Universal Dependencies for 19th-Century Nheengatu from the Lower Amazon Region

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

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

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

Abstract. *We present the morphosyntactic annotation of Nheengatu as spoken in the 19th century in the Lower Amazon region. The annotated data expand the UD_Nheengatu-CompLin treebank, the first for Nheengatu in the Universal Dependencies project, by incorporating forms and syntactic patterns characteristic of that region and time. We describe the corpus source, the orthographic normalization process, and the main annotation strategies used. So far, 345 sentences have been annotated, with 310 already integrated into the current version of the treebank. This historical data annotation enhances the lexical and morphosyntactic coverage, supporting the documentation and computational modeling of Nheengatu.*

1. Introduction

The Amazonian General Language (*Língua Geral Amazônica*, LGA) emerged in the 17th century from the Tupinambá variety of Tupi spoken in Maranhão, shaped by colonial contact and missionary activity in northern Brazil [Borges 1996, Rodrigues 1996, Freire 2011, Rodrigues and Cabral 2011, Moore 2014]. Later known as Nheengatu—literally "good language"—the LGA is first attested under this name in [Seixas 1853].

During the 18th and 19th centuries, Nheengatu spread throughout the Amazon basin, reaching as far as Venezuela and Colombia, and became the most spoken language in the Brazilian Amazon [Navarro et al. 2017]. Despite its historical prominence, Nheengatu is now facing challenges to its intergenerational transmission [Navarro 2012]. According to [Eberhard et al. 2025], approximately 14,000 people still speak Nheengatu today, with around 6,000 speakers located in Brazil.

As a contribution to ongoing efforts to construct computational resources and tools for endangered languages [Galves et al. 2017, Vasquez et al. 2018, Thomas 2019, de Alencar 2021, Tyers and Henderson 2021, Park et al. 2021, Rueter et al. 2021, Martín Rodríguez et al. 2022, Zariquiey et al. 2022, Sandalo and Galves 2023, da Silva and Pardo 2024, Pinhanez et al. 2024, Pugh and Tyers 2024, Santos et al. 2024], the UD_Nheengatu-CompLin treebank was introduced in version 2.11 of the Universal Dependencies (UD) collection in November 2022 [de Alencar 2024a, de Alencar 2024b]. It is the first and, to date, only syntactic treebank available for this language.

The morphosyntactic annotation of Nheengatu sentences in UD_Nheengatu-CompLin is performed using Yauti, a rule-based analyzer developed for the language [de Alencar 2023]. Since incorporating a 19th-century variety of Nheengatu spoken in the Lower Amazon River region into the treebank, some annotation decisions and modifications to the Yauti tool have been necessary.

This paper describes the annotation workflow for this historical *corpus*, from spelling normalization to morphosyntactic annotation, highlighting the key decisions, adjustments, and improvements made during the process.

2. The Lower Amazon Nheengatu

The primary source for the *corpus* used in this study is [Hartt 1938], a posthumous work by Canadian-American geologist and ethnographer Charles Frederick Hartt (1840–1878). In this work, Hartt adopts a fieldwork-based approach, focusing on the Lower Amazon region. The variety Hartt documented in this region exhibits some archaic morphosyntactic features characteristic of its time and location. Among these are the first-person pronominal prefix *xa-*[1] and the imperative prefix *e-* for the second-person singular, an archaism dating back to Old Tupi [de Almeida Navarro 2005].

Most notably, Hartt recorded a series of first- and second-person dative pronouns, derived from suffixing of the enclitic postpositon *bọ* [Hartt 1872, Navarro 2015, Avila 2021]. Although largely replaced by the postposition *arama* in most Nheengatu varieties at the time, Hartt reported residual usage of *bọ* in the dialect he observed [Hartt 1872, p. 67]. Example 1 illustrates both the imperative prefix *e-* and the first-person dative pronoun.

(1) *E-purú ne kisawa ixéu.*
2SG.IMP-lend your hammock 1SG.DAT
'Lend me your hammock!' [Hartt 1938, p. 380]

These features have been integrated into the most recent version of Yauti for proper analysis and will be discussed in more detail in the following section.

3. Annotation workflow

3.1. Spelling normalization

The historical variety of Nheengatu documented by [Hartt 1938] presents distinct phonological and morphosyntactic characteristics that diverge from the varieties previously annotated in the UD_Nheengatu-CompLin treebank. The main challenge prior to annotation was orthographic normalization of Hartt's original texts. To address this, we adopted a conservative approach based on the central guideline of preserving the original form as much as possible. Our primary reference was [Avila 2021], which includes previously adapted versions of some of Hartt's sentences. These were incorporated as metadata under the attributes **`text_sec`** (secondary text) and **`text_sec_source`** (secondary text source), as exemplified in Figure 1.

However, re-adaptation was often necessary to enhance fidelity to the historical forms attested by Hartt. Our approach differs from Avila's by generally retaining segmental and syllabic structures closer to the original, applying only minimal modernization

[1]The variant *xa-* is frequently attested in 19th-century records from different locations [de Magalhães 1876, Rodrigues 1890, Studart 1926], but it is no longer used in contemporary Nheengatu of the Upper Rio Negro. This form represents an innovation that is not found in Old Tupi or in 18th-century LGA records. We thank an anonymous reviewer for suggesting this clarification.

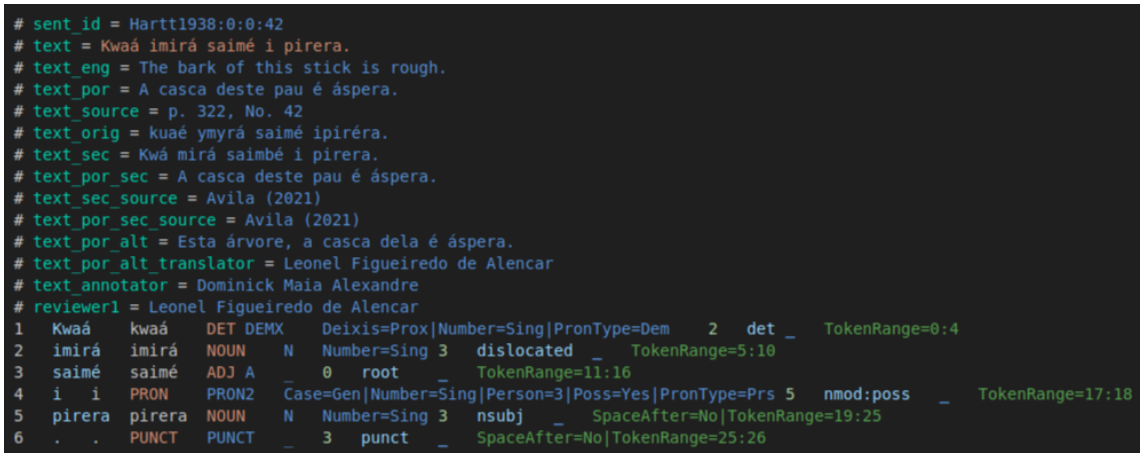

Figure 1. Annotation and metadata for Hartt's sentence No. 42 in UD_Nheengatu-CompLin.

for internal consistency. Compare, for example, the sentence *xamonó seygára táua kyty amú irané oyuyr aráma*, which [Avila 2021, p. 243] adapts as *Amundú se igara tawa kití, amú-wirandé uyuíri arama.*, with our version in (2).

(2) *Xamunú* *se* *igara* *tawa* *kití* *amú* *wirané* *uyuíri*
1SG.ACT:send my canoe village to another tomorrow 2SG.ACT:go_back
arama.
to
'I send my canoe to the village to return the day after tomorrow.' [Hartt 1938, p. 344]

In our adaptations, we preserve the original punctuation, refraining from inserting a comma between the two clauses of the example under discussion, unlike [Avila 2021]. We limit ourselves to capitalizing the initial letter of sentences, which [Hartt 1938] consistently renders in lowercase. In (2), we retain the first-person singular prefix *xa-* in the verb form *xamunú*, which contrasts with the contemporary form *amundú*. We also maintain the syncope of the voiced plosive /d/ in both *xamunú* and *wirané*, following Hartt's phonological representation and recognizing this as a productive process in the 19th-century variety. Finally, we generally do not hyphenate phrases such as *amú wirané* 'day after tomorrow', which [Avila 2021] treats as compounds. In UD theory, such expressions are not considered compounds, as they exhibit no morphosyntactic idiosyncrasy compared to regular modifier–noun constructions [de Marneffe et al. 2024].

Once orthographic adaptation is completed, the resulting sentences are submitted for morphosyntactic annotation, as described in the following section.

3.2. Morphosyntactic annotation

We annotated our *corpus* using Yauti, a rule-based morphosyntactic analyzer developed specifically for Nheengatu [de Alencar 2023], built on the CoNLL-U Parser library.[2] In a Python environment, sentences are input as strings, and Yauti outputs an analysis in

[2] https://pypi.org/project/conllu/

the CoNLL-U format [de Marneffe et al. 2024], specifying lemmas, part-of-speech tags, morphological features, syntactic heads, and dependency labels, among other information. All annotations undergo manual review to ensure linguistic accuracy and consistency. In Figure 2, we show how to import the Yauti module and call the function used to analyze sentences from [Hartt 1938].

```
>>> import Yauti
>>> sent = '''169 - yauára ikyrymáua uaé opurusuú.
169 - Yawara i kirimawa waá upurusuú.
169 - Yawara kirimbawa waá upurusuú. (Hartt, 331, adap.) - O cachorro valente morde gente.
169 - O cachorro valente morde gente.'''
>>> Yauti.parseExampleHartt(sent,331,annotator='Dominick Maia Alexandre')
# sent_id = Hartt1938:0:169:169
# text = Yawara i kirimawa waá upurusuú.
# text_por = O cachorro valente morde gente.
# text_source = p. 331, No. 169
# text_orig = yauára ikyrymáua uaé opurusuú.
# text_eng = TODO
# text_orig_transcriber = Antônio Levy Melo Nogueira
# text_por_modernizer = Antônio Levy Melo Nogueira
# text_sec = Yawara kirimbawa waá upurusuú.
# text_por_sec = O cachorro valente morde gente.
# text_sec_source = Avila (2021)
# text_por_sec_source = Avila (2021)
# inputline = Yawara i kirimawa waá upurusuú.
# text_annotator = Dominick Maia Alexandre
1       Yawara  yawara  NOUN    N       Number=Sing     3       nsubj   _       TokenRange=0:6
2       i       i       CCONJ   CCONJ   _       3       cc      _       TokenRange=7:8
2       i       i       PRON    PRON2   Case=Gen|Number=Sing|Person=3|PronType=Prs      0       _       _       TokenRange=7:8
3       kirimawa        kirimawa        ADJ     A       _       0       _       _       TokenRange=9:17
3       kirimawa        kirimawa        ADV     ADVA    AdvType=Man     0       advmod  _       TokenRange=9:17
3       kirimawa        kirimawa        VERB    V2      Mood=Ind|VerbForm=Fin   0       advcl   _       TokenRange=9:17
4       waá     waá     PRON    REL     Number=Sing|PronType=Rel        3       nsubj   _       TokenRange=18:21
4       waá     waá     SCONJ   SCONJ   _       3       mark    _       TokenRange=18:21
5       upurusuú        upurusuú        _       _       _       0       _       _       SpaceAfter=No|TokenRange=22:30
6       .       .       PUNCT   PUNCT   _       1       punct   _       SpaceAfter=No|TokenRange=30:31
```

Figure 2. Demonstration of Yauti usage.

```
>>> sent = '''169 - yauára ikyrymáua uaé opurusuú.
169 - Yawara i/pron2 kirimawa/v2 waá/rel upurusuú.
169 - Yawara kirimbawa waá upurusuú. (Hartt, 331, adap.) - O cachorro valente morde gente.
169 - O cachorro valente morde gente.'''
>>> Yauti.parseExampleHartt(sent,331,annotator='Dominick Maia Alexandre')
# sent_id = Hartt1938:0:169:169
# text = Yawara i kirimawa waá upurusuú.
# text_por = O cachorro valente morde gente.
# text_source = p. 331, No. 169
# text_orig = yauára ikyrymáua uaé opurusuú.
# text_eng = TODO
# text_orig_transcriber = Antônio Levy Melo Nogueira
# text_por_modernizer = Antônio Levy Melo Nogueira
# text_sec = Yawara kirimbawa waá upurusuú.
# text_por_sec = O cachorro valente morde gente.
# text_sec_source = Avila (2021)
# text_por_sec_source = Avila (2021)
# inputline = Yawara i/pron2 kirimawa/v2 waá/rel upurusuú.
# text_annotator = Dominick Maia Alexandre
1       Yawara  yawara  NOUN    N       Number=Sing     3       nsubj   _       TokenRange=0:6
2       i       i       PRON    PRON2   Case=Gen|Number=Sing|Person=3|PronType=Prs      3       expl    _       TokenRange=7:8
3       kirimawa        kirimawa        VERB    V2      Mood=Ind|VerbForm=Fin   1       acl:relcl       _       TokenRange=9:17
4       waá     waá     PRON    REL     Number=Sing|PronType=Rel        3       nsubj   _       TokenRange=18:21
5       upurusuú        purusuú VERB    V       Mood=Ind|Person=3|VerbForm=Fin  0       root    _       SpaceAfter=No|TokenRange
=22:30
6       .       .       PUNCT   PUNCT   _       5       punct   _       SpaceAfter=No|TokenRange=30:31
```

Figure 3. Disambiguation of the ambiguous forms in `Hartt1938:0:0:42` through manual specification in the input string.

In cases of ambiguity, as exemplified in Figure 2 with *i*, which may be either a pronoun or a conjunction, disambiguation is performed by adding XPOS tags directly in the input string, i.e., `i/pron2`), as shown in Figure 3. The inventory of XPOS tags distinguishes over 80 fine-grained, treebank-specific word classes [de Alencar 2024a, de Alencar 2024b]. For example, following [Navarro 2016, Avila 2021], primary pronouns (`PRON`) are distinguished from secondary pronouns (`PRON2`). Both pronoun types are mapped to `PRON` in the universal inventory of 17 universal part-of-speech categories [de Marneffe et al. 2021]. Words not present in the lexicon are manually added to a glossary. For each lemma in the glossary, Yauti generates its full inflectional paradigm

[de Alencar 2023]. For example, for *munú* 'to send' (Listing 1), Yauti builds all conjugated verb forms, some of which are shown in Listing 2.

Listing 1. Sample entries from Yauti's glossary.

```
[   {     "lemma": "mundú",
            "pos": "v.",
            "gloss": "mandar, enviar"},
    {       "lemma": "munú",
            "pos": "v.",
            "gloss": "var. mundú"}, ]
```

Listing 2. Sample entries from Yauti's full-form lexicon.

```
{ "xamunú": [[ "munú", "V+ARCH+IND+1+SG"]],
    "emunú": [["munú","V+IMP+2+SG"]],
   "pemunú": [["munú", "V+IMPIND+2+PL"]],
  "remunú": [["munú", "V+IMPIND+2+SG"]],
  "amunú": [["munú", "V+IND+1+SG"]],
  "hamunú": [["munú", "V+IND+1+SG"]],
  "umunú": [["munú", "V+IND+3"]],
  "munú": [["munú", "V+NFIN"]] }
```

As reported by [de Alencar 2023], Yauti achieved a labeled attachment score of 73.2 in a previous version of UD_Nheengatu-CompLin with all 1,022 sentences annotated with disambiguating tags and other special tags that enable the tool to analyze unknown words. This indicates that Yauti's performance is neither exhaustive nor error-free, so its output is always reviewed by a human annotator before being incorporated into the treebank [de Alencar 2024a]. Furthermore, human intervention is required to continuously update the lexicon as new forms are encountered in the *corpus*. In the following section, we discuss specific annotation decisions made to handle characteristic features of the Nheengatu variety documented by [Hartt 1938].

3.3. Specific guidelines

The syntactic annotation of Hartt's 19th-century Nheengatu *corpus* required special attention to morphosyntactic constructions that diverge from patterns in contemporary varieties of the language. Below, we outline key phenomena that informed guideline development and lexicon adjustments in the Yauti annotation workflow. These guidelines are grounded in empirical evidence from Hartt's documentation and aim to ensure consistency with both descriptive linguistics and the principles of the Universal Dependencies (UD) framework [de Marneffe et al. 2021].

A notable characteristic of this historical variety is the use of the prefix *e-* to mark second-person singular imperative forms. This prefix is absent in the modern spoken variety of the Upper Rio Negro region, where *re-* is used with both indicative and imperative utterances [da Cruz 2011, Navarro 2016, Avila 2021]. This variation is reflected in the `lexicon.json` file through the inclusion of both forms with appropriate morphological distinctions (Listing 2). The imperative form prefixed with *e-* is automatically recognized and annotated as imperative mood (**`Mood=Imp`**), whereas forms with *re-* require manual disambiguation due to their multifunctionality (**`Mood=Imp,Ind`**). This approach ensures coverage of both imperative formation strategies within the UD schema.

The contrast between the two inflectional prefixes is illustrated in examples (3) and (4). The former is our adaptation of the original example by [Hartt 1938], which reads `emonó payé piám.`, while the latter is the modernization by [Avila 2021].

(3) *E-munú payé piamu.*
2SG.IMP-send shaman for
'Send someone to fetch the shaman.' [Hartt 1938, p. 339]

(4) *Re-mundú [aé] payé piamu.*
2SG.ACT-send him shaman for
'Send [him] to get the shaman.' [Avila 2021, p. 595]

In both examples, *munú* 'to send' appears in an imperative context but is prefixed differently. In (3), the prefix *e-* functions as a marker of the second-person singular imperative, marking the clause as a directive addressed to the addressee. In (4), the prefix *re-*, which in modern Nheengatu typically encodes the indicative mood in the second-person singular, occurs in an utterance with identical pragmatic force. This change may reflect a process of morphological simplification, whereby the Old Tupi imperative prefix *e-*, also attested in other 19th-century Nheengatu varieties, was gradually replaced by the prefix *re-* (*ere-* in Old Tupi), which originally marked indicative mood.

Hartt's documentation also records the use of dative pronominal forms, e.g., *ixéu*, *indéu*, *yanéu* ('to me', 'to you', 'to us'), reflecting a pattern inherited from Old Tupi pronouns performing the same function (*ixébo*, *endébo*, *îandébo*) [Avila 2021, p. 326]. These forms contrast with the more productive construction involving the postposition *arama* (e.g., *ixé arama*, 'to me'). Although Hartt recorded the complete series of dative pronouns in sentences collected in the Lower Amazon during the 1870s, showing frequent use in that region, other contemporary records are scarce, suggesting that these pronouns had limited temporal and geographic distribution in the early Nheengatu period [Avila 2021, p. 326].

To capture this variation, we included dative pronouns in the Yauti lexicon and annotated them using the dependency relation **`iobj`** (indirect object), which captures recipient-like arguments traditionally associated with the dative case, as illustrated in Figure 4.

The prefix *xa-* appears consistently in Hartt's data as a marker of first-person singular subject agreement on verbs, with just a single occurrence of *ha-* attested in the portion of the corpus examined so far. By contrast, the prefix *a-* is the sole form found in the modern Upper Rio Negro variety [da Cruz 2011, p. 133]. As exemplified in Listing 2, the lexical database includes verb forms with all three prefixes, provided with features that enable Yauti to automatically annotate first-person forms across different varieties, while signaling the *xa-* forms with `Style=Arch`.

In example (5), the main verb *xakitika* 'I grate' and the auxiliary *xaikú* 'I am' both instantiate the prefix *xa-*, indicating first-person singular agreement.

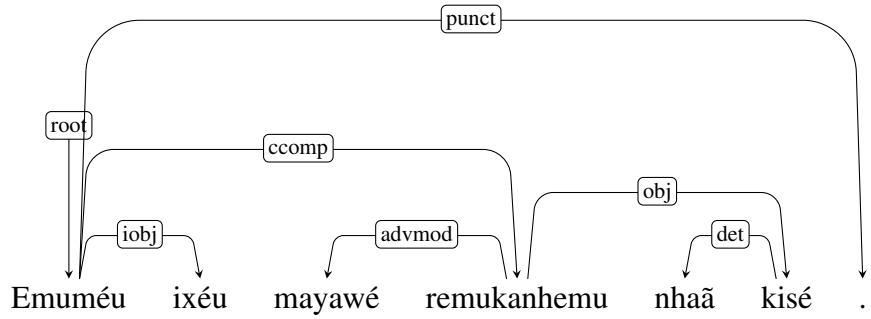

Figure 4. Dependency relations of *Emuméu ixéu mayawé remukanhemu nhaã kisé.* 'Tell me how you lost that knife.' [Hartt 1938, p. 331]

(5) *Xakitika xaikú mangarataya.*
2SG.ACT:grate 2SG.ACT:be ginger
'I'm grating ginger.' [Hartt 1938]

Some verbal forms in Hartt's records lack overt morphological markers of person, tense, or mood. These appear most often in imperative or permissive contexts, and their interpretation depends on context and the Portuguese translation provided by [Hartt 1938]. For such forms, we maintain their original orthography in the annotation but encode expected canonical forms using the **`CorrectForm`**, **`StandardForm`**, or **`StandardMood`** attributes in the `MISC` column, as illustrated in Figure 5,[3] following the UD guidelines [de Marneffe et al. 2024].

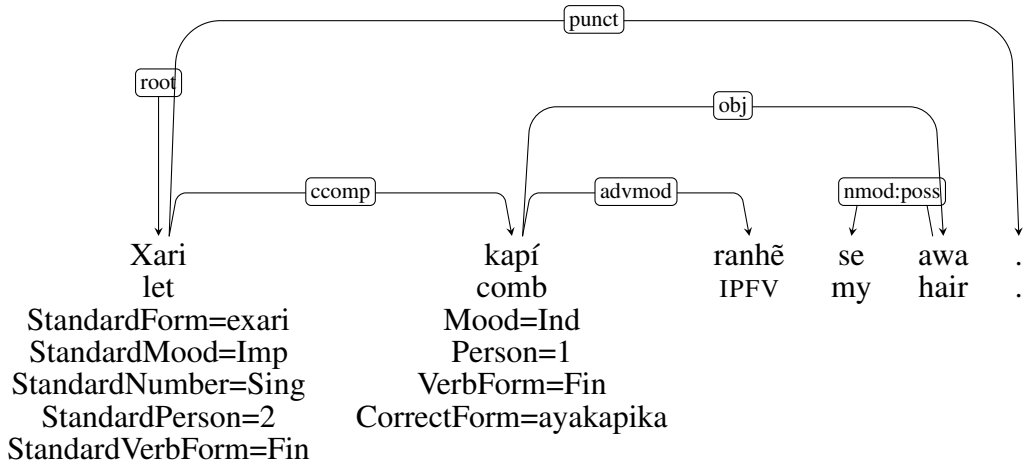

Figure 5. Dependency analysis with morphological and `MISC` features for *Xari kapí ranhẽ se awa.* 'Let me comb my hair.' [Hartt 1938, p. 334]

[3] An anonymous reviewer proposes an alternative interpretation of the example shown in Figure 5, hypothesizing a transcription error in [Hartt 1938]. The reviewer reads the sentence as *saiakapy ranhé seáua* and suggests the corresponding adaptation *xa-yakapi(ka) ranhẽ se awa* (1SG.ACT-comb IPFV my hair, 'I'm still going to comb my hair').

4. Final remarks

This study contributes to the expansion and refinement of the UD_Nheengatu-CompLin treebank through the morphosyntactic annotation of a historical 19th-century variety of Nheengatu. The annotated *corpus*, extracted from [Hartt 1938], provides valuable data on grammatical structures and lexical forms that are no longer productive in contemporary usage, thereby broadening the descriptive and typological scope of the treebank.

To date, 345 of the 919 sentences identified in Hartt's work have been annotated, of which 310, totaling 2,031 words, are included in the current development version of UD_Nheengatu-CompLin. All sentences with the identifier `Hartt1938` have undergone official UD validation through the `validate.py` script, the primary tool for assessing a treebank's eligibility for inclusion in a UD release. These 310 sentences triggered only 12 Udapi "bugs", well below the threshold of 203 (1 per 10 words). All bugs are of the `degree-upos` type, arising from the assignment of the `DEGREE` feature to nouns or stative verbs—the latter functioning as adjectives in languages such as Portuguese or English. Udapi [Popel et al. 2017] is a key component of the UD rating system, which assigns 0–5 stars to treebanks. Annotation of the remaining sentences is ongoing. Figure 6 summarizes the revision status of the 310 sentences, showing the number of sentences reviewed once, twice, or still pending review by an additional human annotator.

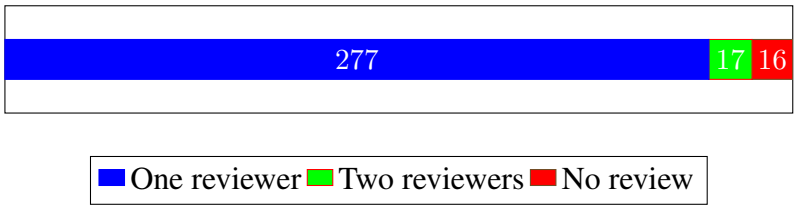

Figure 6. Revision status of `Hartt1938` sentences (n = 310) in the `yrl_complin-ud-train.conllu` file

Additionally, 82 new lexical entries have been added to the `glossary.json` file of the Yauti annotation tool. These entries correspond to lemmas, such as verbs, adjectives, nouns, particles, and adverbs, and represent the number of new words introduced during the annotation of this historical variety. In parallel, the `lexicon.json` file was expanded with all relevant inflected forms, especially for verbs, ensuring that each new lemma is fully represented across its morphological paradigm.

All updates, discussions, and further developments related to the annotation process are being conducted transparently in the project's official GitHub repository[4], where the community can follow the ongoing refinement of this resource and participate in the discussions.

By integrating these older forms into the treebank, this work enhances the typological and diachronic coverage of Nheengatu within the Universal Dependencies framework and supports ongoing efforts in linguistic documentation and the development of NLP tools for under-resourced and endangered languages.

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

References

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

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

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

da Silva, D. P. G. and Pardo, T. A. S. (2024). Grammar induction for Brazilian indigenous languages. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *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.

de Alencar, L. F. (2021). Uma gramática computacional de um fragmento do nheengatu / A computational grammar for a fragment of nheengatu. *Revista de Estudos da Linguagem*, 29(3):1717–1777.

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

de Alencar, L. F. (2024b). 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, PROPOR 2024, Santiago de Compostela, Galicia/Spain, 12-15 March, 2024*, volume 2, pages 37–54, Stroudsburg, PA, USA. Association for Computational Linguistics.

de Almeida Navarro, E. (2005). *Método Moderno de Tupi Antigo: a Língua do Brasil dos Primeiros Séculos*. Global, São Paulo, 3 edition.

de Magalhães, J. V. C. (1876). *O selvagem*. Typographia da Reforma, Rio de Janeiro.

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., Zeldes, A., and Zeman, D. (2024). Universal Dependencies Guidelines. Accessed: 2025-08-06.

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

Eberhard, D. M., Simons, G. F., and Fennig, C. D., editors (2025). *Ethnologue: Languages of the World*. SIL International, Dallas, 28 edition.

Freire, J. R. B. (2011). *Rio Babel: A história das línguas na Amazônia*. EdUERJ, Rio de Janeiro, 2 edition.

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

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

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].

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

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

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

Navarro, E. d. A. (2015). *Dicionário tupi antigo, a língua indígena clássica do Brasil: vocabulário português-tupi e dicionário tupi-português, tupinismos no português do Brasil, etimologias de topônimos e antropônimos de origem tupi*. Global.

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

Navarro, E. d. A., Ávila, M. T., and Trevisan, R. G. (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.

Park, H. H., Schwartz, L., and Tyers, F. M. (2021). Expanding universal dependencies for polysynthetic languages: A case of st. lawrence island yupik. In *Proceedings of the 1st Workshop on NLP for Indigenous Languages of the Americas (AmericasNLP)*, Online. Association for Computational Linguistics.

Pinhanez, C., Cavalin, P., and Nogima, J. (2024). Human evaluation of the usefulness of fine-tuned English translators for the Guarani mbya and nheengatu indigenous languages. In Gamallo, P., Claro, D., Teixeira, A., Real, L., Garcia, M., Oliveira, H. G., and Amaro, R., editors, *Proceedings of the 16th International Conference on Computational Processing of Portuguese - Vol. 2*, pages 32–36, Santiago de Compostela, Galicia/Spain. Association for Computational Lingustics.

Popel, M., Žabokrtský, Z., and Vojtek, M. (2017). Udapi: Universal API for Universal Dependencies. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies (UDW 2017)*, pages 96–101, Gothenburg, Sweden. Association for Computational Linguistics.

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, A. D. (1996). As línguas gerais sul-americanas. *Papia*, 4(2):6–18.

Rodrigues, A. D. and Cabral, A. S. A. C. (2011). A contribution to the linguistic history of the Língua Geral Amazônica. *ALFA: Revista de Linguística*, 55(2).

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

Rueter, J. et al. (2021). Apurinã Universal Dependencies treebank. In Mager, M. et al., editors, *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 28–33, Online. Association for Computational Linguistics.

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

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

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

Studart, J. (1926). Ligeiras noções de língua geral. *Revista do Instituto do Ceará*, 40:26–38.

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.

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

Vasquez, A. et al. (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.

Zariquiey, R., Alvarado, C., Echevarría, X., Gomez, L., Gonzales, R., Illescas, M., Oporto, S., Blum, F., Oncevay, A., and Vera, J. (2022). Building an endangered language resource in the classroom: Universal Dependencies for kakataibo. 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 3840–3851, Marseille, France. European Language Resources Association.