Appendix for *Beyond Hill and Valley: A Sociogeographic Analysis of Argument Coding Complexity in the Eastern Himalayan Region*

Appendix A

This appendix describes the sample, data collection and coding procedure in further detail.

Languages Sampled

Table 1 below details the sample. The sample was initially collected proportionally to size of subgroup, but after excluding languages for lack of adequate data sources, languages were included by convenience. However, this still results in a fairly genealogically diverse sample, as can be seen by the variety in subgroup and node below subgroup, given in the table below. The sample numbers to 32 languages. Data for K'man (glottocode:gema1234) and Asakian (glottocode:chak1270) were collected, but excluded for lack of adequate linguistic and ethnographic information to sufficiently code all variables with confidence.

Table 1: List of references for linguistic information

Language	Glottocode	Subgroup (node below top-level)	Node below subgroup	Citation
Rabha	rabh1238	Kochic	Bodo-Garo	Joseph 2007
Bodo-Mech	bodo1269	Bodo-Mech-Kachari	Bodo-Garo	Brahma 2015, Haokip 2018
Garo	garo1247	Bodo-Garo	Bodo-Garo	Burling 2003
Tiwa	tiwa1253	Sal (Brahmaputran)	Bodo-Garo	Dawson 2020a, 2020b
Turung	sing1264	Sal (Brahmaputran)	Jingpho-Asakian	Morey, 2010
Khiamniungan	khia1236	Sal (Brahmaputran)	Patkaian	Van Dam and Tham, p.c.
Hakhun Tangsa	noct1238	Sal (Brahmaputran)	Patkaian	Boro 2017
Hills Karbi	karb1241	Kuki-Chin-Naga	Karbic	Konnerth, 2014
Manipuri	mani1292	Kuki-Chin-Naga	Manipuri	Chelliah, 1997
Ao Naga	mong1332	Kuki-Chin-Naga	Angami-Ao	Coupe, 2007
Angami	anga1288	Kuki-Chin-Naga	Angami-Ao	Kevichusa-Ezung, 2014
Suansu	suan1234	Kuki-Chin-Naga	Tangkhul-Maring	Ivani, 2023
Zeme	zeme1240	Kuki-Chin-Naga	Zemeic	Chanu, 2017
Mizo	lush1249	Kuki-Chin-Naga	Kuki-Chin	Chhangte, 1993
Thado Chin	thad1238	Kuki-Chin-Naga	Kuki-Chin	Haokip, 2014
Lamkang	lamk1238	Kuki-Chin-Naga	Kuki-Chin	Thounaojam and Chelliah, 2007

Daai Chin	daai1236	Kuki-Chin-Naga	Kuki-Chin	So-Hartmann, 2009
Bori-Karko	bori1243	Macro-Tani	Eastern Tani	Lalrempuii, 2005
Milang	mila1245	Macro-Tani	Koro-Holon	Modi, 2017
Mising	miny1240	Macro-Tani	Mising-Padam-Miri-Minyong	Prasad, 1991
Galo	galo1242	Macro-Tani	Subansiri	Post, 2007
Kera'a	idum1241	Digarish	Digarish	Peck and Reinöhl, p.c.
Hrusso Aka	hrus1242	Hruso	Unclear	D'Souza, 2021
Tibetan	tibe1272	Bodic	Central Tibetan	DeLancey, 2003
Brokpa/Brokpake	brok1248	Bodic	Southern Tibetic	Wangdi, 2021
Bjokapakha	bjok1234	Bodic	Tshanglic	Grollmann, 2020
Tamangic	tama1367	Bodic	Ghale-Tamangic	Owen-Smith, 2015
Limbu	limb1266	Himalayish	Kiranti	van Driem, 1987
Lepcha	lepc1244	Himalayish	Himalayish	Plaisier, 2007
Puroik	puro1234	Kho-Bwa	Kho-Bwa	Lierberherr, 2017
Miji	miji1239	Miji	Miji	Weedall, 2021
Trung/Drung	drun1238	Nungish	Gunong	Perlin, 2020

The following table links ethnographic and other literature used to code sociocultural variables to the languages. Sometimes, linguistic sources (such as grammars) were re-used if they were the best available source on sociocultural information. Oftentimes, linguistic sources were used in conjunction with other sources to fully code sociocultural variables.

Table 2: List of references for sociocultural, agricultural and geographic information.

Language	**Glottocode**	**Subgroup (node below top-level)**	**Node below subgroup**	**Citation**
Rabha	rabh1238	Kochic	Bodo-Garo	Goswamy et al, 2024
Bodo-Mech	bodo1269	Bodo-Mech-Kachari	Bodo-Garo	Boro, 2024
Garo	garo1247	Bodo-Garo	Bodo-Garo	Burling, 1963
Tiwa	tiwa1253	Sal (Brahmaputran)	Bodo-Garo	Doloi et al, 2024
Turung	sing1264	Sal (Brahmaputran)	Jingpho-Asakian	Morey, 2010
Khiamniungan	khia1236	Sal (Brahmaputran)	Patkaian	Van Dam and Tham, p.c.
Hakhun Tangsa	noct1238	Sal (Brahmaputran)	Patkaian	Boro 2017
Hills Karbi	karb1241	Kuki-Chin-Naga	Karbic	Hansepi and Laisram, 2022
Manipuri	mani1292	Kuki-Chin-Naga	Manipuri	Meitei and Sharma, 2023
Ao Naga	mong1332	Kuki-Chin-Naga	Angami-Ao	Mills, 1926
Angami	anga1288	Kuki-Chin-Naga	Angami-Ao	Hutton, 1921
Suansu	suan1234	Kuki-Chin-Naga	Tangkhul-Maring	Singh and Choudhary, 2015
Zeme	zeme1240	Kuki-Chin-Naga	Zemeic	Longkumer, 2007
Mizo	lush1249	Kuki-Chin-Naga	Kuki-Chin	Chhangte, 1993
Thado Chin	thad1238	Kuki-Chin-Naga	Kuki-Chin	Lalthansangsa, 2018
Lamkang	lamk1238	Kuki-Chin-Naga	Kuki-Chin	Thounaojam and Chelliah, 2007

Daai Chin	daai1236	Kuki-Chin-Naga	Kuki-Chin	Chan and Takeda, 2016
Bori-Karko	bori1243	Macro-Tani	Eastern Tani	Raj, 2010
Milang	mila1245	Macro-Tani	Koro-Holon	Modi, 2017
Mising	miny1240	Macro-Tani	Mising-Padam-Miri-Minyong	Chutia, 2020
Galo	galo1242	Macro-Tani	Subansiri	Post, 2007
Kera'a	idum1241	Digarish	Digarish	Peck and Reinöhl, p.c.
Hrusso Aka	hrus1242	Hruso	Unclear	Nimachow and Dai, 2011
Tibetan	tibe1272	Bodic	Central Tibetan	DeLancey, 2003
Brokpa/Brokpake	brok1248	Bodic	Southern Tibetic	Wangdi, 2021
Bjokapakha	bjok1234	Bodic	Tshanglic	Akamatsu, 2015
Tamangic	tama1367	Bodic	Ghale-Tamangic	Owen-Smith, 2015
Limbu	limb1266	Himalayish	Kiranti	Limbu, 2013
Lepcha	lepc1244	Himalayish	Himalayish	Thapa and Allay, 2021
Puroik	puro1234	Kho-Bwa	Kho-Bwa	Ramjuk, 2018
Miji	miji1239	Miji	Miji	Weedall, 2021
Trung/Drung	drun1238	Nungish	Gunong	Perlin, 2020

Full bibliographic references for all data sources can be found at the end of this document.
For languages where the citation says ‘p.c.’, the researcher reached out to academics known to be working on the language. They were asked to fill out a questionnaire, given below. The questionnaire also reflects my own methodology when coding information from a grammar or ethnographic data source.

Questionnaire

Lang: (glottocode), filled out by: (name)

1. Do markers on the noun distinguish core grammatical relations? (Are there case markers?) **If no, ignore Q2 -Q8**
2. Is the only argument of a one-place (intransitive) predicate marked? If so, by what forms? (affixes or adpositions)
3. Is an agentive argument of a one-place (intransitive) predicate marked? If so, by what forms? (affixes or adpositions)
4. Is a patientive argument of a one-place (intransitive) predicate marked? If so, by what forms? (affixes or adpositions)
5. Is the agentive argument of a two-place (transitive) predicate marked? If so, by what forms? (affixes or adpositions)
6. Is the patientive argument of a two-place (transitive) predicate marked? If so, by what forms? (affixes or adpositions)
7. Is the marker for the agentive argument of a transitive predicate optional? If so, under what conditions? (if known)
8. Is the marker for the patientive argument of a transitive predicate optional? If so, under what conditions? (if known)
9. Are there verbal indexes for core grammatical relations in affirmative declarative clauses? If yes, which macro-roles (of *intransitive subject, agent,* and *undergoer*) are indexed? If yes and different in different TAME categories, answer for category with most distinctions. **If no, ignore Q10-Q18.**

10. Is the only argument of a one-place (intransitive) predicate indexed on the verb? If so, what persons and numbers can be indexed on the verb?
11. Is the agentive argument of a two-place (transitive) predicate indexed on the verb? If so, what persons and numbers?
12. Is the patientive argument of a two-place (transitive) predicate indexed on the verb? If so, what persons and numbers?
13. Are the indexes for intransitive argument the same as the ones for the agentive transitive argument?
14. Are the indexes for intransitive argument the same as the ones for the patientive transitive argument?
15. What is the maximal paradigm of agent and undergoer combinations for the transitive clause that cause some morphological marking on the verb?
16. Is there a hierarchy that determines the expression of verbal indexes?
17. Is there an inverse marker on the verb?
18. Are there any TAME requirements for the verbal indexes?
19. What is the pragmatically unmarked constituent order in the transitive clause?
20. Are other constituent orders in the transitive clause possible?
21. Any outstanding comments?

Some geographic/cultural questions, if possible:

1. What is the largest settlement where this language is primarily spoken?
2. How many speakers are there?
3. Is the language spoken as an L2 by any other speech communities?
4. What is the type of agriculture traditionally practiced? (intensive/irrigated, extensive/shifting/slash-and-burn/*jhum*, horticulture/gathering, or casual) (based on https://d-place.org/parameters/EA028#1/30/152)
5. What are/were the jurisdictional levels beyond the local community? (None/autonomous bands or villages, petty chiefdom, larger chiefdom, state, complex state) (based on https://d-place.org/parameters/EA033#1/30/153)
6. Any outstanding comments?

Appendix B:Statistical Analyses

1. Introduction

All code and results for the statistical analyses can be found in the R notebook *synthesis.rmd*, at https://github.com/patrickdas/ehr-argument-coding

Table 3 below details the analyses conducted in this study. Some of these analyses which were not fully detailed in the paper for reasons of space are given in full below.

Table 3: Details of all statistical models used in this study.

Model type	Response variable	Predictor variable	Random effect	N
Logistic regression	Hill/Valley	Elevation	None	31
Logistic regression	Hill/Valley	Terrain roughness	None	31
Logistic regression	Hill/Valley	Elevation, Terrain roughness	None	31
Mixed effects	Complexity	Hill/Valley	Subfamily	31

Mixed effects	Complexity	Agricultural intensity	Subfamily	31
Mixed effects	Complexity	L2 Status	Subfamily	31
Mixed effects	Complexity	Population Size	Subfamily	31
Mixed effects	Complexity	Political Organization	Subfamily	31

2. Ecological Analyses

Three logistic regression models were fitted to test whether ecological variables predict Hill/Valley societal classification, using the sensitivity dataset (N = 31, excluding Limbu). The dependent variable was a binary classification where Hill = 1 and Valley/Split = 0.

Table 4: Model formula: hill_binary ~ scaled_altitude + scaled_stdev_slope, family = binomial. N = 31 languages

Parameter	Estimate	Std. Error	z-value	p-value
Intercept	0.471	0.372	1.266	0.206
Elevation (scaled)	0.268	0.445	0.603	0.546
Terrain roughness (scaled)	-0.216	0.432	-0.501	0.617

Model fit: AIC = 46.95, Null deviance = 41.38 (df = 30), Residual deviance = 40.95 (df = 28), R^2 = 0.018, N = 31

Table 5: Model formula: hill_binary ~ scaled_altitude, family = binomial. N = 31 languages

Parameter	Estimate	Std. Error	z-value	p-value
Intercept	0.466	0.371	1.258	0.208
Elevation (scaled)	0.159	0.385	0.411	0.681

Model fit: AIC = 45.21, Null deviance = 41.38 (df = 30), Residual deviance = 41.21 (df = 29), R^2 = 0.008, N = 31

Table 6: Model formula: hill_binary ~ scaled_stdev_slope, family = binomial. N = 31 languages

Parameter	Estimate	Std. Error	z-value	p-value
Intercept	0.460	0.369	1.247	0.213
Terrain roughness (scaled)	-0.085	0.372	-0.228	0.820

Model fit: AIC = 45.33, Null deviance = 41.38 (df = 30), Residual deviance = 41.33 (df = 29), R^2 = 0.002, N = 31

2.2 Does coding Split as either Hill/Valley affect prediction by ecological variables?

In an alternate analysis, I checked whether coding Split languages as Hill instead affects the prediction by ecological variables. The results show that this reclassification does not improve ecological prediction. While the models show better overall fit (lower AIC values), the ecological predictors themselves remain non-significant across all models. Elevation and terrain roughness still fail to predict societal classification (all $p > 0.4$), reinforcing the conclusion that the ecological foundations of the Hill/Valley framework are not supported in this dataset.

Table 7: Model formula: hill_valley_binary ~ scaled_altitude + scaled_stdev_slope, family = binomial. N = 31 languages

Parameter	Estimate	Std. Error	z-value	p-value
Intercept	1.284	0.450	2.855	0.004**
Elevation (scaled)	0.313	0.585	0.534	0.593
Terrain roughness (scaled)	0.157	0.479	0.327	0.743

Model fit: AIC = 38.33, Null deviance = 33.12 (df = 30), Residual deviance = 32.33 (df = 28), N = 31

Table 8: Model formula: hill_valley_binary ~ scaled_altitude, family = binomial. N = 31 languages

Parameter	Estimate	Std. Error	z-value	p-value
Intercept	1.283	0.450	2.853	0.004**
Elevation (scaled)	0.401	0.522	0.767	0.443

Model fit: AIC = 36.44, Null deviance = 33.12 (df = 30), Residual deviance = 32.44 (df = 29), N = 31

Table 9: Model formula: hill_valley_binary ~ scaled_stdev_slope, family = binomial. N = 31 languages

Parameter	Estimate	Std. Error	z-value	p-value
Intercept	1.256	0.438	2.867	0.004**
Terrain roughness (scaled)	0.292	0.420	0.694	0.487

Model fit: AIC = 36.64, Null deviance = 33.12 (df = 30), Residual deviance = 32.64 (df = 29), N = 31

Sociocultural Analyses

Predicting complexity via L2_status

Linear mixed-effects models were fit with L2_status as a fixed effect (binary: 1 = present, 0 = not) and a random effect for subfamily. The tables below present the estimated coefficients, standard errors, degrees of freedom, t-values, p-values, and R^2 values for each model.

Table 10: Model formula: scaled_cell_complexity ~ L2 + (1 | subfamily)N = 31 languages; 10 subfamilies

Term	Estimate (β)	Std. Error	df	*t*	*p*
Intercept	0.078	0.196	12.72	0.398	0.697
L2	–0.582	0.279	27.93	–2.084	**0.046**

R^2: Marginal = 0.096; Conditional = 0.210

Table 11: Model formula: scaled_form_complexity ~ L2 + (1 | subfamily)N = 31 languages; 10 subfamilies

Term	Estimate (β)	Std. Error	df	*t*	*p*
Intercept	–0.032	0.191	15.54	–0.170	0.867
L2	–0.402	0.265	28.19	–1.516	0.141

R^2: Marginal = 0.066; Conditional = 0.178

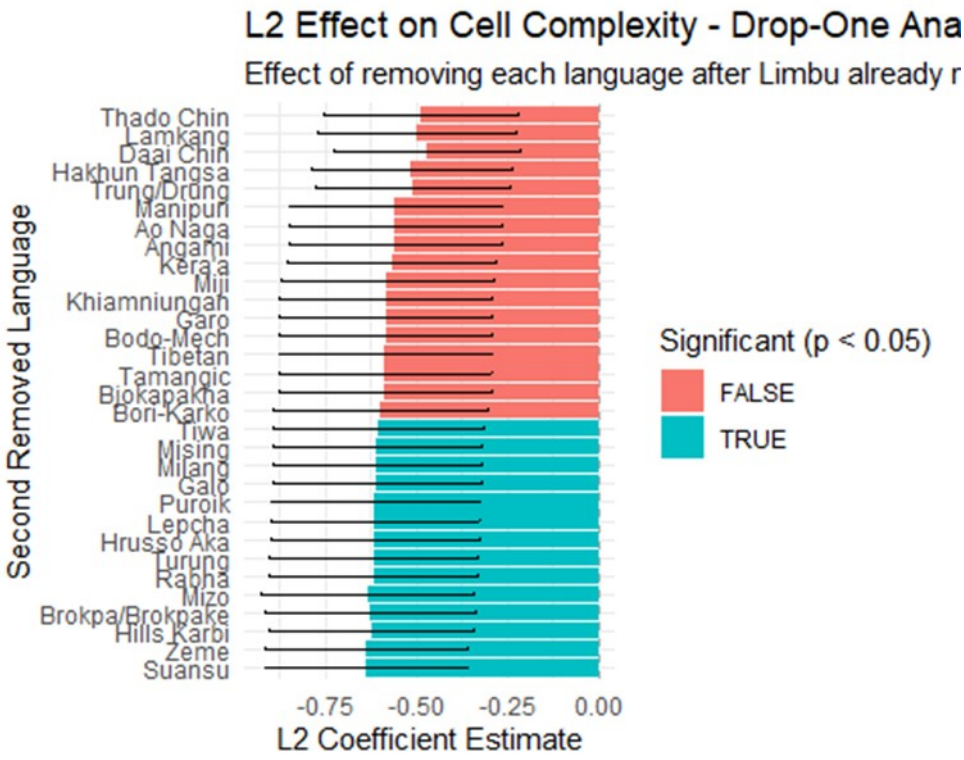
Table 12: Model formula: scaled_case_marking_complexity ~ L2 + (1 | subfamily)N = 31 languages; 10 subfamilies

Term	Estimate (β)	Std. Error	df	*t*	*p*
Intercept	–0.184	0.251	8.36	–0.731	0.485
L2	0.471	0.367	27.23	1.281	0.211

R²: Marginal = 0.040; Conditional = 0.061

Testing robustness of effect of L2_status on cell complexity

To test whether the significant relationship found between L2_status and cell complexity was truly robust, I wrote a function to loop through the revised dataset (excluding Limbu) to exclude one language at a time, to see whether the relationship maintained. The code can be found in the associated R-markdown document (*synthesis-code*) on the github, while I provide a graphical representation of the results of this process here.



The median p-value is 0.55, while the coefficient estimates range from -0.638 to -0.469. The result is significant in 14 out of 31 iterations, which may seem small, but due to the small sample size, a real result may not have fared much better.

Predicting complexity via agricultural intensity

Linear mixed-effects models were fit with **agricultural intensity** as a fixed effect (binary: 1 = Intensive/Irrigated, 0 = Other) and a random effect for subfamily. The tables below present the estimated coefficients, standard errors, degrees of freedom, *t*-values, *p*-values, and R^2 values for each model.

Table 13: Model formula: scaled_cell_complexity ~ agriculture_binary + (1 | subfamily)N = 31 languages; 10 subfamilies

Term	Estimate (β)	Std. Error	df	*t*	*p*
Intercept	–0.054	0.185	8.09	–0.292	0.777
Agricultural intensity	–0.533	0.381	27.24	–1.398	0.173

R^2: Marginal = 0.057; Conditional = 0.158

Table 14: Model formula: scaled_form_complexity ~ agriculture_binary + (1 | subfamily)N = 31 languages; 10 subfamilies

Term	Estimate (β)	Std. Error	df	*t*	*p*
Intercept	–0.104	0.177	10.94	–0.588	0.568
Agricultural intensity	–0.506	0.350	27.72	–1.446	0.159

R^2: Marginal = 0.059; Conditional = 0.185

Table 15: Model formula: scaled_case_marking_complexity ~ agriculture_binary + (1 | subfamily)N = 31 languages; 10 subfamilies

Term	Estimate (β)	Std. Error	df	*t*	*p*
Intercept	–0.046	0.207	5.39	–0.222	0.832
Agricultural intensity	+0.247	0.501	27.21	+0.494	0.626

R^2: Marginal = 0.008; Conditional = 0.022

Predicting complexity via speaker population

Linear mixed-effects models were fit with **scaled_population** as a fixed effect (binary: 1 = Intensive/Irrigated, 0 = Other) and a random effect for subfamily. The tables below present the estimated coefficients, standard errors, degrees of freedom, *t*-values, *p*-values, and R^2 values for each model.

Table 16: Model formula: scaled_cell_complexity ~ scaled_population + (1 | subfamily)N = 31 languages; 10 subfamilies

Term	Estimate (β)	Std. Error	df	*t*	*p*
Intercept	–0.026	0.183	8.32	–0.143	0.889
Population size	–0.162	0.147	27.71	–1.104	0.286

R^2: Marginal = 0.027; Conditional = 0.134

Table 17: Model formula: scaled_form_complexity ~ scaled_population + (1 | subfamily)N = 31 languages; 10 subfamilies

Term	Estimate (β)	Std. Error	df	*t*	*p*
Intercept	–0.050	0.175	9.90	–0.283	0.784
Population size	–0.071	0.133	27.46	–0.535	0.598

R^2: Marginal = 0.009; Conditional = 0.114

Table 18: Model formula: scaled_case_marking_complexity ~ scaled_population + (1 | subfamily)N = 31 languages; 10 subfamilies

Term	Estimate (β)	Std. Error	df	*t*	*p*
Intercept	–0.119	0.206	5.67	–0.577	0.586
Population size	0.154	0.220	26.70	0.699	0.491

R^2: Marginal = 0.014; Conditional = 0.042

Predicting complexity via political organization

Linear mixed-effects models were fit with **political_organization** as a fixed effect (binary: 1 = State, 0 = Acephalous/other) and a random effect for subfamily. The tables below present the estimated coefficients, standard errors, degrees of freedom, *t*-values, *p*-values, and R^2 values for each model.

Table 19: Model formula: scaled_cell_complexity ~ political_organization + (1 | subfamily). N = 31 languages; 10 subfamilies

Term	Estimate (β)	Std. Error	df	*t*	*p*
Intercept	–0.072	0.188	8.06	–0.384	0.710
political_organization	–0.489	0.429	26.83	–1.140	0.264

R^2: Marginal = 0.045; Conditional = 0.176

Table 20: Model formula: scaled_form_complexity ~ political_organization + (1 | subfamily)N = 31 languages; 10 subfamilies

Term	Estimate (β)	Std. Error	df	*t*	*p*
Intercept	–0.080	0.177	11.16	–0.455	0.658
political_organization	–0.482	0.393	26.71	–1.226	0.230

R^2: Marginal = 0.045; Conditional = 0.176

Table 21: Model formula: scaled_case_marking_complexity ~ political_organization + (1 | subfamily)N = 31 languages; 10 subfamilies

Term	Estimate (β)	Std. Error	df	*t*	*p*
Intercept	–0.076	0.208	5.66	–0.367	0.727
political_organization	+0.049	0.521	26.94	+0.093	0.926

R^2: Marginal = 0.000; Conditional = 0.031

Bibliographic References

(to be cleaned and added)