# On the representation of Austronesian voice systems in Universal Dependencies

Universal Dependencies (Zeman et al., 2024) is a multilingual, cross-lingually consistently annotated corpus of dependency parsed sentences in many languages, assembled with the aim of enabling, among other things, corpus-based comparative and typological study at scale. A central desideratum of the UD annotation scheme is to be universal and consistent between languages, so that like-for-like comparisons can be made.

Despite constituting a large portion of the world's documented languages, inclusion of Austronesian languages in Universal Dependencies is currently quite limited, with, as of version 2.15, only Indonesian, Cebuano, Javanese, and Tagalog represented — the former with a large training, development and test set; the latter with only a small test set.

Austronesian languages that exhibit symmetrical-voice systems have long been problematic for various syntactic theories because they do not seem to privilege any voice in a way that is characteristic either of accusative- or of ergative-aligned languages; that is to say, all voices are transitive with no demotion of argument (see Chen & McDonnell 2019). In some symmetrical-voice languages, an agent or patient argument can serve as a so-called "pivot" and thus trigger agreement, which is marked by verbal morphology; in other types, there can be even more voices, involving other kinds of arguments as pivots (e.g., location and instrument arguments). Normally, in the literature, the two-voice systems are characteristic of the so-called "Indonesian type", and the four-voice systems are characteristic of the so-called "Philippine type"; some languages, such as Gorontalo (Sulawesi, Indonesia) can be found in the middle of the spectrum.

Representing all of these voice systems in Universal Dependencies presents a challenge. At present, the UD_Tagalog-TRG corpus is the only one to represent this voice system There is a need for expansion of UD to more Austronesian languages and for a survey of the kind of voice systems that we can expect to encounter when annotating these languages.

In this talk, we will make the following contributions: (1) A test-sized corpus of parsed Gorontalo sentences, comprising both elicited and naturally occurring examples, and showcasing the rich voice system in this language; (2) A preliminary survey of voice systems, including annotated examples from other Austronesian languages, showcasing the difference in parses and feature annotations between the languages; (3) Presentation of a syntactic relation and feature set that applies as widely and parsimoniously as possible to Austronesian languages; and (4) Recommendations for how to annotate Austronesian languages. We hope to lay the groundwork for future annotation efforts for Austronesian languages in UD and cross-linguistically consistent corpora for quantitative typological study of these languages.

**References**

Chen, Victoria and Bradley McDonnell. 2019. Western Austronesian Voice. Annual Review of Linguistics 5(1): 173–195.

Zeman, Daniel; et al., 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, http://hdl.handle.net/11234/1-5787.