What Can Be Learned from Sparse Voice Distributions? A Bayesian Data Analysis of Verb-Voice Associations in Tao/Yami, a Philippine-Type Language

Kaiying Kevin Lin[1], Patrick O'Neil[1] and Joshua Hartshorne[2]

[1]Institute of Linguistics, Academia Sinica (Taiwan); [2] Communication Sciences & Disorders, MGH Institute of Health Professions

INTRODUCTION

- Yami is a Philippine-type Austronesian language with a rich voice-marking system (Agent, Patient, Locative, Instrumental).

1)

a. Actor Voice (subject is Salang) (Huang, 2014)

k-om-an	so	wakay	si	Salang.
<AV>eat	OBL	sweet-potato	NOM	Salang.

"Salang wants to eat a sweet potato."

b. Patient Voice (subject is the sweet potato) (Huang, 2014)

kan-en	na	ni	Salang	o	wakay.
eat-PV	3.S.GEN	GEN	Salang	NOM	sweet-potato

"The sweet potato was eaten by Salang."

c. Locative Voice (subject is the seashore) (Her & Deng, 2012)

ya	ko	pi-akan-an	so	among	o	pasalan	ya.
Aux	I.GEN	<LV>eat	OBL	fish	NOM	shore	Aux

'This seashore is where I eat fish.'

d. Instrumental Voice (subject is the knife) (Her & Deng, 2012)

ya	ko	ya-kan	so	among	o	ipangan	ya.
Aux	I.GEN	IV-eat	OBL	fish	NOM	knife	Aux

'I eat fish with the knife.'

- Understanding verb–voice preferences helps reveal how grammar encodes semantic roles.
- Goal: Model voice preferences statistically and link them to verb semantics.

Previous Studies

- Reference grammar (Rau & Dong, 2017) describes distributional tendencies of voices in Yami.

Voice	Verb types
Agent voice	Stative verbs, verbs with only one nominal (intransitive verbs), position verbs, and verbs which highlights agents (e.g., drink, eat)
Patient voice	Verbs that require definite direct objects, verbs with telic events
Locative voice	Verbs that highlight sources, goals and recipients, reasons, themes of perception and cognitive verbs
Instrument voice	Verbs that involve benefactives or instruments in events

- Huang (2017): Applied Foley's (2005) extended macro-role hierarchy to Yami.
  - Actor role assigned top-down in the hierarchy (volitional performer → movement → stationary → causally affected).
  - Undergoer role assigned bottom-up (change-of-state/state → causally affected).
  - Some AV affixes appear with undergoer subjects → AV also at the bottom of hierarchy.

2)

(simplified version from Huang, 2017)

Actor	
volitional performer	AV affixes
causing an event or change-of-state sentience	AV affixes
movement	AV affixes
stationary	AV affixes
causally affected	IV, LV affixes
undergoing a change in state or being a state	AV, PV affixes
Undergoer	

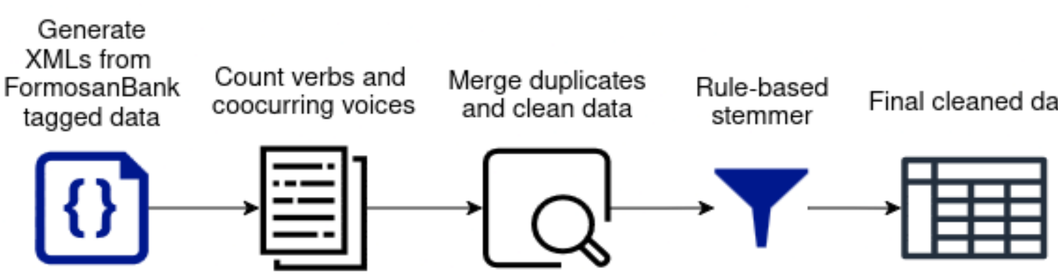
- In Philippine-type languages, Patient Voice (PV) is often reported as dominant (e.g., Garcia & Kidd, 2020).
- No prior quantitative modeling of Yami voice preferences using Bayesian hierarchical methods.

Selected Reference

- Hartshorne, J. K., Le Ferrand, É., Sung, L.-M., & Prud'hommeaux, E. (2024). FormosanBank and why you should use it. Poster presented at Architectures and Mechanisms in Language Processing (AMLaP), Rotterdam.in our data challenges previous PV-dominance claims for Philippine-type languages.
- Pinker, S. (1989). Learnability and Cognition: The Acquisition of Argument Structure.Approach combines probabilistic modeling and semantic analysis, adaptable to other languages.
- Huang, Wan-Tin(2017)。Functions of Yami Verbal Affixes [論雅美語動詞詞綴的功能]。PhD Dissertation. National Cheng-Chi University[國立政治大學]. https://hdl.handle.net/11296/ctjyhn。
- Huang, S. W. T. (2014). Tao Inflection or Both? Argument realisations and related constructions in Austronesian languages: papers from 12-ICAL, Volume 2. eds. I.W. Arka and N.L.K.M. Indrawati, pp 175-195. Asia-Pacific Linguistics.
- Rau, D. V., & Dong, M. N. (2017). Introduction to Tao grammar [達悟語語法概論]. Council of indigenous people [原住民族委員會].
- Her, O. S., Deng, D. H., Butt, M., & King, T. H. (2012). Lexical mapping in Yami verbs. LFG12, Bali.

Methods & Materials

- Dataset: Voice-tagged verbs from Yami corpus in Formosan Bank. (Hartshorne et al., 2024)
- Data preprocessing and lemmatization:



- Two methods:

1. Bayesian hierarchical model:
   - Assumes verbs may favor some voices, but not necessarily.
   - The model learns whether there exist preferences from data.
   - Hierarchical: verbs share an overall distributional pattern.
   - Rare verbs "borrow strength" from frequent ones.
   - Output: probabilities of each verb appearing in each voice.
2. K-means clustering applied to posterior means of voice probabilities.

- Goal: Identify semantic themes in clusters.

Results

- Low concentration parameter (mean ≈ 0.206) → verbs tend to appear ~75% of the time in a single voice. (low c values).
  - AV more common than PV in our dataset (contrast with previous claims).
- More concentrated than Zipfian (C. Yang 2013) — Yami verbs show stronger single-voice dominance than expected from general language/ morphology patterns

Rank	Modeled frequency distribution	Zipfian frequency distribution
1	0.7505	0.48
2	0.1945	0.24
3	0.0476	0.16
4	0.0073	0.12

- K-means (k=4) yielded clusters:
  - Four dominated by a single voice.
- Weak semantic correlations with voice preference, and exceptions exist:
  - AV clusters → intransitives, motion verbs.
  - PV clusters → transitive, object-manipulation verbs.
  - LV cluster → location-related and some cognitive verbs.
  - IV cluster → Psychological verbs in the past, verbs involving transfer

cluster	Dominant Voice	# verbs	Semantic Theme	Examples	% AV	% PV	% IV	% LV
1	AV	37	Motion verbs, intransitive meanings	go, be at, return, say, run, later on	78	7	7	7
2	PV	25	Transitive meanings	take, eat, drag, look, find	19	66	6	9
3	IV	9	psychological verbs in the past events, verbs involving transfer	call, give, angry, hurt, say, worry	7	8	75	11
4	LV	8	perception verbs, verbs involving locations	know, unsatisfied, experience, end, fight, enter	19	9	5	67

Conclusion

- Bayesian modeling quantifies voice preferences and confirms some voice preferences of verbs
- AV prominence in our data challenges previous PV-dominance claims for Philippine-type languages.
- Voice–semantics link is weakly supported: verbs with similar meanings seem to share voice preferences, but exceptions exist.
- Children might be able to employ Semantic bootstrapping (Pinker, 1989) to acquire voice preferences/constraints.
- Future research requires larger corpora to confirm if the correlation between verbal semantics and voice preferences applies to more verbs in general.