Apurinã Universal Dependencies Treebank

Jack Rueter[1], Marília Fernanda Pereira de Freitas[2], Sidney da Silva Facundes[2], Mika Hämäläinen[1] and Niko Partanen[1]
[1]University of Helsinki
[2]Universidade Federal do Pará
[1]firstname.lastname@helsinki.fi
[2]{mfpf,sidi}@ufpa.br

Abstract

This paper presents and discusses the first Universal Dependencies treebank for the Apurinã language. The treebank contains 76 fully annotated sentences, applies 14 parts-of-speech, as well as seven augmented or new features – some of which are unique to Apurinã. The construction of the treebank has also served as an opportunity to develop finite-state description of the language and facilitate the transfer of open-source infrastructure possibilities to an endangered language of the Amazon. The source materials used in the initial treebank represent fieldwork practices where not all tokens of all sentences are equally annotated. For this reason, establishing regular annotation practices for the entire Apurinã treebank is an ongoing project.

1 Introduction

Apurinã (ISO code apu) is an endangered language spoken in the Amazon Basin. The language has around 2,000 native speakers and it is definitely endangered according to the UNESCO classification (Moseley, 2010). This paper is dedicated to describing the first ever Universal Dependencies (UD) treebank for Apurinã[1]. We describe how the treebank was created, and what exact decisions were made in different parts of the process.

The UD project (Zeman et al., 2020) has the goal of collecting syntactically annotated corpora containing information about lemmas, parts-of-speech, morphology and dependencies in such a fashion that the annotation conventions are shared across languages, although there may be inconsistencies between languages (see Rueter and Partanen 2019). As the number of South American languages represented in the Universal Dependencies project has grown rapidly in the last years (see i.e. Vasquez et al., 2018; Thomas, 2019), the descriptions of individual treebanks are thereby also a very valuable resource that helps to maintain consistency in the treebanks of this complex linguistic regions.

The advantage of UD treebanks is that they can be used directly in many neural NLP applications such as parsers (Qi et al., 2020) and part-of-speech taggers (Kim et al., 2017). Although the endangered languages have a very different starting point in comparison with large languages (Hämäläinen, 2021), there has been recent work (Lim et al., 2018; Ens et al., 2019; Hämäläinen and Wiechetek, 2020; Alnajjar, 2021) showcasing good results on a variety of tasks even for the few endangered languages that have a UD treebank.

The fact that UD treebanks can be used with neural models to build higher level NLP tools is one of the key motivations for us to build this resource for Apurinã. In addition to NLP research, UD treebanks have been used in many purely linguistically motivated research papers (Croft et al., 2017; Levshina, 2017, 2019; Sinnemäki and Haakana, 2020). We believe such developments will only grow stronger, and believe that easily available treebanks in the UD project, covering continuously better the world's linguistic diversity, will continue widening their role as suitable and valuable tools for both descriptive linguistic research and computational linguistics. This goal will be achievable only by creating an open discussion about the conventions and choices done in different treebanks, which can be adjusted and refined at the later stage. This study aims to provide such description about Apurinã treebank. An example of a UD annotated sentence in Apurinã can be seen in Figure 1.

2 Modelling the Apurinã Language in UD

The Apurinã language has a rich morphology with regular correlation between numerous formatives and semantic categories. One challenge in the conversion from fieldwork/typology style annotation to that used in the UD project is to choose what

[1] https://github.com/UniversalDependencies/UD_Apurina-UFPA

Proceedings of the First Workshop on Natural Language Processing for Indigenous Languages of the Americas, pages 28–33
June 11, 2021. ©2021 Association for Computational Linguistics

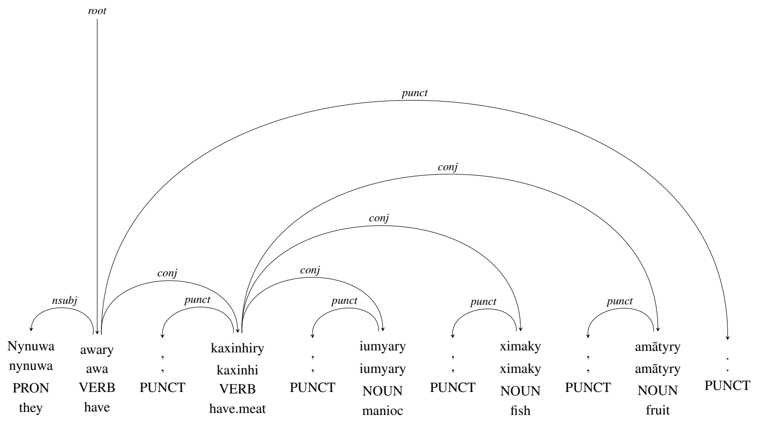

Figure 1: An example of a UD tree for an Apurinã sentence meaning *'They had it, had meat, manioc, fish, fruit'*.

features should or can be highlighted with specific transferability to other UD projects and which ones should only be represented as language specific morphology.

The task has also been contemplated from a finite-state perspective, where regular inflection plays a decisive role in determining lemma and regular inflection strategies. Finite-state description also entails the use of the open-source GiellaLT infrastructure (Norwegian Arctic University, Tromsø) (Moshagen et al., 2014), which introduces a large number of mutual tag definitions and practices that can be applied to Apurinã with ample analogy from the morphologically challenging Uralic and other languages of the Circum-Polar region.

Solutions for dealing with the categories of case, number, person and gender are available in the GiellaLT infrastructure. Extensions, however, have been required for Apurinã in the categories of number, person and gender. Unlike some Indo-European and Uralic languages, the category of gender must also be applied to the subjects and objects of verbs; subject and object marking for number (see Facundes et al. 2021) and person categories could have been adapted directly from description work in the Erzya (Rueter and Tyers, 2018) and Moksha (Rueter, 2018) UD treebanks.

2.1 Case

The Feature of CASE, for example, permeates many of the individual language projects, and some attempts are made to align case documentation with principles adapted in the Unimorph project (Kirov et al., 2018). In the instance of Apurinã, parallel case categories have been adapted with names familiar to those used in work with languages of the Uralic language family. This was done principally because the team involved in the annotation was most familiar with this language family: at the same time the Uralic UD annotations, especially for the minority languages, are already closely adapted to the UD project at large. Whether such generalizations work is also one test for the cross-linguistic suitability of the current annotation model.

The concept of case in Apurinã is most salient in oblique marking. While the subject, object and adposition complements show no special marking, there are at least six oblique marker to deal with (Facundes, 2000, 385–390). The labeling of these cases also underlines a problem not new to UD, namely, every language research tradition tends to apply its own terms for similar functions. Apurinã, as in the Uralic languages, shows evidence of case-like formatives associated not only with nominals but verbs, as well. In the first version of the Apurinã UD treebank, the formative case name pairs have been assigned as follows: *munhi* = Dat (dative, allative, goal), *kata* = Com (comitative, associative), *ã* = Loc (locative, instrumental), *Ø* = Nom (nominative). Subsequent work in the dataset will introduce the additional case formative *sawaky* = Temp (temporal), and show the extent of shared morphology across parts-of-speech.

2.2 Possession

One complexity of Apurinã morphology is encountered in the expression of possession. While the possessor of a noun may be indicated morphologically on the possessum, it is not obligatory. A preceding personal pronoun, for example, also serves as a marker of possession, to which the morphology of the possessum reacts and shows indication of being possessed. Hence, there are four basic categories that can be expressed on the possessum:

person, number and gender of the possessor, on the one hand, and indication of whether the entity is a possessum or not, on the other. These categories are expressed as feature and value pairs in the UD project:

- Gender[psor]=Masc|Fem
- Number[psor]=Plur|Sing
- Person[psor]=1|2|3
- Possessed=Yes|No

While matters of gender, number and person are directly attested in the morphology of the possessum, the feature POSSESSED identifies the individual noun as to whether there is or is not marking indicating that it is possessed. This particular issue of research is dealt with extensively in Freitas, 2017.

Apurinã nouns can be split into four groups on the basis of how their morphology is affected by possession. There are nouns that never take possession or possessive affixes. Such nouns include proper names (Freitas, 2017, 179–180). The remaining nouns, however, take possessive affixes, on the one hand, and additional marking to indicate whether the word is possessed or not. First, there are nouns, such as kinship terms, that virtually always appear with possessive affixes and no morphology to indicate that they are possessed. These nouns may only be construed as not possessed in some verbal incorporations where the noun is non-specific by nature. A formative *-txi* is present to indicate the noun is not possessed. Other words in this group, including terms for body parts and individual belongings, for example, take the *-txi* formative to indicate the item is not possessed more freely, e.g. *kywy* 'head (possessed)' vs *kywĩtxi* 'head (possessed)' (Freitas, 2017, 163-171; Facundes, 2000, 199-204,228-236). Second, there are noun categories that take the formatives *-ne*, *-te* and *-re1* to indicate the item is possessed, but they, in contrast, have no morphology to indicate that the item is not possessed. Third, there is group of nouns which actually mark both the possessed with the formative *-re2* and the non-possessed with the formative *-ry2*. This alternation is described in Facundes, 2000, and explicitly Freitas, 2017, (112-123) (see Table 1)

The Apurinã treebank solution has been to introduce the **possessed** feature with **Yes** and **No** values. Nouns that cannot be possessed are simply left without the feature Possessed.

	Possessed	Not Possessed	translation
body part	kywy	kywĩ-txi	'head'
person	sytu-re	sytu	'woman'
other	kuta-re2	kuta-ry2	'basket'

Table 1: Marking of possessed feature

2.3 Intransitive descriptive verbs

Apurinã verbs can bear morphology indicating subject and object, be that simultaneously or separately. What is interesting, however, is that a specific subclass of intransitive descriptive verbs attest to the use of object marking to indicate congruence with the subject (Facundes, 2000, 278–283). There are, in fact, certain verbs that distinguish object and subject marking strategies for the same intransitive verbs, such that subject marking indicates a short temporal frame, and object marking indicates permanency (cf. Chagas, 2007; Freitas, 2017, 70–71).

The solution here has been to refer to object-looking morphology with subject congruence as subject marking:

- Gender[subj]=Fem|Masc
- Number[subj]=Plur|Sing
- Person[subj]=1|2|3

To cope, an additional feature value set has been introduced to distinguish verbs of the intransitive descriptive (Vid) nature, and this subset is subsequently split on the on basis of whether the formative entails object-identical *Vido* or subject-identical marking *Vids*.

2.4 Derivations

Fieldwork annotations of certain derivational morphology are minimalistic, and their conversion in the UD treebank calls for more specific representation. Whereas some formatives have been referred to using the same terms, e.g. nominalizer, gerund, we have been obliged to elaborate. Only one feature has been provided for Derivation, Proprietive (*ka-*). The proprietive construction is one of many annotated as **atrib** in the fieldwork materials.

2.5 Lemmatization

The Apurinã language is spoken in 18 indigenous communities of the Purus basin (Lima Padovani et al., 2019). Grammar descriptions from Facundes, 2000 to Freitas, 2017 demonstrate a change in orthographic development, on the one hand, and actual variation in forms of the same words in relation to geographic location, on the other. Materials in the treebank alone show some vacillation with regard to stem-initial *h* and word-internal *e* vs *i*. Since the orthographic standard is still in a developmental state, lemma forms have been chosen on a basis of whether they occur in the manuscript dictionary (Lima-Padovani and Facundes, 2016) or not, and a preference for longer word forms, i.e., *h*-initial stems are forwarded, since it easier to drop a letter in the description than to automatically insert one. Thus the form *hãty* 'one' is given as a lemma instead of its variant *ãty* (as given in the dictionary), and *herãkatxi* (given as a variant) is forwarded as a lemma over both *erãkatxi* and *erẽkatxi* (given in the examples of the alphabet), *arẽkatxi*. The high vowel *i* is preferred over the middle *e* such that *tiwitxi* 'thing' is given as a lemma for the forms *teetxi* and *tiitxi*. Fortunately, work with Apurinã variation is continuing (Lima Padovani et al., 2019), and an updated version of the Apurinã-Portuguese dictionary is forthcoming.

3 Treebanks in figures

There were 76 valid and dependency-annotated sentences in the first release. Broken into figures, these sentences contain 574 tokens and a 454 word count, which can be further broken down into features, parts-of-speech and dependency relations.

The most salient features are *Case* (101), *Gender* (96), *Number* (73), but the newly introduced *Gender[obj]* (47) is also well attested. The *Case* feature owes its prominence to the presence of all nouns not marked for oblique cases, i.e. *Nom*; this leaves a total of 25 obliques (see Table 2).

Feature	№	Feature	№
AdvType=Tim	1	Number[obj]=Plur,Sing	1
Aspect=Prog	1	Number[obj]=Sing	51
Case=Com	4	Number[psor]=Sing	10
Case=Dat	7	Number[subj]=Plur	1
Case=Loc	11	Number[subj]=Sing	7
Case=Nom	76	Person=3	53
Case=Temp	3	Person[obj]=3	52
Derivation=Proprietive	2	Person[psor]=3	8
Gender=Fem	14	Person[subj]=3	8
Gender=Masc	82	Possessed=No	27
Gender[obj]=Masc	47	Possessed=Yes	8
Gender[psor]=Fem	3	PronType=Prs	53
Gender[psor]=Masc	11	VerbForm=Conv	2
Gender[subj]=Masc	8	VerbForm=Vnoun	9
Number=Plur	16	VerbType=Vido	2
Number=Sing	57		

Table 2: Features

The most prominent parts-of-speech the NOUN (170) and VERB (137) classes, followed by PRON (59) and ADV (39), whereas two instances of the same unknown word *pekana* outnumber the ADJ, CCONJ and PROPN, each at one (see Table 3).

PoS	№	PoS	№	PoS	№
ADJ	1	DET	11	PROPN	1
ADP	3	NOUN	170	SCONJ	3
ADV	39	NUM	9	VERB	137
AUX	6	PART	13	X	2
CCONJ	1	PRON	59		

Table 3: Part-of-speech Figures

An important dependency relation (*deprel*) is *nsubj* (83), which is made possible through the extensive use of the *conj* relation. Language-specific *deprels* have extensions such as: *lmod* = locative modifier, *neg* = negation, *poss* = possession, *relcl* = relative clause *tcl* = temporal clause and *tmod* = temporal modifier (see Table 4).

deprel	№	deprel	№	deprel	№
acl	10	mark	3	advmod:lmod	1
advcl	5	nmod	18	advmod:neg	13
advmod	22	nsubj	83	advmod:tmod	13
aux	5	nummod	9	nmod:poss	2
case	3	obj	63	nsubj:cop	2
cc	3	obl	15	obj:agent	1
conj	48	root	76	obl:lmod	19
dep	2	xcomp	1	obl:tmod	4
det	24	acl:relcl	5		
csubj	2	advcl:tcl	2		

Table 4: Dependency relations

4 Future work

Due to the size and orientation of the dataset some features of the Apurinã language have been neglected. It will also be a challenge to apply recent studies in noun incorporation annotation for UD in Tyers and Mishchenkova, 2020 to what Facundes and Freitas, 2015 describe for Apurinã noun and classifier incorporation.

Another obvious goal for further work is to make Apurinã treebank so large that it can be split into train, test and dev portions. The goal to expand the treebank is connected to the availability of resources. Currently the sentences used in the treebank come mainly from the grammatical descriptions. As a language documentation corpus exists[2], an important consideration is whether the treebank sentences could be more closely connected to audio and video recordings as well, and, of course, the main corpora in Belém, as multimodal resources are valuable in language documentation.

[2] https://elar.soas.ac.uk/Collection/MPI1029704

References

Khalid Alnajjar. 2021. When word embeddings become endangered. In Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar, editors, *Multilingual Facilitation*. Rootroo Ltd.

Angela Fabíola Alves. Chagas. 2007. *Aspectos Semântico, Morfológicos e Morfossintáticos das Palavras Descritivas Apurinã*. Belém, Pará. Belém, Pará: Programa de Pós-graduação em Letras – Mestrado em Estudos Linguísticos da Universidade Federal do Pará (Dissertação de Mestrado), 2007.

W. Croft, D. Nordquist, Katherine Looney, and Michael Regan. 2017. Linguistic Typology meets Universal Dependencies. In *TLT*.

Jeff Ens, Mika Hämäläinen, Jack Rueter, and Philippe Pasquier. 2019. Morphosyntactic disambiguation in an endangered language setting. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 345–349.

Sidney da Silva Facundes. 2000. *The Language of the Apurinã People of Brazil (Maipure/Arawak)*. SUNY Buffalo, New York.

Sidney da Silva Facundes, Marília Fernanda Pereira de Freitas, and Bruna Fernanda Soares de Lima-Padovani. 2021. Number expression in apurinã (arawák). In Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar, editors, *Multilingual Facilitation*. Rootroo Ltd.

Sidney da Silva Facundes and Marília Fernanda Pereira de Freitas. 2015. De compostos nominais produtivos a um sistema incipiente de classificação nominal em Apurinã (Aruák). *Revista Moara – Edição 43*, vol. 2 – jul - dez 2015, Estudos Linguísticos:23–50.

Marília Fernanda Pereira de Freitas. 2017. *A Posse em Apurinã: descrição de construções atributivas e predicativas em comparação com outras línguas Aruák*. Universidade Federal Do Pará Programa De Pós-Graduação Em Letras Curso De Doutorado Em Letras – Estudos Linguísticos.

Mika Hämäläinen and Linda Wiechetek. 2020. Morphological disambiguation of South Sámi with FSTs and neural networks. In *Proceedings of the 1st Joint Workshop on Spoken Language Technologies for Under-resourced languages (SLTU) and Collaboration and Computing for Under-Resourced Languages (CCURL)*, pages 36–40.

Mika Hämäläinen. 2021. Endangered languages are not low-resourced! In Mika Hämäläinen, Niko Partanen, and Khalid Alnajjar, editors, *Multilingual Facilitation*. Rootroo Ltd.

Joo-Kyung Kim, Young-Bum Kim, Ruhi Sarikaya, and Eric Fosler-Lussier. 2017. Cross-lingual transfer learning for POS tagging without cross-lingual resources. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2832–2838, Copenhagen, Denmark. Association for Computational Linguistics.

Christo Kirov, Ryan Cotterell, John Sylak-Glassman, Géraldine Walther, Ekaterina Vylomova, Patrick Xia, Manaal Faruqui, Sabrina J Mielke, Arya D McCarthy, Sandra Kübler, et al. 2018. Unimorph 2.0: Universal morphology. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Natalia Levshina. 2017. Communicative efficiency and syntactic predictability: A cross-linguistic study based on the Universal Dependencies corpora. In *Proceedings of the NoDaLiDa 2017 Workshop on Universal Dependencies, 22 May, Gothenburg Sweden*, 135, pages 72–78. Linköping University Electronic Press.

Natalia Levshina. 2019. Token-based typology and word order entropy: A study based on Universal Dependencies. *Linguistic Typology*, 23(3):533–572.

KyungTae Lim, Niko Partanen, and Thierry Poibeau. 2018. Multilingual dependency parsing for low-resource languages: Case studies on North Saami and Komi-Zyrian. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Bruna Fernanda S. de Lima-Padovani and Sidi Facundes. 2016. *Dicionário pedagógico Apurinã-Português*. Amazonas – Manaus. Esta publicação foi produzida com recursos do FNDE em parceria com a UFPA, FOCIMP e CIMI Todos os direitos autorais são reservados às comunidades indígenas Apurinã.

Bruna Fernanda Soares de Lima Padovani, Rayssa Rodrigues da Silva, and Sidney da Silva Facundes. 2019. Levantamentos da variação linguística em três domínios do complexo dialetal Apurinã (ARÚAK). *Entreletras, Araguaína*, page 161–179.

Christopher Moseley, editor. 2010. *Atlas of the World′s Languages in Danger*, 3rd edition. UNESCO Publishing. Online version: http://www.unesco.org/languages-atlas/.

Sjur Moshagen, Jack Rueter, Tommi Pirinen, Trond Trosterud, and Francis M. Tyers. 2014. Open-Source Infrastructures for Collaborative Work on Under-Resourced Languages. The LREC 2014 Workshop "CCURL 2014 - Collaboration and Computing for Under-Resourced Languages in the Linked Open Data Era".

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. 2020. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*.

Jack Rueter. 2018. rueter/erme-ud-moksha: Erme ud moksha v1.0. In *Zenodo*. 10.5281/zenodo.1156112.

Jack Rueter and Niko Partanen. 2019. Survey of Uralic Universal Dependencies development. In *Third Workshop on Universal Dependencies (UDW, SyntaxFest 2019)*, United States. The Association for Computational Linguistics.

Jack Michael Rueter and Francis M. Tyers. 2018. Towards an open-source universal-dependency treebank for Erzya. International Workshop for Computational Linguistics of Uralic Languages, IWCLUL ; Conference date: 08-01-2018 Through 09-01-2018.

Kaius Sinnemäki and Viljami Haakana. 2020. Variation in Universal Dependencies annotation: A token-based typological case study on adpossessive constructions. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 158–167, Barcelona, Spain (Online). Association for Computational Linguistics.

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

Francis Tyers and Karina Mishchenkova. 2020. Dependency annotation of noun incorporation in polysynthetic languages. In *Proceedings of the Fourth Workshop on Universal Dependencies (UDW 2020)*, pages 195–204.

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.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Ahrenberg, et al. 2020. Universal dependencies 2.7. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.