Extending Enhanced Universal Dependencies to Nheengatu: Semi-Automated Control and Raising Annotation

Dominick Maia Alexandre[1], **Leonel Figueiredo de Alencar**[1,2]

[1]Centro de Humanidades – Universidade Federal do Ceará (UFC)
Av. da Universidade, 2683 – 60020-181 – Fortaleza – CE – Brazil

[2]Instituto de Estudos da Linguagem – Universidade Estadual de Campinas (Unicamp)
R. Sérgio Buarque de Holanda, 571 – 13083-859 – Campinas – SP – Brazil

`dominick@letras.ufc.br, leonel.de.alencar@ufc.br`

Abstract. *This paper presents the first Enhanced Universal Dependencies annotation effort for a Brazilian Indigenous language, focusing on XCOMP constructions in Nheengatu, an endangered Tupian language. We describe a semi-automatic pipeline for the UD_Nheengatu-CompLin treebank to annotate control and raising predicates and propagated subjects. The method combines theoretical discussions on XCOMP constructions with corpus-based analysis of Nheengatu syntax to create rule-based enhancement heuristics. The work also provides annotation guidelines and expands EUD resources for Indigenous and low-resource languages within the Universal Dependencies framework.*

1. Introduction

Universal Dependencies (UD) has become a widely adopted framework for cross-linguistic morphosyntactic annotation, encompassing, as of release 2.17, 339 treebanks representing 186 languages [Nivre et al. 2016, Church and Liberman 2021, Zeman et al. 2025]. Beyond basic dependency trees, the framework defines an optional enhanced annotation layer, known as Enhanced Universal Dependencies (EUD), which is designed to make implicit syntactic relations explicit in dependency graphs [Schuster and Manning 2016].

EUD graphs cover 6 types of enhancements: (i) empty nodes for elided predicates; (ii) propagation of incoming dependencies to conjuncts; (iii) propagation of outgoing dependencies from conjuncts; (iv) additional subject dependencies for control and raising constructions; (v) coreference in relative clause constructions; and (vi) case or mark information for adpositions and conjunctions [Universal Dependencies 2026].

Although EUD annotation has expanded for high-resource languages such as Portuguese [Pagano et al. 2023, Souza et al. 2024, Sanches Duran et al. 2025], it remains uncommon in low-resource UD treebanks. In this context, Nheengatu, an endangered Tupian language spoken mainly in the Upper Rio Negro region of Brazil [Navarro 2012], is represented by UD_Nheengatu-CompLin, the only syntactically annotated corpus of the language within the UD framework [de Alencar 2023, de Alencar 2024, de Alencar 2025, Alencar et al. 2025]. Until now, however, the treebank has lacked guidelines for EUD annotation.

To address this gap, we propose the first EUD annotation effort for a Brazilian Indigenous language, focusing on XCOMP constructions in Nheengatu. Several morphosyntactic phenomena in the language pose challenges for computational modeling, including pro-drop, clause chaining, and argument sharing, whose underlying syntactic relations are not recoverable from basic dependency trees alone. Moreover, unlike the canonical non-finite XCOMP patterns described in the literature, verbs heading XCOMP clauses in Nheengatu are finite and inflected for person. Our approach combines theoretical analyses of control and raising with a rule-based enhancement pipeline implemented in Python using the `conllu` library. Based on a corpus analysis of Nheengatu syntax, we formalize subject-control, object-control, raising-like configurations, and propagated subjects in the EUD layer, following the official UD guidelines.

The remainder of this paper is organized as follows. Section 2 contextualizes our work within Amerindian-language UD treebanks and prior EUD research. Section 3 discusses XCOMP constructions in Nheengatu. Section 4 presents the enhancement pipeline and heuristics. Section 5 reports the annotation results and error analysis, and Section 6 provides concluding remarks and future directions.

2. Related Work

As of the Universal Dependencies release 2.17 (November 15, 2025), the UD collection comprises 25 treebanks covering 24 Amerindian languages, including two treebanks for Mbya-Guarani [Thomas 2019]. Among these resources, UD_Nheengatu-CompLin is the second largest in terms of token count, surpassed only by UD_Bororo-BDT.

These 25 treebanks constitute only 7.4% of the 339 treebanks and less than 1% of the 37 million tokens of the UD collection [Zeman et al. 2025]. Among them, only two treebanks include EUD annotations: Western Sierra Puebla Nahuatl [Pugh et al. 2022] and Highland Puebla Nahuatl [Pugh and Tyers 2024].

Amerindian UD treebanks also vary in corpus depth (e.g., use of POS tags, features, genre, size), as reflected in the official UD evaluation metrics. Most receive between 0 and 2.5 stars[1], indicating ongoing development and uneven syntactic coverage. The UD_Nheengatu-CompLin, however, has a rating of 3.5 stars, the highest among Amerindian-language UD treebanks. Despite its small size (2,839 sentences), its 3.5-star rating matches that of UD_Portuguese-DANTEStocks [Di Felippo et al. 2021] and approaches the 4-star rating of UD_Portuguese-Porttinari [Duran et al. 2023], both of which represent a high-resource language.

Previous work by [Guillaume and Perrier 2021] presented a rule-based graph rewriting system that converted basic UD graphs into EUD graphs for 17 languages. For five high-resource languages, the authors derived language-specific lexicons of control and raising predicates to support their annotation. Besides the fact that Nheengatu has not previously received EUD, its existing lexical resources also lack a control/raising lexicon. As EUD annotation for Nheengatu expands, using the rule-based approach of [Guillaume and Perrier 2021] may become a viable alternative, a strategy that has already proved effective for Brazilian Portuguese [Sanches Duran et al. 2025].

For Nheengatu, the need for EUD arises from morphosyntactic properties that challenge syntactic representation, particularly overt subjects in open clausal complements, which constitute the starting point of the EUD annotation proposed in this

[1]This distribution was obtained automatically through the analysis of the `eval.log` files of all 25 Amerindian-language treebanks available in UD v2.17.

work. Their annotation involves not only canonical verbal XCOMP constructions but also finite embedded predicates, a phenomenon previously discussed for Portuguese by [de Alencar and Rademaker 2022].

3. Open Clausal Complements in Nheengatu

The concept and terminology of XCOMP as an *open clausal complement* originate in Lexical-Functional Grammar (LFG), particularly in the distinction between functional and anaphoric control formalized by [Bresnan et al. 2015]. In LFG, an XCOMP is defined as an open complement that lacks an overt subject, without direct evidence for an empty NP position within the complement clause [Kroeger 2004, p. 111]. Such constructions occur only when specified in the lexical entry of the matrix predicate, whose verbs are commonly referred to as *control predicates*.

Universal Dependencies adopts this view by analyzing XCOMP constructions as subjectless core arguments in obligatory control relations, typically controlled by the matrix object or subject [de Marneffe et al. 2025].

While Head-Driven Phrase Structure Grammar (HPSG), like LFG, analyzes XCOMP as a core argument whose subject is determined externally, the two frameworks differ in how they represent control and clause relations. LFG models XCOMP through functional control in f-structure, while HPSG represents these relations through structure-sharing and valency features. HPSG also makes a distinction between control and raising constructions, whereas LFG traditionally treats both through *functional control*.

Our analysis of UD_Nheengatu-CompLin identified four main XCOMP construction types: subject-control, object-control, raising-like, and subject propagation.

3.1. Subject Control

In subject-control constructions, the understood subject of the embedded predicate is obligatorily coreferential with the matrix subject; the EUD layer makes this explicit via an enhanced `nsubj` arc from the `xcomp` head to the matrix subject node [Schuster and Manning 2016]. Three semantic classes of subject-control predicates in Nheengatu were identified: desiderative predicates, motion predicates with a purposive interpretation, and emotive predicates.

Desiderative predicates. Following [Abeillé 2024], control predicates can be classified according to the semantic class of the matrix verb, with volitional, desiderative, and evaluative verbs typically triggering subject control. Figure 1 illustrates a prototypical desiderative construction, in which the matrix subject *Adana* controls the understood subject of the embedded predicate *mendari* ('marry').

Motion predicates with purposive interpretation. A productive subclass of subject-control constructions involves motion predicates, represented in our corpus by *sú* ('to go'), which accounts for 11 occurrences. In these constructions, the motion verb functions as a matrix head, selecting a dependent predicate denoting purpose, goal, or sequential action associated with the movement event.

Figure 2 illustrates this. The matrix nominal subject *Payé* ('shaman') controls the understood subject position of the embedded predicate *pusanú* ('to heal'). Although both verbs bear independent person-marking prefixes (3SG), a common pattern in Nheengatu,

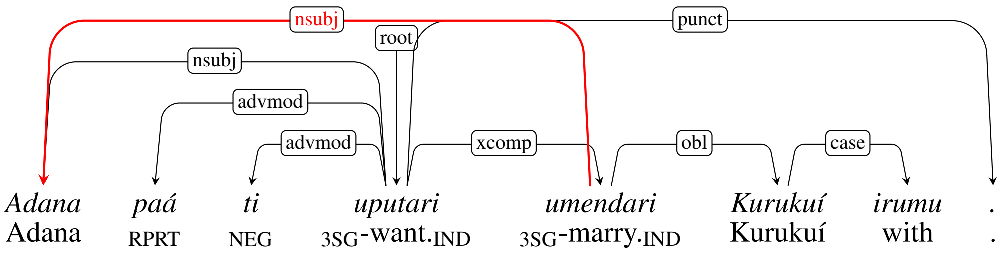

Figure 1. Enhanced dependencies for Casasnovas2006:12:15:181: 'Adana did not want to marry Kurukuí.' [Casasnovas 2006, p. 99]

they form a unified event in which the embedded predicate cannot license its own NP subject.

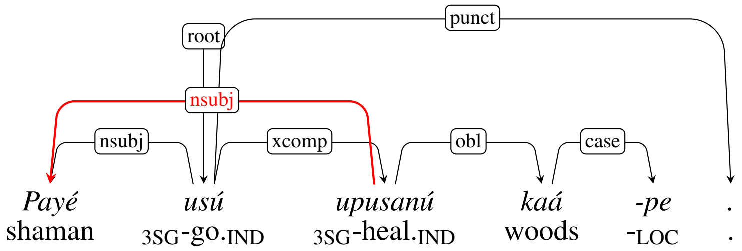

Figure 2. Enhanced dependencies for Hartt1938:0:0:509: 'The shaman went to heal in the woods'. [Hartt 1938, p. 356]

Emotive and preference predicates. A third subclass of subject control is formed by psychological and preferential predicates, exemplified by verbs such as *yumutari* ('like/enjoy'). Unlike desiderative predicates, which assign a volitional Agent role to the matrix subject, emotive predicates assign an Experiencer role: the matrix subject is an entity for whom the embedded event constitutes an object of affect or preference. Despite this semantic difference, the syntactic behavior is the same: the embedded predicate is an open clausal complement whose subject is obligatorily controlled by the matrix subject.

An example is shown in Figure 3. The matrix subject *Yawara* ('dog') functions as the Experiencer of the emotive predicate and obligatorily controls the understood subject of the embedded predicate *yenú* ('lie down').

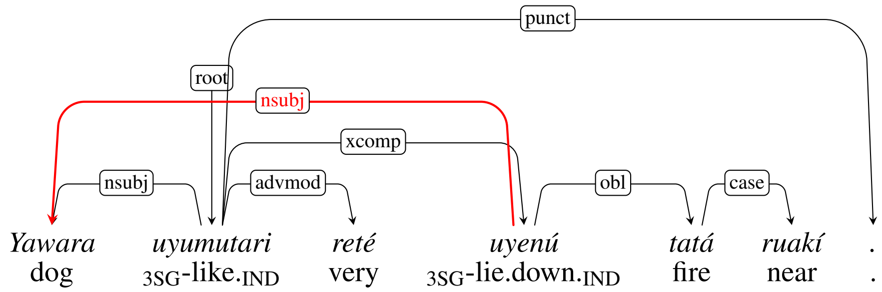

Figure 3. Enhanced dependencies for Hartt1938:0:0:141: 'The dog very much likes to lie down near the fire.' [Hartt 1938, p. 329]

Subject control also extends beyond canonical verbal predicates to inchoative constructions. The stative predicate *pitá* ('to become') selects adjectival or nominal

XCOMP complements controlled by the matrix subject, as in *upitá sasiára* ('[He] was sad').[2] When the matrix subject is phonologically null (pro-drop), no enhanced `nsubj` arc can be instantiated in the EUD layer, since the controller is recoverable only through verbal agreement morphology.

3.2. Object Control

In object-control constructions, the understood subject of the embedded predicate is coreferential with the matrix object. In the EUD representation, the same nominal node receives both a basic `obj` relation and an enhanced `nsubj` relation. Three main semantic subclasses of object-control predicates are attested in the corpus.

Manipulative predicates such as *purandú* ('ask'), *mundú* ('order/send'), and *mungitá* ('advise') denote events in which the matrix subject induces the matrix object to perform the embedded action. In Figure 4, *Jesus* functions both as the `obj` of *upurandú* and as the understood subject of the embedded predicate *uxari* ('let'), formalized in the EUD layer through an enhanced `nsubj` relation.

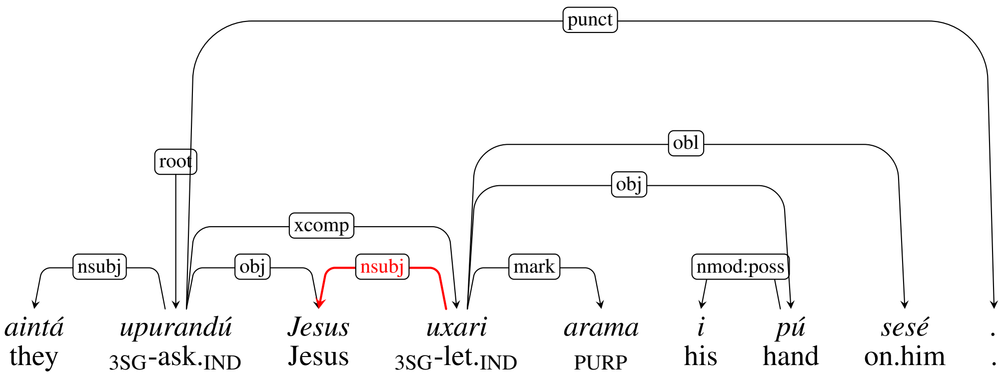

Figure 4. Enhanced dependencies for NTLN2019:5:1:37: 'they begged Jesus to touch him.'

Perception predicates such as *maã* ('see'), *sendú* ('hear'), and *saã* ('feel') form another subclass of object-control. In these, the matrix object denotes the perceived entity while simultaneously functioning as the understood subject of the embedded predicate. This pattern is illustrated in (1), where the object *indé* ('you') is interpreted as the understood subject of the predicate *amurupí* ('be different').

(1) *Wirandé xa-maã ranhẽ indé amurupí.*
yesterday 1SG-see.IND IMPF you.ACC differently
'Yesterday I still saw you differently.' [de Amorim 1928, p. 99]

Naming predicates such as *musera* ('name'), *seruka* ('name/call'), and *senúi* ('call') select a matrix object corresponding to the entity being named and a nominal `xcomp` complement expressing the assigned designation. Figure 5 illustrates this pattern with the predicate *musera* ('name'), where the object *aé* ('it') is interpreted as the understood subject of the nominal predicate *Carminha*.

[2] See sentence `Casasnovas2006:1:2:2` in UD_Nheengatu-CompLin.

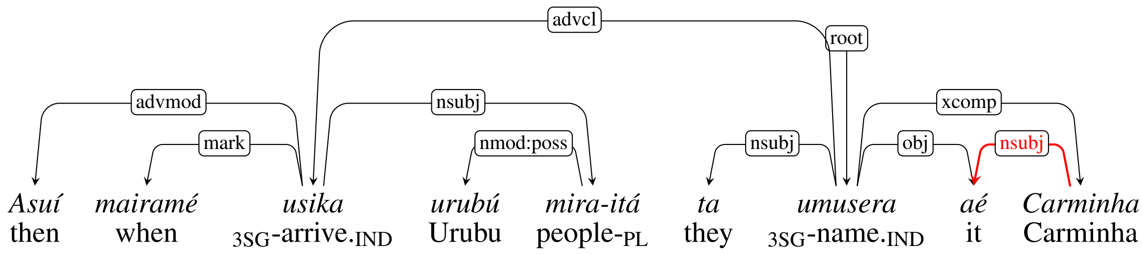

Figure 5. Enhanced dependencies for Avila2021:40:2:487: 'Then, when the people of the Urubu ethnicity arrived, they named it "Carminha".'

3.3. Raising-like predicates

Raising predicates do not assign an independent thematic role to the raised argument; its semantic role is determined by the embedded predicate [Kroeger 2004]. Although control and raising predicates are both annotated as `xcomp` in basic UD, we keep this distinction to facilitate rule refinement and future pipeline development. Since Nheengatu lacks non-finite verb forms, EUD annotation relies more heavily on lexical information.

The most frequent raising-like context is the inceptive aspectual predicate *yupirú* ('start'): the matrix subject's semantic role derives entirely from the embedded predicate, and *yupirú* itself contributes only onset aspectual information without introducing a participant role. Figure 6 illustrates this pattern.

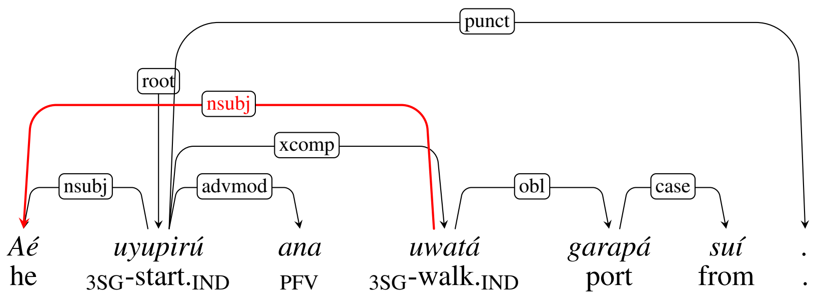

Figure 6. Enhanced dependencies for Navarro2016:0:0:150: 'He started walking from the port.'

3.4. Propagation of implicit subjects

A fourth category involves XCOMP controllers not found in the immediately local matrix domain, requiring recovery across coordination or paratactic boundaries. This pattern is frequent in the corpus: Nheengatu's use of clause-chaining via `parataxis` frequently produces chains in which a single overt subject controls embedded predicates several nodes away. Example (2) illustrates this: the noun phrase *Kurumĩ* ('boy') functions as the understood subject of the distant predicate *uyenú* ('lie down'). Making this kind of implicit argument sharing explicit is one of the enhancements the EUD layer was designed to capture [Schuster and Manning 2016].

(2) *Kurumĩ u-sú renundé u-inú suíwa rakapira upé urarí , arirí*
boy 3SG-go.IND before 3SG-put.IND his.arrow tip -LOC curare , then
u-sú u-yenú kuruatá pupé inayawasú aresé .
3SG-go.IND 3SG-lie.IND spathe inside inajá.AUG on .

'The boy first put poison on the tip of the arrow, then went to lie down on the spathe of the large inajá.' [Rodrigues 1890, p. 56]

4. Methodology

The pipeline consists of two scripts operating over CoNLL-U files: a quantitative feature-extraction script (`xcomp_context.py`) and a rule-based graph-enrichment script (`xcomp_enhancer.py`).[3]

4.1. Corpus profiling and feature extraction

We processed a pre-filtered subset containing only sentences with at least one `xcomp` relation. For each target token, the script extracted: (i) the UPOS tag and lemma of the matrix predicate; (ii) the UPOS tag and lemma of the embedded complement; (iii) the syntactic role of the matrix predicate (e.g., `root`, `parataxis`, `conj`); and (iv) the frequency of structural and lexical patterns (e.g., `VERB`→`xcomp(ADJ)`, *putari*→`xcomp(`*mendari*`)`).

The object-control predicate inventory comprises 17 lemmas and was derived from a frequency analysis of the 188 sentences containing `xcomp` relations. Predicates were then manually classified if they satisfied two criteria: (i) they occurred in at least one treebank instance with both a matrix object and an XCOMP complement, and (ii) the matrix object was interpreted as the implicit subject of the embedded predicate.

4.2. Controller resolution heuristics

The annotation script converts the frequency patterns identified in the XCOMP context analysis into a rule-based procedure guided by two main principles. First, **controller assignment is lexically conditioned**: a manually curated inventory (`OBJ_CTRL_VERBS`) identifies predicates that license object control, while all remaining predicates are treated as subject-control predicates, following the assumption that control is a lexical property of the matrix predicate [Bresnan et al. 2015, Kroeger 2004]. Second, **argument sharing is resolved either locally or through subject propagation**. When no overt controller is found in the local clause, the algorithm traverses `parataxis` and `conj` relations to identify controllers in higher structures, capturing recurrent pro-drop and clause-chaining patterns in Nheengatu. Table 1 summarizes the heuristic rules implemented in `xcomp_enhancer.py`.

Successful control resolution appends a dual-edge structure (`CTRL_ID:nsubj|HEAD_ID:xcomp`) to the DEPS field while preserving the original basic dependency tree. Cases that remain unresolved or already contain overt embedded subjects are automatically routed to a diagnostics log (`_skipped.tsv`) for manual review.

4.3. Manual Review

Following automatic enhancement, all annotations were manually validated using diagnostic logs for modified and unresolved cases. Manual review focused primarily on: (i) false negatives involving non-canonical object-control structures; (ii) ambiguous controllers in coordinated matrix predicates; (iii) complex `parataxis` chains; and (iv) pro-drop constructions with phonologically null controllers.

[3] All data and scripts used in this work are available at: `https://github.com/dominickmaia/EUDforNheengatu`.

Rule	Stage	Condition	Result
R1	Phase 1	Head lemma ∈ `OBJ_CTRL_VERBS` and overt `obj`/`iobj` present	Assign object as controller
R2	Phase 1	Overt `nsubj` of head (if R1 does not apply)	Assign subject as controller
R3	Phase 2	Dominating node (via `parataxis`/`conj`) has overt `nsubj` compatible with `agrees()`	Propagate subject as controller
R4	Phase 2	`advcl`/`conj` sibling of dominating node has `nsubj` compatible with `agrees()`	Propagate subject as controller
R5	Phase 2	A previously annotated `xcomp` is reached; controller inherited if compatible with `agrees()`	Propagate subject as controller
R6	Climbing	Current node is `advcl` with an *arama* `mark` dependent	Permit ascent
R7	Climbing	Current node deprel ∈ {`ccomp`, `acl`, `acl:relcl`}	Block ascent (clausal barrier)
R8	Pre-check	`xcomp` head already has an overt `nsubj` dependent	Skip (subject already present)
R9	Pre-check	No controller found after R1–R5	Skip (pro-drop)

Table 1. Heuristic rule mapping in `xcomp_enhancer.py`.

5. Results

The pipeline achieved 96.7% precision. Most errors involved incorrect controller propagation from higher clauses.

Type	Scope	Correct	Incorrect	Total	Precision
Subject control	Local	70	1	71	98.6%
Subject control	Propagated	14	3	17	82.4%
Object control	Local	33	0	33	100.0%
Total		**117**	**4**	**121**	**96.7%**

Table 2. Automatic annotation precision by control type and structural scope.

Among the 197 `xcomp` tokens, 71 (36.0%) involve subject control, 33 (16.8%) object control, and 17 (8.6%) long-distance propagation. The remaining 76 cases (38.6%) do not receive an enhanced `nsubj` relation because the controller is omitted (pro-drop) or requires discourse-level coreference beyond the scope of the current pipeline.

5.1. Error Analysis

The first error involves incorrect controller propagation across a `parataxis` chain. In sentence (3), the algorithm assigned *apigá* ('man') as the controller of the adjectival predicate *sasiára* ('sad') because it selected a compatible third-person argument while traversing the higher clause structure. However, this construction is better analyzed as object control: after the event *unupá aé* ('hit her'), the object pronoun *aé* ('her') functions as the understood subject of the following predicates *uyaxiú* ('cry') and *upitá sasiára* ('remain sad'). The script treats object-control separately through a manually curated inventory of object-control predicates, but this predicate was not included in that list, and the correct controller was not identified during propagation because both candidate arguments shared third-person features.

(3) *Ape , paá , nhaã apigá i piaíwa , u-nupá aé , yawé arã , paá , u-yaxiú*
then , RPRT , that man 3SG angry , 3SG-hit.IND her , so of , RPRT , 3SG-cry.IND
, u-pitá sasiára .
, 3SG-remain.IND sad .
'The man got angry and hit her, she cried and was sad.' [Casasnovas 2006, p. 89]

The second error involves incorrect controller propagation in a pro-drop configuration. In sentence (4), the algorithm incorrectly assigned *yakaré* ('alligator') as the controller of the adjectival predicate *pitunawasú* ('dark/darkness') because it selected the nearest structurally compatible third-person argument while traversing the `parataxis` chain. In this case, however, the subject of *upitá pitunawasú* ('get dark') is phonologically null, and no enhanced `nsubj` relation should be instantiated in the EUD layer.

(4) *Yawé waá , ta u-maramunha pukusá , yakaré u-suú tatá , u-puri i*
so REL , 3PL 3SG-fight.IND while , alligator 3SG-bite.IND fire , 3SG-jump.IND 3SG
irumu paraná kití , ape u-pitá pitunawasú yandé arã iké ara upé , umbaá
with river to , then 3SG-remain.IND darkness 1PL for here world in , NEG
mayé ya-maã arã .
how 1PL-see.IND PURP .
'So, while they were fighting, the alligator bit the fire and jumped into the river with it, then it got dark for us here in the world, we couldn't see.' [Avila 2021, p. 258]

The third error reveals a gap in the object-control rule. In (5), the algorithm propagated *aintá* ('they', token 5), the subject of the unrelated adjunct clause, as the controller of the `xcomp` predicate *itá* ('stone'). The correct one, however, is *aintá* ('them', token 11), the object of *remutatatinga* ('smoke'), the verb that governs the adjunct clause headed by *uyeréu* ('turn into'). Because this object is not a direct dependent of *uyeréu* but of the clause governing it, R1 does not recognize it as a local controller and fails to resolve it in Phase 1. Addressing it requires extending R1 to recognize objects of a governing clause as local controllers in adjunct clauses.

(5) *Kuíri , ti arama aintá u-musarái ne resé , re-mutatatinga aintá sikantá -pe*
now , NEG PURP 3PL 3SG-mock.IND 2SG of , 2SG-smoke.IMP 3PL resin -LOC
u-yeréu arama itá .
3SG-turn.IND PURP stone .
'Now, lest they mock you, smoke them with resin to turn them to stone.' [de Amorim 1928, p. 294]

The fourth error differs from the previous cases because it originated from an inconsistency in the original treebank annotation rather than from the enhancement procedure itself. In (6), the noun *Kurasí* ('sun') was incorrectly annotated as the `nsubj` of the parataxis predicate *asú* ('I go'), instead of being linked to the root predicate *iwaté* ('high'). As a result, the algorithm propagated *Kurasí* as the controller of the `xcomp` predicate *apurakí* ('work'). Since the head predicate is morphologically marked for first person (*a-*), a stricter person-agreement check would have blocked the incorrect propagation. The error nevertheless proved useful for identifying and revising the original annotation.

(6) *Kurasí iwaté ã , a-sú -putari a-purakí .*
sun high PFV , 1SG-go.IND -want 1SG-work.IND .
'The sun is already up, I want to go to work.' [Avila 2021, p. 350]

6. Conclusions and Future Work

The proposed pipeline achieved 96.7% precision, showing that EUD is feasible for a low-resource Indigenous language without previous EUD resources or specialized parsing tools. The pipeline also functioned as a diagnostic tool for treebank revision, identifying some pre-existing annotation inconsistencies that might otherwise have remained unnoticed. These included cases where `xcomp` dependents incorrectly contained overt `nsubj` dependents, as well as errors involving the attachment of `xcomp` relations to the wrong predicate head, improving the internal consistency of the treebank itself.

This work shows that EUD annotation for endangered and low-resource languages can be developed through rule-based methods combining corpus-based heuristics and manual validation. Future work includes enriching the JSON glossary used by the Yauti tool with valency information, integrating the enhancer into the Yauti automatic UD annotation tool [de Alencar 2023], and extending the approach to additional EUD phenomena beyond `xcomp` relations. The resulting guidelines also provide a starting point for future inter-annotator agreement experiments and for expanding EUD resources for Indigenous languages within the UD ecosystem.

We also plan to investigate the adaptation of the rule-based graph rewriting system of [Guillaume and Perrier 2021] to Nheengatu by using language-specific lexical information for control and raising predicates. This will allow us to compare this approach with the Python-based pipeline presented here in terms of annotation quality, computational complexity, processing speed, and ease of integration with the existing annotation tool, Yauti.

Limitations

The reported frequencies of object-control predicates reflect only the current version of the treebank, and as new sentences are annotated, both predicate frequencies and the inventory may change.

Furthermore, the pipeline was developed and applied by a single annotator experienced in UD annotation for Nheengatu, who also manually reviewed the 121 automatically generated EUD annotations and analyzed the annotation errors. The verb classifications were reviewed by a second experienced annotator. However, the annotation guidelines were not validated through an inter-annotator agreement experiment, and their consistency across independent annotators remains to be assessed.

Ethics Statement

All data used in this study come from the UD_Nheengatu-CompLin treebank, distributed as part of Universal Dependencies v2.17 [Zeman et al. 2025]. No new data involving human participants was collected. We recognize that computational research on Indigenous languages involves responsibilities toward the communities connected to these languages, and we support efforts to develop accessible and reusable NLP resources beyond academic contexts.

Generative AI disclosure

We used generative AI tools, including Claude and Gemini, to help write the Python scripts, assist with LaTeX editing and formatting, correct grammatical errors, and polish the overall clarity of the text. All linguistic analysis, annotation decisions, rule creation, error analysis, and final code verification were carried out by the authors.

Acknowledgements

We thank the anonymous reviewers for their insightful feedback and suggestions, which helped improve the quality of this paper.

References

Abeillé, A. (2024). Control and raising. In Müller, S., Abeillé, A., Borsley, R. D., and Koenig, J.-P., editors, *Head-Driven Phrase Structure Grammar: The Handbook*, number 9 in Empirically Oriented Theoretical Morphology and Syntax, pages 519–570. Language Science Press, Berlin, 2 edition.

Alencar, L. F. d., Silva, H. L. B., Gurgel, J. L., and Alexandre, D. M. (2025). *Per aspera ad astra*: Improving the automatic evaluation of a Universal Dependencies treebank for a low-resourced language. *Revista de Estudos da Linguagem*. In press.

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

Bresnan, J., Asudeh, A., Toivonen, I., and Wechsler, S. (2015). *Lexical-functional syntax*, volume 16. John Wiley & Sons.

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

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

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

de Alencar, L. F. (2024). A Universal Dependencies treebank for Nheengatu. In Gamallo, P., Claro, D., Teixeira, A. J. S., Real, L., García, M., Oliveira, H. G., and Amaro, R., editors, *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.

de Alencar, L. F. (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.

de Alencar, L. F. and Rademaker, A. (2022). Modelação da valência verbal numa gramática computacional do português no formalismo hpsg. *Domínios de Lingu@gem*, 16(4):1339–1400.

de Amorim, A. B. (1928). Lendas em Nheêngatu e em Portuguez. *Revista do Instituto Historico e Geographico Brasileiro*, 154(100):9–475. Tomo 100, vol. 154 (2º de 1926).

de Marneffe, M.-C., Ginter, F., Goldberg, Y., Hajič, J., Manning, C., McDonald, R., Nivre, J., Petrov, S., Pyysalo, S., Schuster, S., Silveira, N., Tsarfaty, R., Tyers, F., and Zeman, D. (2025). xcomp: open clausal complement. `https://universaldependencies.org/u/dep/xcomp.html`.

Di Felippo, A., Postali, C., Ceregatto, G., Gazana, L., Silva, E., Roman, N., and Pardo, T. (2021). Descrião preliminar do corpus DANTEStocks: Diretrizes de segmentação para anotação segundo Universal Dependencies. In Ruiz, E. E. S. and Torrent, T. T., editors, *Proceedings of the 13th Brazilian Symposium in Information and Human Language Technology*, pages 335–343, Porto Alegre, Brazil. Association for Computational Linguistics.

Duran, M., Lopes, L., das Graças Nunes, M., and Pardo, T. (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.

Guillaume, B. and Perrier, G. (2021). Graph rewriting for enhanced Universal Dependencies. In Oepen, S., Sagae, K., Tsarfaty, R., Bouma, G., Seddah, D., and Zeman, D., editors, *Proceedings of the 17th International Conference on Parsing Technologies and the IWPT 2021 Shared Task on Parsing into Enhanced Universal Dependencies (IWPT 2021)*, pages 175–183, Online. Association for Computational Linguistics.

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

Kroeger, P. (2004). *Analyzing Syntax: A Lexical-Functional Approach*. Cambridge University Press, Cambridge.

Navarro, E. d. A. (2012). O último refúgio da língua geral no Brasil. *Estudos Avançados*, 26(76):245–254.

Nivre, J. et al. (2016). Universal Dependencies v1: A multilingual treebank collection. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1659–1666, Portorož, Slovenia. European Language Resources Association (ELRA).

Pagano, A. S., Duran, M. S., and Pardo, T. A. S. (2023). Enhanced dependencies para o português brasileiro. In Pardo, T. A. S., Duran, M. S., and Lopes, L., editors, *Proceedings of the 2nd Edition of the Universal Dependencies Brazilian Festival*, pages 461–470, Belo Horizonte, Brazil. Association for Computational Linguistics.

Pugh, R., Huerta Mendez, M., Sasaki, M., and Tyers, F. (2022). Universal Dependencies for western sierra Puebla Nahuatl. In Calzolari, N., Béchet, F., Blache, P., Choukri, K., Cieri, C., Declerck, T., Goggi, S., Isahara, H., Maegaard, B., Mariani, J., Mazo, H., Odijk, J., and Piperidis, S., editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 5011–5020, Marseille, France. European Language Resources Association.

Pugh, R. and Tyers, F. (2024). A Universal Dependencies treebank for Highland Puebla Nahuatl. In Duh, K., Gomez, H., and Bethard, S., editors, *Proceedings of the 2024*

Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), pages 1393–1403, Mexico City, Mexico. Association for Computational Linguistics.

Rodrigues, J. B. (1890). *Poranduba amazonense ou kochiyma-uara porandub, 1872-1887*. Typ. de G. Leuzinger & Filhos, Rio de Janeiro.

Sanches Duran, M., de Souza, E. A., das Graças Volpe Nunes, M., Pagano, A. S., and Pardo, T. A. S. (2025). Extending the enhanced Universal Dependencies – addressing subjects in pro-drop languages. In Bouma, G. and Çöltekin, Ç., editors, *Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025)*, pages 143–152, Ljubljana, Slovenia. Association for Computational Linguistics.

Schuster, S. and Manning, C. D. (2016). Enhanced English Universal Dependencies: An improved representation for natural language understanding tasks. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2371–2378, Portorož, Slovenia. European Language Resources Association (ELRA).

Souza, E., Duran, M., Nunes, M. d. G. V., Sampaio, G., Belasco, G., and Pardo, T. (2024). Automatic annotation of enhanced Universal Dependencies for Brazilian Portuguese. In Claro, D. B. and Pagano, A., editors, *Proceedings of the 15th Brazilian Symposium in Information and Human Language Technology*, pages 22–31, Belém do Pará, Brazil. Association for Computational Linguistics.

Thomas, G. (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.

Universal Dependencies (2026). Enhanced Universal Dependencies syntax overview. `https://universaldependencies.org/u/overview/enhanced-syntax.html`. Accessed: 2026-05-01.

Zeman, D. et al. (2025). Universal dependencies 2.17. Accessed: 2026-05-14.