Towards better annotation practices for symmetrical voice in Universal Dependencies

Andrew Thomas Dyer
Language Science and Technology
Saarland University
Saarbrücken, Germany
andrew.dyer@uni-saarland.de

Colleen Alena O'Brien
Department of Asian Studies
Palacký University Olomouc
Olomouc, Czechia
obrienca@hawaii.edu

Abstract

Austronesian languages exhibit features that are challenging for Universal Dependencies: most notably, the symmetric voice system, whereby agent, patient, and instrumental arguments (among others) can be the pivot of a transitive structure – complicating the usual assumption that subjects of transitive sentences are semantic agents, and objects semantic patients. To showcase our ideas of how to address the representation of such systems in Universal Dependencies, we introduce a small treebank of sentences from texts and elicitation sessions in Gorontalo, an Austronesian language of Sulawesi (Indonesia), which exhibits a Philippine-type voice system. We discuss the annotation guidelines for this language, and the extensions of the Universal Dependencies guidelines that are needed to accommodate this and other Austronesian languages.

1 Introduction

In Universal Dependencies (Zeman et al., 2024), as in linguistics in general, the representation of morphosyntactic alignment of core arguments is heavily influenced by the grammatical traditions of Indo-European languages – particularly Latin. In such languages there is a clear syntactic subject, which corresponds in an unmarked (active voice) sentence to the semantic agent (Andrews, 2007). Such sentences can be passivized, so that the construction becomes intransitive, the semantic patient is the subject, and the agent is demoted to the status of an oblique, or dropped completely.

Austronesian languages, especially Philippine-type languages, challenge this view by allowing arguments with various semantic roles to become the *pivot* of the sentence: the trigger of agreement, as shown by verbal morphology, and the syntactically privileged argument (Blust, 2013; Chen and Mcdonnell, 2019). The pivot of a transitive sentence can be actors, patients, instruments, locations, and many more. In many of these languages, no voice seems to be more privileged or less marked than any other; nor does the actor voice seem derived from the patient voice nor vice versa. Analysis of these languages is problematic, and there is great debate regarding whether to label them as ergative-absolutive (Aldridge, 2012); nominative-accusative (Chen, 2025); or a separative category (Foley, 2008), among other classifications.

The grammatical relation of subject has been noted to be problematic in several languages (Croft, 2003; Haspelmath and Sims, 2010), and in Austronesian languages other terms such as *pivot* or *focus* are often used instead. Nevertheless, the approach in Austronesian language treebanks so far has been to use the existing labels of UD for Austronesian languages – specifically, by representing the pivot argument of a sentence as the subject, regardless of the argument type.

This is a decent compromise, and it allows the Austronesian voice system to fall into the universal framework of UD. However, there are problems with this approach to annotation in the treebanks that exist so far, particularly for voices other than the actor voice.

In this paper we will:

- Compare different approaches in the existing language treebanks.
- Describe some of the potential problems with the current annotation choices for voices and core arguments in Austronesian languages.
- Suggest some improvements that can be consistently applied across languages.
- Showcase some of these decisions in our treebank for a new language: Gorontalo.

2 Related work

Several Austronesian treebanks of varying sizes already exist in Universal Dependencies, and we use them as precedent for our annotation. Table 1

Proceedings of the Eighth Workshop on Universal Dependencies (UDW, SyntaxFest 2025), pages 137–142
August 27, 2025 ©2025 Association for Computational Linguistics

shows the currently existing treebanks.

Aside from the Austronesian treebanks themselves, de Marneffe et al. (2021) discuss the representation of symmetric voice in their description of the Universal Dependencies framework. We are not aware of other studies or opinion pieces on this topic.

Language	Treebanks
Indonesian	PUD, GSD, CSUI
Javanese	CSUI
Cebuano	GJA
Tagalog	TRG, Ugnayan

Table 1: A table of currently existing Austronesian language treebanks in Universal Dependencies.

3 Voice and Core Arguments in Universal Dependencies

3.1 General

Most of the well-resourced languages in Universal Dependencies use two main voices: active and passive. In a passive-marked sentence, the syntactic subject instead corresponds to the semantic patient, while the semantic agent becomes an optional oblique argument.

In annotation of active voice sentences, the agent-subject is given the bare dependency label *nsubj*, and the patient-object *obj* or *iobj*. In passive-voice clauses, the passive patient-subject takes the label *nsubj:pass*, showing that it is the subject of a passive construction. The agent is treated as an oblique, but to denote that the oblique is still an agent, the label *obl:agent* is used.

On the morphological level, the feature `Voice` is applied to the VERB or AUX that marks the voice, with the values *Act* or *Pass*. Since the active voice is considered unmarked in the majority of languages, the *Act* value of the `Voice` feature is often not applied, with only passive voice Pass applied when a verb or auxiliary is inflected for passive voice.

3.2 Austronesian languages

Austronesian languages in Universal Dependencies have been annotated to conform with this system as closely as possible. They usually treat the pivot argument as a subject, whether that be an actor, patient, instrument, or any other kind of argument.

Verbs use the `Voice`[1] feature for their voice inflections. Actor voice and patient voice are aliased to active voice and passive voice (*Act* and *Pass* respectively). Three additional `Voice` values are used only in Austronesian languages: *Ifoc* (instrument focus), *Lfoc* (location focus), and *Bfoc* (beneficiary focus). Table 2 shows the representation of these voices in the current Austronesian languages.

Voice	Full name	Used in
Act	Active voice	ind, jav, ceb, tgl
Pass	Passive voice	ind, jav, ceb, tgl
Ifoc	Instrument focus	ceb, tgl
Lfoc	Location focus	ceb, tgl
Bfoc	Beneficiary focus	tgl

Table 2: The distinct voice values contained in the current UD Austronesian languages: Indonesian (ind), Javanese (jav), Cebuano (ceb), and Tagalog (tgl).

The annotation of the syntactic relations between pivot and non-pivot arguments varies by treebank.

Indonesian and Javanese

Indonesian and Javanese in Universal Dependencies have only two voices attested, corresponding to actor and patient.[2] This makes them unproblematic for Universal Dependencies voice representation: actor voice is aliased with active voice; patient with passive. Active voice sentences have the labels of *nsubj* and *obj* respectively. In the Indonesian GSD (McDonald et al., 2013) treebank, patient-subjects are usually given the dependency label *nsubj:pass*, and the actor-object is given the label *obj*. In Indonesian CSUI (Alfina et al., 2020) and Javanese CSUI (Alfina et al., 2023), the labels *nsubj:pass* and *obl:agent* respectively are used, analogously with the handling of passives in European languages.

Cebuano

The Cebuano GJA treebank (Arañes, 2022) treats the pivot of a transitive sentence as the subject, whichever voice it is expressed in, with the other core arguments being treated as objects. Verbs are annotated with the `Voice` feature, but core argument dependencies are annotated with plain labels. Thus, in a patient voice sentence, an actor-object

[1] `https://universaldependencies.org/u/feat/Voice.html`

[2] The term *Undergoer* is sometimes preferred for the non-agent voice, as it can perform a broad set of semantic roles.

would just have the label of *obj* without any reference to its semantic role. Likewise in instrument or location voice sentences, other core arguments would have object labels, including patients and actors.

Tagalog

The Tagalog TRG treebank (Samson, 2018) has five voices attested. Pivots are treated as the subject (*nsubj*), and non-pivot arguments as objects (*obj*/*iobj*). However, this treebank often (but inconsistently) uses sub-deprels to differentiate between non-agent subject types, sometimes (but not always) using sub-labels such as *nsubj:pass* (patient), *nsubj:ifoc* (instrumental), etc for nominal subjects. Core non-pivot arguments take object labels, with sub-labels such as *obj:agent* in non-actor voice sentences. However, active voice transitive sentences are treated as unmarked, only using the standard labels without sub-labels.

3.3 Our criticisms

Our biggest point of contention is that the use of Indo-European labels/annotations for Austronesian languages muddles the actual syntax of these languages. We find with the current annotation norm of using active and passive as aliases for actor and patient voice problematic, because these voices are not actually equivalent. Non-pivot actors of a patient-voice sentence are *not* demoted to obliques the way they are in passive voice, but remain core arguments of the verb (Chen, 2025), and there are ongoing debates about the transitivity of actor- and patient-voice clauses.

The labels could be misleading to users familiar with the passive in Indo-European languages, who may expect similar properties in what is labelled as passive in Austronesian languages: namely that the agent of a passive sentence is optional and oblique, whereas in reality patient voice sentences still usually expect an actor to be mentioned as a core argument.

Though some treebanks annotate the non-pivot agent as an object with the labels *obj* or *iobj*, retaining its core status, this is still confusing, because semantic agents are not generally thought of as objects of transitive clauses, nor do they necessarily function as grammatical objects in Philippine-type languages, in the same way syntactic objects do in Indo-European languages.

The annotation in the Tagalog TRG treebank is closer to what we would consider ideal, as it also uses sub-labels to explicitly note the semantic role of the syntactic relations: for example, using *obj:agent* to denote an actor-object. It is only selectively applied, though: bare labels are used for actor voice transitives, implicitly considering this an unmarked voice.

Our other point of contention is the `Voice` feature values in Universal Dependencies. The three extra `Voice` values – *Ifoc*, *Lfoc*, and *Bfoc* – are useful, but the naming is confusing. Focus is a common term in Austronesian linguistics to refer to the morphosyntactic pivot of a clause, but in general linguistics it is also understood as a term in information structure, which is not its actual role (Himmelmann, 2002). For the sake of both universality and clarity, we suggest that it would be better to stick to the term Voice for the purpose of clarity/precision.

These problems are acknowledged by de Marneffe et al. (2021) in their outline of Universal Dependencies linguistic theory, but UD continues to reuse the active and passive labels and voice features for convenience. This is understandable, but it still privileges Indo-European terminology above that of a large proportion of the world's languages, which can have implications for usability of these treebanks in e.g., cross-linguistic typological studies.

4 Annotation

We took these issues into account when designing our annotation for Gorontalo. Our hope is to combine the best practices from the existing treebanks and to make it as clear as possible for users of Universal Dependencies.

Annotation was performed using Arborator (Guibon et al., 2020).

4.1 Gorontalo language

Gorontalo is spoken in Gorontalo Province, northern Sulawesi, Indonesia. It is a member of the Gorontalo-Mongondow branch of the Greater Central Philippine subgroup of Malayo-Polynesian (Usup, 1986). Its word order of *pivot verb non-pivot* is typical of Indonesian languages, but it has a Philippine-type voice system, with at least three attested voices.

Gorontalo has complex verbal morphology (mainly prefixes and infixes) to indicate voice, aspect, mood, intentionality, and more. There are gender-based case markers for proper nouns in all voices (Author et al., To appear), and for non-pivot core arguments in patient voice. The language also has oblique markers, clitics for first and second person, and some derivational morphology.

4.2 The data

At present, the sentences contained within the treebank are elicited sentences from sessions with a native Gorontalo speaker. These sentences were specifically selected to demonstrate the voice and alignment system of Gorontalo, and how this system would be parsed under Universal Dependencies.

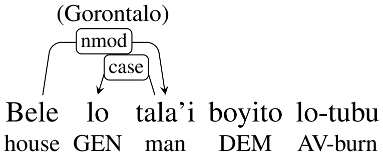
Future additions to the treebank will come from fieldwork data. The data consists of about nine hours of conversations between two or three Gorontalo speakers at a time, with 25 total speakers, in which one person asks the other person questions about their life, local knowledge (regarding topics such as fishing or farming), and traditional beliefs (such as taboos or communication with spirits).

4.3 Annotation decisions

4.3.1 Parts of speech and morphology

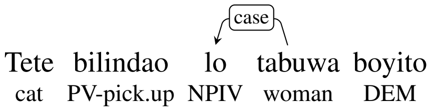
As a Philippine-type language, Gorontalo shares many features with Cebuano and Tagalog, and we have matched much of our morphological annotation with that of the Tagalog TRG corpus in particular.

Pivot and non-pivot markers are adpositions (*ADP*) annotated with the Case feature, with values *Nom* (nominative) and *Gen* (genitive) respectively. This follows the precedent in the Tagalog corpus. While the use of *Nom* is motivated by analogy with subjects, *Gen* is more unusual, and deserves some explanation here. In many Philippine-type languages, including Gorontalo, the non-pivot marker doubles as a genitive marker in genitive phrases. Figs. 1 and 2 show examples of this dual usage in Gorontalo.



"The house of the man burned down." (Actor voice)

Figure 1: In this example, the non-pivot marker *lo* is functioning as a genitive marker in a nominal modification.



"The woman picked up the cat." (Patient voice)

Figure 2: In this example, the same marker *lo* is instead functioning as a non-pivot marker.

In the case of a proper name, pivot/non-pivot markers may also encode Gender as *Masc* or *Fem*.

Pronouns may also be inflected for pivot or non-pivot status, expressed here with Case as *Nom* and *Gen*.

We introduce new values for the Voice feature on verbs. Gorontalo has actor-, patient- and instrument-voices[3]. The first of these is unproblematically aliased with *Act* (active), there being reasonable enough overlap between the two concepts. On the other hand, we replace *Pass* (passive) with *Pat* (patient): a value that would be new to UD and would make clear this is patient voice, not passive voice. Finally, we make a single-letter change to the already existing *Ifoc* (instrument focus) to make *Ivoc* (instrument voice), eliminating the ambiguity with information structural terms.

4.3.2 Syntax

As with parts of speech and morphology, we base our syntactic annotation of core arguments on precedent in Tagalog TRG, annotating the pivot NP as a subject and all others as variations of object. We also append sub-labels to specify types of argument. However, we do so for *all* arguments in transitive sentences, clarifying the semantic roles of both subjects and objects in each instance.

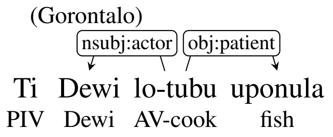
The sub-labels we use are also modified. We replace the sub-label *:pass* (passive) with *:patient*, once again making it clear that this is not to be confused with a passive construction. Where in TRG they replicate voice labels such as *:ifoc*, we spell out the word *instrument* for the instrument semantic role expressed by instrumental voice.

Our motivation for doing this is two-fold. Firstly, we make it easier for any user searching the treebank to query specific syntactic and semantic role combinations, such as *nsubj:instrument* or *obj:patient*. Secondly, we avoid privileging any one voice over another as the unmarked type.

[3] And possibly location voice, though this is less certain.

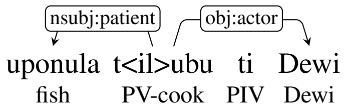
4.4 Annotation examples

We show some examples of annotation in Figs. 3 - 5.



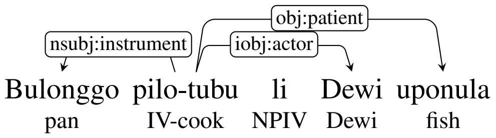
"Dewi cooked the fish." (Actor voice)

Figure 3: A sentence in actor voice in Gorontalo.



"Dewi cooked the fish." (Patient voice)

Figure 4: A sentence in patient voice in Gorontalo. Note infixing of the voice-inflecting morpheme within the verb *tubu* (cook).


obj:patient
nsubj:instrument
iobj:actor
Bulonggo pilo-tubu li Dewi uponula
pan IV-cook NPIV Dewi fish

"Dewi cooked the fish in the pan." (Instrument voice)

Figure 5: A sentence in instrument voice in Gorontalo.

5 Discussion

We have tried to be conservative with the annotations we make, keeping with the conventions of preceding treebanks and restricting the number of additions we need to make to the UD features and relations. However, we could go further in adding more features and labels specifically for Austronesian languages.

For example, we could add *Piv* (pivot) and *Npiv* (non-pivot) respectively as `Case` values for the pivot markers. This would remove the oddness of using *Gen* to describe the role of non-pivot core arguments, while also showing that non-pivots can perform multiple roles. Alternatively, we could keep the *Gen* value for clear instances of nominal modification as in Fig. 1.

6 Conclusion

We have presented a summary of the current state of symmetrical voice and alignment annotation for Austronesian languages, and our criticisms and suggestions for improvements.

In short, we have four main suggestions:

- Replace *Pass* with *Pat* in the `Voice` feature.
- Replace *_foc* with `_voc` in the `Voice` feature.
- Replace the dependency sub-label *pass* with *patient*.
- Obligatorily use semantic role sub-labels for subject and object arguments in transitive sentences.

We make available our preliminary work on a treebank for Gorontalo that demonstrates the changes we would like to see in annotation for Austronesian languages in general[4].

Our next steps are to continue to annotate more data for eventual release in Universal Dependencies. As more data is annotated, we will compare and contrast approaches to more Austronesian linguistic phenomena in UD.

We hope that these proposals will be discussed by the Universal Dependencies community and any contributors of new Austronesian languages, so that Austronesian languages can be included in a way that properly describes them and de-centres Indo-European terminology.

Acknowledgments

Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) – Project-ID 232722074 – SFB 1102.

This work received support from the CA21167 COST action UniDive, funded by COST (European Cooperation in Science and Technology).

This work was supported by the Fulbright Commission and National Endowment of the Humanities (USA).

All elicitations are from sessions with Novi Usu, a native speaker of Gorontalo, and a lecturer at Gorontalo State University. who is also collaborating on the treebank.

This collection was made possible by the cooperation of the Gorontalo community of Gorontalo Province, for whom this language is their cultural heritage.

[4] `https://github.com/andidyer/UTS_Gorontalo_Sanggala/tree/main`

Ethics

Texts were collected as part of a Fulbright-funded project and were transcribed by Gorontalo speakers studying at the local university. All transcribing assistants were paid per hour of transcription with their hourly rate based on teacher salaries in Gorontalo Province. All elicitations are from sessions with a native speaker of Gorontalo who is also collaborating on the treebank.

Limitations

The Gorontalo language is still being documented and described, and there remain some phenomena in the language that lack analysis, some of which appear in this treebank. As the language undergoes more description, some of the annotation will likely change.

References

Edith Aldridge. 2012. Antipassive and ergativity in tagalog. *Lingua*.

Ika Alfina, Indra Budi, and Heru Suhartanto. 2020. Tree rotations for dependency trees: Converting the head-directionality of noun phrases. *Journal of Computer Science*, 16(11):1585–1597.

Ika Alfina, Arlisa Yuliawati, Dipta Tanaya, Arawinda Dinakaramani, and Daniel Zeman. 2023. A Gold Standard Dataset for Javanese Tokenization, POS Tagging, Morphological Feature Tagging, and Dependency Parsing.

Avery D. Andrews. 2007. *The major functions of the noun phrase*, page 132–223. Cambridge University Press.

Glyd Jun Arañes. 2022. The gja cebuano treebank: Creating a cebuano universal dependencies treebank. Master's thesis, University of Eastern Finland.

Robert Blust. 2013. *The Austronesian languages*. Asia-Pacific Linguistics, School of Culture, History and Language, College of Asia and the Pacific, The Australian National University.

Victoria Chen. 2025. The syntax of philippine-type alignment: Insights from case-marking. *Nat Lang Linguist Theory*.

Victoria Chen and Bradley Mcdonnell. 2019. Western Austronesian Voice. *Annual Review of Linguistics*.

William Croft. 2003. *Typology and Universals*. Cambridge University Press.

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

William Foley. 2008. The place of philippine languages in a typology of voice systems. *Voice and grammatical relations in Austronesian languages*.

Gaël Guibon, Marine Courtin, Kim Gerdes, and Bruno Guillaume. 2020. When collaborative treebank curation meets graph grammars. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5291–5300, Marseille, France. European Language Resources Association.

Martin Haspelmath and Andrea Sims. 2010. *Understanding Morphology (2nd ed.)*. Routledge.

Nikolaus P. Himmelmann. 2002. Voice in western austronesian: an update. *The history and typology of western Austronesian voice systems*.

Ryan McDonald, Joakim Nivre, Yvonne Quirmbach-Brundage, Yoav Goldberg, Dipanjan Das, Kuzman Ganchev, Keith Hall, Slav Petrov, Hao Zhang, Oscar Täckström, Claudia Bedini, Núria Bertomeu Castelló, and Jungmee Lee. 2013. Universal Dependency annotation for multilingual parsing. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 92–97, Sofia, Bulgaria. Association for Computational Linguistics.

Stephanie Dawn Samson. 2018. A treebank prototype of tagalog. Master's thesis, University of Tübingen.

Hunggu Tadjuddin Usup. 1986. *Rekonstruksi Proto-Bahasa Gorontalo-Mongondow [Proto-Gorontalo-Mongondow Language Reconstruction]*. Ph.D. thesis, Universitas Indonesia.

Daniel Zeman, Joakim Nivre, Mitchell Abrams, Elia Ackermann, Noëmi Aepli, Hamid Aghaei, Željko Agić, Amir Ahmadi, Lars Ahrenberg, Chika Kennedy Ajede, Arofat Akhundjanova, Furkan Akkurt, Gabrielė Aleksandravičiūtė, Ika Alfina, Avner Algom, Khalid Alnajjar, Chiara Alzetta, Erik Andersen, Matthew Andrews, and 633 others. 2024. Universal dependencies 2.15. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.