Enhancing a Nheengatu Morphosyntactic Analyzer for Word Formation and Non-standard Language

Leonel Figueiredo de Alencar[1]

[1]Universidade Federal do Ceará (UFC), Brazil

Av. da Universidade 2683 – 60.020-181 – Fortaleza – CE – Brazil

leonel.de.alencar@ufc.br

Abstract. *UD_Nheengatu-CompLin is the highest-rated and second-largest Amerindian language treebank in Universal Dependencies. It is annotated with Yauti, an analyzer for Nheengatu that uses special tags to handle unknown words. This paper presents a major revision of the special tag mechanism, extending coverage to phenomena such as reduplication, typos, and stylistic variation. A multi-level validator was implemented and passed 416 test cases. All 231 treebank sentences with special tags parsed, yielding a macro F_1 score of 0.92 and both weighted and micro F_1 scores of 0.96 in feature assignment.*

1. Introduction

With 319 treebanks for 179 languages, Universal Dependencies (UD) has become the *de facto* standard for cross-linguistic morphosyntactic annotation [de Marneffe et al. 2021, Zeman 2023, Zeman et al. 2025]. A wide range of applications in both linguistics and NLP highlights its relevance [Futrell et al. 2015, Kondratyuk and Straka 2019, Li 2025].

The UD collection contains 25 treebanks for 24 Amerindian languages, 14 of which are from Brazil, including Nheengatu—a descendant of Tupinambá, adopted as a *lingua franca* by colonizers in the 16$^{\text{th}}$–17$^{\text{th}}$ centuries, and now considered endangered [Rodrigues 1996, Freire 2011, Navarro et al. 2017, Eberhard et al. 2025]. UD_Nheengatu-CompLin debuted in UD v2.11, becoming the largest among Amerindian treebanks from v2.13 to v2.15, and now ranking second [de Alencar 2024b, de Alencar 2024a]. While the ratings of all other Amerindian treebanks range between 0 and 2.0 stars in UD v2.16, UD_Nheengatu-CompLin holds 4.0—matched by only two of the six Portuguese treebanks. UD_Nheengatu-CompLin has steadily improved across UD releases, rising from 2.0 stars in v2.14 to 3.15 in v2.15, while many other treebanks have either stagnated or declined. Star ratings are computed at every UD release, based on factors such as genre diversity, annotation quality, lemmatization, size, validation status, and the ratio of words per "bug", among others.

On 2025-06-24, UD_Nheengatu-CompLin contained 22,421 tokens and 2,227 sentences from a broad spectrum of sources spanning nearly two centuries. This paper presents recent enhancements to the treebank's annotation methodology, focusing on challenges posed by variation in 19$^{\text{th}}$-century Nheengatu. In particular, we describe a substantial extension of the *special tags* mechanism introduced by [de Alencar 2023], used by human annotators to tag tokens that represent strings unrecognized by Yauti, the morphosyntactic analyzer employed in building the treebank.

Section 2 outlines the annotation methodology. Section 3 details the extensions to the special tag notation for handling historical variants and typographical errors. Section 4 discusses the multi-level validation of special tags and presents evaluation results. Finally, Section 5 sums up the contributions and identifies directions for future work.

2. Treebank Annotation Methodology

UD_Nheengatu-CompLin is annotated using Yauti [de Alencar 2023]. For an input sentence, it produces a UD analysis in the CoNLL-U format [de Marneffe et al. 2024a]. This consists of a ten-column table, of which the first eight are shown in Table 1. The ninth column, reserved for enhanced dependencies [Schuster and Manning 2016], is typically empty in UD treebanks. The tenth column contains miscellaneous annotations (MISC).

(1) *Pituna ukiri uikú ií ripí-pe.*
night 3:sleep 3:be water botton=in
'Night was sleeping at the bottom of the water.' [de Magalhães 1876, p. 164]

ID	FORM	LEMMA	UPOS	XPOS	FEATS	HEAD	DEPREL
1	Pituna	pituna	NOUN	N	Number=Sing	2	nsubj
2	ukiri	kiri	VERB	V	Mood=Ind|Person=3|...	0	root
3	uikú	ikú	AUX	AUXFS	Mood=Ind|Person=3|...	2	aux
4	ií	ií	NOUN	N	Number=Sing	5	nmod:poss
5-6	ripí-pe	_	_	_	_	_	_
5	ripí	tipí	NOUN	N	Number=Sing|Rel=Cont	2	obl
6	pe	upé	ADP	ADP	AdpType=Post|Clitic=Yes	5	case
7	.	.	PUNCT	PUNCT	-	2	punct

Table 1. First eight columns of the CoNLL-U analysis for (1) produced by Yauti, with simplified feature values.

Yauti is a knowledge-based tool built on the Python CoNLL-U Parser.[1] It uses a full-form lexicon to assign part of speech and morphological features to tokens. Listing 1 shows sample entries for *apigawa* 'man' and *kiri* 'to sleep'. Fused words such as *ripí-pe* in (1) are not included in the lexicon and are instead handled by the tokenizer. Using the information available for each token, syntactic rules determine its `HEAD` and `DEPREL`. While Yauti analyzes some syntactically non-trivial sentences correctly, such as (1), it incurs mistakes at the different annotation levels. If all tokens in a sentence are unambiguous and known to Yauti, most errors relate to `HEAD` and `DEPREL`.

Listing 1. Sample entries from Yauti's full-form lexicon.

```
{"akiri": [["kiri", "V+IND+1+SG"]],"ukiri": [["kiri", "V+IND+3"]],
"apigawa": [["apigawa","A"],["apigawa","N+SG"]],}
```

Unknown words tend to produce degraded results with validation errors, e.g., non-trees and/or empty `UPOS`, `HEAD`, and `DEPREL`. Ambiguous words pose another problem. Yauti only differentiates pronouns from determiners and main verbs from auxiliaries. Figure 1 depicts Yauti's unsuccessful analysis for (2). The tool creates a token with the same `ID` for each of the senses of *apigawa* 'man' in Listing 1, breaking CoNLL-U conformance. Since the lexicon does not include an entry for *yaitiwa* 'thicket', the corresponding token has empty UPOS and DEPREL, causing validation failure.

[1] `https://pypi.org/project/conllu/`

(2) *Apigawa usú usikari yaitiwa wírupi rupí.*
man 3:go 3:search thicket beneath through
'The man went looking underneath the thicket.' [de Magalhães 1876, p. 213]

```
>>> s='Apigawa usú usikari yaitiwa wírupi rupí. (p. 213) O homem foi procurar por baixo do cerrado. - Apgáua    oçô
ocikári iaitíua uírpe   rupí.'''
>>> Yauti.parseExample(s,'Magalhaes1876',10,37,272,annotator='Leonel Figueiredo de Alencar')
# sent_id = Magalhaes1876:10:37:272
# text = Apigawa usú usikari yaitiwa wírupi rupí.
# text_eng = TODO
# text_por = O homem foi procurar por baixo do cerrado.
# text_source = p. 213
# text_orig = Apgáua    oçô     ocikári iaitíua uírpe   rupí.
# text_annotator = Leonel Figueiredo de Alencar
# inputline = Apigawa usú usikari yaitiwa wírupi rupí.
1       Apigawa apigawa ADJ     A       _       3       _       _       TokenRange=0:7
1       Apigawa apigawa NOUN    N       Number=Sing     3       nsubj   _       TokenRange=0:7
2       usú     sú      AUX     AUXFR   Mood=Ind|Person=3|VerbForm=Fin  3       aux     _       TokenRange=8:11
3       usikari sikari  VERB    V       Mood=Ind|Person=3|VerbForm=Fin  0       root    _       TokenRange=12:19
4       yaitiwa yaitiwa NOUN    N       Number=Sing     3       obj     _       TokenRange=20:27
5-6     wírupi  _       _       _       _       _       _       _       TokenRange=28:34
5       wira    wira    NOUN    N       Number=Sing     3       obl     _       _
6       upé     upé     ADP     ADP     AdpType=Post|Clitic=Yes 5       case    _       _
7       rupí    rupí    ADP     ADP     AdpType=Post    5       case    _       SpaceAfter=No|TokenRange=35:39
8       .       .       PUNCT   PUNCT   _       3       punct   _       SpaceAfter=No|TokenRange=39:40
```

Figure 1. Analysis of (2) with Yauti.

Semi-automatic rule-based tools have proven useful in various annotation scenarios, e.g., [Faria et al. 2024, Freitas and de Souza 2024, Santos and Mota 2010]. Due to the limitations just described, Yauti is employed interactively. The human annotator prepares the input string for `parseExample` or a similar function, supplying a sentence in modernized spelling, page number, Portuguese translation, and the original text (Figure 1). The annotator executes the function and checks the output for ambiguous and unknown words. To handle ambiguity, Yauti accepts pos-tagged input strings. After appending the `n` lower cased `XPOS` tag for nouns, separated by a slash, to ambiguous *apigawa* in (2), Yauti builds a well-formed tree (Figure 2). This analysis, however, has a fatal UD validation error: the `DEPREL` of *yaitiwa* 'thicket' is empty. Once the lexicon is updated with an entry for this word, Yauti assigns it the direct object relation, building a tree that is indeed valid, but incorrect because the word functions as a genitive complement to the following noun. The human annotator is supposed to correct such errors. If the error goes unnoticed, it should be corrected by the reviewer.

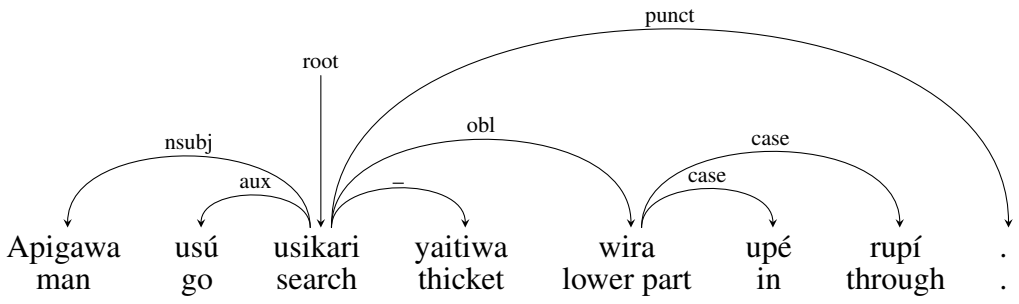

Figure 2. Yauti's dependency tree for (2) after disambiguation of *apigawa*.

Updating the lexicon, however, does not solve all cases of unknown words. Consider *u* in (3), lemmatized only as a conjunction in [Avila 2021]. Creating a lexicon entry for such *ad hoc*, non-conventionalized forms would be of little use. Instead, they are processed with *special tags*, which instruct Yauti to perform specific actions on a token. In Figure 3, the human annotator edits the input string, tagging the unknown form as

`u/=intj`. This triggers a function that creates a `Token` with the properties of lexicalized interjections. Note that the tagged input string is stored in the custom `inputline` metadata attribute, which has proven useful for various documentation purposes—such as tracing annotator decisions, revising annotations, and training new annotators.

(3) *Yautí usuaxara: — U!*
tortoise 3:answer u
'The tortoise replied: "U!"'

```
>>> s='''Yautí usuaxara: — U/=intj! (p. 210) O jabuti respondeu: — U! - Iáutí   oçuaxára:      — U!'''
>>> Yauti.parseExample(s,'Magalhaes1876',10,5,240, annotator='Leonel Figueiredo de Alencar')
# sent_id = Magalhaes1876:10:5:240
# text = Yautí usuaxara: — U!
# text_eng = TODO
# text_por = O jabuti respondeu: — U!
# text_source = p. 210
# text_orig = Iáutí     oçuaxára:       — U!
# text_annotator = Leonel Figueiredo de Alencar
# inputline = Yautí usuaxara: — U/=intj!
1       Yautí   yautí   NOUN    N       Number=Sing     2       nsubj   _       TokenRange=0:5
2       usuaxara        suaxara VERB    V       Mood=Ind|Person=3|VerbForm=Fin  0       root    _       SpaceAfter=
No|TokenRange=6:14
3       :       :       PUNCT   PUNCT   _       2       punct   _       TokenRange=14:15
4       —       —       PUNCT   PUNCT   _       2       punct   _       TokenRange=15:16
5       U       u       INTJ    INTJ    _       2       discourse       _       SpaceAfter=No|TokenRange=17:18
6       !       !       PUNCT   PUNCT   _       2       punct   _       SpaceAfter=No|TokenRange=18:19
```

Figure 3. Using the special tag `=intj` to analyze the interjection *u* in (3).

The special tag syntax supports colon-separated keyword parameters of the form `arg|val`, where `arg` is a key and `val` is its value, as illustrated in (4) and (5). The form *pô* is a Portuguese interjection, and *segurari* is a coinage based on Portuguese *segurar* 'to hold'; neither is lemmatized in [Avila 2021]. The arguments `o` and `s` specify the original language and the source form (if different from the value of `FORM`) of a non-native word embedded in a Nheengatu sentence. From this information, Yauti builds the `MISC` features `Orig=enganar` and `OrigLang=pt` for (5) [de Marneffe et al. 2024b]. Note that Nheengatu phonology and morphology are interwoven with the Portuguese verb *segurar* 'hold' in (5). Such code-mixing is especially common in Christian religious texts [Aguiar 1898, do Brasil 2019] and among contemporary bilingual speakers [Taylor 1985, Moore et al. 1994, da Cruz 2011]. Similar tags handle other non-Nheengatu elements.

(4) *"Pô/=intj:o|pt!", unheẽ yepé/art apigá/n maxí.* '"Gee!" said a leper man.' [da Cruz 2011, p. 261]

(5) *Tenki resegurari/=v:o|pt:s|segurar* 'You have to hold' [da Cruz 2011, p. 514]

Another application for special tags is word formation. One typical example is reduplication, a morphological process that is very common in Amazonia and elsewhere [Dixon and Aikhenvald 1999, Gómez and van der Voort 2014, Inkelas and Downing 2015, Beck 2017]. In principle, all Nheengatu verbs and adjectives can undergo reduplication, whereby a stem or part thereof is repeated, conveying iteration, high intensity, etc. [da Cruz 2011, Navarro 2016]. Yauti analyzes fully reduplicated active verbs without human intervention if, following [Avila 2021], a hyphen separates the reduplicant from the base, as in *tuká-tuká* 'to beat repeatedly'. However, [Avila 2021] does not apply this convention consistently: the hyphen is missing in some lemmas, as in *weráwerá* 'to shine intensely or continuously' and *saãsaã* 'to try repeatedly'.

Partial reduplication in Nheengatu is quite challenging to parse automatically. It can be both prefixal and suffixal (6), as well as infixal (7). It involves prosodic structure [da Cruz 2011, da Cruz 2014], which is not easily recoverable from the spelling.

```
>>> Yauti.pp('Aiwana aintá/pron uwari i/pron2 ara-pe, umusasaka/=red:l|2:p|2 pawa/total sukwera/abs.')
1       Aiwana  aiwana  ADV     ADVT    AdvType=Tim     3       advmod  _       TokenRange=0:6
2       aintá   aintá   PRON    PRON    Case=Acc,Nom|Number=Plur|Person=3|PronType=Prs  3       nsubj   _       Tok
enRange=7:12
3       uwari   wari    VERB    V       Mood=Ind|Person=3|VerbForm=Fin  0       root    _       TokenRange=13:18
4       i       i       PRON    PRON2   Case=Gen|Number=Sing|Person=3|Poss=Yes|PronType=Prs     5       nmod:poss
_       TokenRange=19:20
5-6     ara-pe  _       _       _       _       _       SpaceAfter=No|TokenRange=21:27
5       ara     ara     NOUN    N       Number=Sing     3       obl     _       _
6       pe      upé     ADP     ADP     AdpType=Post|Clitic=Yes 5       case    _       _
7       ,       ,       PUNCT   PUNCT   _       8       punct   _       TokenRange=27:28
8       umusasaka       musaka  VERB    V       Mood=Ind|Person=3|Red=Yes|VerbForm=Fin  3       parataxis       _
TokenRange=29:38
9       pawa    pawa    PART    TOTAL   Aspect=Compl    8       advmod  _       TokenRange=39:43
10      sukwera sukwera NOUN    N       Number=Sing|Rel=Abs     8       obj     _       SpaceAfter=No|TokenRange=44
:51
11      .       .       PUNCT   PUNCT   _       3       punct   _       SpaceAfter=No|TokenRange=51:52
```

Figure 4. Yauti's analysis of infixal reduplication in (7).

(6) *pikũi-kũi* dig-RED or *piku-pikũi* RED-dig 'to dig a lot' [Avila 2021, p. 598]

(7) *Aiwana aintá u-wari i ara-pe, u-mu-sa-saka pawa s-ukwera*
then they 3-fall him top-on 3-CAUS-RED-detach TOT NCONT-flesh
'Then they fell upon him and tore his flesh apart.' [Rodrigues 1890, p. 37]

To handle partial reduplication, we extended the inventory of special tags with `=red`, which requires the specification of the length of the reduplicant as a value of the parameter `l`. It is incumbent on the human annotator to attach this special tag with the appropriate parameters to each partially reduplicated word. In Figure 4, `p|2` and `l|2` indicate that the reduplicant starts at position 2 in the stem and has length 2. These values are passed to the `handlePartialRedup` function, which lemmatizes the verb and constructs its morphological features, including the feature-value pair `Red=Yes`, which signals reduplication in some UD treebanks for South American Indigenous languages.

The `handlePartialRedup` function performs lemmatization from left to right by default, stripping `length` characters from the stem, starting at position 0 by default or at the position specified by the human annotator in case of infixal reduplication. Suffixal reduplication is handled with the key-value pair `u|t`. This is converted to `suffix=True` and passed to `handlePartialRedup`, which performs lemmatization from right to left using the value of `length`. Non-hyphenated full reduplication constitues a special instance of partial reduplication, e.g., `asaãsaã/=red:l|3` 'I try repeatedly'. The special tag `=red` can also pass other arguments to `handlePartialRedup`. For example, to parse *purapuranga* 'very beautiful', the annotator attaches the tag `=red:l|4:x|a`, instructing Yauti with `x|a` to assign *puranga* 'beautiful' the adjective `XPOS` tag.[2] Since reduplication is more common with verbs in Nheengatu, the `xpos` argument defaults to `'V'`.

Figure 5 exemplifies `=red` with additional parameters. In this sentence, *ganari*, coined after Portuguese *enganar* 'to deceive' and not listed in [Avila 2021], is reduplicated. The human annotator is therefore instructed to treat it as non-native, specifying the original language and source form with `o|pt` and `s|enganar`.

3. Annotation of Non-standard Language

This section describes the formal mechanisms implemented to annotate different kinds of non-standard language, thereby preserving the characteristics of the source texts—which can be relevant for training robust parsers—while enabling structured manipulation of

[2]This word also functions as a manner adverb in Nheengatu [Avila 2021, p. 640].

```
>>> s='''Ti maã/cond aganaganari/=red:l|4:o|pt:s|enganar. (Example No. 1029 Wr, p. 503) Não enganaria. - ti=maã a-g
ana~ganai'''
>>> Yauti.parseExample(s,'Cruz2011',0,0,119,annotator='Leonel Figueiredo de Alencar')
# sent_id = Cruz2011:0:0:119
# text = Ti maã aganaganari.
# text_eng = TODO
# text_por = Não enganaria.
# text_source = Example No. 1029 Wr, p. 503
# text_orig = ti=maã a-gana~ganai
# text_annotator = Leonel Figueiredo de Alencar
# inputline = Ti maã/cond aganaganari/=red:l|4:o|pt:s|enganar.
1       Ti      ti      PART    NEG     PartType=Neg|Polarity=Neg       3       advmod  _       TokenRange=0:2
2       maã     maã     PART    COND    Modality=Cond|PartType=Mod      3       advmod  _       TokenRange=3:6
3       aganaganari     ganari  VERB    V       Mood=Ind|Number=Sing|Person=1|Red=Yes|VerbForm=Fin      0       roo
t       _       Orig=enganar|OrigLang=pt|SpaceAfter=No|TokenRange=7:18
4       .       .       PUNCT   PUNCT   _       3       punct   _       SpaceAfter=No|TokenRange=18:19
```

Figure 5. Analysis of prefixal reduplication of *ganari* 'deceive'.

the treebank, such as retrieving occurrences of specific forms or lemmas. We first explain how Yauti deals with unknown words resulting from typographical errors (*typos*) in the source text. Typos are marked with `Typo=Yes` and fall into three main categories: (i) misspelled words, (ii) wrongly merged words, and (iii) wrongly split words [de Marneffe et al. 2024c]. We then describe the technique developed to process the morphosyntactic particularities of historical texts. They also represent deviations from the standard language but are annotated with `Style=Arch` [de Marneffe et al. 2024c].

In (8), *kunhambukú* is a type (i) spelling error handled by the special tag `=typo`. This tag requires the correct form and optionally accepts `x` for `XPOS` and `n` for another special tag. Example (9) shows the latter: *mumunéu* should be corrected to *muamunéu* 'to dress', but since the verb is in the middle voice, handled via `=mid`, the argument `n|mid` instructs Yauti to remove the prefix *yu* from *xayumuamunéu*, allowing correct lemmatization and feature assignment (Figure 6).

(8) *i remixira usuaxara kunhambukú/=typo:c|kunhamukú:x|n marika upé*
his roast 3:answer woman belly in
'the roast replied from the girl's belly' [Rodrigues 1890, p. 65]

(9) *Xasú xayumumunéu/=typo:c|xayumuamunéu:n|mid.*
1s:go 1s:MID:dress
'I will get dressed.' [Hartt 1938, p. 334]

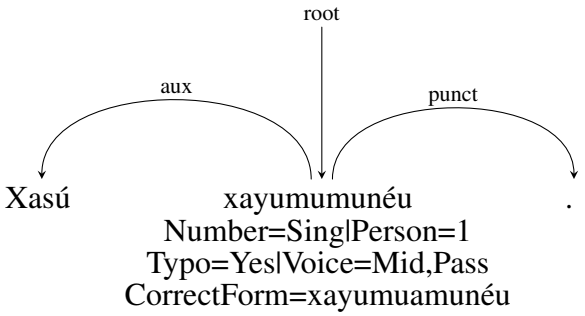

Figure 6. Dependency tree of (9) with sample features.

In (10), *Intíawá* results from merging *intí* 'not' and *awá* 'someone', exemplifying type (ii) error. The special tag `=spl` triggers splitting the token into two syntactic words w_1 and w_2, whereby w_2 is assigned to the `w` parameter. The `x|ind` argument

disambiguates the split-off word as an indefinite pronoun. Conformant to the UD guidelines, Yauti inserts `Typo=Yes` in the `FEATS` field and `CorrectSpaceAfter=Yes` and `SpaceAfter=No` in the `MISC` field of *intí*. This enriched annotation allows for both the correction of the spelling error and the accurate syntactic representation of the sentence while preserving the characteristics of the source text. The special tag `=spl` takes other optional arguments. In (11), the second-class pronoun *i* [Navarro 2016, Avila 2021] is fused with the postposition *irũ* 'with'. Splitting the two words results in *i* and *rũ*. Since the latter form constitutes a nonword, the human annotator supplies `c|irũ`, disambiguating *irũ* with `x|adp`. The argument-value pair `h|pron2` disambiguates *i*.

(10) *Intíawá/=spl:w|awá:x|ind ukwáu aintá nheenga [...].*
no:one 3:know their language
'No one knew their language [...].' [de Amorim 1928, p. 196]

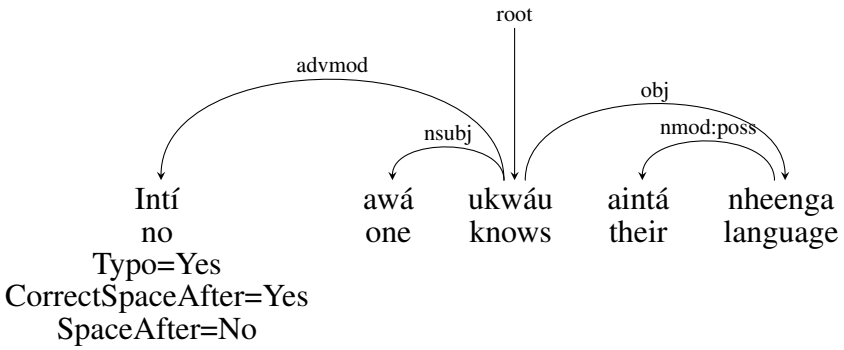

Figure 7. Dependency tree of (10) with sample features.

(11) *Intí xawatá-kwáu irũ/=spl:w|rũ:h|pron2:c|irũ:x|adp*
not 1s:walk-can with
'I can't go for a walk with him' [Hartt 1938, p. 325]

Type (iii) spelling error comprises examples like *repi sika* in (12), where a blank occurs inside the second person singular form of *pikisa* 'to catch'. Yauti handles this error by combining the special tags `=typo` and `=x`, which the human annotator appends on *repi* and *sika*, respectively. In the former case, it performs exactly like in type (i) errors, while in the latter, it assigns the word the dummy `UPOS` tag `X` and links it to the preceding node via `goeswith`. Figure 8 depicts the corresponding tree after the corrections by the human annotator, which only affected the auxiliary and the pronoun.

(12) *Kũi ana repi/=typo:c|repisika sika/=x aintá [...].*
[2S:IMP]go PFV 2s:catch sika them
'Go get them now [...].' [de Amorim 1928, p. 444]

The formalism also supports annotating wrongly split words whose second part displays a typo. In (13), *etá* is detached from the noun and misspelled as *aetá*. Assigning this form the `UPOS` tag `X` via `=typo` instructs Yauti to treat it as a split-off segment linked to the previous word via `goeswith`, placing its correct form in the `MISC` field.

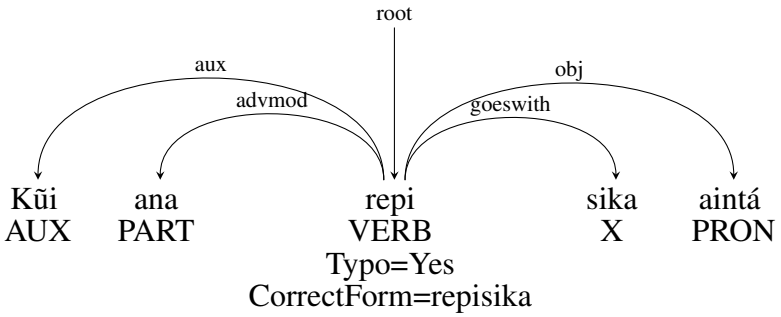

Figure 8. Dependency tree of (12).

(13) *aintá remiré/=typo:c|rimirikú-etá aetá/=typo:c|-etá:x|x suí*
their wife-PL from
'from their wives' [de Amorim 1928, p. 23]

Particularities that Nheengatu lost as it converged more closely to Portuguese can make 19th-century texts challenging for the modern speaker [Rodrigues 1986, Moore et al. 1994, Borges 1996, Rodrigues and Cabral 2011]. Simply modernizing ancient texts would facilitate understanding. However, this would compromise the treebank's usefulness for historical linguistic analysis. The `Style=Arch` feature and the associated formal apparatus enable meeting both requirements to a great extent. In (14), *puitá* 'to become' lacks the agreement prefix *u*. In his modernization, [Avila 2021, p. 382] not only supplies the missing agreement marker but also substitutes modern lemma *pitá* for historical *puitá*. The special tag `=mf` instructs Yauti to parse both the original *puitá* and its modernization. The modern form *upitá* is passed as the value of the `m` argument. Yauti constructs a `Token` object for *puitá* with the corresponding lemma and morphological features, enriched with `Style=Arch`. The diverging properties of *upitá*, such as lemma *pitá* and `Person=3`, are displayed in the `MISC` field, prefixed with `Modern` [de Marneffe et al. 2024b]. Optional arguments enable analyzing much more complex examples, as shown in (15), where the non-lexicalized verb *salvari*, coined after Portuguese *salvar* 'to save', lacks subject agreement and appears in the middle voice.

(14) *i pewa katú puitá/=mf:m|upitá*
it flat very become
'it became totally flat' [Rodrigues 1890, p. 123]

(15) *Yepé mira yusalvari-kwáu/=mf:m|uyusalvari:n|mid:o|pt:s|salvar [...]?*
a person MID:save-can
'Can a person be saved [...]?' [Aguiar 1898, p. 35]

4. Validation and Evaluation of the Special Tags

A significant deficiency of Yauti before the improvements described in this paper was that it did not validate the special tags. Invalid tags—whether structurally ill-formed, containing nonexistent function or argument names, or carrying values of incompatible types—were either silently ignored or caused Python runtime errors that were difficult to debug, particularly for less experienced annotators and reviewers. To improve annotation

consistency and facilitate revision, we implemented a multi-level tag validation mechanism, loosely inspired by the Universal Dependencies (UD) validation system.

The first level is syntactic validation. It is based on an EBNF grammar defining the language of special tags. Using the Python Lark library,[3] this grammar is compiled into a parser that generates a syntax tree for any well-formed tag according to the grammar; otherwise, it raises an `UnexpectedToken` exception. Each generated tree is transformed into a Python dictionary with keys for the different components of the tag, such as the function name, a dictionary containing the argument-value pairs, and so on. The next step is semantic validation. Functions and arguments are checked against function signatures specifying the sets of required and optional arguments (Listing 2). Afterward, string values intended to represent booleans or integers are converted into their corresponding Python types to ensure type correctness during execution. Specific exceptions are raised upon failure of any of these procedures.

Listing 2. Sample function signatures for semantic validation of special tags.

```
SIGNATURES = {'intj': {'required': set(), 'optional': {'o', 's'},},
'mf': { 'required': {'m'},'optional': {'h', 'n', 'o', 'r', 's', 'x'},},
'mid': {'required': set(),'optional': {'o', 's'},},
'red': {'required': {'l'},'optional': {'a', 'o', 'p', 's', 'u', 'x'},},
'spl': {'required': {'w'},'optional': {'b', 'c', 'h', 'x'},},
'typo': {'required': {'c'},'optional': {'n', 'x'},},}
```

At runtime, each special tag invokes a Yauti function that manipulates the tagged token. For instance, the tag `=red` triggers the function `handlePartialRedup`. The arguments specified in the tag are passed to the function. Before invoking the function, however, additional constraints are checked for specific parameters. The value of `o` must be a valid 2-letter or 3-letter language code according to the Python `iso639` library, while `h` and `x` must belong to the inventory of `XPOS` tags. Finally, inside the function, the correct form and modern form specified as values of the `c` and `m` keys, if present, are checked against Yauti's lexicon entries.

We implemented tests for different levels of validation. The first test set targets syntactic validation and includes 21 well-formed tags and 34 ill-formed ones. As expected, the former were accepted by the syntactic validation function, while the latter were rejected. To evaluate semantic validation, we constructed two sets: 35 valid tags (hereafter PSEM) and 40 invalid tags (NSEM). Running `pytest` on both sets yielded the expected outcomes: all valid tags passed, and all invalid ones were rejected. Next, we created a combined test set by unioning PSEM with BTAGS, the set of 181 distinct combinations of function and arguments extracted from 344 total instances in UD_Nheengatu-CompLin. BTAGS includes examples of 19 out of the 20 tags defined in the function signatures, the most frequent being `=mf` and `=typo` (see Figure 9). We then augmented NSEM by injecting errors into BTAGS using various invalidation strategies—e.g., removing required arguments, inserting unsupported arguments, or replacing the function name with an undefined one. After this step, our final test set included 416 cases. Running `pytest` confirmed that all tests were handled correctly by the validation logic.

The 344 instances of special tags occur in 231 sentences. To evaluate their

[3] `https://github.com/lark-parser/lark`

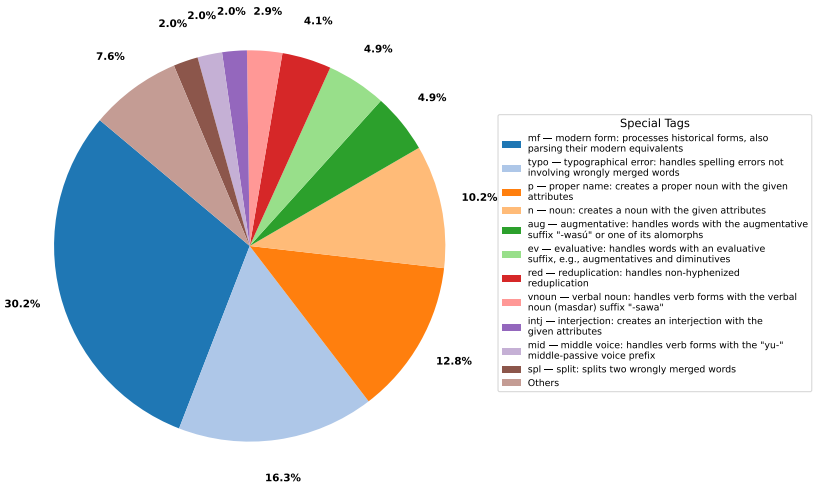

Figure 9. Relative frequency of special tags in UD_Nheengatu-CompLin on 2025-06-24 ($N = 344$)

functionality in realistic scenarios, we ran Yauti's `parseSentence` on the respective `inputlines`. Two cases failed due to validation errors. One involved a missing argument (`a|t`) for `=ev`, the function for evaluative suffixes [Rio-Torto 2015], which prevented proper lemmatization of *wirawasumirĩ* 'little hawk' from *wirawasú* 'hawk'. The other exposed a bug in `handlePartialRedup`. After correction, all sentences parsed. Using Scikit-learn [Pedregosa et al. 2011], we computed macro F_1, weighted F_1, and micro F_1 scores for the features of the tokens with special tags in this sample, obtaining 0.92, 0.96, and 0.96. A cursory inspection of the individual scores and their confusion matrices revealed errors in the treebank or in the functions triggered by the special tags. These results will serve as the basis for a series of issues on the corresponding repositories.

5. Final Remarks

In this paper, we presented significant enhancements to the mechanism of special tags in Yauti, the tool used to annotate UD_Nheengatu-CompLin. First introduced in [de Alencar 2023], special tags enable the parsing of words not represented in Yauti's lexicon or handled by the tokenizer. We extended the coverage of derivational morphology, introducing tags for reduplication, middle-passive voice, evaluative suffixes, and more. We implemented UD-conformant annotation of typographical errors and stylistic variations. The treatment of complex interactions involving typos, derivation, archaic forms, and code-mixing was exemplified. A second contribution was the implementation of a multi-level validation system for special tags, whose correctness was demonstrated through hundreds of test cases.

There is ample room for future improvements. Yauti currently consists of a series of scripts; refactoring it as a modular library would facilitate tighter integration with the validator. The handling of sentences involving multiple interacting phenomena remains somewhat *ad hoc*. A more principled approach could leverage function embedding via nested tagging. Another avenue of research is to automate the treatment of some phenomena currently handled by special tags, thereby reducing human annotation effort—an endeavor that requires more training data or a deeper understanding of the linguistic mechanisms involved.

References

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

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

Beck, D. (2017). The typology of morphological processes: Form and function. In Aikhenvald, A. Y. and Dixon, R. M. W., editors, *The Cambridge Handbook of Linguistic Typology*, pages 325–360. Cambridge University Press, Cambridge.

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 Cruz, A. (2014). Reduplication in nheengatu. In Gómez, G. G. and van der Voort, H., editors, *Reduplication in Indigenous Languages of South America*, pages 115–141. Brill, Leiden, The Netherlands.

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 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 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., and Zeman, D. (2024a). CoNLL-U format. Accessed: 2024-01-09.

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. (2024b). MISC attributes in CoNLL-U. Accessed: 2025-06-07.

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. (2024c). Typos and other errors in underlying text. Accessed: 2025-06-07.

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

Dixon, R. M. W. and Aikhenvald, A. Y. (1999). *The Amazonian Languages*. Cambridge Language Surveys. Cambridge University Press, Cambridge.

do Brasil, M. N. T., editor (2019). *Novo Testamento na língua Nyengatu*. Sociedade Bíblica do Brasil, Barueri, SP, 2nd edition. Original work published in 1973.

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

Faria, P., Galves, C., and Magro, C. (2024). Syntactic annotation for portuguese corpora: standards, parsers, and search interfaces. *Language Resources and Evaluation*, 58(1):301–346.

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

Freitas, C. and de Souza, E. (2024). A study on methods for revising dependency treebanks: in search of gold. *Language Resources and Evaluation*, 58(1):111–131.

Futrell, R., Mahowald, K., and Gibson, E. (2015). Quantifying word order freedom in dependency corpora. In Nivre, J. and Hajičová, E., editors, *Proceedings of the Third International Conference on Dependency Linguistics (Depling 2015)*, pages 91–100, Uppsala, Sweden. Uppsala University, Uppsala, Sweden.

Gómez, G. G. and van der Voort, H., editors (2014). *Reduplication in Indigenous Languages of South America*. Brill, Leiden, The Netherlands.

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].

Inkelas, S. and Downing, L. J. (2015). What is reduplication? typology and analysis part 1/2: The typology of reduplication. *Language and Linguistics Compass*, 9(12):502–515.

Kondratyuk, D. and Straka, M. (2019). 75 languages, 1 model: Parsing Universal Dependencies universally. In Inui, K., Jiang, J., Ng, V., and Wan, X., editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 2779–2795, Hong Kong, China. Association for Computational Linguistics.

Li, W. (2025). Evaluating the effectiveness of linguistic knowledge in pretrained language models: A case study of universal dependencies.

Moore, D., Facundes, S., and Pires, N. (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.

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.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Rio-Torto, G. (2015). Formação de avaliativos. In Rio-Torto, G., Rodrigues, A. S., Pereira, I., Pereira, R., and Ribeiro, S., editors, *Gramática Derivacional do Português*, pages 357–389. Imprensa da Universidade de Coimbra, Coimbra, 2 edition.

Rodrigues, A. D. (1986). *Línguas brasileiras: Para o conhecimento das línguas indígenas*. Loyola, São Paulo.

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.

Santos, D. and Mota, C. (2010). Experiments in human-computer cooperation for the semantic annotation of Portuguese corpora. In Calzolari, N., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., and Tapias, D., editors, *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*, Valletta, Malta. European Language Resources Association (ELRA).

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 2016)*.

Taylor, G. (1985). Apontamentos sobre o nheengatu falado no rio negro, brasil. *Amérindia: revue d'ethnolinguistique amérindienne*, 10:5–23.

Zeman, D. (2023). *Cross-Language Harmonization of Linguistic Resources*. Institute of Formal and Applied Linguistics (ÚFAL), Prague. Habilitation thesis.

Zeman, D., Nivre, J., Abrams, M., Ackermann, E., Adolphe, J., Aepli, N., Aghaei, H., Agić, Ž., Ahmadi, A., Ahrenberg, L., Ajede, C. K., Akhundjanova, A., Akkurt, F., Aleksandravičiūtė, G., Alfina, I., Algom, A., Alnajjar, K., Alzetta, C., Anastasopoulos, A., Andersen, E., Andrews, M., Antonsen, L., Aoyama, T., Aplonova, K., Aquino, A., Aragon, C., Aranes, G., Aranzabe, M. J., Arıcan, B. N., Arnardóttir, H̄., Arutie, G., Arwidarasti, J. N., Asahara, M., Ásgeirsdóttir, K., Aslan, D. B., Asmazoğlu, C., Ateyah, L., Atmaca, F., Attia, M., Atutxa, A., Augustinus, L., Avelãs, M., Badmaeva, E., Bajorat, J., Balasubramani, K., Ballesteros, M., Banerjee, E., Bank, S., Barbosa, B. K. d. S., Barbu Mititelu, V., Barkarson, S., Basile, R., Basmov, V., Batchelor, C., Bauer, J., Bedir, S. T., Behzad, S., Belieni, J., Bémová, A., Bengoetxea, K., Benli, I., Ben Moshe, Y., Benzerrak, M., Berg, A., Berk, G., Bhat, R. A., Biagetti, E., Bick, E., Bielinskienė, A., Bilgin Taşdemir, E. F., Binici, H., Bjarnadóttir, K., Blaschke, V., Blokland, R., Böbel, N., Bobicev, V., Boizou, L., Bompolas, S., Bonilla, J., Borges Völker, E., Börstell,

C., Bosco, C., Bouma, G., Bowman, S., Boyd, A., Braggaar, A., Branco, A., Bras, M., Brokaitė, K., Bu, L., Buráňová, E., Burchardt, A., Cabeza, C., Cáceres Arandia, N., Campos, M., Candito, M., Caron, B., Caron, G., Carvalheiro, C., Carvalho, R., Cassidy, L., Castro, M. C., Castro, S., Cavalcanti, T., Cebiroğlu Eryiğit, G., Cecchini, F. M., Celano, G. G. A., Çepani, A., Čéplö, S., Cesur, N., Cetin, S., Çetinoğlu, Ö., Chalub, F., Chamila, L., Chamoreau, C., Chauhan, S., Chen, Y., Chi, E., Chika, T., Cho, Y., Choi, J., Chontaeva, B., Chun, J., Chung, J., Cignarella, A. T., Cinková, S., Collomb, A., Çöltekin, Ç., Connor, M., Corbetta, C., Corbetta, D., Costa, F., Courtin, M., Crabbé, B., Cristescu, M., Cvetkoski, V., Dahan, N., Dale, I. L., Daniel, P., Daoudi, K., Dash, B., Dash, S. R., Davidson, E., de Alencar, L. F., Dehouck, M., de Laurentiis, M., de Marneffe, M.-C., Demir, A., de Paiva, V., Derin, M. O., de Souza, E., Diaz de Ilarraza, A., Díaz Hernández, R. A., Dickerson, C., Di Felippo, A., Dinakaramani, A., Di Nuovo, E., Dione, B., Dirix, P., Do, H., Dobrovoljc, K., Döhmer, C., Doyle, A., Dozat, T., Droganova, K., Duran, M. S., Dwivedi, P., Ebert, C., Eckhoff, H., Eguchi, M., Eiche, S., Eiselen, R., Eli, M., Elkahky, A., Ephrem, B., Erina, O., Erjavec, T., Esher, L., Eslami, S., Essaidi, F., Etienne, A., Evelyn, W., Facundes, S., Farkas, R., Faryad, J., Favero, F., Ferdaousi, J., Fernanda, M., Fernandez Alcalde, H., Fethi, A., Foster, J., Francioni, B., Fransen, T., Freitas, C., Fujita, K., Gajdošová, K., Galbraith, D., Galy, E., Gamba, F., Garcia, M., García-Miguel, J. M., Gärdenfors, M., Gaustad, T., Genç, E. E., Gerardi, F. F., Gerdes, K., Gessler, L., Ginter, F., Godoy, G., Goenaga, I., Gojenola, K., Gökırmak, M., Goldberg, Y., Goldin, G., Gómez Guinovart, X., González Saavedra, B., Griciūtė, B., Grioni, M., Grobol, L., Grūzītis, N., Guillaume, B., Guiller, K., Guillot-Barbance, C., Güngör, T., Gurevich, V., Habash, N., Hafsteinsson, H., Hahn, M., Hajič, J., Hajič jr., J., Hajičová, E., Hämäläinen, M., Hà Mỹ, L., Han, N.-R., Hanifmuti, M. Y., Harada, T., Hardwick, S., Harris, K., Hassert, N., Haug, D., Havelka, J., Heinecke, J., Hellwig, O., Hennig, F., Hladká, B., Hlaváčová, J., Hociung, F., Hoefels, D., Hohle, P., Howell, N., Huang, Y., Huerta Mendez, M., Hwang, J., Ikeda, T., Iliadou, I., Ingason, A. K., Ion, R., Irimia, E., Ishola, Ọ., Islamaj, A., Ito, K., Iurescia, F., Ivani, J. K., Jagodzińska, S., Jannat, S., Jelínek, T., Jha, A., Jiang, K., Job, S., Jobanputra, M., Johannsen, A., Jónsdóttir, H., Jørgensen, F., Ju, Z., Juutinen, M., Kaşıkara, H., Kabaeva, N., Kahane, S., Kanayama, H., Kanerva, J., Kara, N., Karahóǧa, R., Kárník, J., Kåsen, A., Kayadelen, T., Kengatharaiyer, S., Kettnerová, V., Kharatyan, L., Kirchner, J., Klementieva, E., Klyachko, E., Kocharov, P., Köhn, A., Köksal, A., Kolářová, V., Kopacewicz, K., Korkiakangas, T., Köse, M., Koshevoy, A., Kote, N., Kotsyba, N., Kovačić, B., Kovalevskaitė, J., Kowner, E., Krek, S., Krishnamurthy, P., Kübler, S., Kučová, L., Kuqi, A., Kuyrukçu, O., Kuzgun, A., Kwak, S., Kyle, K., Laan, K., Laippala, V., Lambertino, L., Landau, I., Lando, T., Larasati, S. D., Larrivée, P., Lavrentiev, A., Lee, J., Lê Hồng, P., Lenci, A., Lertpradit, S., Leung, H., Levina, M., Levine, L., Li, C. Y., Li, J., Li, K., Li, Y., Li, Y., Lim, K., Lima Padovani, B., Lin, Y.-J. J., Lindén, K., Liu, Y. J., Liu, Z., Ljubešić, N., Lobzhanidze, I., Loginova, O., Lopatková, M., Lopes, L., Luftiu, E., Lukashevskyi, A., Lusito, S., Lutgen, A.-M., Luthfi, A., Luukko, M., Lyashevskaya, O., Lynn, T., Macketanz, V., Mahamdi, M., Maillard, J., Makarchuk, I., Makazhanov, A., Mambrini, F., Mandl, M., Manning, C., Manurung, R., Marşan, B., Mărănduc, C., Mareček, D., Marheinecke, K., Markantonatou, S., Martínez Alonso, H., Martín Rodríguez, L., Martins, A., Martins, C., Mašek, J., Matsuda, H., Matsumoto, Y., Mazzei, A., McDonald, R., McGuinness,

S., Mehta, M., Ménard, P. A., Mendonça, G., Merhav, H., Merzhevich, T., Meurer, P., Miekka, N., Mikulová, M., Milano, E., Miletić, A., Miller, A., Min, J., Minerbi, Y., Mírovský, J., Mischenkova, K., Missilä, A., Mititelu, C., Mitrofan, M., Miyao, Y., Mohapatra, B., Mojiri Foroushani, A., Molnár, J., Moloodi, A., Montemagni, S., More, A., Moreno Romero, L., Moretti, G., Mori, S., Morioka, T., Moro, S., Mortensen, B., Moskalevskyi, B., Muischnek, K., Munro, R., Murawaki, Y., Mus, N., Müürisep, K., Nainwani, P., Nakhlé, M., Navarro Horñiacek, J. I., Nedoluzhko, A., Nešpore-Bērzkalne, G., Nevaci, M., Nguyễn Thị, L., Nguyễn Thị Minh, H., Nikaido, Y., Nikolaev, V., Nitisaroj, R., Norrman, V., Nourian, A., Novák, M., Nunes, M. d. G. V., Nurmi, H., Ojala, S., Ojha, A. K., Óladóttir, H., Olúòkun, A., Omura, M., Onwuegbuzia, E., Ordan, N., Osenova, P., Östling, R., Ott, A., Øvrelid, L., Oya, M., Özateş, Ş. B., Özçelik, M., Özgür, A., Öztürk Başaran, B., Paccosi, T., Pajas, P., Palmero Aprosio, A., Panevová, J., Panova, A., Pardo, T. A. S., Parida, S., Park, H. H., Partanen, N., Pascual, E., Passarotti, M., Patejuk, A., Paulino-Passos, G., Pedonese, G., Peeters, O., Peljak-Łapińska, A., Peng, S., Peng, S. L., Pereira, R., Pereira, S., Perez, C.-A., Perkova, N., Perrier, G., Petrov, S., Petrova, D., Peverelli, A., Phelan, J., Pierre-Louis, C., Piitulainen, J., Pinter, Y., Pinto, C., Pintucci, R., Pirinen, T. A., Pitler, E., Plamada, M., Plank, B., Plum, A., Poibeau, T., Ponomareva, L., Popel, M., Poujade, C., Pretkalniņa, L., Pretorius, R., Prévost, S., Prokopidis, P., Przepiórkowski, A., Pugh, R., Puolakainen, T., Purschke, C., Pyysalo, S., Qi, P., Querido, A., Rääbis, A., Rabinovich, E., Rademaker, A., Rahman, M.-u., Rahoman, M., Rama, T., Ramasamy, L., Ramisch, C., Ramos, J., Rashel, F., Rasooli, M. S., Ravishankar, V., Real, L., Rebeja, P., Reddy, S., Regnault, M., Rehm, G., Riabi, A., Riabov, I., Rießler, M., Rimkutė, E., Rinaldi, L., Rituma, L., Rizqiyah, P., Rocha, L., Rögnvaldsson, E., Roksandic, I., Roman, N. T., Romanenko, M., Romanova, N., Rosa, R., Roșca, V., Roulon, P., Rovati, D., Rozonoyer, B., Rudina, O., Rueter, J., Ruffolo, P., Rúnarsson, K., Rushiti, R., Sadde, S., Safari, P., Sahala, A., Sahoo, K., Sahoo, S., Saleh, S., Salomoni, A., Samardžić, T., Sampanis, K., Samson, S., Sánchez-Rodríguez, X., Sanguinetti, M., Sanıyar, E., Särg, D., Sartor, M., Sarymsakova, A., Sasaki, M., Saulīte, B., Savary, A., Sawanakunanon, Y., Saxena, S., Scannell, K., Scarlata, S., Schang, E., Schneider, N., Schuster, S., Schwartz, L., Seddah, D., Seeker, W., Sellmer, S., Seraji, M., Ševčíková, M., Sgall, P., Shahzadi, S., Shen, M., Shimada, A., Shin, G.-H., Shirasu, H., Shishkina, Y., Shohibussirri, M., Shvedova, M., Sibille, J., Siewert, J., Sigurðsson, E. F., Silva, J., Silveira, A., Silveira, N., Silveira, S., Simi, M., Simionescu, R., Simkó, K., Šimková, M., Símonarson, H. B., Simov, K., Sitchinava, D., Sither, T., Smith, A., Soares-Bastos, I., Solberg, P. E., Sollberger, D., Sonnenhauser, B., Sourov, S., Speransky, N., Sprugnoli, R., Stamou, V., Steingrímsson, S., Stella, A., Štěpánek, J., Štěpánková, B., Stephen, A., Straka, M., Strass, O., Strickland, E., Strnadová, J., Suhr, A., Sulestio, Y. L., Sulubacak, U., Sung, H., Suzuki, S., Swanson, D., Szántó, Z., Taguchi, C., Taji, D., Talamo, L., Tamburini, F., Tan, M. A. C., Tanaka, T., Tanaya, D., Tavoni, M., Teker, N., Tella, S., Tellier, I., Testori, M., Thomas, G., Tıraş, T. E., Tollersrud, T., Tonelli, S., Torga, L., Toribio, L., Toska, M., Trosterud, T., Trukhina, A., Tsarfaty, R., Tulchynska, K., Türk, U., Tyers, F., Hó̄rðarson, S., Hó̄rsteinsson, V., Uematsu, S., Untilov, R., Urešová, Z., Uria, L., Uszkoreit, H., Utka, A., Vagnoni, E., Vajjala, S., Vak, S., Vakirtzian, S., van der Goot, R., Vanhove, M., van Niekerk, D., van Noord, G., Varga, V., Vedenina, U., Venturi, G., Vergez-Couret, M., Vidová Hladká, B.,

Villemonte de la Clergerie, E., Vincze, V., Vissamsetty, A., Vlasova, N., Vligouridou, E., Wakasa, A., Wallenberg, J. C., Wallin, L., Walsh, A., Wang, J., Washington, J. N., Weissweiler, L., Wendt, M., Widmer, P., Wigderson, S., Wijono, S. H., Wille, V. B., Williams, S., Winkler, M., Wintner, S., Wirén, M., Wittern, C., Witzlack-Makarevich, A., Woldemariam, T., Wong, T.-s., Wróblewska, A., Wu, Q., Yako, M., Yamashita, K., Yamazaki, N., Yan, C., Yang, X., Yasuoka, K., Yavrumyan, M. M., Yenice, A. B., Yılandiloğlu, E., Yıldız, O. T., Yu, Z., Yuliawati, A., Žabokrtský, Z., Zahra, S., Zeldes, A., Zhou, H., Zhu, H., Zhu, Y., Zhuravleva, A., Ziane, R., and Znotiņš, A. (2025). Universal dependencies 2.16. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.