectations, differences in the parsing pipeline appear to play a minimal role, as re-running the earlier experiment with UDPipe 1.4.0 yields results that are essentially equivalent to those obtained with UDPipe 1.2.

As shown in the right panel of Figure 6, Experiment 2 consistently achieves higher LAS scores across all folds, and these differences remain stable when using UDPipe 1.4.0, remaining statistically significant (Mann–Whitney $U = 98.0$, $p = 0.00033$).

6 Frequent parsing errors

The most frequent confusion involves the core argument relations *nsubj* and *obj* (Figure 7). Across the test splits, 169 gold *nsubj* relations were predicted as *obj*, and 140 gold *obj* relations as *nsubj*. Figure 8 illustrates the former: the parser analyzes the subject *mbira-itá* ‘my children’ of *ukiri* ‘sleep’ as the object of the preceding subordinate verb *asika* ‘arrive’. Figure 9 shows the correct analysis.

The prediction in Figure 8 is not entirely implausible. In sentences with two verbal clauses, a noun phrase occurring between the verbs may, in principle, be interpreted either as the object of the first verb or as the subject of the second. In addition, Nheengatu is a pro-drop language, allowing the subject position to remain unexpressed, which makes an analysis in which *mbira-itá* does not attach to *ukiri* structurally possible. However, sentence (2) is not genuinely ambiguous, since the embedded verb is intransitive and does not license an object.

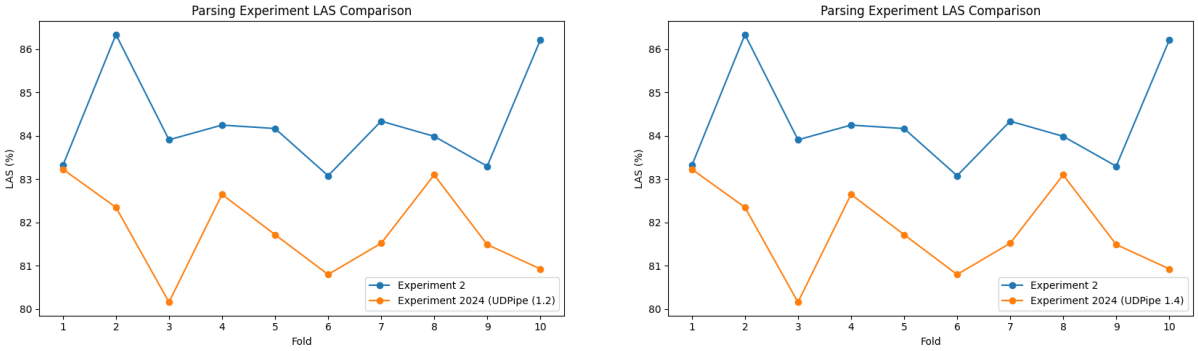

Figure 6: LAS comparison between Experiment 2 and the 2024 experiment using different UDPipe versions (v1.2 on the left, v1.4.0 on the right).

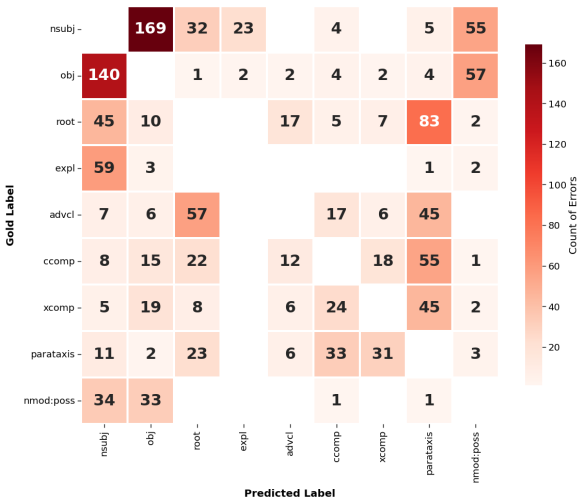

Figure 7: Heatmap of the most frequent UD Relations confusions (Raw Counts)

(2) *Asika* *ramé* *se* *mbira-itá*
1SG.arrive when 1SG.GEN child.PL
ukiri *ana* *uikú.*
3SG.sleep PFV 3SG.be

'When I arrive, my children are already sleeping.' (Moore et al., 1994, p. 110)

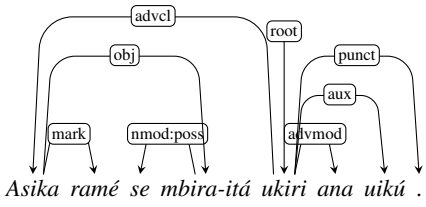

Figure 8: Incorrect dependency tree for (2).

While this error also occurs in shorter, typically monoclausal sentences, where the parser fails to recognize postverbal subjects (e.g., of unaccusative verbs), 78.2% of cases occur in sentences ranging from nine to 49 tokens. Longer sentences typically involve multiple coordinated, juxtaposed, or embedded predicates (e.g., in relative, adverbial, and complement clauses). Together, these patterns suggest that the parser struggles both to identify clause boundaries and to take verb valency into account.

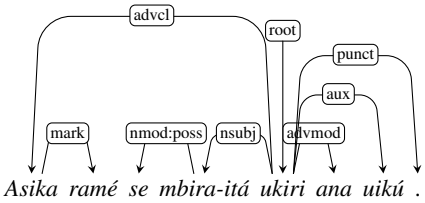

Figure 9: Gold dependency tree for (2).

The heatmap also reveals a frequent confusion between *root* and *parataxis*, reflecting the parser's difficulty in identifying the main predicate in complex sentences. Similarly, clausal relations such as *advcl*, *ccomp*, and *xcomp* are often predicted as *parataxis* or *root*. Table 5 summarizes the corresponding metrics and error distribution. Attachment-only errors are more frequent than label-only errors, suggesting that incorrect head assignment is a more common source of error than label misclassifications alone. In addition, a substantial number of cases (1,434) involve errors in both head and label, indicating that some errors arise in more complex sentences rather than from isolated misclassifications.

7 Summary of Results

Across evaluation settings, Experiment 2 yields higher parsing accuracy and reduced variability across folds. Expanding the UD_Nheengatu-CompLin treebank leads to consistent improve-

Metric	Value
Tokens compared	26,785
UAS	88.11%
Label accuracy	90.82%
LAS	84.28%
Attachment-only errors	1,750
Label-only errors	1,026
Head+label errors	1,434

Table 5: Global parsing performance and error distribution across all test splits.

ments in accuracy and stability. Error analysis indicates that errors concentrate in core arguments (e.g., *nsubj* vs. *obj*) and clause-level relations (e.g., *root*, *parataxis*, *advcl*).

Although the expected 10% gain was not observed, the results point to the importance of incremental treebank expansion combined with clear annotation guidelines, internal consistency, and review by a second annotator when developing resources for a low-resource language.

8 Final Remarks

This study investigated how the expansion of the UD_Nheengatu-CompLin treebank affects dependency parsing performance under controlled (gold tokenization and gold tags) and fully automatic conditions. By extending the treebank with historical nineteenth-century material, we assessed how increased data volume and linguistic diversity interact with different stages of a modern parsing pipeline for a low-resource language.

Across evaluation settings, the expanded treebank is associated with more stable parsing performance, reflected in lower variance across cross-validation folds. Under gold-input conditions, the differences between experiments are statistically significant, suggesting that increased annotated data supports more consistent syntactic predictions when tokenization and morphosyntactic annotation are fixed. In fully automatic parsing, average accuracy differences are small, but Experiment 2 shows lower variability across folds, indicating more stable model behavior.

The inclusion of historical data introduced greater linguistic heterogeneity, including orthographic variation and less frequent morphosyntactic patterns. While this diversity expanded syntactic coverage, it also increased the difficulty of morphosyntactic tagging. These findings underscore the importance of interpreting parsing results in relation to both corpus composition and processing conditions, particularly for low-resource Indigenous languages.

Some limitations of this study point to directions for future work. For instance, genre distribution in the treebank was not controlled or analyzed. A more detailed evaluation by genre would require prior classification of sentences, which is not currently available and is challenging due to the heterogeneous nature of the sources (e.g., narrative texts, grammatical examples, and mixed materials).

We plan to further expand the UD_Nheengatu-CompLin treebank, along with a more fine-grained analysis of dependency relation labels that remain challenging for the parsing pipeline, focusing on error patterns at the level of specific relations in order to identify systematic sources of ambiguity and inform refinements to annotation guidelines or modeling strategies. Another direction for future work is the evaluation of alternative parsing architectures.

Beyond Nheengatu, the methodology adopted in this study was designed to be reproducible, and the release of updated resources and evaluation scripts enables the same approach to be applied to other Indigenous and low-resource languages within the Universal Dependencies framework, supporting transparent resource development and comparable evaluation of parsing performance across minority languages.

Acknowledgments

This work was supported by the Brazilian CAPES Foundation and FAPESP (Grant No. 22/09158-5, DACILAT project at UNICAMP). We thank the anonymous reviewers for their helpful suggestions. We acknowledge the use of large language models (ChatGPT, Gemini, and Grammarly) for grammar and style revision, as well as for coding support.

References

Alexandra Y. Aikhenvald and R. M. W. Dixon. 2001. Introduction. In Alexandra Y. Aikhenvald and R. M. W. Dixon, editors, *Areal diffusion and genetic inheritance: Problems in comparative linguistics*, pages 1–26. Oxford University Press, Oxford.

Leonel Figueiredo de Alencar. 2024. UD_Nheengatu-CompLin: o corpus sintaticamente anotado do nheengatu da coleção Universal Dependencies. In *Anais*

Eletrônicos do XVI Encontro de Linguística de Corpus e da XII Escola Brasileira de Linguística Computacional, volume 1, pages 105–109, Brasília. Associação Brasileira de Linguística de Corpus.

Dominick Maia Alexandre and Leonel Figueiredo de Alencar. 2025. Universal Dependencies for 19th-Century Nheengatu from the Lower Amazon Region. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 588–598, Porto Alegre, RS, Brasil. SBC.

Marcel Twardowsky Avila. 2021. *Proposta de dicionário nheengatu-português*. Ph.D. thesis, Faculdade de Filosofia, Letras e Ciências Humanas da Universidade de São Paulo.

S. Bird. 2020. Decolonising speech and language technology. In *COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference*, COLING 2020 - 28th International Conference on Computational Linguistics, Proceedings of the Conference, pages 3504–3519. Association for Computational Linguistics (ACL).

Frederic Blum. 2022. Evaluating zero-shot transfers and multilingual models for dependency parsing and POS tagging within the low-resource language family tupían. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 1–9, Dublin, Ireland. Association for Computational Linguistics.

Luiz Carlos Borges. 1996. O nheengatú: uma língua amazônica. *Papia*, 4(2):44–55.

Kenneth Church and Mark Liberman. 2021. The future of computational linguistics: On beyond alchemy. *Frontiers in Artificial Intelligence*, 4:1–18.

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

Diego Pedro Gonçalves da Silva and Thiago Alexandre Salgueiro Pardo. 2024. Grammar induction for Brazilian indigenous languages. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 2*, pages 64–72, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Leonel Figueiredo de Alencar. 2023. Yauti: A tool for morphosyntactic analysis of Nheengatu within the Universal Dependencies framework. In *Anais do XIV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 135–145, Porto Alegre, RS, Brasil. SBC.

Leonel Figueiredo de Alencar. 2024a. A Universal Dependencies Treebank for Nheengatu. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, volume 2, pages 37–54, Santiago de Compostela, Galicia, Spain. Association for Computational Linguistics.

Leonel Figueiredo de Alencar. 2024b. Aspectos da construção de um corpus sintaticamente anotado do nheengatu no modelo Dependências Universais. *Texto Livre*, 17:e52653.

Leonel Figueiredo de Alencar. 2025. Enhancing a Nheengatu Morphosyntactic Analyzer for Word Formation and Non-standard Language. In *Anais do XVI Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana*, pages 13–28, Porto Alegre, RS, Brasil. SBC.

Magali Duran, Lucelene Lopes, Maria das Graças Nunes, and Thiago Pardo. 2023. 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*, pages 115–124, Porto Alegre, RS, Brasil. SBC.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2025. *Ethnologue: Languages of the World*, 28 edition. SIL International, Dallas.

Charles Frederick Hartt. 1872. Notes on the Lingoa Geral or Modern Tupi of the Amazonas. *Transactions of the American Philological Association*, 3:58–76.

Charles Frederick Hartt. 1938. Notas sobre a língua geral, ou tupí moderno do Amazonas. *Anais da Biblioteca Nacional do Rio de Janeiro*, LI:305–390. [1929].

Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The State and Fate of Linguistic Diversity and Inclusion in the NLP World. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293, Online. Association for Computational Linguistics.

Lucelene Lopes and Thiago Pardo. 2024. Towards Portparser - a highly accurate parsing system for Brazilian Portuguese following the Universal Dependencies framework. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 1*, pages 401–410, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Lucelene Lopes, Thiago Pardo, and Magali Duran. 2024. Syntactic parsing: where are we going? In *Anais do XV Simpósio Brasileiro de Tecnologia da Informação e da Linguagem Humana (STIL 2024)*, pages 67–74, Porto Alegre, RS, Brasil. SBC.

Manuel Mager, Ximena Gutierrez-Vasques, Gerardo Sierra, and Ivan Meza-Ruiz. 2018. Challenges of language technologies for the indigenous languages of the Americas. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 55–69, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Henry B. Mann and Donald R. Whitney. 1947. On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. *The Annals of Mathematical Statistics*, 18(1):50–60.

Lorena Martín Rodríguez and 1 others. 2022. Tupían language ressources: Data, tools, analyses. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 48–58, Marseille, France. European Language Resources Association.

Denny Moore. 2014. Historical development of Nheengatu (Língua Geral Amazônica). In Salikoko S. Mufwene, editor, *Iberian Imperialism and Language Evolution in Latin America*, pages 108–142. University of Chicago Press, Chicago.

Denny Moore, Sidney Facundes, and Nádia Pires. 1994. 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*, pages 93–118, Berkeley, CA. [University of California]. Acesso em: 26 jul. 2024.

Eduardo de Almeida Navarro. 2012. O último refúgio da língua geral no Brasil. *Estudos Avançados*, 26(76):245–254.

Eduardo de Almeida Navarro, Marcel Twardowsky Ávila, and Rodrigo Godinho Trevisan. 2017. O Nheengatu, entre a vida e a morte: A tradução literária como possível instrumento de sua revitalização lexical. *Revista Letras Raras*, 6(2):9–29.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Robert Pugh, Marivel Huerta Mendez, Mitsuya Sasaki, and Francis Tyers. 2022. Universal Dependencies for Western Sierra Puebla Nahuatl. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5011–5020, Marseille, France. European Language Resources Association.

Aryon Dall'Igna Rodrigues. 1986. *Línguas Brasileiras: Para o conhecimento das línguas indígenas*. Loyola, São Paulo. Vários quadros numerados e outros sem numeração.

Elvis Souza, Aline Silveira, Tatiana Cavalcanti, Maria Castro, and Claudia Freitas. 2021. PetroGold – Corpus padrão ouro para o dominio do petroleo. In *Proceedings of the 13th Brazilian Symposium in Information and Human Language Technology*, pages 29–38, Porto Alegre, Brazil. Association for Computational Linguistics.

Ermanno Stradelli. 2014. *Vocabulário português-nheengatu, nheengatu-português*. Ateliê Editorial, Cotia, SP. Original work published in 1929.

Milan Straka, Jan Hajič, and Jana Straková. 2016. UDPipe: Trainable pipeline for processing CoNLL-U files performing tokenization, morphological analysis, POS tagging and parsing. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 4290–4297, Portorož, Slovenia. European Language Resources Association (ELRA).

Guillaume Thomas. 2019. Universal Dependencies for Mbyá Guaraní. In *Proceedings of the Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, pages 70–77, Paris, France. Association for Computational Linguistics.

Francis M. Tyers and Robert Henderson. 2021. A corpus of K'iche' annotated for morphosyntactic structure. In *Proceedings of the First Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*.

Alonso Vasquez and 1 others. 2018. Toward Universal Dependencies for Shipibo-konibo. In *Proceedings of the Second Workshop on Universal Dependencies (UDW 2018)*, pages 151–161, Brussels, Belgium. Association for Computational Linguistics.

Pauli Virtanen, Ralf Gommers, Travis E. Oliphant, Matt Haberland, Tyler Reddy, David Cournapeau, Evgeni Burovski, Pearu Peterson, Warren Weckesser, Jonathan Bright, Stéfan J. van der Walt, Matthew Brett, Joshua Wilson, K. Jarrod Millman, Nikolay Mayorov, Andrew R. J. Nelson, Eric Jones, Robert Kern, Eric Larson, and 16 others. 2020. SciPy 1.0: Fundamental Algorithms for Scientific Computing in Python. *Nature Methods*, 17:261–272.