A Universal Dependencies Treebank for Nheengatu

Leonel Figueiredo de Alencar
Universidade Federal do Ceará
Av. da Universidade, 2683 - 60020-180 Fortaleza, Brazil
leonel.de.alencar@ufc.br

Abstract

We present UD_Nheengatu-CompLin, the inaugural treebank for Nheengatu, an endangered Indigenous language of Brazil with limited digital resources. This treebank stands as the largest among Indigenous American languages in version 2.13 of the Universal Dependencies collection. The developmental version comprises 1,336 trees, encompassing 13,246 tokens and 13,374 words. In a 10-fold cross-validation experiment using UDPipe 1.2, parsing with gold tokenization and gold tags achieved a labeled attachment score (LAS) of 81.17 ± 1.02, outperforming Yauti, the rule-based analyzer employed for sentence annotation.

1 Introduction

Universal Dependencies (henceforth UD) provides a framework for consistent morphosyntactic annotation across languages of different families, aiming at both linguistic typology and natural language processing (Nivre et al., 2016; de Marneffe et al., 2021). The UD collection has grown from 10 treebanks of 10 European languages in version 1.0 of January 2015 to 259 treebanks of 148 languages from all continents in version 2.13 of November 2023 (Zeman et al., 2023). However, the enormous diversity of Indigenous languages in the Americas is still underrepresented despite the efforts in the last five years.[1]

This paper introduces UD_Nheengatu-CompLin, the first UD treebank for Nheengatu (ISO 639-3: `yrl`), an endangered Indigenous language of Brazil, also known as Modern Tupi and *Língua Geral Amazônica* (hereafter LGA). Although it made its debut in UD v2.11 on November 15, 2022, with only 196 trees, it has since expanded significantly. With 1,239 trees totaling 12,621 tokens, it stands as the largest treebank for an Indigenous American language in UD v2.13. To our knowledge, no analogous resource for Nheengatu exists. It is made available under a CC BY-SA license.

2 Related work

Wagner et al. (2016) adapted the UD annotation guidelines to Arapaho, an Algonquian language spoken in Wyoming, USA. Shipibo-Konibo, however, seems to have been the first Indigenous American language with a treebank under the UD framework (Vasquez et al., 2018).[2] There followed Mbya Guarani (Thomas, 2019), Yupik (Park et al., 2021), K'iche' (Tyers and Henderson, 2021), Apurinã (Rueter et al., 2021), Nahuatl (Pugh et al., 2022), Tupinamba, and ten other languages, mostly Tupian of Brazil (Martín Rodríguez et al., 2022; Santos et al., 2024). As expected of treebanks for low-resource languages, they are "opportunistic corpora" (McEnery and Hardie, 2012, p. 11) with no reported inter-annotator agreement.

Parsing experiments with these treebanks showed that performance is heavily dependent on factors like gold part-of-speech (POS) tags and training data size. For instance, parsing Shipibo-Konibo with gold POS tags yielded a labeled attachment score (LAS) of 81.25 ± 3.45, while parsing raw text resulted in a score of 30.39 ± 1.34, indicating a significant drop in performance (Vasquez et al., 2018). This is not surprising given the small size of the treebank with only 407 trees and 2,706 tokens. Similarly, for the Nahuatl treebank, which had a larger size of 10,356 tokens and 939 trees, UDPipe 1 (Straka et al., 2016; Straka and Straková, 2017) was used to obtain a LAS score of 68.1 ± 2.0 with normalized text (Pugh et al., 2022).

Nheengatu, with a Digital Language Support Level of only 0.07 (Simons et al., 2022; Eberhard

[1] Following the recommendations in The Chicago Manual of Style Online (2024) and elsewhere, we capitalize *Indigenous* in the sense used in this paper.

[2] This treebank has never been part of any release of the UD collection. Instead, the UD homepage lists it among the "possible future extensions".

et al., 2023), is among many minority languages impacted by the digital divide, despite recent initiatives. For example, da Rocha D'Angelis et al. (2021) discusses the localization of a smartphone operating system for Nheengatu. However, this system does not provide any text input enhancement technologies, e.g., word completion, spelling correction, etc. After summarizing previous directly related work, de Alencar (2023) proposes a tool called Yauti for the UD annotation of Nheengatu. Cavalin et al. (2023) included Nheengatu in a study of language identification.

3 Nheengatu, the "good language"

Nheengatu originated in the 17th century in Maranhão from Tupinamba, one of the many varieties of Tupi, which was dominant along the Brazilian coast in the 16th century (Edelweiss, 1969; Borges, 1996; Rodrigues, 1996; Freire, 2011; Rodrigues and Cabral, 2011; Navarro, 2012; Finbow, 2023). The Portuguese colonizers adopted Tupi as *língua geral*, i.e. lingua franca, of which other varieties besides the LGA developed (de Lurdes Zanoli, 2022; Leite, 2013). Description and teaching of Tupi, e.g., Anchieta (1595); Figueira (1621), were incumbent on Jesuits (Edelweiss, 1969; de Almeida Navarro, 2009; Altman, 2022). Not Portuguese, but Tupi was Brazil's de facto first national language (Drumond, 1964). It was widespread among black Africans as well as Europeans and their descendants of Indigenous women, some of these mixed families attaining high economic status and social prestige (Moore, 2014). Seixas (1853) is the earliest known usage of the term *Nheengatu* 'good language' to designate the LGA.

A Royal Charter of 1689 made Tupi the official language of the State of Maranhão and Grão-Pará until an analogous document in 1727 prohibited it in favor of the Portuguese language (Moore, 2014). However, as D'Angelis (2023) points out, the mere existence of a document stating a preference for a particular language does not necessarily guarantee its widespread adoption. In fact, by 1750, except for some colonial administrators who came from Portugal, the LGA was still the predominant language spoken throughout the colony (Moore, 2014). It continuously spread along the Amazon River and its tributaries, like the Rio Negro, eventually reaching Colombia and Venezuela. In the middle of the 19th century, the LGA was the most widely spoken language in the Brazilian Amazon, including larger cities such as Belém. Documentation of Nheengatu boomed from this time until the early 20th century (Altman, 2022). On the one hand, emperor Pedro II promoted field research on Nheengatu, which resulted in the publication of oral Nheengatu literature and grammars, e.g., de Magalhães (1876). On the other, Nheengatu was part of the curriculum of the Seminary of Belém, and Church representatives produced teaching materials (Seixas, 1853; Aguiar, 1898; Costa, 1909).

The *Cabanagem* revolt (1835-1845) and mass immigration from the Northeast starting in 1877, among other factors, triggered Nheengatu's continual decline (Navarro et al., 2017). Today, as a first language, it is limited to São Gabriel da Cachoeira in the Upper Rio Negro, where it is co-official, having replaced the original non-Tupi Arawak languages of the Bare, Baniwa, and Warekena, whose languages are extinct or moribund (Eberhard et al., 2023). Nheengatu itself, with reportedly 6000 speakers in Brazil and 8000 in Colombia, where it ranks 6b and 7 on the EGIDS scale, respectively, is severely endangered, being "nearly extinct" in Venezuela, with 8b status and "[v]ery few, if any, speakers left" (Eberhard et al., 2023). Nheengatu as a contact language has also dramatically diminished (Finbow, 2020). Fortunately, diverse revitalization initiatives, e.g., in the Middle Amazon River (Lima Schwade, 2021) and the Lower Tapajós River, have targeted Nheengatu (Silva Meirelles, 2020). Besides, Indigenous people whose original languages have long gone extinct, from places as far away from the Amazon region as the Ceará State, are learning Nheengatu to affirm their ethnic identity (Filho, 2010). In 2021, the Monsenhor Tabosa municipality in Ceará adopted "Tupinheengatu" as a co-official language (Government, 2021).

All this background places Nheengatu in a unique position among the approximately 150 Indigenous languages that are still alive in Brazil, according to Storto (2019). Unlike any other, Nheengatu is supra-ethnic and has never been a tribal language (Borges, 1996; Navarro, 2012). Its influence on Brazilian Portuguese is unparalleled (de Souza Martins, 2012, 2014). Not only that, but Nheengatu has also had a significant impact on intellectuals of the stature of Mario de Andrade, Villa-Lobos, and Guimarães Rosa (Avila and Trevisan, 2015; Campoi, 2015; Pucci, 2017; Toni and Fresca, 2022). Moore (2014, p. 108) states: "Nheengatu has a notable charm. People delight in learning it and regard it with affection." Indeed, since the last decade, non-Indigenous learners have contributed significantly to the stock of texts in Nheengatu, e.g., by translating literary classics such as Graciliano Ramos, Saint-Exupéry, and Tolstoy (Avila, 2016; Trevisan, 2017; Costa, 2019). August 2023 marked a significant milestone: the Federal Supreme Court and the National Council of Justice published a translation of the Brazilian Constitution into Nheengatu (Lucchesi et al., 2023), making it the first Indigenous language to receive such an honor.

4 Overview of the treebank

Sentence lengths in the UD_Nheengatu-CompLin treebank range from 2 to approximately 50 words (Figure 1), with a mean and median of 10.01 and 8.0 words, respectively, and a standard deviation of 6.72, reflecting the richness found in Nheengatu texts, as represented in Table 1.

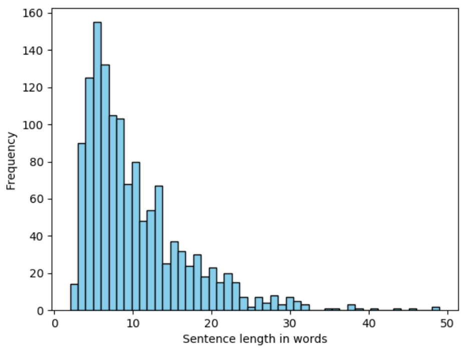

Figure 1: Frequency distribution of sentence length in the treebank.

Table 2 presents statistics for the current development version, calculated using UD's `conllu-stats.pl` script. It includes corresponding data from UD_Guajajara-TuDeT (Gerardi et al., 2022; Martín Rodríguez et al., 2022; Santos et al., 2024), the second-largest treebank for a Brazilian Indigenous language in the UD collection. Guajajara also pertains to the Tupian family. The Nheengatu treebank surpasses the Guajajara in most dimensions (Section 4.4).

The second column of Table 1 contains the first field of the sentence identifier (Section 4.1).[3] About half of the sentences stem from Avila (2021),

Freq.	Source	Rel. Freq.
705	Avila2021	0.5273
216	Navarro2016	0.1617
86	Magalhaes1876	0.0644
61	Cruz2011	0.0457
56	Alencar2021	0.0420
48	NTLN2019	0.0360
46	Rodrigues1890	0.0345
38	MooreFP1994	0.0285
23	Casasnovas2006	0.0173
23	Amorim1928	0.0173
16	Sympson1877	0.0120
7	TerraPreta2013	0.0052
3	Aguiar1909	0.0022
2	Stradelli2014	0.0015
2	Melgueiro2022	0.0015
2	DLGA2019	0.0015
1	Seixas1853	0.0007
1	Hartt1938	0.0007
1336	Total	1.0000

Table 1: Frequency of treebank examples per bibliographical source.

on which we have mostly based the selection of sources. With circa 8,000 lemmas and 4,000 unique examples, this is certainly the most comprehensive dictionary of a Brazilian Indigenous language, perhaps only rivaled by Navarro's (2015) dictionary of Ancient Tupi. The entries have a rich microstructure covering semantic, grammatical, and etymological aspects, anchored in a wide-coverage research of practically all known sources of Nheengatu from the 18th to the 21st century.

Making up 16% of the treebank, the second largest group of sentences derives from de Almeida Navarro (2016). This is a self-contained coursebook with 13 lessons containing both constructed and authentic contemporary as well as historical texts, accompanied by didactic translations into Portuguese. The lessons follow a grammatical progression that facilitates the annotation. The treebank presently covers almost all examples up to the 4th lesson. Sympson (1877); Casasnovas (2006) are two other important coursebooks (Table 1).

The 3rd portion of the treebank derives from de Magalhães (1876), perhaps the most influential oeuvre of 19th-century Nheengatu literature. Rodrigues (1890); de Amorim (1928) contain analogous collections of fables and myths. da Cruz

[3]NTLN2019 = do Brasil (2019), MooreFP1994 = Moore et al. (1994), TerraPreta2013 = Bird et al. (2013), Stradelli2014 = Stradelli (1929, 2014), DLGA2019 = Muller et al. (2019).

Treebank	Sentences	Words	Lemmas	Forms	Fusions	Features	Dependency Relations
Nheengatu	1336	13374	1244	1707	89	71	36
Guajajara	1182	9160	593	1314	138	72	29

Table 2: Comparison of statistics between UD_Nheengatu-CompLin and UD_Guajajara-TuDeT.

(2011) makes up the 4th portion. This is the most comprehensive description of the phonology and grammar of 21st-century Nheengatu as spoken by the Bare, Baniwa, and Warekena in the Upper Rio Negro. The 5th treebank portion consists of a sample from the test set of constructed sentences expressing a qualifying predication, as described in de Alencar (2021). Diverse studies have shown the importance of biblical texts for NLP (McCarthy et al., 2020; Liu et al., 2021). An indispensable textual resource documenting late 20th-century Nheengatu is the New Testament translation (do Brasil, 2019), of which the treebank features 92 sentences. 44 stem from Avila (2021). We manually extracted and adapted the other 48 sentences (Table 1), such a limited number being due to annotation difficulties. The treebank contains all 38 sentences from Moore et al. (1994), a concise but fairly complete description of Nheengatu phonology and grammar based on the transcribed speech of two native speakers from the Upper Rio Negro. The examples show to what extent Nheengatu changed structurally towards Portuguese and to what extent it remained true to Tubinamba.

The treebank only contains a few examples from textual materials by Indigenous writers, e.g., Bird et al. (2013); Filho and Neto (2016); da Silva et al. (2021); Yamã et al. (2021); Melgueiro (2022) (Table 1). Incorporating more extensive passages beyond what would be considered fair use requires permission from authors. We are already contacting some of them about this.

Our ultimate goal with the treebank is to acknowledge the linguistic significance, cultural richness, and social relevance of Nheengatu, encompassing all the texts from the 19th and early 20th centuries that are in the public domain, e.g., Seixas (1853); Hartt (1872); de Magalhães (1876); Sympson (1877); Rodrigues (1890); Aguiar (1898); Costa (1909); de Amorim (1928); Stradelli (1929); Hartt (1938). Apart from copyright restrictions, contemporary texts pose greater challenges to morphosyntactic annotation in the context of UD due to a lack of interlinear glossing or suitable translations. They also exhibit strong orthographic variation and grammatical or lexical idiosyncrasies. We aim to overcome these challenges by involving Indigenous speakers with a background in linguistics in the sentence annotation workflow.

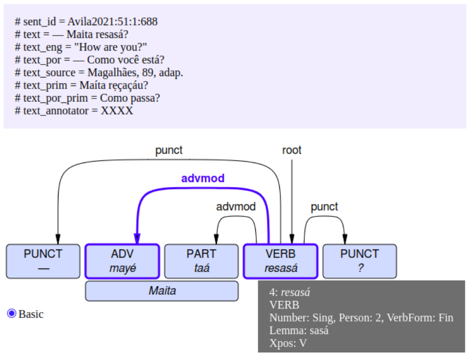

Figure 2: Dependency tree of (1) highlighting the features of the verb form *resasá*.

4.1 Metadata

UD does not specify a rigid scheme for metadata. The official `validate.py` script only requires the CoNLL-U files of a treebank to have two attributes: `text` and `sent_id` (Conllu). Therefore, one encounters great variability in the types and names of metadata attributes in the validated treebanks of the UD collection. In our treebank, sentences additionally have the obligatory attributes `text_eng`, `text_por`, `text_source` and `text_annotator`, which encode the English and Portuguese translations, the source of the sentence, and the annotator (Figure 2).[4] Unless otherwise noted, we use, if available, the translation provided in the same publication we extracted the Nheengatu example from, translating it to English or Portuguese as appropriate.

`sent_id` is a unique sentence identifier, consisting of four colon-separated pieces of information, namely, (i) an abbreviation keying to the publication the sentence stems from, (ii) an integer identifying a complete text or a continuous text fragment

[4]The graph of Figure 2 was produced with `https://urd2.let.rug.nl/~kleiweg/conllu/`.

within the publication, (iii) a sequencing index for the sentence in this text, and (iv) a count number for the sentences from the same source. Examples (1)–(3) help clarify this. In (1) (respectively Figure 2) and (2), `Avila2021` refers to Avila (2021).[5] The third and second field in (1) and (2) identify the the first two sentences of the 51st text fragment of the treebank stemming from Avila (2021), which are the 688th and 689th sentence from this source. In (3), `1:2` designates the second sentence of the first myth of de Magalhães (1876).

(1) — *Maita resasá?* 'How are you?' (Avila2021:51:1:688) (de Magalhães, 1876, p. 89)

(2) — *Se katu**ntu**.* 'I'm just fine.' (Avila2021:51:2:689) (de Magalhães, 1876, p. 89)

(3) *Pituna ukiri uikú ií ripí-**pe**.* 'The night was sleeping at the bottom of the water.' (Magalhaes1876:1:2:2)

In case of isolated sentences, the second and third fields are set to 0, see (4)–(6). Example (4) actually stems from a story but is cited without any additional context.

(4) *Yepé paá uwapika igaraganti**me**, amú uwapika yakuma**me**.* 'It is said that one was sitting in the bow of the canoe, the other was sitting in the stern.' (Avila2021:0:0:342) (Casasnovas, 2006, p. 75)

(5) *Setimã pinima pá.* 'Her leg is all painted up.' (Cruz2011:0:0:41)

(6) *Aikú suakí.* 'I'm close to her.' (Navarro2016:0:0:203)

The `text_source` attribute includes various types of information that help to locate the example within the original publication. The treebank sentences from Avila (2021) simply reproduce the string in the form of a bibliographic key and a page number that accompanies the dictionary examples. For instance, the primary source of the sentence in Figure 2 is de Magalhães (1876, p. 89).

To facilitate treebank usage for a wide range of purposes, we provide additional metadata. We limit our discussion here to `text_orig` and `text_prim`. Both convey the verbatim text of an example when it differs from the value of `text`. The `text_orig` attribute applies to an example extracted from the source identified in the `sent_id` attribute (Figure 3), while `text_prim` indicates that the source in the `sent_id` attribute is not primary (Figure 2). A total of 37.43% of the treebank sentences have one or both of these attributes, which can be relevant for training or evaluating a language detector or a spelling converter.

4.2 Annotation methodology

The construction of a treebank for one of the Indigenous languages of Brazil is particularly challenging. A total of approximately 150 languages compete for human resources to perform this task. An annotator must be familiar not only with the lexicon and grammar structure of the particular language but also with the annotation framework. It seems that the challenge has not been so appealing to the Brazilian NLP and computational linguistics communities. The treebanks referred to in Section 2 owe their existence to the participation of foreign researchers or institutions.

At first sight, it looks like UD theory only requires high school-level knowledge of traditional concepts such as parts of speech and syntactic relations, e.g., subject, object, and indirect object. Such a simplistic view will soon vanish once one starts annotating complex sentences from authentic texts and delves into the UD documentation, where one comes across non-trivial concepts such as "open clausal complement" (`xcomp`), "depictive predicate", etc. Familiar-sounding concepts such as "indirect object", "apposition", or "adverbial clause" are employed in UD in a technical sense whose understanding demands a background in syntactic theory. UD's inventory of 17 parts of speech includes categories such as particles that are not part of the traditional descriptions of Portuguese, which are generally limited to up to ten categories (Cunha and Cintra, 1985; Macambira, 1999).

The Nheengatu treebank has been annotated by a team of three non-Indigenous annotators, consisting of a senior linguist (SLIN) and two undergrad students — of whom one (EULIN) is much more experienced in the annotation task than the other (UULIN). All three are foreign-language learners of Nheengatu. SLIN and EULIN roughly possess the grammatical and lexical knowledge of de Almeida Navarro's (2016) coursebook. UULIN is less familiar with the language but has some knowledge of Ancient Tupi. SLIN is acquainted

[5]Boldface indicates the morphemes the tokenizer splits off, as explained in Section 4.3.

```
>>> import Yauti
>>> s='''Tapiíra unhehẽ: – Aramé/advj aikú asú. (p. 182) A anta falou: – Então estou-me indo. -
Tapiíra onhehẽ: – Aramé a ikộ xa çộ.'''
>>> Yauti.parseExample(s,'Magalhaes1876',2,40,83,annotator='XXXX')
# sent_id = Magalhaes1876:2:40:83
# text = Tapiíra unhehẽ: – Aramé aikú asú.
# text_eng = The tapir said: – Then I'm leaving.
# text_por = A anta falou: – Então estou-me indo.
# text_source = p. 182
# text_orig = Tapiíra onhehẽ: – Aramé a ikộ xa çộ.
# text_annotator = XXXX
1	Tapiíra	tapiíra	NOUN	N	Number=Sing	6	nsubj	_	TokenRange=0:7
2	unhehẽ	unhehẽ	_	_	_	6	_	_	SpaceAfter=No|TokenRange
=8:14
3	:	:	PUNCT	PUNCT	_	6	punct	_	TokenRange=14:15
4	–	–	PUNCT	PUNCT	_	6	punct	_	TokenRange=15:16
5	Aramé	aramé	ADV	ADVJ	AdvType=Cau	6	advmod	_	TokenRange=17:22
6	aikú	ikú	VERB	V	Number=Sing|Person=1|VerbForm=Fin	0	root
_	TokenRange=23:27
7	asú	sú	VERB	V	Number=Sing|Person=1|VerbForm=Fin	6	parataxi
s	_	SpaceAfter=No|TokenRange=28:31
8	.	.	PUNCT	PUNCT	_	6	punct	_	SpaceAfter=No|TokenRange
=31:32
```

Figure 3: Analysis of a novel example with Yauti.

with most lexical and grammatical descriptions of Nheengatu, e.g., de Magalhães (1876); Sympson (1877); Stradelli (1929); Moore et al. (1994); Casasnovas (2006); da Cruz (2011); Moore (2014); de Almeida Navarro (2016); Avila (2021).

We adopted the following labor division: EULIN and UULIN annotated, respectively, 165 and 45 sentences from de Almeida Navarro (2016), all of which SLIN revised, totaling 15.7% of the treebank. EULIN and UULIN also revised 46 and 29, respectively, of each other's sentences. SLIN annotated all 1,126 remaining sentences, i.e., 84.3% of the treebank.

All sentences were first annotated with Yauti (de Alencar, 2023) and, in case of errors, manually corrected. Typically, the annotation workflow roughly consisted of the following steps: (i) select an example for annotation; (ii) format the example; (iii) apply Yauti to the formatted example; (iv) check the resulting CoNLL-U output for any remaining ambiguities and unknown words; (v) if necessary, update Yauti's glossary and manually annotate the example with XPOS tags or token creation functions, as described in de Alencar (2023); (vi) reapply Yauti on the example; (vii) manually correct any errors; (viii) insert the serialized CoNLL-U data in the treebank file; (ix) run `validate.py` on the file and correct any detected errors. Figure 3 exemplifies the annotation of an example. The `advj` XPOS tag enables disambiguation. Yauti fails to recognize the second word due to a spelling mistake, the correct form being *unheẽ* 'it says'. Yauti also renders the Portuguese translation into English by means of Google Translate using the `deep_translator` library.[6]

[6]https://pypi.org/project/deep-translator/

4.3 Spelling normalization, tokenization, and lemmatization

One of the factors that hinder the development of computational tools and resources for minority languages is the lack of orthography standardization (Mager et al., 2018; Ebrahimi et al., 2023). This problem especially affects both historical and contemporary Nheengatu, an exclusively oral language until very recently.[7] On the one hand, each of the researchers that have collected oral stories, recorded dialogues, or produced vocabularies and grammar descriptions since the 19th century coined their own spelling system, e.g., Seixas (1853); de Magalhães (1876); Sympson (1877); Rodrigues (1890); Aguiar (1898); Costa (1909); de Amorim (1928); Stradelli (1929). On the other hand, ethnic, cultural, and religious heterogeneity and geographical dispersion of the speaker communities have prevented agreement on a common system or at least a reduced number of standards. As Avila (2021) observes, not only does each publication use its own orthography, but there is often variation within a single publication. Contemporary Nheengatu has far more than the four spelling systems identified by D'Angelis (2023). We looked up seven common words, e.g., pronouns and forms of *munhã* 'to make', across 20 publications, about half of which were by Indigenous writers, and found out that none coincides in all spellings. For example, *yam*, *yã*, and *nyã* are variants of demonstrative *nhaã* 'that' in some recent publications.

Orthographic variation in Nheengatu texts results from differences not only in the mapping of phonemes onto graphemes, possibly related to di-

[7]Avila's (2021) bibliography only includes Indigenous writers from the early 2000s onward.

alectal pronunciations, but also in word segmentation. Person and number are marked by prefixes, of which there are two series, namely, the *active*, *dynamic* or *verbal* (IP_A) and the *inactive*, *stative* or *nominal* (IP_E) (Moore et al., 1994; da Cruz, 2011; Moore, 2014; Finbow, 2020). In both historical and contemporary texts, these prefixes are sometimes spelled together, sometimes separately from their heads.[8] For example, *semayã* 'my mother' in one text corresponds to *se mãya*, *çe mãya* and *sé manha* in other texts. This sort of variation impacts many other morphemes, with the additional complication of the use of a hyphen as a separator in some texts.

To make the construction of the treebank manageable, we decided to adopt Avila's (2021) orthographic system (henceforth AVO) due to its practical advantages. First, it provides the most comprehensive description of the language, particularly regarding the lexicon, facilitating the lexical lookup of words in the treebank. Second, Yauti heavily relies on Avila's (2021) lexical and grammatical information. Third, AVO closely aligns with de Almeida Navarro's (2016) orthography, allowing those teaching or learning the language with this coursebook to easily consult the treebank. The treebank already includes 216 examples directly extracted from de Almeida Navarro (2016). Finally, AVO shares many commonalities with orthographies in use by speaker communities.

Following de Almeida Navarro (2016), Avila (2021) treats the syllabic IP_E prefixes, e.g., 1st and 3rd person singular *se* and *i* in (2) and (7), respectively, as *second class pronouns*, separating them from their heads, an approach also adopted by speakers of so-called Traditional Nheengatu (Yamã et al., 2021). In (2), the IP_E functions as an agreement marker of the stative verb *katú* 'to be fine', while it is a pronoun realizing the internal argument of the noun *resá* ''eyes" in (8) and of the postposition *irumu* ''with" in (7) and (9). It seems that, to properly reflect the role of the IP_E as an inflectional morpheme, *se katú* 'I'm fine' in (2) should be treated as a single syntactic word. While syntactic words with an internal white space are, in principle, permitted, they are discouraged by the UD guidelines (Universal Dependencies). This has led us to uniformly adopt de Almeida Navarro's (2016) and Avila's (2021) approach, tokenizing syllabic IP_E prefixes as separate syntactic words in all situations. By contrast, both authors treat the relational non-contiguity prefix R^2 and its head as a single syntactic word, e.g., *setimã* 'her leg' and *suakí* 'near her' in (5) and (6), which we also adhere to, despite the functional parallelism with the *i* IP_E, e.g., (7).

(7) *Makití i manha usú, usú i irumu.* 'Where his mother went, he went with her.' (Avila2021:14:2:158) (Rodrigues, 1890, p. 233)

(8) *Kunhã uyumuseẽ-**kwáu** ixé arama, aé umurí-**kwáu** tẽ ixé, se resá ti amuyeréu aintá i xupé, amukití aintá uikú.* "A woman can sweeten herself for me, she can even please me, my eyes don't turn to her, they are turned to the other side." (Avila2021:0:0:87) (de Amorim, 1928, p. 366)

(9) *— Resú-**putari** se irumu?* "'Do you want to go with me?"' (Avila2021:53:1:696)

In a few cases, Yauti automatically splits tokens into distinct syntactic words. In (1), the content question particle *taá* fuses with the interrogative adverb *mayé* 'how'. In (2), we have an enclitic adverb (de Almeida Navarro, 2016). Sentences (3) and (4) exemplify the clitic alomorphs of postposition *upé* 'in'. In (8) and (9), the capability and volition auxiliaries *kwáu* ''can" and *putari* 'to want' incorporate into the main verb (da Cruz, 2011).

Avila (2021, p. 145) goes beyond a mere spelling adaptation of usage examples from the literature. He often normalizes historical variants to align with the contemporary form in Upper Rio Negro Nheengatu. For instance, in (1), the original form *reçaçáu* transforms into *resasá*, although his dictionary also registers historical variant *sasáu*. Additionally, he adjusts original punctuation to adhere to current Portuguese conventions and undertakes various interventions to enhance readability for contemporary speakers.

In the general case, Yauti automatically carries out lemmatization. It strips off the plural suffix from nouns and pronouns and the person-number prefixes from conjugated active verbs, filling in the 3rd CoNLL-U column with the appropriate lemma and encoding the morphosyntactic properties of the affix as features in the sixth column (Figures 2 and 3). Yauti's capabilities in this domain, however, are still restricted to inflectional morphology. To parse derivational morphology, e.g., evaluative, collective, privative, and aspectual suffixes, it is

[8] This variation affects IP_E prefixes more often.

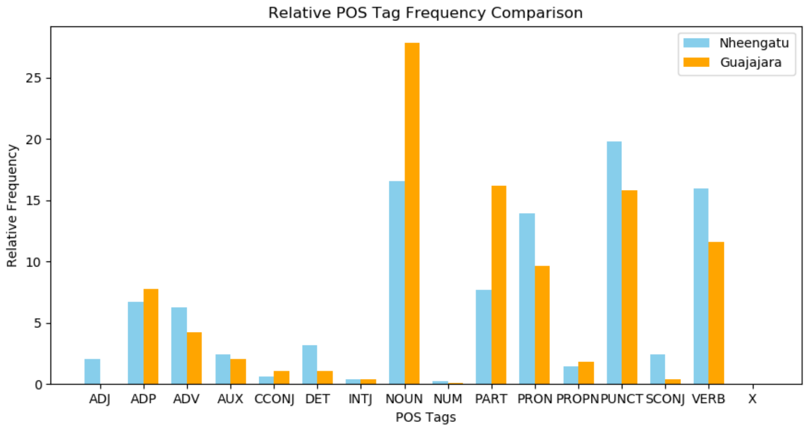

Figure 4: Relative frequency of parts of speech in UD_Nheengatu-CompLin and UD_Guajajara-TuDeT.

necessary to annotate the examples with special tags (de Alencar, 2023).

4.4 Aspects of the annotation

Of the 17 universal parts-of-speech tags (UPOS), only `SYM` and `X` have not been used because no sentence with such words has yet occurred. Usually, the tag assigned to a particular word in the treebank corresponds to Avila's (2021) taxonomy, which mostly matches the one of da Cruz (2011). Discussing the various word classification proposals for Nheengatu is beyond the scope of this paper. We focus on property concept words (Peng, 2016), which da Cruz (2011) classifies as stative verbs. Following Moore (2014); Avila (2021), we treat these words as adjectives when they do not inflect for person and number.

Figure 4 compares the relative frequency of UPOS in UD's Nheengatu and Guajajara treebanks. Except for adjectives, absent in the Guajajara treebank, the two tagsets coincide. The Nheengatu treebank has one feature less but much more dependency relations than the Guajajara (Table 2), which lacks, e.g., `acl:relcl`, `amod`, `cop`, `csubj`, `nmod:poss`, and `nsubj`. The two treebanks share 17 feature names, e.g., `Rel`, `Red`, and `Person[psor]` for relational prefixes, reduplication, and possessor's person. `Clusivity` is one of the 15 feature names of Guajajara that are missing in Nheengatu, which failed to inherit this property from Tupinamba (Rodrigues, 1990, 2013). On the other hand, `Clitic`, `Compound`, `Definite`, `Deixis`, `Derivation`, `Number[psor]` and `PartType` are some of the 14 feature names of Nheengatu that are absent in Guajajara. In the UD collection, only Nheengatu possesses `Number[grnd]=Sing` and `Person[grnd]=3`, which encode the corresponding features of the internal argument of a postposition, i.e., the *landmark* or *ground* (Tosco, 2006), when it is expressed by the relational prefix R^2, as in (6). We speculate that some of the discrepancies between the Nheengatu and Guajajara treebanks might be due to the changes the former underwent as lingua franca.

5 Parsing experiment

In this section, we report on a 10-fold cross-validation experiment with UDPipe (Straka et al., 2016; Straka and Straková, 2017). Our purpose was to assess the usefulness of the treebank for parsing, to bootstrap sentence annotation.

```
udpipe --train model training_file
udpipe --tokenize --tokenizer=ranges ↩
    ↪--accuracy --tag --parse  model ↩
    ↪test_file
udpipe --accuracy --parse  model ↩
    ↪test_file
```

Listing 1: Commands for training and testing the models.

Although UDPipe 2.0 attains better parsing results (Straka, 2018; Straka et al., 2019), due to time constraints, we limited ourselves in the experiment to the light-weight UDPipe 1.2 (Straka et al., 2016; Straka and Straková, 2017). Using the `KFold` function from the scikit-learn library (Pedregosa et al., 2011) with `shuffle=True` and `random_state=42` for reproducibility, we divided the treebank sentences into ten equal-sized folds, training and testing ten times, each time with a different fold as the test set and the remaining nine folds as the training set. We used the commands in Listing 1 for training and evaluating the models, which pretty much correspond to the default settings (Straka, 2023).

While Listing 1:2 treats the test data as raw text, also performing tokenization and tagging, Listing 1:3 takes into account the gold tokenization with the gold tags. In each of the 10 executions of these commands, UDPipe 1.2. aggregates the performance results into reports like the ones in Appendix A. With Listing 1:2, accuracy in tokenization, tagging, and parsing is computed using the F1 score, i.e. the harmonic mean of precision and recall (Straka et al., 2016; Zeman et al., 2017). Tables 3 and 4 exhibit the averages of the F1 scores and standard deviations for these three dimensions computed with the NumPy library's `mean` and `std` functions from the values of the reports generated by the ten runs of Listing 1:2. Tokenization encompasses not only splitting up text into sentences and these, in turn, into surface tokens but also two other tasks, namely, the identification of multiword tokens and syntactic words. Besides UPOS, tagging involves correctly assigning language-specific part-of-speech tags (XPOS) (Appendix B), morphological features (FEATS), and lemmas. Parsing is assessed in terms of the unlabeled attachment (UAS) and labeled attachment (LAS) scores. Table 5 presents the average UAS and LAS scores with standard deviations for parsing from gold tokenization with gold tags, computed over ten executions of Listing 1:3 as previously described. UDPipe 1.2 outperforms the rule-based Yauti morphosyntactic analyzer, which attained 80.0 and 73.2, respectively, in an analogous setting (de Alencar, 2023).

Tokenization Metric	**F1 Score (%)**
Tokenizer tokens	94.376 ± 1.19
Tokenizer multiword tokens	86.187 ± 10.28
Tokenizer words	94.279 ± 1.20
Tokenizer sentences	66.102 ± 4.53

Table 3: Tokenization results.

Tagging/Parsing Metric	**F1 Score (%)**
Tagging - UPOS	89.039 ± 1.11
Tagging - XPOS	88.16 ± 1.17
Tagging - FEATS	87.289 ± 1.17
Tagging - Lemmas	91.598 ± 1.42
Parsing - UAS	70.466 ± 1.77
Parsing - LAS	64.506 ± 1.85

Table 4: Tagging, UAS, and LAS F1 scores for parsing raw text.

	UAS (%)	**LAS (%)**
Average $\pm$ SD	86.30 ± 0.96	81.17 ± 1.02

Table 5: Parsing from gold tokenization with gold tags. SD = standard deviation.

6 Final remarks

We are continually revising the annotated sentences and adding new ones. We will train a model with UDPipe 2.0 to assess its impact on accelerating annotation. Given the growth rate of the UD_Nheengatu-CompLin treebank, we anticipate reaching 1800 sentences by the next UD release on May 15, 2024. A further interesting question to pursue is understanding whether the discrepancies from the other Tupian treebanks stem from Nheengatu history or theoretical preferences.

Acknowledgements

Diverse people and institutions have contributed to the version of the UD_Nheengatu-CompLin treebank presented in this paper. The Ceará State Foundation to Support Scientific and Technological Development (*Fundação Cearense de Apoio ao Desenvolvimento Científico e Tecnológico* — Funcap) provided scientific initiation scholarships for Vivianne Anselmo Nascimento and Dominick Maia Alexandre, who annotated and revised numerous treebank sentences. Thanks are especially due to Dominick Maia Alexandre for her diligent effort in annotating or revising more than two hundred sentences. We would like to express our immense gratitude to Eduardo de Almeida Navarro for allowing us to include the text materials from his coursebook (de Almeida Navarro, 2016) in our treebank. We are also indebted to the Federal University of Amazonas Press (*Editora da Universidade Federal do Amazonas* — UFAM), particularly to its director, Sérgio Freire, for granting permission to incorporate texts from Casasnovas (2006) into the treebank. We owe Marcel Twardowsky Avila a huge debt of gratitude. He kindly made an XML version of his Nheengatu dictionary (Avila, 2021) available to us. He generously shared his philological expertise on Nheengatu with us, clarifying many lexical and grammatical questions about the language and providing an adaptation of the myth "How the night appeared" (de Magalhães, 1876). We also thank the ongoing contributions that will mostly be reflected in the next release of the UD collection. Dominick Maia Alexandre and Juliana Lopes Gurgel have been carefully reviewing the annotation of a random sample of 200 sentences. Since January 2024, Juliana Lopes Gurgel has been annotating a great number of texts as part of her engagement as a scholarship holder with the DACILAT project, funded by the São Paulo State Research Support Foundation (*Fundação de Amparo à Pesquisa do Estado de São Paulo* – FAPESP) under Process No. 22/09158-5.

The final version of the paper has immensely benefited from the comments and suggestions of the two anonymous reviewers, to whom we are deeply thankful. We also acknowledge the use of AI-based tools Grammarly, Quillbot, and ChatGPT3.5 for the improvement of spelling, grammar, vocabulary, and style. We also utilized ChatGPT3.5 to speed up the writing of Python, Bash, and Latex code. AI suggestions were carefully examined, tested, and often corrected by us, whereby we take full responsibility for the form and content of the paper.

References

Costa Aguiar. 1898. *Doutrina christã destinada aos naturaes do amazonas em nhihingatu' com traducção portugueza em face*. Pap. e Tip. Pacheco, Silva & C., Petrópolis.

Maria Cristina Fernandes Salles Altman. 2022. As partes da oração na tradição gramatical do Tupinambá / Nheengatu. *Limite. Revista de Estudios Portugueses y de la Lusofonia*, 6:11–51.

José de Anchieta. 1595. *Arte de Grammatica da Lingoa mais usada na costa do Brasil*. Antonio de Mariz, Coimbra.

Marcel Twardowsky Avila. 2016. Estudo e prática da tradução da obra infantil "A terra dos meninos pelados", de Graciliano Ramos, do português para o Nheengatu. Dissertação de mestrado, Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo. Acesso em: 2023-12-18.

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

Marcel Twardowsky Avila and Rodrigo Godinho Trevisan. 2015. Jaguanhenhém: um estudo sobre a linguagem do Iauaretê. *Magma*, 22(12):297–335.

Steven Bird, Katie Gelbart, and Isaac McAlister, editors. 2013. *Fábulas de Terra Preta: Uma coletânea bilíngue*. sine nomine, Manaus.

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

Juliana Flávia de Assis Lorenção Campoi. 2015. A literatura brasileira em nheengatu: uma construção de narrativas no século XIX. Mestrado em literatura brasileira, Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo, São Paulo. Accessed: 2023-12-19.

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

Paulo Cavalin, Pedro Domingues, Julio Nogima, and Claudio Pinhanez. 2023. Understanding native language identification for Brazilian indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 12–18, Toronto, Canada. Association for Computational Linguistics.

Conllu. 2022. Conll-u format. `https://universaldependencies.org/u/overview/tokenization.html`. Accessed: 2024-01-09.

Adriano Luis Costa. 2019. Do português ao nheengatu: tradução da obra "De quanta terra precisa o homem?", de Leon Tolstoi. Dissertação de mestrado, Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo.

D. Frederico Costa. 1909. *Carta pastoral de D. Frederico Costa bispo do Amazonas a seus amados diocesanos*. Typ. Minerva, Fortaleza.

Celso Cunha and Lindley Cintra. 1985. *Nova Gramática do Português Contemporâneo*, 2 edition. Nova Fronteira, Rio de Janeiro.

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

Wilmar da Rocha D'Angelis, Mateus Coimbra de Oliveira, and Michéli Carolíni de Deus Lima Schwade. 2021. Acesso ao mundo digital ou acesso digital ao mundo? *Revista Digital de Políticas Linguísticas*, 15:134–158.

Florêncio Cordeiro da Silva, Aline da Cruz, and Ademar dos Santos Lima, editors. 2021. *Mayé yamuyã bũgu: Uma abordagem sociolinguística sobre a origem do bongo*. Dom Modesto, Blumenau.

Leonel Figueiredo de Alencar. 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.

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

Eduardo de Almeida Navarro. 2009. Anchieta, literato y humanista. *Língua e Literatura*, 29:177–191.

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

Antonio Brandão de Amorim. 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).

Maria de Lurdes Zanoli. 2022. *O nheengatu de São Paulo (língua geral ou língua brasílica): para uma reconstrução da área linguística das capitanias de São Vicente e de São Paulo*. Ph.D. thesis, Universidade de São Paulo, São Paulo. Tese (Doutorado).

José Vieira Couto de Magalhães. 1876. *O selvagem*. Typographia da Reforma, Rio de Janeiro.

Marie-Catherine de Marneffe, Christopher D. Manning, Joakim Nivre, and Daniel Zeman. 2021. Universal Dependencies. *Computational Linguistics*, 47(2):255–308.

José de Souza Martins. 2012. Você fala nheengatu? O Estado de S. Paulo. Page C6.

José de Souza Martins. 2014. Book flap. In Ermanno Stradelli, editor, *Vocabulário português-nheengatu, nheengatu-português*. Ateliê Editorial, Cotia.

Missão Novas Tribos do Brasil, editor. 2019. *Novo Testamento na língua Nyengatu*, 2nd edition. Sociedade Bíblica do Brasil, Barueri, SP. Original work published in 1973.

Carlos Drumond. 1964. Das tupi, die erste Nationalsprache Brasiliens. *Staden-Jahrbuch*, XI/XII:19–29.

Wilmar da Rocha D'Angelis. 2023. A língua Nheengatu e suas ortografias: questões técnicas e de política linguística. *LIAMES: Línguas Indígenas Americanas*, 23(00):e023004.

David M. Eberhard, Gary F. Simons, and Charles D. Fennig, editors. 2023. *Ethnologue: Languages of the World*, twenty-sixth edition. SIL International, Dallas.

Abteen Ebrahimi, Manuel Mager, Shruti Rijhwani, Enora Rice, Arturo Oncevay, Claudia Baltazar, María Cortés, Cynthia Montaño, John E. Ortega, Rolando Coto-solano, Hilaria Cruz, Alexis Palmer, and Katharina Kann. 2023. Findings of the AmericasNLP 2023 shared task on machine translation into indigenous languages. In *Proceedings of the Workshop on Natural Language Processing for Indigenous Languages of the Americas (AmericasNLP)*, pages 206–219, Toronto, Canada. Association for Computational Linguistics.

Frederico G. Edelweiss. 1969. *Estudos Tupis e Tupi-Guaranis: confrontos e revisões*. Livraria Brasiliana Editora, Rio de Janeiro.

Luis Figueira. 1621. *Arte da lingva brasilica*. Manoel da Silva.

Florêncio Almeida Baz Filho and Antônio Fernandes Góes Neto, editors. 2016. *Nheengatu Tapajowara: Livro do Projeto de Extensão Curso de Nheengatu UFOPA/GCI*, 2 edition. SELO Gráfica Editora, Santarém, PA.

Florêncio Almeida Vaz Filho. 2010. *A Emergência étnica dos povos indígenas do baixo Rio Tapajós, Amazônia*. Ph.D. thesis, Universidade Federal da Bahia, Salvador. Tese (Doutorado). Programa de Pós-Graduação em Ciências Sociais, Faculdade de Filosofia e Ciências Humanas. Área de concentração em Antropologia.

Thomas Finbow. 2020. Nheengatu Dâw: A preliminary study of the phonetic, phonological and morphosyntactic aspects of a case of Tupi-Guarani and Nadahup Contact in the Upper Rio Negro. *Cadernos De Linguística*, 1(3):01–21.

Thomas Finbow. 2023. The nature and emergence of the Língua Geral Amazônica according to Mufwene's Language Ecology Model. *Revista do GEL*, 19(2):75–112.

José Ribamar Bessa Freire. 2011. *Rio Babel: A história das línguas na Amazônia*, second edition. EdUERJ, Rio de Janeiro.

Fabrício Ferraz Gerardi, Stanislav Reichert, Carolina Aragon, Lorena Martín-Rodríguez, Gustavo Godoy, and Tatiana Merzhevich. 2022. TuDeT: Tupían Dependency Treebank (v0.4).

Ceará Government. 2021. Língua nativa de povos indígenas é adotada como cooficial de Monsenhor Tabosa. `https://bit.ly/3RPPqgc`. Accessed: 2023-12-19.

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

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

Fabiana Raquel Leite. 2013. A Língua Geral Paulista e o "Vocabulário Elementar da Língua Geral Brasílica". Master's thesis, Universidade Estadual de Campinas, Instituto de Estudos da Linguagem, Campinas, SP. Dissertação (mestrado).

Michéli Carolíni de Deus Lima Schwade. 2021. *"Tupi" do Rio Andirá: o Nheengatu no Médio Rio Amazonas*. Tese (doutorado), Universidade Estadual de Campinas, Instituto de Estudos da Linguagem, Campinas, SP. Accessed: 17 dez. 2023.

Ling Liu, Zach Ryan, and Mans Hulden. 2021. The usefulness of Bibles in low-resource machine translation. In *Proceedings of the 4th Workshop on the Use of Computational Methods in the Study of Endangered Languages Volume 1 (Papers)*, pages 44–50, Online. Association for Computational Linguistics.

Marco Lucchesi, José Ribamar Bessa Freira, Luis Geraldo Sant'Ana Lanfredi, Andréa Jane Silva de Medeiros, and Luanna Marley, editors. 2023. *Mundu Sa Turusu Waá : Ubêuwa Mayé Míra Itá Uikú Arãma Purãga Iké Braziu Upé*. Supremo Tribunal Federal, Conselho Nacional de Justiça, Brasília.

José Rebouças Macambira. 1999. *Estrutura Morfosintática do Português*, 9 edition. Pioneira, São Paulo.

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

Lorena Martín Rodríguez, Tatiana Merzhevich, Wellington Silva, Tiago Tresoldi, Carolina Aragon, and Fabrício F. Gerardi. 2022. Tupían language ressources: Data, tools, analyses. In *Proceedings of the 1st Annual Meeting of the ELRA/ISCA Special Interest Group on Under-Resourced Languages*, pages 48–58, Marseille, France. European Language Resources Association.

Arya D. McCarthy, Rachel Wicks, Dylan Lewis, Aaron Mueller, Winston Wu, Oliver Adams, Garrett Nicolai, Matt Post, and David Yarowsky. 2020. The Johns Hopkins University Bible corpus: 1600+ tongues for typological exploration. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 2884–2892, Marseille, France. European Language Resources Association.

Tony McEnery and Andrew Hardie. 2012. *Corpus Linguistics: Method, Theory and Practice*. Cambridge University Press, Cambridge.

Edilson Martins Melgueiro. 2022. *O Nheengatu de Stradelli aos dias atuais: uma contribuição aos estudos lexicais de línguas Tupí-Guaraní em perspectiva diacrônica*. Ph.D. thesis, Universidade de Brasília.

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

Denny Moore, Sidney Facundes, and Nádia Pires. 1994. Nheengatu (Língua Geral Amazônica), its history, and the effects of language contact. In *Proceedings of the Meeting of the Society for the Study of the Indigenous languages of the Americas, July 2-4, 1993 and the Hokan-Penutian workshop, July 3, 1993*, Report / Survey of California and other Indian Languages ; 8, pages 93–118, Berkeley, CA. [University of California].

Jean-Claude Muller, Wolf Dietrich, Ruth Monserrat, Cândida Barros, Karl-Heinz Arenz, and Gabriel Prudente, editors. 2019. *Dicionário de Língua Geral Amazônica*. Universitätsverlag Potsdam – Museu Paraense Emílio Goeldi, Potsdam – Belém/Pará. Primeira transcrição por Gabriel Prudente. Edição diplomática, revisada e ampliada com comentários e anexos por Wolf Dietrich, Ruth Monserrat e Jean-Claude Muller.

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

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

Eduardo de Almeida Navarro. 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.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Yoav Goldberg, Jan Hajič, Christopher D. Manning, Ryan McDonald, Slav Petrov, Sampo Pyysalo, Natalia Silveira, Reut Tsarfaty, and Daniel Zeman. 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).

Hyunji Hayley Park, Lane Schwartz, and Francis Tyers. 2021. Expanding Universal Dependencies for polysynthetic languages: A case of St. Lawrence Island Yupik. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 131–142, Online. Association for Computational Linguistics.

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

Siyao Peng. 2016. Property concept words in six Amazonian languages. Master's thesis, Faculty of Humanities, Leiden University, Linguistics (MA).

M. D. Pucci. 2017. Influência da voz indígena na música brasileira. *Música Popular em Revista*, 4(2):5–30. Accessed: 19 dez. 2023.

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

Aryon D. Rodrigues. 1990. You and I = neither you nor I: The personal system of Tupinambá (Tupi-Guaraní). In Doris L. Payne, editor, *Amazonian Linguistics: Studies in Lowland South American Languages*, pages 393–405. University of Texas Press, Austin.

Aryon Dall'Igna Rodrigues. 1996. As línguas gerais sul-americanas. *Papia*, 4(2):6–18.

Aryon Dall'Igna Rodrigues. 2013. Some cases of regrammaticalization in Tupí-Guaraní languages. *Revista Brasileira de Linguística Antropológica*, 2(2):231–240.

Ayron Dall'Igna Rodrigues and Ana Suelly Arruda Câmara Cabral. 2011. A contribution to the linguistic history of the língua geral amazônica. *ALFA: Revista de Linguística*, 55(2).

João Barbosa Rodrigues. 1890. *Poranduba amazonense ou kochiyma-uara porandub, 1872-1887*. Typ. de G. Leuzinger & Filhos, Rio de Janeiro.

Jack Rueter, Marília Fernanda Pereira de Freitas, Sidney Da Silva Facundes, Mika Hämäläinen, and Niko Partanen. 2021. Apurinã Universal Dependencies treebank. In *Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas*, pages 28–33, Online. Association for Computational Linguistics.

Luana Luiza Santos, Carolina Coelho Aragon, and Fabrício Gerardi. 2024. Línguas minoritárias e anotações sintáticas de corpora: experiências de pesquisa na iniciação científica. *Letras de hoje*, 59(1):1–9. Published: 2024-01-10.

Manoel Justiniano de Seixas. 1853. *Vocabulario da lingua indigena geral para o uso do Seminario Episcopal do Pará*. Typ. de Mattos e Compª., Pará.

Sâmela Ramos da Silva Silva Meirelles. 2020. *A reinscrição de uma língua destituída: o Nheengatu no Baixo Tapajós*. Ph.D. thesis, Universidade Estadual de Campinas, Instituto de Estudos da Linguagem. Accessed: 17 dez. 2023.

Gary F. Simons, Abbey L. L. Thomas, and Chad K. K. White. 2022. Assessing digital language support on a global scale. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 4299–4305, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Luciana Raccanello Storto. 2019. *Línguas indígenas: tradição, universais e diversidade*. Mercado de Letras, Campinas, SP.

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

Ermano Stradelli. 1929. Vocabularios da lingua geral portuguez-nheêngatú e nheêngatú-portuguez, precedidos de um esboço de Grammatica nheênga-umbuêsáua mirî e seguidos de contos em lingua geral nheêngatú poranduua. *Revista do Instituto Historico e Geographico Brasileiro*, 158(104):9–768.

Milan Straka. 2018. UDPipe 2.0 prototype at CoNLL 2018 UD shared task. In *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 197–207, Brussels, Belgium. Association for Computational Linguistics.

Milan Straka. 2023. *UDPipe 1 User's Manual*.

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

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

Milan Straka, Jana Straková, and Jan Hajic. 2019. UDPipe at SIGMORPHON 2019: Contextualized embeddings, regularization with morphological categories, corpora merging. In *Proceedings of the 16th Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 95–103, Florence, Italy. Association for Computational Linguistics.

Pedro Luiz Sympson. 1877. *Grammatica da lingua brazilica geral, fallada pelos aborigines das provincias do Pará e Amazonas*. Typographia do Commercio do Amazonas, Manaus.

The Chicago Manual of Style Online. 2024. Capitalization. `https://www.chicagomanualofstyle.org/qanda/data/faq/topics/Capitalization/faq0106.html`. Accessed: February 6, 2024.

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

Flávia Camargo Toni and Camila Fresca. 2022. Natureza e modernismo: Mário de Andrade e Villa-Lobos antes da Semana. *Estudos Avançados*, 36(104):143–183.

Mauro Tosco. 2006. Towards a geometry of adpositional systems: A preliminary investigation of Gawwada. In Pier Giorgio Borbone, Alessandro Mengozzi, and Mauro Tosco, editors, *Loquentes Linguis: Studi Linguistici e Orientali in Onore di Fabrizio Pennacchietti*, pages 695–702. Harrassowitz, Wiesbaden.

Rodrigo Godinho Trevisan. 2017. Tradução comentada da obra “Le Petit Prince”', de Antoine de Saint-Exupéry, do francês ao nheengatu. Dissertação de mestrado, Faculdade de Filosofia, Letras e Ciências Humanas, Universidade de São Paulo. Acesso em: 2023-12-18.

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

Universal Dependencies. 2023. Tokenization and word segmentation. https://universaldependencies.org/u/overview/tokenization.html. Accessed: 2023-12-25.

Alonso Vasquez, Renzo Ego Aguirre, Candy Angulo, John Miller, Claudia Villanueva, Željko Agić, Roberto Zariquiey, and Arturo Oncevay. 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.

Irina Wagner, Andrew Cowell, and Jena D. Hwang. 2016. Applying Universal Dependency to the Arapaho language. In *Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016)*, pages 171–179, Berlin, Germany. Association for Computational Linguistics.

Yaguarê Yamã, Elias Yaguakãng, Egídia Reis, and Mario José. 2021. *Dicionário e estudo de nheengatu tradicional*, 2 edition. Cintra, São Paulo.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Salih Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Erik Andersen, Lene Antonsen, Tatsuya Aoyama, Katya Aplonova, Angelina Aquino, Carolina Aragon, Glyd Aranes, Maria Jesus Aranzabe, Bilge Nas Arıcan, H͡órunn Arnardóttir, Gashaw Arutie, Jessica Naraiswari Arwidarasti, Masayuki Asahara, Katla Ásgeirsdóttir, Deniz Baran Aslan, Cengiz Asmazoğlu, Luma Ateyah, Furkan Atmaca, Mohammed Attia, Aitziber Atutxa, Liesbeth Augustinus, Mariana Avelãs, Elena Badmaeva, Keerthana Balasubramani, Miguel Ballesteros, Esha Banerjee, Sebastian Bank, Verginica Barbu Mititelu, Starkaður Barkarson, Rodolfo Basile, Victoria Basmov, Colin Batchelor, John Bauer, Seyyit Talha Bedir, Shabnam Behzad, Juan Belieni, Kepa Bengoetxea, İbrahim Benli, Yifat Ben Moshe, Gözde Berk, Riyaz Ahmad Bhat, Erica Biagetti, Eckhard Bick, Agnė Bielinskienė, Kristín Bjarnadóttir, Rogier Blokland, Victoria Bobicev, Loïc Boizou, Emanuel Borges Völker, Carl Börstell, Cristina Bosco, Gosse Bouma, Sam Bowman, Adriane Boyd, Anouck Braggaar, António Branco, Kristina Brokaitė, Aljoscha Burchardt, Marisa Campos, Marie Candito, Bernard Caron, Gauthier Caron, Catarina Carvalheiro, Rita Carvalho, Lauren Cassidy, Maria Clara Castro, Sérgio Castro, Tatiana Cavalcanti, Gülşen Cebiroğlu Eryiğit, Flavio Massimiliano Cecchini, Giuseppe G. A. Celano, Slavomír Čéplö, Neslihan Cesur, Savas Cetin, Özlem Çetinoğlu, Fabricio Chalub, Liyanage Chamila, Shweta Chauhan, Ethan Chi, Taishi Chika, Yongseok Cho, Jinho Choi, Jayeol Chun, Juyeon Chung, Alessandra T. Cignarella, Silvie Cinková, Aurélie Collomb, Çağrı Çöltekin, Miriam Connor, Claudia Corbetta, Daniela Corbetta, Francisco Costa, Marine Courtin, Benoît Crabbé, Mihaela Cristescu, Vladimir Cvetkoski, Ingerid Løyning Dale, Philemon Daniel, Elizabeth Davidson, Leonel Figueiredo de Alencar, Mathieu Dehouck, Martina de Laurentiis, Marie-Catherine de Marneffe, Valeria de Paiva, Mehmet Oguz Derin, Elvis de Souza, Arantza Diaz de Ilarraza, Carly Dickerson, Arawinda Dinakaramani, Elisa Di Nuovo, Bamba Dione, Peter Dirix, Kaja Dobrovoljc, Adrian Doyle, Timothy Dozat, Kira Droganova, Magali Sanches Duran, Puneet Dwivedi, Christian Ebert, Hanne Eckhoff, Masaki Eguchi, Sandra Eiche, Marhaba Eli, Ali Elkahky, Binyam Ephrem, Olga Erina, Tomaž Erjavec, Farah Essaidi, Aline Etienne, Wograine Evelyn, Sidney Facundes, Richárd Farkas, Federica Favero, Jannatul Ferdaousi, Marília Fernanda, Hector Fernandez Alcalde, Amal Fethi, Jennifer Foster, Theodorus Fransen, Cláudia Freitas, Kazunori Fujita, Katarína Gajdošová, Daniel Galbraith, Federica Gamba, Marcos Garcia, Moa Gärdenfors, Fabrício Ferraz Gerardi, Kim Gerdes, Luke Gessler, Filip Ginter, Gustavo Godoy, Iakes Goenaga, Koldo Gojenola, Memduh Gökırmak, Yoav Goldberg, Xavier Gómez Guinovart, Berta González Saavedra, Bernadeta Griciūtė, Matias Grioni, Loïc Grobol, Normunds Grūzītis, Bruno Guillaume, Kirian Guiller, Céline Guillot-Barbance, Tunga Güngör, Nizar Habash, Hinrik Hafsteinsson, Jan Hajič, Jan Hajič jr., Mika Hämäläinen, Linh Hà Mỹ, Na-Rae Han, Muhammad Yudistira Hanifmuti, Takahiro Harada, Sam Hardwick, Kim Harris, Dag Haug, Johannes Heinecke, Oliver Hellwig, Felix Hennig, Barbora Hladká, Jaroslava Hlaváčová, Florinel Hociung, Petter Hohle, Yidi Huang, Marivel Huerta Mendez, Jena Hwang, Takumi Ikeda, Anton Karl Ingason, Radu Ion, Elena Irimia, Ọlájídé Ishola, Artan Islamaj, Kaoru Ito, Sandra Jagodzińska, Siratun Jannat, Tomáš Jelínek, Apoorva Jha, Katharine Jiang, Anders Johannsen, Hildur Jónsdóttir, Fredrik Jørgensen, Markus Juutinen, Hüner Kaşıkara, Nadezhda Kabaeva, Sylvain Kahane, Hiroshi Kanayama, Jenna Kanerva, Neslihan Kara, Ritván Karahóğa, Andre Kåsen, Tolga Kayadelen, Sarveswaran Kengatharaiyer, Václava Kettnerová, Lilit Kharatyan, Jesse Kirchner, Elena Klementieva, Elena Klyachko, Petr Kocharov, Arne Köhn, Abdullatif Köksal, Kamil Kopacewicz, Timo Korkiakangas, Mehmet Köse, Alexey Koshevoy, Natalia Kotsyba, Jolanta Kovalevskaitė, Simon Krek, Parameswari Krishnamurthy, Sandra Kübler, Adrian Kuqi, Oğuzhan Kuyrukçu, Aslı Kuzgun, Sookyoung Kwak, Kris Kyle, Käbi Laan, Veronika Laippala, Lorenzo Lambertino, Tatiana Lando, Septina Dian Larasati, Alexei Lavrentiev, John Lee, Phuong Lê Hồng, Alessandro Lenci, Saran Lertpradit, Herman Leung, Maria Levina, Lauren Levine, Cheuk Ying Li, Josie Li, Keying Li, Yixuan Li, Yuan Li, KyungTae Lim, Bruna Lima Padovani, Yi-Ju Jessica Lin, Krister Lindén, Yang Janet Liu, Nikola Ljubešić, Irina Lobzhanidze, Olga Loginova, Lucelene Lopes, Stefano Lusito, Andry Luthfi, Mikko Luukko, Olga Lyashevskaya, Teresa Lynn, Vivien Macketanz, Menel Mahamdi, Jean Maillard, Ilya Makarchuk, Aibek Makazhanov, Michael Mandl, Christopher Manning, Ruli Manurung, Büşra Marşan, Cătălina Mărănduc, David Mareček, Katrin Marheinecke, Stella Markantonatou, Héctor Martínez Alonso, Lorena Martín Rodríguez, André Martins, Cláudia Martins, Jan Mašek, Hiroshi Matsuda, Yuji Matsumoto, Alessandro Mazzei, Ryan McDonald, Sarah McGuinness, Gustavo Mendonça, Tatiana Merzhevich, Niko Miekka, Aaron Miller, Karina Mischenkova, Anna Missilä, Cătălin Mititelu, Maria Mitrofan, Yusuke Miyao, AmirHossein Mojiri Foroushani, Judit Molnár, Amirsaeid Moloodi, Simonetta Montemagni, Amir More, Laura Moreno Romero, Giovanni Moretti, Shinsuke Mori, Tomohiko Morioka, Shigeki Moro, Bjartur Mortensen, Bohdan Moskalevskyi, Kadri Muischnek, Robert Munro, Yugo Murawaki, Kaili Müürisep, Pinkey Nainwani, Mariam Nakhlé, Juan Ignacio Navarro Horñiacek, Anna Nedoluzhko, Gunta Nešpore-Bērzkalne, Manuela Nevaci, Luong Nguyễn Thị, Huyền Nguyễn Thị Minh, Yoshihiro Nikaido, Vitaly Nikolaev, Rattima Nitisaroj, Alireza Nourian, Maria das Graças Volpe Nunes, Hanna Nurmi, Stina Ojala, Atul Kr. Ojha, Hulda Óladóttir, Adédayọ Olúòkun, Mai Omura, Emeka Onwuegbuzia, Noam Ordan, Petya Osenova, Robert Östling, Lilja Øvrelid, Şaziye Betül Özateş, Merve Özçelik, Arzucan Özgür, Balkız Öztürk Başaran, Teresa Paccosi, Alessio Palmero Aprosio, Anastasia Panova, Thiago Alexandre Salgueiro Pardo, Hyunji Hayley Park, Niko Partanen, Elena Pascual, Marco Passarotti, Agnieszka Patejuk, Guilherme Paulino-Passos, Giulia Pedonese, Angelika Peljak-Łapińska, Siyao Peng, Siyao Logan Peng, Rita Pereira, Sílvia Pereira, Cenel-Augusto Perez, Natalia Perkova, Guy Perrier, Slav Petrov, Daria Petrova, Andrea Peverelli, Jason Phelan, Claudel Pierre-Louis, Jussi Piitulainen, Yuval Pinter, Clara Pinto, Rodrigo Pintucci, Tommi A Pirinen, Emily Pitler, Magdalena Plamada, Barbara Plank, Thierry Poibeau, Larisa Ponomareva, Martin Popel, Lauma Pretkalniņa, Sophie Prévost, Prokopis Prokopidis, Adam Przepiórkowski, Robert Pugh, Tiina Puolakainen, Sampo Pyysalo, Peng Qi, Andreia Querido, Andriela Rääbis, Alexandre Rademaker, Mizanur Rahoman, Taraka Rama, Loganathan Ramasamy, Carlos Ramisch, Joana Ramos, Fam Rashel, Mohammad Sadegh Rasooli, Vinit Ravishankar, Livy Real, Petru Rebeja, Siva Reddy, Mathilde Regnault, Georg Rehm, Arij Riabi, Ivan Riabov, Michael Rießler, Erika Rimkutė, Larissa Rinaldi, Laura Rituma, Putri Rizqiyah, Luisa Rocha, Eiríkur Rögnvaldsson, Ivan Roksandic, Mykhailo Romanenko, Rudolf Rosa, Valentin Roșca, Davide Rovati, Ben Rozonoyer, Olga Rudina, Jack Rueter, Kristján Rúnarsson, Shoval Sadde, Pegah Safari, Aleksi Sahala, Shadi Saleh, Alessio Salomoni, Tanja Samardžić, Stephanie Samson, Manuela Sanguinetti, Ezgi Sanıyar, Dage Särg, Marta Sartor, Mitsuya Sasaki, Baiba Saulīte, Agata Savary, Yanin Sawanakunanon, Shefali Saxena, Kevin Scannell, Salvatore Scarlata, Emmanuel Schang, Nathan Schneider, Sebastian Schuster, Lane Schwartz, Djamé Seddah, Wolfgang Seeker, Mojgan Seraji, Syeda Shahzadi, Mo Shen, Atsuko Shimada, Hiroyuki Shirasu, Yana Shishkina, Muh Shohibussirri, Maria Shvedova, Janine Siewert, Einar Freyr Sigurðsson, João Silva, Aline Silveira, Natalia Silveira, Sara Silveira, Maria Simi, Radu Simionescu, Katalin Simkó, Mária Šimková, Haukur Barri Símonarson, Kiril Simov, Dmitri Sitchinava, Ted Sither, Maria Skachedubova, Aaron Smith, Isabela Soares-Bastos, Per Erik Solberg, Barbara Sonnenhauser, Shafi Sourov, Rachele Sprugnoli, Vivian Stamou, Steinh͡ór Steingrímsson, Antonio Stella, Abishek Stephen, Milan Straka, Emmett Strickland, Jana Strnadová, Alane Suhr, Yogi Lesmana Sulestio, Umut Sulubacak, Shingo Suzuki, Daniel Swanson, Zsolt Szántó, Chihiro Taguchi, Dima Taji, Fabio Tamburini, Mary Ann C. Tan, Takaaki Tanaka, Dipta Tanaya, Mirko Tavoni, Samson Tella, Isabelle Tellier, Marinella Testori, Guillaume Thomas, Sara Tonelli, Liisi Torga, Marsida Toska, Trond Trosterud, Anna Trukhina, Reut Tsarfaty, Utku Türk, Francis Tyers, Sveinbjörn H͡órðarson, Vilhjálmur H͡orsteinsson, Sumire Uematsu, Roman Untilov, Zdeňka Urešová, Larraitz Uria, Hans Uszkoreit, Andrius Utka, Elena Vagnoni, Sowmya Vajjala, Socrates Vak, Rob van der Goot, Martine Vanhove, Daniel van Niekerk, Gertjan van Noord, Viktor Varga, Uliana Vedenina, Giulia Venturi, Eric Villemonte de la Clergerie, Veronika Vincze, Natalia Vlasova, Aya Wakasa, Joel C. Wallenberg, Lars Wallin, Abigail Walsh, Jonathan North Washington, Maximilan Wendt, Paul Widmer, Shira Wigderson, Sri Hartati Wijono, Vanessa Berwanger Wille, Seyi Williams, Mats Wirén, Christian Wittern, Tsegay Woldemariam, Tak-sum Wong, Alina Wróblewska, Qishen Wu, Mary Yako, Kayo Yamashita, Naoki Yamazaki, Chunxiao Yan, Koichi Yasuoka, Marat M. Yavrumyan, Arife Betül Yenice, Olcay Taner Yıldız, Zhuoran Yu, Arlisa Yuliawati, Zdeněk Žabokrtský, Shorouq Zahra, Amir Zeldes, He Zhou, Hanzhi Zhu, Yilun Zhu, Anna Zhuravleva, and Rayan Ziane. 2023. Universal dependencies 2.13. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Daniel Zeman, Martin Popel, Milan Straka, Jan Hajič, Joakim Nivre, Filip Ginter, Juhani Luotolahti, Sampo Pyysalo, Slav Petrov, Martin Potthast, Francis Tyers, Elena Badmaeva, Memduh Gokirmak, Anna Nedoluzhko, Silvie Cinková, Jan Hajič jr., Jaroslava Hlaváčová, Václava Kettnerová, Zdeňka Urešová, Jenna Kanerva, Stina Ojala, Anna Missilä, Christopher D. Manning, Sebastian Schuster, Siva Reddy, Dima Taji, Nizar Habash, Herman Leung, Marie-Catherine de Marneffe, Manuela Sanguinetti, Maria Simi, Hiroshi Kanayama, Valeria de Paiva, Kira Droganova, Héctor Martínez Alonso, Çağrı Çöltekin, Umut Sulubacak, Hans Uszkoreit, Vivien Macketanz, Aljoscha Burchardt, Kim Harris, Katrin Marheinecke, Georg Rehm, Tolga Kayadelen, Mohammed Attia, Ali Elkahky, Zhuoran Yu, Emily Pitler, Saran Lertpradit, Michael Mandl, Jesse Kirchner, Hector Fernandez Alcalde, Jana Strnadová, Esha Banerjee, Ruli Manurung, Antonio Stella, Atsuko Shimada, Sookyoung Kwak, Gustavo Mendonça, Tatiana Lando, Rattima Nitisaroj, and Josie Li. 2017. CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 1–19, Vancouver, Canada. Association for Computational Linguistics.

A Example parsing reports

Listings 2 and 3 display the commands of the first run of the ten-fold cross-validation, which generate the reports reproduced further below.

```
udpipe --tokenize --tokenizer=ranges ↩
    ↪--accuracy --tag --parse model1.↩
    ↪output test1.conllu
```

Listing 2: First run of the 10-fold cross-validation: parsing from raw text.

```
Number of SpaceAfter=No features in gold data: 373
Tokenizer tokens - system: 1277, gold: 1313, precision: 96.01%, recall: 93.37%, f1: 94.67%
Tokenizer multiword tokens - system: 10, gold: 11, precision: 90.00%, recall: 81.82%, f1: 85.71%
Tokenizer words - system: 1287, gold: 1324, precision: 95.96%, recall: 93.28%, f1: 94.60%
Tokenizer sentences - system: 104, gold: 134, precision: 77.88%, recall: 60.45%, f1: 68.07%
Tagging from plain text (CoNLL17 F1 score) - gold forms: 1324, upostag: 90.00%, xpostag: 89.70%, feats: 88.70%, alltags: 87.09%, lemmas: 92.38%
Parsing from plain text with computed tags (CoNLL17 F1 score) - gold forms: 1324, UAS: 73.54%, LAS: 67.79%
```

```
udpipe --accuracy --parse model1.↩
    ↪output test1.conllu
```

Listing 3: First run of the 10-fold cross-validation: Parsing with gold tokenization and gold tags.

```
Parsing from gold tokenization with gold tags - forms: 1324, UAS: 87.39%, LAS: 83.23%
```

B Language-specific tagset (XPOS)

XPOS	Abbreviation	Abbreviation expansion
A	adj.	first class adjective
A2	adj. 2ª cl.	second class adjective
ADP	postp.	postposition
ADV	adv.	adverb
ADVA	adv. manner	adverb of manner
ADVC	adv. loc.	locative adverb
ADVD	adv. dem.	demonstrative adverb
ADVDI	adv. dem. dist.	distal demonstrative adverb
ADVDX	adv. dem. prox.	proximal demonstrative adverb
ADVG	adv. gr.	degree adverb
ADVJ	adv. conj.	causal conjunctional adverb
ADVL	adv. rel.	relative adverb
ADVLA	adv. rel. man.	manner relative adverb
ADVLC	adv. rel. loc.	locative relative adverb
ADVLT	adv. rel. temp.	temporal relative adverb

Table 6: XPOS tags (part 1).

Tables 6, 7 and 8 explain UD_Nheengatu-CompLin's language-specific part-of-speech tags (XPOS) as employed in Yauti's full-form lexicon. The second column reproduces the Portuguese abbreviations for word classes of Yauti's glossary, which are fully translated into English in the third column.

XPOS	Abbreviation	Abbreviation expansion
ADVM	adv. mod.	modal adverb
ADVNC	adv. ind. loc.	indefinite locative adverb
ADVNT	adv. ind. temp.	indefinite temporal adverb
ADVO	adv. ord.	ordinal adverb
ADVP	adv. conj. opos.	concessive conjunctional adverb
ADVR	adv. interr.	interrogative adverb
ADVRA	adv. interr. man.	manner interrogative adverb
ADVRC	adv. interr. loc.	locative interrogative adverb
ADVRT	adv. interr. temp.	temporal interrogative adverb
ADVRU	adv. interr. caus.	causal interrogative adverb
ADVS	adv. intens.	intensity adverb
ADVT	adv. temp.	temporal adverb
AFF	part. afirm.	affirmation particle
ART	art. indef.	indefinite article
ASSUM	part. assum.	assumption particle
AUXFR	aux. flex. pre.	preverbal inflected auxiliary
AUXFS	aux. flex. post.	postverbal inflected auxiliary
AUXN	aux. non-flex.	noninflected auxiliary
CARD	num. card.	cardinal numeral
CCONJ	cconj.	coordinating conjunction
CERT	part. cert.	certainty particle
CLADP	postp. encl.	enclitic postposition
CLADV	adv. encl.	enclitic adverb
COND	part. cond.	conditional particle
CONJ	conj.	conjunction
CONS	part. cons.	consent particle
COP	cop.	copula verb
CQ	part. interr. cont.	content question particle
DEM	pron. dem.	demonstrative pronoun
DEMS	pron. dem. dist.	distal demonstrative pronoun
DEMSN	pron. dem. dist. non-flex.	noninflected distal demonstrative pronoun
DEMX	pron. dem. prox.	proximal demonstrative pronoun
EMP	pron. enf.	emphasis pronoun
EXST	part. exist.	existential particle
FOC	part. focus	focus particle
FRUST	part. frust.	frustrative particle
FUT	part. fut.	future particle
IMPF	part. imperf.	imperfective particle
IND	pron. indef.	indefinite pronoun
INDQ	pron. quant.	indefinite quantifier pronoun
INT	pron. interr.	interrogative pronoun
INTJ	interj.	interjection
MOD	part. mod.	modal particle
N	s.	common noun
NEC	part. neces.	necessity deontic particle

Table 7: XPOS tags (part 2).

Tag	Abbreviation	Abbreviation expansion
NEG	part. neg.	negation particle
NEGI	part. neg. imp.	negative imperative particle
ORD	num. ord.	ordinal numeral
PART	part.	particle
PFV	part. perf.	perfective particle
PQ	part. interr. pol.	polar question particle
PREC	part. prec.	precative particle
PREF	pref.	prefix
PREP	prep.	preposition
PRET	part. pret.	past tense particle
PRON	pron.	first class pronoun
PRON2	pron. 2ª cl.	second class pronoun
PROPN	s. próprio	proper noun
PROTST	part. prot.	protestative particle
PRSV	part. pres.	presentative particle
REL	pron. rel.	relative pronoun
RELF	pron. rel. livre	free relative pronoun
RPRT	part. report.	reportative particle
SCONJ	sconj.	postverbal subordinating conjunction
SCONJR	sconj. pre.	preverbal subordinating conjunction
SUFF	suf.	suffix
TOT	pron. quant. univ.	universal quantifier pronoun
TOTAL	part. tot.	totalitive particle
V	v.	first class verb
V2	v. 2ª cl.	second class verb
V3	v. 3ª cl.	third class verb
VSUFF	v. suff.	noninflectionable suffixal verb

Table 8: XPOS tags (part 3).