Universal Dependencies for the Amahuaca language

Candy Angulo
University at Buffalo

Pilar Valenzuela
Chapman University

Roberto Zariquiey
Pontificia Universidad Católica del Perú

Abstract

This paper presents the creation of a Universal Dependency (UD) treebank for Amahuaca (Peru), marking the first UD treebank within the Headwaters subbranch of the Panoan family, spoken mostly in Peru and Brazil. While the UD guidelines provided a general framework for our annotations, language-specific decisions were necessary due to the rich morphology of the Amahuaca language. The paper also describes specific constructions to initiate a discussion on several general UD annotation guidelines, particularly those concerning clitics and morpheme-level dependencies.

1 Introduction

This paper describes the methodology employed in the creation of the UD treebank for the language. On the one hand, this treebank aims to enhance the future development of an NLP toolkit for this language as well as contribute to its revitalization. On the other hand, this work aims also to contribute to the discussion on how to integrate polysynthetic languages into the lexically oriented framework of Universal Dependencies (UD). Following Park et al. (2021), we argue that adopting a morpheme-level framework is indispensable due to the morphosyntax of Amahuaca. Specifically, it is crucial to accurately capture the intricate morphological relationships and dependencies within the language, particularly considering the unique characteristics of clitic behavior and their interaction with other morphemes. By focusing on morpheme-level annotations, we aim to provide a clearer understanding of the syntactic structure and the grammatical functions of various elements. This approach facilitates a deeper exploration of the language's complexity, ultimately contributing to more effective natural language processing applications and linguistic analysis.

The structure of the paper is as follows: Section 2 provides a brief overview of some notable features of the Amahuaca language. Section 3 explains the reasons behind our choice of morpheme-level analysis and presents the dependency relations found. Section 4 details the data collection process as well as the composition of the corpus. The following sections present the POS tags and the dependency relations. Section 7 focuses on the comparison between the morpheme-level annotation scheme and the word-level annotation scheme.

2 The Amahuaca language

The Amahuaca people are primarily concentrated in some provinces of the Ucayali region, in Peru. In the Atalaya province, they reside in the basins of the Yurúa River (Yurúa district), Inuya and Mapuya Rivers (Raymondi district), and Sepahua River (Sepahua district). In the Purús province, they occupy a community in the Purús River basin, within the district of the same name. Some settlements in the Upper Inuya and Mapuya regions host Amahuaca populations in "initial contact situations." For more information on Amahuaca society and culture, see Dole (1998) and Hewlett (2014). As mentioned before, this language is endangered, with approximately 400 speakers, most of whom are over 40 years old, and children are no longer learning it.

Amahuaca is a language, characterized by rich morphology. While there are works that describe this language (see Sparing-Chávez 2012, Clem 2019), we base the analysis on Valenzuela et al. (in prep.), which focuses more on the behavior of clitics in the language. Similar to other Panoan languages (for more information about Shipibo-Konibo and Kakataibo languages, see Valenzuela 2003, Zariquiey 2018), this language is characterized by being postpositional and predominantly agglutinative. A notable feature of the language is the absence of deverbal derivation and the use of auxiliaries to convert a noun into a verb; consequently, some nouns may carry verbal inflection markers. We will discuss this point in more detail later.

The language primarily follows a basic constituent order of SOV, but this order is flexible. Constructions like (1) can be found, where the subject *michito chaho* ‘black cat’ precedes the object ‘Paco’, and the verb is at the end carrying the inflectional clitic.

1. Mishito chahonmun Paco ratuuxonu.

mishito chaho=n=mun Paco ratuu=xo=nu

cat black=A=FOC Paco scare=PFV.3=DECL

‘The black cat scared Paco.’

However, sentences with final subjects are found, as shown in (2). The subject *vaku maxko* ‘little baby’ appears at the end, and the verb *oyo* ‘suck’ precedes it. What is interesting about this free word order behavior is that the inflectional morphology is not always attached to the verbal root. Additionally, when S or A is not in the unmarked position, it loses its case marking and takes the form of the copy pronoun. This language is characterized by the presence of doubling pronouns in constructions with transitive verbs.

2. Jaton jaha chochomun oyoni vaku maxkokinu.

jaton jaha=n chocho=mun oyo=niko vaku maxko=ki=nu

3SG.POSS mother=GEN breast=FOC suck=ENDEAR baby=IPFV.2/3=DECL

‘The babies are sucking their mothers' breasts.’

Comparing (1) and (2), it can be observed that the clitic =*ki*, which encodes an aspectual meaning, in the first sentence is attached to the verbal root *ratuu* ‘to scare’, but in the second sentence, it is attached to the noun phrase *vaku maxko* ‘baby’.

3 Morpheme-level annotation scheme

Universal Dependencies (UD) traditionally employs a word-level annotation scheme (Nivre et al. 2017, 2020), which works well for many languages with relatively straightforward morphological structures. Shipibo-Konibo (2018) and Kakataibo, other Panoan languages, have UD treebanks. Consequently, we have based our guidelines for Amahuaca on these resources. However, Amahuaca's rich morphological system and the significant role of clitics require a different approach. After reviewing studies on handling phenomena in polysynthetic languages, such as noun incorporation (Tyers & Mishchenkova, 2020), as well as more general works like Park et al. (2021) and Çöltekin (2016), we decided to follow the direction of morpheme-level annotations proposed in the second paper, as will be explained later.

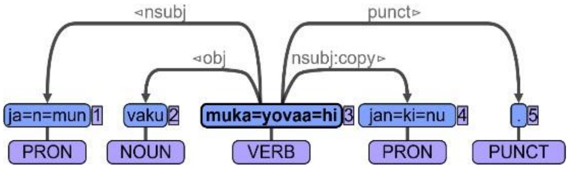
Table 1. Amahuaca bound morphemes behavior.

Morpheme Type	Within a Constituent	At the edge of phrases	Fixed Position	Selective for Host	Without Host
Case markers	NO	YES	YES	YES	NO
=*mun*	possible	usually	usually	NO	NO
Perfective aspect	YES	NO	YES	NO	NO
Degree of remoteness	usually	Possible	NO	NO	NO
Person markers	YES	NO	YES	NO	NO
Declarative markers	NO	YES	usually	NO	NO
=*kiha*	possible	usually	usually	NO	possible
Switch-reference markers	NO	YES	usually	usually	possible

Unlike Shipibo-Konibo, Amahuaca morphemes sometimes do not require an open-class word as a host for their pronunciation. Additionally, clitics can attach to various parts of speech and carry important grammatical information such as tense, aspect, mood, and case. These clitics often do not function as standalone words but as bound morphemes that modify the meaning and function of their host words. Table 1 summarizes the behavior of such bound morphemes.

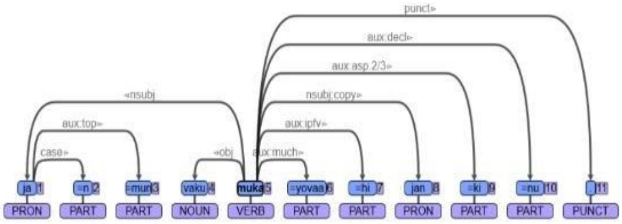
Firstly, case markers are selective for a host and are attached to them. However, the topic marker =*mun* can appear without a host, but it must follow another clitic; if it appears alone or in the first position, it is not allowed. This restriction applies to aspect, tense, and mood markers. But, switch-reference markers, the hearsay marker =*kiha*, as well as degree of remoteness markers, can appear without an open-class word as a host. “Within a constituent” refers to being inside a phrase, which could be a noun or verb phrase. “At the edge of phrases” means at the end of a syntactic constituent. “Fixed position” indicates if the clitic always occupies the same position in relation to the host. For example, case markers always come immediately after the nucleus of the constituent they modify (whether it is just a noun or a noun phrase). “Selective for a host” indicates if it can serve as the base morpheme where a clitic can be attached. Finally, "without host" means that it cannot appear without a host. From our perspective, a word-level annotation would fail to capture the dependency relationships between these clitics and their hosts accurately. Compare the representations of the sentence *Janmun vaku mukayovaahi janhkinu*. ‘He is laughing a lot at his baby’. (3) corresponds to the lexicalist representation, namely Word-level. As observed, only 4 dependency relations are shown. While it captures the grammatical relations of subject and object as a transitive sentence, it overlooks the fact that inflection does not occur entirely on the verb – only the aspectual marker =*hi*- is shown, but it fails to indicate its complement =*ki*, which attaches to the doubling subject *jan.*

3. Word-level representation



(4) shows the dependency relation at a morpheme-level, a total of 11. This analysis adequately captures the fact that the aspectual marker =*ki* attaches to the doubling subject. Even though =*ki* corresponds to a grammatical person, it works together with the aspectual marker =*hi*, because the latter clitic requires to be with a person marker within the same clause. In other words, if there is no =*ki*, the sentence would be agrammatical.

4. Morpheme-level representation



In this section, we presented examples that demonstrate the necessity of morpheme-level annotation. We show how clitics interact with other morphemes and how their roles are more clearly defined in a morpheme-level framework. This approach not only provides a more accurate representation of Amahuaca syntax but also helps in understanding the language's morphological richness. In Section 6, we will explore in greater depth the clitics and their corresponding dependency relations that we have assigned to them.

4 Corpus

The annotated corpus for Amahuaca consists of sentences that were translated from Spanish into Amahuaca. This work is part of a broader initiative to compare treebanks of various Peruvian Amazonian languages, including Amahuaca. Each language was assigned a set of 60 sentences, resulting in a total of 420 sentences to translate. Three native speakers of Amahuaca participated by translating all 420 sentences into their language, after which each translation was reviewed with them. Of these, 202 sentences have been manually annotated for Amahuaca, while the remaining sentences are still awaiting verification. Our corpus contains two sets of the same 202 sentences but annotated from different perspectives. The first one, corresponding to Word-level, has 1028 words, while the second one, corresponding to Morph-level, has 1928. For annotations following the word-level notation paradigm, the process was not done manually. Instead, a Python script was used to automatically attach all "words" in the original text that start with "=" to the preceding word. Finally, manually, it was necessary to double-check the number assigned for each dependency relations.

5 POS Tags

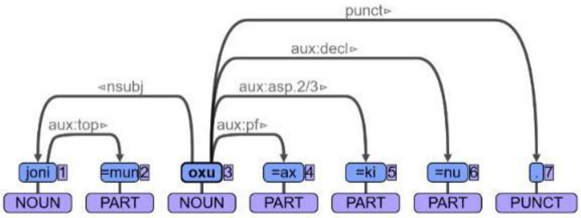
The difference between word-level and morpheme-level POS tags is illustrated in Table 2. We should note that the primary distinction between the two schemes lies in the PART category. This is expected since clitics, which are often overlooked in word-level annotations, have been explicitly labeled as PART in the morpheme- level scheme.

Table 2. POS Frequency

POS	Word-level	Morph-level
NOUN	268	268
ADJ	41	41
PART	-	889
PROPN	31	31
VERB	201	201
PUNCT	210	210
PRON	169	169
DET	70	70
ADV	29	33
NUM	5	5

While the language has clear nominal and verbal bases, it is important to note that there is no deverbal derivation, so nouns may carry "verbal" morphology, as seen in (5), where *oxu* 'moon' has no morphological derivation, but it means 'turn into the moon'. In these cases, we maintain the grammatical category of the base, as it is a property of the language.

5. The man turned into the moon.



6 Dependency Relations

Our annotation scheme utilizes 56 types of dependency relations. Generally, we have adhered to the guidelines provided by UD, except for cases involving clitics. In the morpheme-level scheme, there are a total of 1,927 dependency relations, while for the word- level scheme, there are 1,031.

Table 3. Clitic and its dependency relation label Frequency

Clitic	Dependency relation Label	Frequency
=*mun*	aux:top	183
=*nu*	aux:decl	178
=*n*	case	93
=*ki*	aux:2/3	70
=*xo*	aux:pfv.3	67
=*hi*	aux:ipfv	36
=*ku*	aux:pfv.1/2	22
=*x*	case	17
=*ka*	aux:1	15
=*ra*	aux:int	14

While Universal Dependencies (UD) aims to provide "a universal inventory of categories and guidelines to facilitate consistent annotation of similar constructions across languages" (Nivre et al., 2017), it also accommodates language- specific subtype relation labels when necessary. Following Vásquez et al. (2018), we have chosen to treat clitics as distinct syntactic entities. Consequently, connections between words and clitics are regarded as syntactic and annotated using the appropriate dependency structure. In fact, Amahuaca grammatical elements, specifically clitics, exhibit such a free distribution that they resemble words. We employ the label "aux" for non-nominal clitics, as illustrated in Table 3. Except for =*n* and =*x*, which are clitics for cases, the other more frequent clitics are non-nominal: topic (*aux:top*), declarative (*aux:decl*), verbal persons (*aux:2/3*, *aux:1*), perfective (*aux:pfv.3*), imperfective (*aux:ipfv.2/3*), and interrogative (*aux:int*).

We considered introducing new subtype relation labels corresponding to verbal inflection, mood, and focus clitics. However, to ensure that the label reflects the syntactic meaning of the dependency relation, we decided to use "aux" followed by the gloss corresponding to the clitic. For example, if it is =*nu*, marking declarative mood, the corresponding label would be "aux:decl". Additionally, we found it necessary to include the "nsubj:copy" relation due to the doubling pronouns mentioned earlier in preceding sections (See (4)).

7 Conclusions

This paper presented the results obtained from the manually annotated corpus following both a morpheme-level and a word-level annotation schema for the Amahuaca language. As explained in detail in Section 3, annotating according to a morpheme-level schema is more convenient for Amahuaca, a language with rich morphology characterized by complex morphosyntactic relations among morphemes and interactions with clitics. For instance, in the sentence *Janmun jan ruratixon machitoxon nixohnu*, meaning "He made machetes and axes," where =ni, the temporal clitic, functions as the root of the sentence, this interaction would not be adequately captured in a word-level analysis.

The evaluation of accuracy between these two schemas using UDPipe remains pending, allowing for a comparison of whether there is a significant difference between them. While the morpheme-level annotation may require more linguistic resources, such as a morphological analyzer and morphological segmentation, it provides a deeper insight into the language and has the potential to improve automatic parsing. Ultimately, it is expected that a morpheme-level syntactic dependency annotation may be a more effective way to represent polysynthetic languages within the framework of Universal Dependencies.

References

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.

Cristopher Hewlett. 2014. *History, kinship and comunidad: learning to live together amongst Amahuaca people on the Inuya River in the Peruvian Amazon* (Doctoral dissertation, University of St Andrews).

Çağrı **Ç**öltekin. 2016. (When) do we need inflectional groups? In *Proceedings of The 1st International Conference on Turkic Computational Linguistics*, page (to appear).

Emily Clem. 2019. Amahuaca ergative as agreement with multiple heads. *Natural Language & Linguistic Theory,* 37, 785-823.

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, Barcelona, Spain (Online). Association for Computational Linguistics.

Gertrude Dole. 1998. Los amahuaca. *Guía etnográfica de la Alta Amazonía*, 3, 125-273.

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.

Joakim Nivre, Daniel Zeman, Filip Ginter, and Francis Tyers. 2017. Universal Dependencies. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Tutorial Abstracts*, Valencia, Spain. Association for Computational Linguistics.

Joakim Nivre, Marie-Catherine de Marneffe, Filip Ginter, Jan Hajič, Christopher D. Manning, Sampo Pyysalo, Sebastian Schuster, Francis Tyers, and Daniel Zeman. 2020. Universal Dependencies v2: An Evergrowing Multilingual Treebank Collection. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4034–4043, Marseille, France. European Language Resources Association.

Margarethe Sparing-Chávez. 2012. Aspects of Amahuaca grammar: An endangered language of the Amazon basin. *Dallas: SIL International.*

Pilar Valenzuela. 2003. *Transitivity in shipibo- konibo grammar*. University of Oregon.

Pilar Valenzuela, Roberto Zariquiey and Candy Angulo. In preparation. *A grammar sketch of Amahuaca (Pano, Peru)*

Roberto Zariquiey, Claudia Alvarado, Ximena Echevarría, Luisa Gomez, Rosa Gonzales, Mariana Illescas, Sabina Oporto, Frederic Blum, Arturo Oncevay, and Javier Vera. 2022. Building an Endangered Language Resource in the Classroom: Universal Dependencies for Kakataibo. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3840–3851, Marseille, France. European Language Resources Association.