Austroasiatic Linguistics

In honour of Gérard Diffloth (1939-2023)

Edited by
Paul Sidwell

Publication data:

Austroasiatic Linguistics, in honor of Gérard Diffloth (1939-2023)

ISBN 978-616-398-980-2

Editor Paul Sidwell

Publisher Myanmar Center, Faculty of Humanities
Chiang Mai University
Thailand

PDF version, 24 October 2024

Gérard in forest, 22 May 2010, Heumensoord, The Netherlands, photo by N. J. Enfield

Gérard in restaurant with Som Wongjaroen Somruan, Roger Blench and N. J. Enfield, 12 January 2009, Siem Reap, Cambodia, photo by N. J. Enfield

Gérard alone working at table, 25 August 2006, Nakai District, Laos, photo by N. J. Enfield.

Gérard with Nick working at table, 5 May 2007, Nijmegen, The Netherlands, photo by Som Wongjaroen.

Contents

Abbreviations

Generally glossing of examples follows the *Notational Conventions for Austroasiatic Linguistics* at: https:// icaal.net/notational-conventions-for-austroasiatic-linguistics unless otherwise noted by authors.

Editor's Preface

It was with considerable shock that we learned of Gérard Diffloth's passing in mid-August of 2023. While generally the Austroasiatic studies community had known that our colleague, already in his eighties, had had serious health problems, such news always comes as a blow. Subsequently that October, at the 11th meeting of the *International Conference on Austroasiatic Linguistics*, there was no hesitation at the suggestion that Gérard should be memorialized with an edited volume produced under ICAAL auspices.

Gérard was an active participant in ICAAL since its first meeting in Hawaii in January of 1973, and had played a key role in the revival of ICAAL in the 3rd Millennium with the "Pilot Picnic" in Siem Reap in 2006 (a town that had become his home) and the 3rd ICAAL in 2007 (Deccan College, Pune, India). He also put substantial effort into editing papers for the second ICAAL proceedings (also known as SICAL[1]) proceedings, that meeting held in Mysore (India, 1978).

Following this Preface is a bibliography of Gérard's publications. This was compiled to the best of our efforts: not included are conference papers, handouts, or consultancy reports, rather it is focussed on collating formal research publications. There is also a wealth of unpublished draft works, some of which have been widely circulated at times, but such are a matter for the estate. In any case, it is probably premature to speculate on the full impact of Gérard's work on the field of linguistics, as his legacy of insight, mentoring, and as yet unreleased papers will unfold over the coming decades.

Readers of this volume will find two excellent personal tributes to Gérard (by Nathan Badenoch and Nick Enfield). These pieces make clear the special spirit that this scholar brought to his work; Gérard was dedicated to immersion in the subjective experience of language in its natural context, the recording of rich observational detail, and the value of intuitive insight based on deep experience. It was both a principled and a romantic view of linguistic work and worldview that complemented, and occasionally antagonized, the work of others who strive to abstract structure from data and build analytical models that seek to capture an objective understanding of the world. This is a long-standing and essential tension within the arts and sciences generally; real progress comes when researchers across the spectrum are able to fulsomely develop and exercise their skills with the kind of passion and effort that Gérard brought to bear.

It is especially satisfying that for this volume we are able to present a previously unpublished study by the honoree himself: *Proto-Vietic Glottal Features in Kri* (chapter 4). This is a fitting testimony to Gérard's attention to detail; some 45 lexical comparisons between Kri, Ahlau, and Vietnamese are given, illustrating various correlations between the phonation and glottalization of Kri syllables and their Vietnamese cognates. While the phonetic history of Vietnamese and Vietic has been theorized about for many decades, it is only by such precise and detailed description of

[1] Ultimately the proceedings of that meeting were not published as a volume, although fortunately various SICAL papers can be accessed at: https://icaal.net/icaal-2-1978-mysore

otherwise obscure and often difficult-to-recognize phonetic details in languages across the Vietic branch that one can achieve leaps of insight and test our hypotheses in a grounded and reliable way. Undoubtedly, Gérard's many notebooks contain a cornucopia of richly detailed knowledge that awaits a fuller appreciation.

Some of Gérard's unpublished legacy is available for examination at the Cornell University Library. In Chapter 3, Emily Zinger, takes us on a brief tour of the more than 400 folders of papers that were left behind when Gérard moved on from Cornell in 1996. About a quarter of those folders contain notes in Gérard's own hand, and include, "research notes and word lists, early drafts of his books and articles, and correspondence". Two decades ago, I spent a week working through those folders and it is fair to say that a week was far from adequate to fully appreciate the extent of the materials and the insights within.

The bulk of this volume comprises a range of contributions that all relate specifically to one or more Austroasiatic languages from diverse approaches. Four chapters (15 through 18) are write-ups of presentations given at the 11th ICAAL conference, while others were specially written for this memorial volume, with some (chapters 5, 6, 9, 10) touching specifically on themes that were important to Gérard, especially in historical-comparative reconstruction.

Ryan Gehrmann's piece on Proto-Ta'oi Phonology (chapter 5) deals directly with the rhyme glottalization phenomenon that is discussed by Gérard in respect of Kri, and bears directly on the question of whether glottalization or "creak" should be reconstructed for Proto-Austroasiatic, a question that Gérard had puzzled over for decades. Chapter 6 on Proto-Aslian harkens back to Gérard's earliest historical-comparative work and builds on fundamental observations and arguments he made in the 1960s and 70's in order to revise more recent work on Proto-Aslian. In chapter 9 Roger Blench delivers one of his signature big-picture historical models, considering the possibility that Austroasiatic speakers dispersed as far afield as Nepal and/or Borneo, testing the limits of what lexical similarities can suggest about the deep past. And in chapter 10, Mark Alves critiques the Vieto-Katuic Hypothesis, which was discussed by Gérard at the first annual meeting of the Southeast Asian Linguistics Society (Wayne State University, Michigan) in 1991. Other chapters tackle wide-ranging issues in Austroasiatic linguistics.

In Chapter 7, Gogoi, Horo and Anderson take a deep dive into the phonetic details of Mundari vowels. Unlike our honoree who preferred subjective perceptual methods, these scholars apply contemporary computer-assisted instrumental methods to measure and analyze syllable nuclei and support phonological conclusions. This is crucial fundamental work that transcends synchronic description, providing clues into earlier phonetic complexity in Munda.

Ratree Wayland (Chapter 8) takes us on a brief voyage into phonetic interpretation of graphemes by reference to loanword phonology, focusing on Khmer <au>, <ai>. This study reminds us of the importance of philological methods, the role of language contact in linguistic change, and the needs for realism in our interpretation of written and inscriptional materials.

Chapters 11 and 12 relate specifically to Monic languages; Christian Bauer discusses issues in Mon dialectology, while Mathias Jenny reviews the linguistics of Nyah Kur and what this contributes to our understanding of Monic history.

Khasian is the focus of chapters 13 and 14. Sarah Lyngdoh and Rymphang Rynjah compare word order and agreement across Khasian languages, highlighting the extent of VSO order in speech and the interaction of word order with agreement marking. In

her chapter Uma Pappuswamy takes a deep look at gender assignment in Khasi, revealing much about the semantics of this phenomenon, including an array of starkly counter-intuitive details (for example, while most fruits are classed as masculine, strikingly bananas are treated as feminine!).

The last four chapters relate to Vietnamese in diverse ways. Wenjiu Du offers a verbal-semantic typology of motion serial verb constructions in Vietnamese. Hong Duong Do discusses the maintenance and teaching of Vietnamese in Korea and Japan, both home to hundreds of thousands of ethnic Vietnamese. Ngọc Bình Nguyễn investigates the prominence of 'rice' in Vietnamese figurative language across a range of speech types. And finally, Duong Bui Duy and Hien Tran explore the diverse ways that the notion of 'intestine' as a seat of emotions and thought shapes idiom among Vietnamese speakers.

Various colleagues contributed in other ways to the production of this volume, and I would like to thank Mathias Jenny, Mark Alves, Nathan Badenoch, and Ngọc Bình Nguyễn for providing comments on drafts. Also, a special thanks is due to Nick Enfield for providing photographs of the honoree.

I am very pleased that the Myanmar Center of Chiang Mai University agreed to host this memorial volume. One of the main goals of the Myanmar Center CMU, under the leadership of Assist. Prof. Dr. Ampika Rattanapitak, is to promote research and publications on the linguistic and cultural diversity of Myanmar. Austroasiatic languages may be only a small part of the linguistic make-up of present-day Myanmar, but with Mon and Palaungic the country is home to two important branches. The Myanmar Center CMU has been a long-standing and regular host for ICAAL meetings and workshops, both onsite and online. I take this opportunity to extend thanks to the Myanmar Center CMU for its continued support of ICAAL and Austroasiatic studies, as also witnessed by the publication of the present volume.

Paul Sidwell
Nelligen, New South Wales

Gérard Diffloth: Bibliography

Single Author Publications

Diffloth, Gérard. 1968. Proto-Semai Phonology. *Federation Museums Journal* (new series), 13: 65-74.

Diffloth, Gérard. 1972. Notes on expressive meanings, *Papers from the Eighth Regional Meeting of the Chicago Linguistics Society*, 440-7, Chicago.

Diffloth, Gérard. 1972. Ambiguïté morphologique en Semai, *Langues et techniques, nature et société* (offert en hommage à André G. Haudricourt à l'occasion de son 60e anniversaire, Barrau J. et al.), 1: 91-94, Paris: Klincksieck.

Diffloth, Gérard. 1974. Austro-Asiatic Languages. In *Encyclopaedia Britannica*: Chicago/London/Toronto/Geneva Encyclopaedia Britannica Inc. Macropaedia 2:480-484. 15th edition.

Diffloth, Gérard. 1974. April. Body moves in Semai and in French. In *10th Regional Meeting, Chicago Linguistic Society*. pp. 128-138.

Diffloth, Gérard. 1974. The KmMu' principle. In Eric Hamp (ed.) *Parmentier festschrift.* Chicago, UNiversity of Chicago Press.

Diffloth, Gérard. 1975. Les langues Mon-Khmer de Malaisie: classification historique et innovations. *Asie du Sud-Est et Monde Insulindien* 6.4:1-19.

Diffloth, Gérard. 1975. Remarks on the Jah-Hĕt language, in R. Werner, ed., *Jah-He|t of Malaysia, Art and Culture*, University of Malaya Press, Kuala Lumpur.

Diffloth, Gérard. 1976. Mon-Khmer Numerals in Aslian Languages *Linguistics*, vol. 14. 174, pp. 31-38.

Diffloth, Gérard. 1976. *An Appraisal of Benedict's Views on Austroasiatic and Austro-Thai Relations*. *Discussion Paper*, no. 82, Kyoto: Center for Southeast Asian Studies, Kyoto University

Diffloth, Gérard. 1976. *Jah-Hut, an Austroasiatic language of Malaysia*. in N.D. Liem (ed.). *Southeast Asian Linguistic Studies* Vol.2. Canberra, Australian National University, Pacific Linguistics (vol. C-No.42.) pp. 73-118.

Diffloth, Gérard. 1976. *Proto-Mon-Khmer Final Spirants*. Discussion Paper No. 88, The Centre for Southeast Asian Studies, Kyoto: Kyoto University.

Diffloth, Gérard. 1976. Expressives in Semai. *Oceanic linguistics special publications* 13.1:249-264.

Diffloth, Gerard. 1976. Minor-Syllable Vocalism in Senoic Languages. In *Austroasiatic Studies*, edited by Philip N. Jenner et al.. Honolulu, 229-248. The University Press of Hawaii.

Diffloth, Gerard. 1976. Mon-Khmer numerals in Aslian languages, *Austro-Asiatic number systems, Linguistics*, special publication no. 174: 31-38, Mouton, The Hague.

Diffloth, Gerard. 1976. Translation fo a part of A. Moskalev's Grammar of the Chuang language (traduit du russe), *Discussion paper*, no. 91, Center for Southeast Asian Studies, Kyoto University, Kyoto.

Diffloth, Gérard. 1977. Mon-Khmer Initial Palatals and substratumized Austro-Thai. *Mon-Khmer Studies*, 6: 39-57.

Diffloth, Gérard. 1977. *Proto-Waic and the Effects of Register on Vowel Gliding*. Paper read at the Tenth International Conference on Sino-Tibetan Languages and Linguistics, Oct. 14-16, 1977, Georgetown, University of Washington.

Diffloth, Gérard. 1977. Towards a History of Mon-Khmer: Proto-Semai Vowels. *Tônan Ajia Kenkyû (Southeast Asian Studies)* 14.4:463-95.

Diffloth, Gérard. 1979. Aslian languages and Southeast Asian prehistory. *Federation Museums Journal*. 24ns:3-16.

Diffloth, Gérard. 1979. Expressive phonology and prosaic phonology in Mon-Khmer. In *Studies in Taj and Mon-Khmer Phonetics and Phonology in honor of Eugénie J. A. Henderson*, ed. Theraphan L. Thongkum et al, 49–59. Bangkok: Chulalongkorn University Press.

Diffloth, Gérard. 1979. *The Wa Languages.* Linguistics of the Tibeto-Burman Area. Vol. 5.2. Berkeley, University of California.

Diffloth, Gérard. 1980. ภาษาญัฮกุร มอญโบราณ กับอาณาจักรทวารวดี [Nyah Kur, Vieux-Môn et royaume de Dvāravati] (in Thai), *Bulletin of the Faculty of Arts*, 12.1: 54-85, Chulalongkorn University, Bangkok.

Diffloth, Gérard. 1980. The Wa Languages, *Linguistics of the Tibeto-Burman Area*, 5.2., Berkeley, 192 p.

Diffloth, Gérard. 1980. To Taboo Everything At All Times. *Proceedings of the Berkeley Linguistic Society*, 6:157-65.

Diffloth, Gérard. 1981. Reconstructing Dvâravatî Old Mon. In P. Bhumadon, ed., *Recent discoveries from the period of early Indian influence*, Lopburi Museum Publications, Lopburi.

Diffloth, Gérard. 1982. Mon registers: Two, three, four, ...? *Berkeley Linguistics Society* 8.148-147.

Diffloth, Gérard. 1982. Registres, dévoisement, timbres vocaliques: leur histoire en Katouique. *Mon-Khmer Studies* 11:47-82.

Diffloth, Gérard. 1984. *The Dvaravati Old Mon Language and Nyah Kur*. Monic language studies, vol. 1. Bangkok, Thailand: Chulalongkorn University Print. House

Diffloth, Gérard. 1985. The registers of Mon vs. the spectrographist's tones, *UCLA working papers in Phonetics*, Los Angeles.

Diffloth, Gérard. 1987-88 [1990]. *What happened to Austric?* 16-17:1-9.

Diffloth, Gérard. 1989. Proto-Austroasiatic Creaky Voice. *Mon-Khmer Studies* 15:139-154.

Diffloth, Gérard. 1991. Palaungic Vowels in Mon-Khmer Perspective. In *Austroasiatic Languages, Essays in honour of H. L. Shorto*, edited by Jeremy H.C.S. Davidson. 13-28. School of Oriental and African Studies, University of London.

Diffloth, Gérard. 1991. Vietnamese as a Mon-Khmer language, *Papers from the First Annual Meeting of the Southeast Asian Linguistics Society*, M. Ratliff and E. Schiller, eds., pp. 125-139, Arizona State University Press.

Diffloth, Gérard. 1992. On the Bulang (Blang, Phang) Languages. *Mon-Khmer Studies* 18-19:35-43.

Diffloth, Gérard. 1992. Khmer language, *International Encyclopædia of Linguistics*, W. O. Bright, ed., Oxford University Press.

Diffloth, Gerard. 1994. The lexical evidence for Austric so far. *Oceanic Linguistics* 33.2:309-321.

Diffloth, Gerard. 1994. /i:/ big, /a:/ small. In: Hinton L, Nichols J, Ohala JJ, eds. *Sound Symbolism*. Cambridge University Press; pp.107-114.

Diffloth, Gérard. 1996. History of the word 'khmer', *Khmer Studies, Proceedings of the International Conference on Khmer Studies*, Université royale de Phnom Penh, Sorn Samnang, ed., vol. 2, p. 644-652, Phnom Penh.

Diffloth, Gérard. 2001. Les expressifs de Surin et où cela conduit. *BEFEO* 88:261-269, Paris.

Diffloth, Gérard. 2005. Glottalised rimes in Vietic and Katuic. In: *The 6th Pan-Asiatic International Symposium on Linguistics*, 82–94. Hanoi: Nhà Xuất Bản Khoa Học Xã Hội.

Diffloth, Gérard. 2005. The contribution of linguistic palaeontology to the homeland of Austro-asiatic. In: Sagart, Laurent, Roger Blench and Alicia Sanchez-Mazas (eds.). *The Peopling of East Asia: Putting Together Archaeology, Linguistics and Genetics*. London & New York: Routledge/Curzon. 79-82.

Diffloth, Gérard. 2008. Proto-Aslian diphthongs and historical parallels in other Austroasiatic languages. Paper presented at the 18th Meeting of the Southeast Asian Linguistic Society, Universiti Kebangsaan Malaysia, Bangi, Selangor. 22 May, 2008.

Diffloth, Gérard. 2008. Shafer's parallels between Khasi and Sino-Tibetan. *North East Indian Linguistics*, *2*, pp.93-104.

Diffloth, Gérard. 2011. Considerations on the Homeland of Austroasiatic People. In K.S. Nagaraja (ed.) Austro-Asiatic Linguistics: In memory of R. Elangaiyan. (Proceedings of the 3rd International Conference on Austroasiatic Languages). Mysore, Central Institute of Indian Languages. pp.33-34

Diffloth, Gérard. 2011. Kuay in Cambodia: vocabulary with historical comments. UNESCO Phnom Penh. Tuk Tuk Editions.

Diffloth, Gérard. 2011. 13 Austroasiatic word histories: boat, husked rice and taro. In Nick Enfield ed. *Dynamics of human diversity*, Canberra, Pacific Linguistics pp.295-314.

Diffloth, Gérard. 2011. The westward expansion of Chamic influence in Indochina: A view from historical linguistics. In Trần Kỳ Phương & Bruce M. Lockhard (eds.), *The Cham of Vietnam: History, Society and Art*, 348-362. Singapore: NUS Press.

Co-Authored Publications

Diffloth, Gérard & Norman Zide (eds). 1976. Austroasiatic Number Systems. *Linguistics Special Publication* 174.

Diffloth, Gérard. & Norman Zide. 1992. Austro-Asiatic languages. In: William Bright (ed.): *International Encyclopedia of Linguistics*. New York: Oxford University Press. Vol. I:137-42

Enfield, Nick J., Gérard Diffloth. 2009. Phonology and sketch grammar of Kri, a Vietic language of Laos, *Cahiers de Linguistique Asie Orientale* 38.1: v-69

Ly, G., Alard, B., Laurent, R., Lafosse, S., Toupance, B., Monidarin, C., Diffloth, G., Bourdier, F., Evrard, O., Pavard, S. and Chaix, R., 2018. Residence rule flexibility and descent groups dynamics shape uniparental genetic diversities in South East Asia. *American Journal of Physical Anthropology*, *165*(3), pp.480-491.

Ly, G., Laurent, R., Lafosse, S., Monidarin, C., Diffloth, G., Bourdier, F., Evrard, O., Toupance, B., Pavard, S. and Chaix, R., 2019. From matrimonial practices to genetic diversity in Southeast Asian populations: the signature of the matrilineal puzzle. *Philosophical Transactions of the Royal Society B*, *374*(1780), p.20180434.

Diffloth, G., Chamberlain, J.R. and Badenoch, N., 2024. Notation and phonology of the Tri language in Vilabouly: An introduction and tribute to Gérard Diffloth and his fieldwork. *Linguistics of the Tibeto-Burman Area*, 47.1:1-16.

Werner, Roland (with remarks on the Jah-hĕt language by Gérard Diffloth). 1975. Jah-hĕt of Malaysia: art and culture. Kuala Lumpur.

Diffloth, Gérard and Nathan Badenoch. 2015. “Austroasiatic Languages”, in Sybesma, R. P. E., Behr, W., Gu, Y., Handel, Z. J., Huang, C. T. J., & Myers, J. (Eds.). *Encyclopedia of Chinese Language and Linguistics*. Leiden:Brill.

Reviews

Diffloth, Gérard. 1973. Review of Johnston et al. Mon-Khmer studies III, *Language*, 49.1: 233-234.

Diffloth, Gérard. 2008. Review of Sidwell (2005): The Katuic Languages, Classification, Reconstruction and Comparative Lexicon *Diachronica* 26:3, pp. 442–447

Diffloth, Gérard. 2008. Review of Harry Shorto A Mon-Khmer Comparative Dictionary, *Diachronica*, 25(1):137-142.

Contributors

Mark Alves (PhD, University of Hawaii, 2000) is a professor at Montgomery College (Maryland) and serves as Editor-in-Chief of the *Journal of the Southeast Asian Linguistics Society*. His research has explored Vietnamese and Vietic historical phonology and language history; Southeast Asian regional language contact and loanwords spread from Daic, Khmeric, Sinitic, and Indic languages; Southeast Asian ethnohistorical linguistics with reference to historical and archaeological data; Austroasiatic typological linguistics, especially morphology; among others.

Gregory D. S. Anderson, President Living Tongues Institute for Endangered Languages, has published on a range of topics on the Munda languages, including descriptive, historical, typological, sociolinguistic, lexicographic, phonetic, phonological, textual and ethnobotanical materials on individual languages and comparatively, including The Munda Verb, the 2008 volume Munda Languages and the 2022 volume Munda Linguistics: Descriptive, Diachronic and Typological Perspectives. He is currently preparing large studies on Sora and Gtaʔ and a study on phonological domains, phonetics and morphology and the mapping of phono-prosodic structures onto the large grammatical words in seven Munda languages spanning the genetic diversity of the family.

Nathan Badenoch (PhD, Kyoto University, 2006) is Associate Professor in the Department of Global Interdisciplinary Studies at Villanova University. He researches languages and cultures of mainland Southeast Asia, with a particular interest in the poetics of linguistic performance and linguistic encoding of ecological knowledge in Austroasiatic and Tibeto-Burman languages.

Christian Bauer is Professor of Southeast Asian Philology at Humboldt University, Berlin; although nominally retired in 2017, he continues teaching at HU. Fieldwork in Thailand and Burma on Mon and Khmer dialects, inscriptions and manuscripts. He most recently completed cataloguing Mon manuscripts held in the library of the Ecole française d'Extrême-Orient in Paris (Calames, 2023).

Roger Blench (PhD, Social Anthropology, University of Cambridge, 1984) has pursued a twin-track career as a consultant in development, usually in Sub-Saharan Africa and as a linguist and ethnomusicologist in Africa and SE Asia. He is currently a Research Associate at the McDonald Institute for Archaeological Research at the University of Cambridge, a Researcher in the Department of History, University of Jos, Nigeria, and State Co-ordinator for Ethnographic Research in Arunachal Pradesh, Northeast India. He is the Research Director for the Kay Williamson Educational Foundation. He has published and/or edited some twenty-six books.

Hong Duong Do (PhD, Linguistics, Vietnam National University, 2011), is a lecturer in the Faculty of Linguistics at the University of Social Sciences and Humanities, VNU Hanoi. Her research interests include Vietnamese syntax and language education. She received research scholarships in South Korea and Japan, focusing on heritage language maintenance for mixed-heritage children. She has authored numerous research papers on syntax, textbooks for Vietnamese elementary schools, and Vietnamese language teaching materials for foreigners.

Wenjiu Du is a graduate student of linguistics at Goethe University Frankfurt, specializing in typology, syntax, and semantics. His specific research interests include argument structure, complementation, and DP structure, with a focus on languages spoken by China's minority groups, as well as those in Southeast Asia and Africa. He employs an empirically oriented approach, dedicated to describing intriguing linguistic phenomena and exploring their theoretical implications.

Duong Duy Bui (PhD, Department of Vietnamese Studies and Language, University of Social Sciences and Humanities, Vietnam National University) received a postdoctoral fellowship in Department of Anthropology, University of Toronto. His research interests include the history of Vietnamese language, Vietnamese lexicology and grammar. He has held Visiting Professor positions at Wunnan University for Nationalities, GuangDong University of Foreign Study and Tokyo University of Foreign Study.

N. J. Enfield is Professor of Linguistics at the University of Sydney. He has conducted extensive field research on the languages of mainland Southeast Asia, especially those of Laos. He is the author of *A Grammar of Lao* (Mouton, 2007) and *The Languages of Mainland Southeast Asia* (Cambridge, 2021), editor of the volume *Dynamics of Human Diversity: the case of mainland Southeast Asia* (Pacific Linguistics, 2011) and co-author (with Gérard Diffloth) of the article *Phonology and Sketch Grammar of Kri, a Vietic Language of Laos* (*Cahiers de Linguistique – Asie Orientale*, 2009).

Ryan Gehrmann (PhD 2022, University of Edinburgh) is a lecturer in the Linguistics Department, International College, Payap University, Thailand. His research is focused on reconstructing the phonological history of the Austroasiatic language family and on the phonetics, phonology and historical development of tone and register in Southeast Asian languages.

Pamir Gogoi (PhD, Linguistics, University of Florida, 2021) is a Phonetics researcher at the Living Tongues Institute for Endangered Languages currently focusing on the phonetics of Mundari and Santali. She worked as a PhD Intern at Microsoft Research India from 2021 to 2022, contributing to multiple projects related to building language technology for low-resource languages. With her current focus on the phonetic documentation of endangered languages, her academic and professional research reflects a strong commitment to linguistic research for low-resource languages.

Mathias Jenny (PhD, University of Zurich 2005) is a senior researcher at the University of Chiang Mai, Thailand. Specializing in languages and linguistics of Thailand and Myanmar, he has written numerous papers on Southeast Asian areal and descriptive linguistics. He is co-author of *Burmese - a comprehensive* grammar

(Routledge 2016) and co-editor of *The Languages and Linguistics of Mainland Southeast Asia* (de Gruyter 2021) and *The Handbook of Austroasiatic languages* (Brill 2014). He is actively involved in the development and introduction of an orthography for and documentation of Htanaw (Danau), a little-known AA language in Shan State, Myanmar.

Luke Horo (PhD) is a Senior Researcher in phonetics at the Living Tongues Institute for Endangered Languages, focusing on the South Asia region. He is a field researcher with specialized expertise in conducting instrumental studies in phonetic science. His research primarily centers on the description and documentation of Munda languages in India. Dr. Horo earned his PhD from the Indian Institute of Technology Guwahati, where he completed his dissertation, "A Phonetic Description of Assam Sora," in 2018 He has published several articles in Journal of the Acoustical Society of America in 2020 and other phonetics venues on vowels and syllable prominence in the Sora varieties of Assam and Odisha, Santali and Mundari.

Saralin A. Lyngdoh is an Associate Professor in the Department of Linguistics at North-Eastern Hill University (NEHU). She received her PhD in Linguistics from Delhi University. Her doctoral research focused on "Empty Categories in Khasi", exploring the syntactic and theoretical implications of these phenomena within the Khasi language. Dr. Lyngdoh has published extensively on syntax and natural language processing of the Khasi language and is actively involved in projects aimed at documenting and revitalizing endangered languages in Northeast India. Her work has been recognized for its contributions to theoretical linguistics and language preservation.

Nguyen Ngoc Binh (Ph.D Linguistics, Research Institute for Languages and Cultures of Asia, Mahidol University), is affiliated with the Faculty of Linguistics, University of Social Sciences and Humanities, Vietnam National University, Hanoi. His research specialties are Comparative Linguistics, Ethnolinguistics, Kra-Dai languages. His past achievements include: participation in consulting on ethnic minority language policies for the Government of Vietnam, training dozens of Vietnamese and foreign students specializing in the study of ethnic languages of Vietnam.

Umarani Pappuswamy (PhD, University of Mysore1998) is a Professor and Deputy Director at the Central Institute of Indian Languages, India. She heads the Centre of North East Language Development, the Centre of Language Planning & Policy, and Sociolinguistics, along with various other responsibilities. Since 2019, she has been a member of the UNESCO Chair on Language Policies for Multilingualism. Her research spans Documentary Linguistics, Typology, Semantics, Lexicography, Multilingualism and Language Policies, Computational Linguistics, Artificial Intelligence in Education, and Translation Studies.

Rymphang K. Rynjah (PhD) is a researcher and currently employed as a guest lecturer in the Department of Linguistics at North-Eastern Hill University (NEHU). He received his PhD in linguistics from NEHU. His doctoral research, titled "War-Khasi and War-Jaiñtia: A Comparative Syntactic Study", provides an in-depth analysis of the syntactic structures and comparative grammar of these two closely related varieties of Khasi language. Dr. Rynjah's expertise lies in the comparative and historical linguistics of the

Khasi-Jaiñtia group, and he has been instrumental in several fieldwork-based linguistic projects across the Meghalaya region.

Paul Sidwell (PhD, University of Melbourne 1999) retired from the Australian National University in 2016 after serving as Senior Lecturer and Australian Research Council Future Fellow. Specializing in comparative Austroasiatic linguistics, he has authored numerous books and papers in AA reconstruction and related studies and co-edited *The Languages and Linguistics of Mainland Southeast Asia* (de Gruyter 2021) and *The Handbook of Austroasiatic languages* (Brill 2014). He is actively involved in organizing the annual meetings of the *Southeast Asian Linguistics Society* and the *International Conference on Austroasiatic Linguistics*.

Hien Tran (PhD, University of New Mexico, USA) is employed in the Department of Vietnamese Studies and Language, University of Social Sciences and Humanities, Vietnam National University. Her research interests include Cultural models, Emotions and Language acquisition, Metaphor and Metonymy in teaching Vietnamese as a second language.

Ratree Wayland (PhD) is a Professor in the Department of Linguistics at the University of Florida, USA. She received her Ph.D. in Linguistics from Cornell University. Her doctoral research on the acoustic and perceptual investigation of breathy and modal phonation phonemic contrast in an older dialect of Khmer spoken in Chanthaburi province, Thailand, was funded by a Fulbright-Hays Doctoral Dissertation Abroad grant. She is the author of Phonetics: A Practical Introduction (Cambridge University Press, 2018). Her research on cross-language speech learning was funded by the NIH, and her current research on speech as biomarkers of neuromuscular degenerative diseases is funded by an NSF grant.

Emily Zinger serves as the Southeast Asia Digital Librarian for Cornell University. In this role she is project manager for the Southeast Asia Digital Library, a cooperatively managed open access repository of over 10,000 rare and unique archival materials (sea.lib.niu.edu). She also sits on the editorial board of the Journal of Critical Digital Librarianship. Emily has a Master's of Information Studies from McGill University and a Bachelor's in both Psychology and English from the College of William & Mary.

1

In Memoriam: Gérard Diffloth (1939-2023)[1]

Nathan Badenoch

It is hard to imagine someone in the field of Southeast Asian linguistics who was not profoundly influenced by the work of Gérard Diffloth. For anyone who has heard him speak formally about his work, the breadth and depth of his understanding of Austroasiatic languages was spectacular. For those who had the chance to talk with him leisurely about his views on language in Southeast Asia, it was a winding journey through the sub-domains of linguistics into the worlds of entomology, metallurgy, migration, upland farming and more. His sense for how history, language and society have evolved together has illuminated the complexity of this fascinating region for more than six decades.

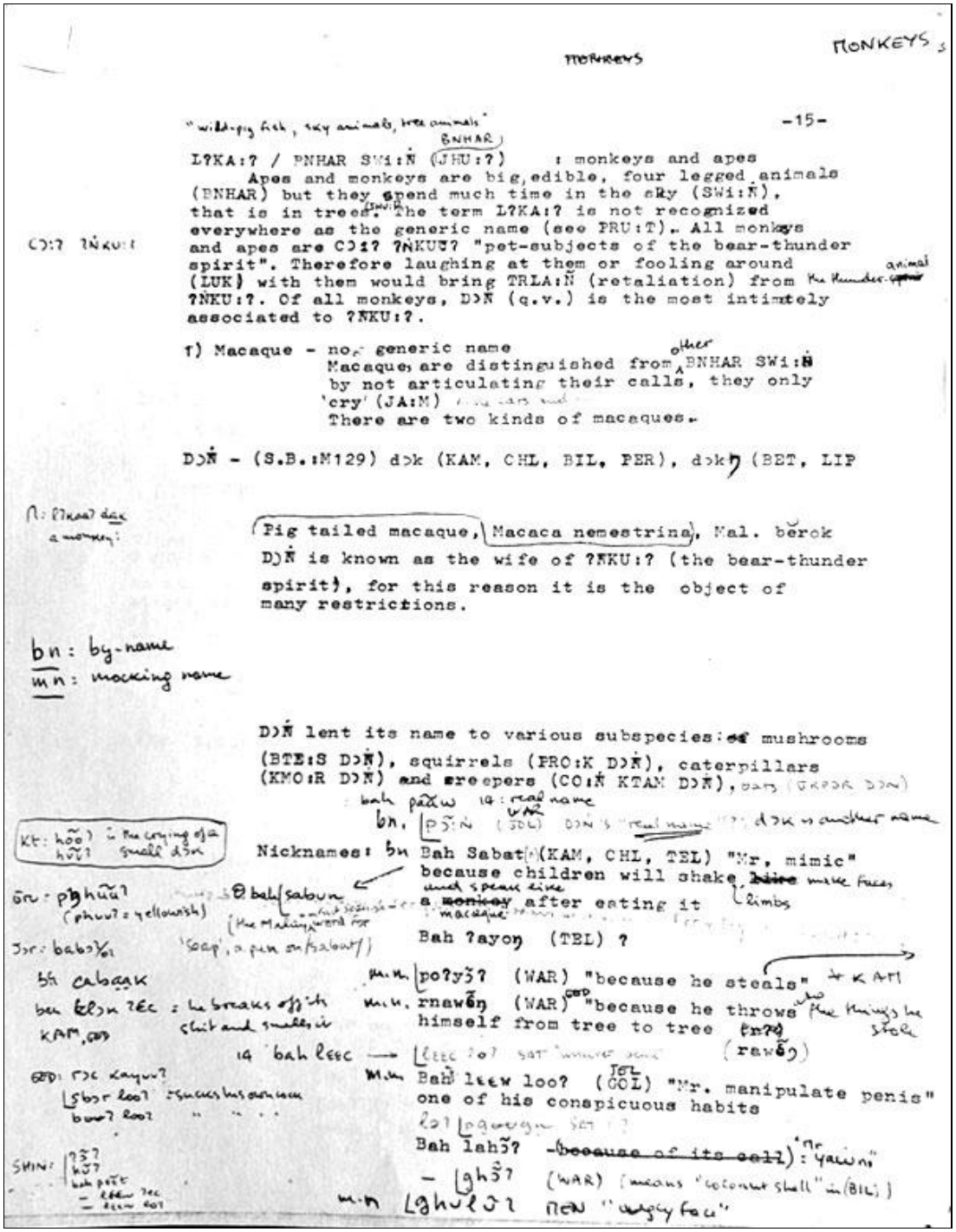
The technical precision, intellectual creativity, and ambitious scope of his historical linguistics approach were intimately informed by his early training in mathematics, journalism, and ethnomusicology. Gérard was born in Chateauroux, France, in 1939, and his education took him to the University of Paris, the École Supérieure de Journalisme de Lille, and the University of California Los Angeles where he wrote on the Dravidian Irula language. The experience of studying and working in the United States was transformative, but even in his later years, he spoke often of how his experience with French and German sociolinguistics during World War II, fascination with Gaulish etymologies of French toponyms and study of Farsi underpinned the development of his linguistic curiosities. After teaching at the University of Chicago he moved to a Southeast Asia position at Cornell University from 1988 to 1996, a historical linguist within that institution's tradition of area studies.

Gérard's work is epitomized for many by such meticulous studies in linguistic history and classification as *The Dvaravati Old Mon Language and Nyah Kur* (1984) and *The Wa Languages* (1980). Others continue to be inspired by pioneering theoretical and conceptual works like "Les expressifs de Surin, et où cela conduit" (1994) and "To Taboo Everything at All Times" (1980). We have not yet seen the full extent of his work on the history of animal names in Austroasiatic, through which he focused his social history on people's experience with co-animates rather than material culture, which he felt to be interesting but susceptible to "civilizational" forces. Names of birds, insects and mammals provided a solid foundation for recreating linguistic history for him, and this is evident in the way in which he applied the comparative method in Austroasiatic. During one of his many stays at the Center for Southeast Asian Studies in Kyoto, he began a talk for Kyoto University linguistics students saying, "I have been asked to speak about my philosophy for conducting fieldwork, but what I would really like to do is tell you about the history of the Khabit word for 'fish', *mʔuə*. It shouldn't be like that, but it is, and it is a wonderful story of history in Southeast Asia". As the co-organizer, I was not surprised. When we first met in 2010, he began by speaking of the social

[1] Originally published in JSEALS 16.2 (2023)

history of the tree monitor *Varanus bengalensis*—"As you know, *trkəət* in Khabit"—in the Austroasiatic world. He only knew that I had been studying that language for a year, but this was an invitation to his Austroasiatic research if one was willing to go off the path of basic vocabulary and descriptive syntax into the larger natural world in which languages are spoken. Gérard's considerations were deeply multidisciplinary yet in his writings he seemed to ignore disciplinary thinking entirely, exemplified in this work around animal names.

Photo 1 caption: From Diffloth's unpublished reconstruction of Aslian proto-zoology based on Semai animal names (Photo credit: Jim Chamberlain)



I would also find myself dropped into a fascinating landscape of poetics and play. The

first time I visited Gérard in Siem Reap later that year, after dinner on the first night he said, "I think you have probably found some strange words in Khabit that mean just how they sound. What do they say about that?" We were soon in the thick of Surin Khmer expressives, with Gérard's partner Wongjaroen Somruan—known by many who knew Gérard as Som—explaining the micro-nuances of meaning, rejecting or accepting his words exploring the morphological paradigms, and pointing out how *ʔntriiiin* 'the sad feeling of lonely silence' must be said with falsetto voice and an elongated vowel. It was nigh impossible for Gérard to speak of expressives without frequent reference to his time with the Semai in Malaysia, where many of the most important foundations of his Austroasiatic work were lain. "Expressives are not said or spoken, they are shot. Like an arrow. That's what they say. Once said, they cannot be taken back. And the speaker doesn't care about what the speaker thought. They are expressives, shot at that particular time." For Gérard, expressives and tabooing animal names were an interlinked part of language in the Semai forest. Experiences with this type of linguistic play fit nicely with his take on Rudy Keller's (1994) idea of an "invisible hand" in linguistic change.

Indeed, Gérard had a fantastic ability to see complex relationships clearly in his data and articulate his ideas so eloquently. But we must pause to recognize his intense commitment to fieldwork–directly with speakers of languages in the settings in which they are spoken, with Som always by his side. He relied on his finely tuned ear, countless black field notebooks and a rainbow of colored pens. He maintained a defiantly analog approach to his field research. He insisted that fieldwork was not a science, and that everyone had to develop their own field style. A more subtle message was that fieldworkers need to be guided by their understanding of the research context and driven by their desire to learn from speakers. Gérard did not work with word lists and voice recorders. However, when he conducted fieldwork, he had an expansive cognitive-historical map in his head that led him windingly through the vast world of native-speaker linguistic knowledge, directly into the depths of linguistic history.

We knew Gérard for his humble approach to collaboration and academic exchange, but nowhere could we see his respect and humility for his field more clearly than in his interactions with the thousands of people who worked with him tirelessly, from northeastern India to Hanoi, from Sipsong Panna to the Bolaven Plateau, from the Aslian highlands to the Nicobar Island coast, and of course his beloved northeastern Cambodia and Surin. Having studied so many languages in so many diverse situations across Asia, Gérard maintained a healthy suspicion of claims to universality, and he employed the highest empirical criteria of evidence for drawing conclusions. At times he was frustratingly stubborn on insisting for ever more evidence in support of a sound change or semantic variation. He loved to follow the possibility of complex phonological innovations, no matter how seemingly far-fetched. Many—most—of these ended in "I don't like it." But one was always left enlightened by the attempt. It was this deep engagement with so many spoken languages that led him to a model of classification in the Austroasiatic world that had many nested sub-branches of significant historical depth. Confidence in drawing out historical relationships was a matter of constant crossing between his theoretical brilliance and rich realizations emanating from direct field experience.

He painstakingly coded his field notes, with different colors used with words to indicate different levels of classification within the family. Based on these notes, he created endless lists of cognates through which he followed the winding roads of phonological change, semantic shift and morphological transformation.

In addition to the Semai, there are several language-speaker groups that held a special place in Gérard's intellectual and personal life. His fascination with the depth of the Khasi languages was one of these—from internal phonological variation and lexical diversity, to expressives and morphology. In researching Khasi, he was challenged by the reality of how literacy can change language, which made him all the more dedicated to discovery in the spoken language of rural communities. Many of our discussions about Bit and related Austroasiatic languages of northern Laos were concluded with "You must go to Khasi!" Another language that dedicated his energies to was the Kuay of the Cambodia-Thailand border area. He went deep into the historical multilingualism of these speakers and their position in the Katuic branch, as well as the tradition of iron-working within the social and economic networks of Angkor. He also turned his efforts to such projects as the UNESCO-sponsored Kuay in Cambodia: vocabulary with historical comments (2011), which provides a historical view on the cultural language of the Kuay. When trying to gather information for this project, Gérard was in a Cambodian village where a marginal dialect had been spoken. The "last speaker" was an old, old woman who was not responding to the basic questions like "how do you say eat rice?" But when Gérard asked her about catching elephants, she sat up and spoke her Kuay variety clearly and lucidly for him for an hour and half. As he often said, fieldwork is about the people you work with and their lives.

The joy that was clearly visible on his face when eliciting words of historical significance with a native speaker was matched only by the endless enthusiasm he had for combing over his data, searching for cognates, developments, and connections. His last years were dedicated to working on Nico-Monic, another historical relationship that was for decades close to his heart drawing together his fieldwork on the Nicobar and Monic languages. Over the years Gérard was welcomed at research institutions, academic gatherings and social events around the world. He also graciously welcomed visitors to his Siem Reap home, where there was always a captivating and stimulating discussion to be had. He loved to share stories of fieldwork and listened to them with excitement as well. Gérard was extraordinarily generous with his ideas and guidance, particularly when approached with questions and conundrums coming directly out of observations made in the field.

In 2014, Gérard told the Kyoto University Center of Southeast Asian Studies Newsletter that "when a language disappears, it is as if a cathedral collapsed or a library was burnt to the ground." We now feel the profound loss of Gérard's passing–his knowledge a cathedral or library its own right–but let us continue to be inspired and motivated by his boundless curiosity, his razor-sharp attention to micro-level detail and his passionate drive to understand the big picture of language and history in Southeast Asia. It is hard to imagine Southeast Asian linguistics without Gérard Diffloth, but his intellectual legacy will influence us long into the future.

References

Center for Southeast Asian Studies. 2014. Language and History in Southeast Asia: An Interview with Gérard Diffloth, *CSEAS Newsletter No. 69*. Kyoto: Center for Southeast Asian Studies.

Diffloth, Gérard. 2001. Les expressifs de Surin et où cela conduit. *BEFEO* 88:261-269, Paris.

Diffloth, Gérard. 1984. *The Dvaravati Old-Mon language and Nyah Kur*. Chulalongkorn University Printing House: Bangkok.

Diffloth, Gérard. 1980. *The Wa languages*. Berkeley: Dept. of Linguistics, University of California.

Diffloth, Gérard. 1980. To taboo everything at all times. *Papers of the Berkeley Linguistics Society* 6:157-65.

Diffloth, Gérard. 2011. *Kuay in Cambodia: vocabulary with historical comments*. UNESCO Phnom Penh. Tuk Tuk Editions.

Rudy Keller. 1994. *On language change: The invisible hand in language*. Routledge: London and New York.

2

Gérard Diffloth: Fieldworker, Thinker, Dreamer

N. J. Enfield

Gérard Diffloth was a distinct brand of scholar. I thought of him as a purist and a free spirit, utterly unbound by the shackles of university life, and joyfully steeped in scholarship: equal parts fieldwork and conceptual analysis, practice and theory. It is true that he spent nearly thirty years in universities, including at some of the world's most prestigious institutions, from UCLA to the University of Chicago to Cornell. But fieldwork was his passion. Fieldwork is what he pursued full-time when he retired at just 57, an age that many professors would consider far too young to leave the academy. This, of course, was no retirement. He would spend the next quarter century immersed in field research on the hundreds of Austroasiatic languages spoken across greater mainland Southeast Asia.

I met Gérard in the late 1990s soon after he had left the US to set up a permanent base in Southeast Asia. His time in the US academic system, mostly among linguists, had frustrated him as it kept him from his favoured sources of language: the living speakers of the hundreds of mostly unwritten languages that were under-resourced and under-represented relative to the large, written languages that linguistics tended to favour. Gérard's fascination with the unsung was evident from the beginning. His 1968 PhD thesis—chaired by Bill Bright at the UCLA Department of Linguistics—analysed phonology and verb morphophonemics in Irula, a Dravidian language of Southern India. He chose to focus on this minority language, spoken by a few thousand people and "virtually unknown" at the time, rather than closely-related Tamil, spoken by many millions and studied for centuries. Why this choice? Irula, he explained, would provide a new and unique lens on linguistic questions, both general and specific. It would have been much easier for him to study Tamil, but Irula had special qualities: it preserved features that Tamil had lost, and it had made certain unique innovations. This would yield new insights. Here are the concluding words of his dissertation:

> *The study of such obscure languages as Irula, even though toilsome may give a standing point from which we may gain new perspectives on better known languages like Tamil.*

This, in a nutshell, was his philosophy. Don't take the convenient, obvious path of studying large, well-known, well-resourced languages. Instead, make the effort to put under-studied languages in the foreground. Invert the standard bias.

Accordingly, in my many conversations with Gérard about the great language ecology of mainland Southeast Asia, he had to beat out of me the idea that languages like Lao, Thai, and Vietnamese were "typical" of the area. He chided me for succumbing to the typologist's equivalent of Teeter's Law, the tendency to see all languages as departures from the language one knows best or learned first.

This concern with over-riding our own biases was typical of Gérard's fierce insistence on academic integrity, that our claims be accurate and well-evidenced. For

example, more than once, I saw him publicly defend the core principles of the historical-comparative method—establishing common ancestry with reference to evidenced regular sound changes—against other research procedures that happened to use the same terminology. A case in point concerned the word *cognate.* Gérard was in the audience when a conference presenter defined "cognates" solely in terms of phonetic similarity between contemporary words in different languages. In the question period he stood and responded with a passionate call not to use the label *cognate* for words whose only known relation was inspectional resemblance. Suggesting that a new term would usefully disambiguate, he offered the neologism *pognate* for these surface-similar word pairs, later musing that *fognate* might be even more apt. This was not conservatism but commitment. When I talked with him about emerging quantitative methods in historical linguistics, I referred to his work as "traditional" historical-comparative linguistics. He bristled at this, saying "What I do is not 'traditional' historical-comparative linguistics, it is merely historical-comparative linguistics".

Gérard's great care in collecting and handling data meant that he insisted on clarity and precision, even when this produced redundancies. In around 2007, Gérard and I worked together on an analysis of the phonology of Kri, a Vietic language spoken in upland Central Laos. This involved many hours poring over my wordlists, over weeks and months, in Southeast Asia and in Europe, listening to my field recordings, and cross-checking against Gérard's database of Austroasiatic forms and reconstructions. Eventually, Gérard would join me on a field trip to Nakai District in Laos to work directly with speakers of Kri. He was especially interested in eliciting words for plants, animals, and other natural kinds. When we started the work, it was naturally with the phonetics of the raw recordings I had made in the field. In Kri, certain finals showed a three-way distinction in terminance, with voiced or modal being the "unmarked" category:

Voiced: [ʐˈɑ̱ː] 'turtle'
Checked: [ʐˈɑ̱ːˀ] 'pig basket'
Devoiced: [ʐˈɑ̱ːʰ] 'dry'

But Gérard didn't like the "unmarked" approach. He insisted that words with voiced terminance, such as the word for 'turtle' above, should be written with an overt mark showing that the word is positively known to have modal terminance, that is, known not to have a glottal stop or devoicing. So, we agreed to write those words with a 'zero' explicitly marked, like this:

[ʐˈɑ̱ːᵚ] 'turtle'

This piece of practical scholarship reduced uncertainty. Such intentional redundancy is not something that a razor-armed phonologist might prefer but it introduced a form of certainty and error-anticipation that characterizes careful scholarship and data-handling. Gérard's chapter in this volume illustrates the practice.

While our Kri phonology collaboration was fundamentally non-historical—we sought to describe the phonological system that Kri speakers acquire—Gérard never veered from an interest in the history of the system. The lesson I took from this was not (only) that the history of language systems is fascinating in its own right. It was that to understand why systems are the way they are, we need to think about how they got that way. For example, had I simply focused on the paradigm of contrasts in the segmental

system, I would have stopped at the simple discovery, based on minimal pairs, that Kri has a three-way contrast in voice onset time in syllable-initial stops: e.g., /ɓ-, p-, p^h-/. But Gérard always looked through his comparative lens. With data from other languages, he first pointed out that most of the Kri words with voiceless-aspirated stops were clearly borrowings. Then, with data from word type frequencies in the Kri lexicon, he pointed out that words with aspirated-stop initials were ten times less type-frequent in the lexicon than words with unaspirated-stop initials. This, he said, was to be expected, given that neither Proto-Vietic nor Proto-Austroasiatic had voiceless-aspirated stops. To be sure, the aspirated stops were part of the synchronic system, but they were a recent innovation.

It was a joy to spend time with Gérard, not just intellectually, but personally. He liked to get outside, to mix with people, to laugh. His infectious delight at making a linguistic discovery was the same infectious delight he would frequently display in everyday life, for example in seeing a new kind of bird, a new park, or a new kind of gesture. While a Frenchman to the core (he would chide me for cutting good cheese the wrong way), Gérard was a villager at heart. He detested air-conditioning and would curse the cold, dry air of hotels and conference venues, wearing his wind jacket indoors in protest.

Gérard's mind was always on puzzles, not just words but their relations to the natural world and to ancient lifeways. He had a special fondness for the creatures and plants whose names in Austroasiatic languages were so important to historical analysis. His head was usually somewhere between the clouds and the deepest past, which made him an intellectual misfit in the context of worldly affairs, especially the travails of university life. This stance was evident from his earliest work. Exactly fifty years ago, as a junior Faculty member at U. Chicago, Gérard presented a paper at the 1974 Chicago Linguistic Society meetings. Typical of his approach of juxtaposing virtually unknown languages against major world languages (recall Irula vs. Tamil), he compared the syntactic devices used in describing "body moves" (e.g., *The child nodded his head* or *He closes his eyes*) in French versus in Semai, an Aslian language of upland peninsular Malaysia. Here are the opening words of that 1974 article:

> *Trying to do semantic field work in the mountain forests of Malaya is beneficial in one way at least: from that vantage point, current theoretical issues take on the appearance of being discussions over "What is the best style for writing abstracts?", and one finds oneself wishing that more time could have been spent on questions like "What do dolphins choose to mean?".*

These enigmatic words from a youthful Gérard Diffloth are the words of a dreamer, a man who would leave the world of big institutions and choose instead to roam the world of small villages, where he could escape the air-conditioned halls and live intellectual life to the fullest.

3

In the Archives with the Gérard Diffloth Papers at Cornell University

Emily Zinger

Gérard Diffloth served as Professor of Linguistics and Asian Studies at Cornell University from 1988 to 1996, leading to the eventual donation of his archives to Cornell's Division of Rare and Manuscript Collections (RMC) in 1998. Totaling seven archival boxes, the equivalent of seven cubic feet of documents, Collection 6313, the Gérard Diffloth papers contains a wealth of information on the scholarship of one of the world's premier researchers of Austroasiatic languages.[1] A small additional portion of Diffloth's papers are housed at the Center for Khmer Studies in Siem Reap, Cambodia.[2]

Diffloth's records at Cornell date from approximately 1963 to 1993. That being said, the vast majority of papers are undated. We create paper evidence of our lives by living. An act not usually done with the interests of future researchers in mind. To quote the collection description, the papers are compiled of "field notes, articles, research findings and support material, subject files, historical studies, bibliographies, dictionaries and other linguistic lists, language origin studies, etymological lexicons, and other papers and records of a professor of Southeast Asian languages."[3] The vast majority of the documents are textual, though a handful of photographs are scattered throughout the folders. Whether these photographs were taken by Diffloth himself or came into his possession over the years is unclear. The occasional presence of these visuals, of the peoples and places of Southeast Asia, round out the thousands of pages of text that a patron of Diffloth's archives encounters. They remind one of the living, speaking people behind his lists of words.

More common than photographs are maps. Nearly all are hand drawn, though the absence of a credit line again casts uncertainty on whether these maps were Diffloth's own creations or resources that he collected from others to inform his research. The fact that many of these maps appear in numerous photocopied quantities does point to the possibility that they were for distribution, to be shared with students or fellow scholars.

Beyond these two anomalous formats the archives are, as mentioned, primarily made up of textual records. Of the over 400 folders within the collection, around a quarter contain handwritten documents. These largely appear to be in Diffloth's own hand. These handwritten records can be further divided into three general categories: Diffloth's research notes and word lists, early drafts of his books and articles, and

1 View the finding aid for this collection: https://rmc.library.cornell.edu/EAD/htmldocs/RMM06313.html

2 Browse this collection by selecting 'Gérard Diffloth' under the 'Collection' tab of the Center's 'Advanced Search' page, https://library.khmerstudies.org/cgi-bin/koha/opac-search.pl

3 Gérard Diffloth papers., Undated. Division of Rare and Manuscript Collections, Cornell University Library.

correspondence. The research notes are arguably the most intriguing of these genres, though archival researchers should be warned that they may be the most difficult to parse. Diffloth took notes on a wide ranging assortment of papers. Many word families are meticulously organized on reams of graph paper. Yet other notes are spread across the backs of academic office work detritus (memos and lecture advertisements), unbound slips of small scratch papers, individual sheets of hotel stationary, the unused space of pages from a day calendar, slightly crumpled napkins, and in one especially charming instance, the actual back of an envelope.

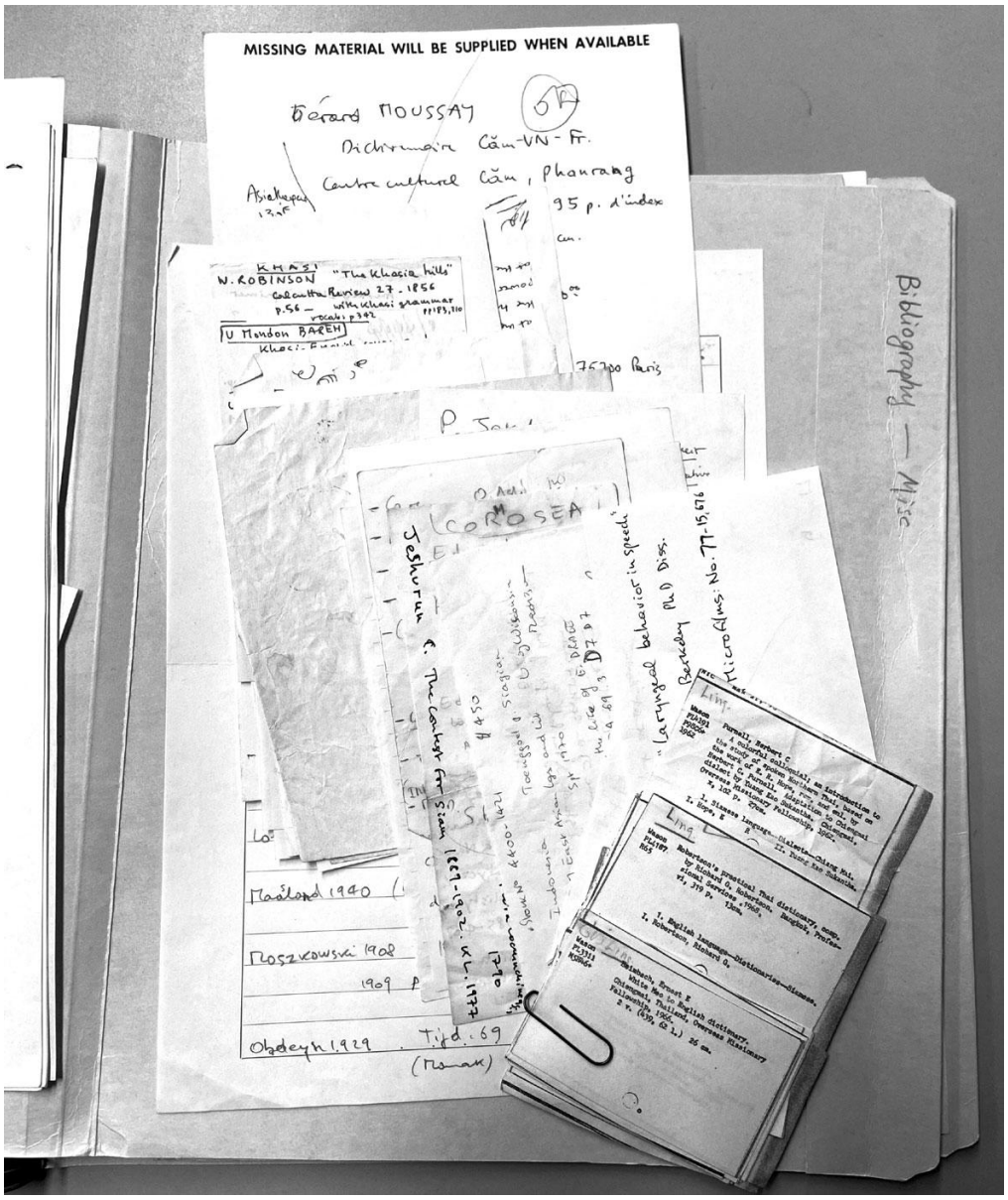


Figure 1: Diffloth's notes, Division of Rare and Manuscript Collections, Cornell University, Box 6, Folder Bibliography - Misc

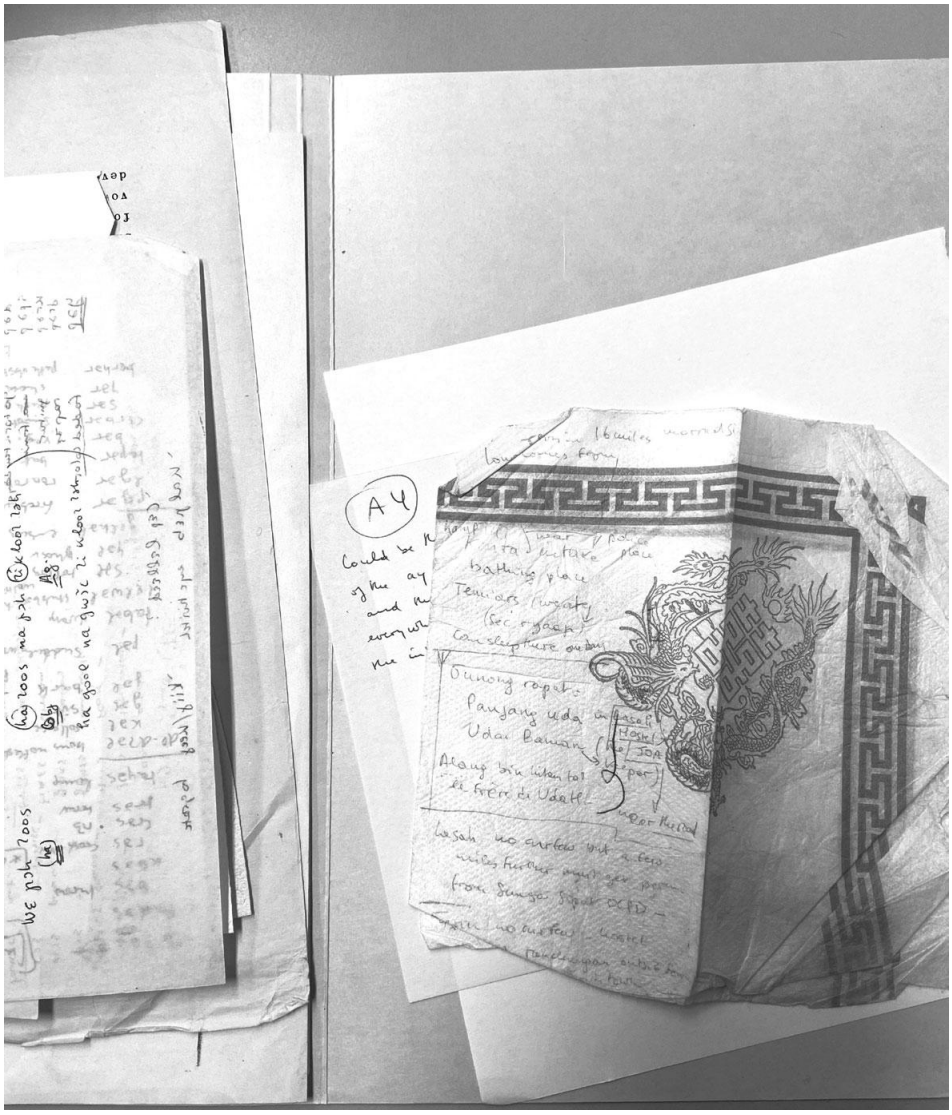


Figure 2: Diffloth's handwritten notes, Division of Rare and Manuscript Collections, Cornell University, Box 3, Folder Miscellaneous
(Photo Credit: Emily Zinger)

What is clear from this complex body of notes is the insistent rate at which Diffloth's mind worked. His research did not stop when he left his field work. From the diversity of papers used as vehicles for his thoughts there appear to be few moments or places where he was not reaching for a blank piece of paper to jot down a realization or question. It is this wealth of documents, not to mention allegiance to the analog, that makes for such rich archives.

The word lists include compilations of terms in Tausug recorded in 1964; verb affixes of Surigaonon from 1969; and words from Waic languages for made objects and verbs, among many others. Altogether, over 40 languages are referenced in the collection, additionally including Akha, Andamanese, Bahnar, Bai, Burmese, Chamic, Chechen, Gayo, Hani, Hanunoo, Hawaiian, Irula, Jiamao, Javanese, Khmer, Khmu, Lahu, Lament, Lao, Madurese, Malay, Mon-Khmer, Naxi, Nicobarese, Palauan, Ponapean, Portuguese, Semai, Taiwanese, Tamil, Temiar, Thai, and Vietnamese. This list, though extensive, even still does not include the numerous dialects and language families that Diffloth read or wrote about during the period covered in these papers.

The most in-depth word lists in the collection are those of Mon-Khmer terms. One miscellaneous example of a handwritten list of words in old Mon-Khmer includes terms for malaria, banana, harvest rice, mosquito, spleen, cold, and sleep. Other folders include drafts of a compilation of Mon-Khmer terms, though here we only have evidence of Part I: Nouns. Included in this portion are sections for fauna, flora, anatomy, society, made objects, natural objects, numerals, pronouns, deictics, and directions. Of those sections, the collection contains the most documentation of the fauna words and multiple drafts of this chapter can be found. Some drafts are handwritten, implying earlier stages of work. A subsequent draft has been copied with Diffloth's own annotations transcribed in the margins of his work. In another typewritten draft of the same chapter words and their meanings have been physically cut out and re-pasted elsewhere into the document providing a tangible map of Diffloth's editing process. Cleaner typed drafts are present as well, completing the picture with a more final stage of the research project.

For much more of Diffloth's work, final products cannot be found in this collection. Only the research which informed those absent publications is present. There are hundreds of pages of words, sometimes portioned out with only one or two lines per page, segmented into drawn boxes and transcribed using a complex system of differently colored inks. In his obituary of Diffloth, Nathan Badenoch paints a picture of the researcher entering the field armed with "a rainbow of colored pens."[4] Here in the archives we see the variated notes born from that rainbow. Unfortunately, the notes have no corresponding key as to the meaning of each color. Perhaps a scholar familiar with the transcribed languages would glean useful insights from decoding Diffloth's system.

Diffloth's collection is not limited to his own research. Numerous folders are full of pre-prints and photocopies of publications written by his colleagues. Some of these include marginalia in Diffloth's hand. Many are publications from conferences that it appears Diffloth attended such as the International Conference on Sino-Tibetan Languages & Linguistics in 1988 and the Symposium on Categorization and Noun Classification in 1983, charting his scholastic travels across the globe. Many others have handwritten notes in their top corners written by the authors for Diffloth himself: "Best wishes" abound, though my personal favorite is signed "From one trickster to another."[5] Reviewing the articles that Diffloth not only received, but felt were important enough to retain and include in his archival donation fills in the intellectual landscape among which his own in-progress research belonged.

Diffloth's central presence in an international circle of linguists and Southeast Asianists is clear not only from these articles, but from his correspondence. There are missives concerning research questions and upcoming conferences that read with a familiarity implying long standing and warm professional relationships. There are also a handful of letters from students, clearly nervous to be reaching out to Gérard Diffloth, one of the "famous researchers in our field," as described by one correspondent, ever thankful for the advice he gave them about their work and careers in his responses

4 Badenoch, Nathan (2023): In Memoriam: Gérard Diffloth (1939-2023). *Journal of the Southeast Asian Linguistics Society* 16(2). i–iv.

5 Matisoff, James A. Trickster and the Village Women: a psychosymbolic discourse analysis of a Lahu picaresque story. Box 5, Folder Lahu. Division of Rare and Manuscript Collections, Cornell University.

absent from this collection.[6] “How kind you are to an unknown student,” one petitioner says.[7] Several letters discuss the particular difficulties of researching Southeast Asian languages. One colleague described their inability to collect the names of fish because the river where she was conducting field work in Malaysia had been dynamited upstream and there were no longer any fish in the waters.[8]

Figure 3. Thai stamps sent to Diffloth from a colleague, Division of Rare and Manuscript Collections, Cornell University, Box 5, Folder Collection of Thai Stamps (Photo Credit: Emily Zinger)

Twined throughout these research updates there are also touchingly personal notes. One such message begins, “Let this letter carry across the Pacific Ocean my heart-felt

6 Lindell, Kristina (1981): Kristina Lindell to Gérard Diffloth. Box 3, Folder Letters - Diffloth - Lindell. Division of Rare and Manuscript Collections, Cornell University.

7 Charoenma, Narumol (1980): Narumol Charoenma to Gérard Diffloth. Box 1, Folder Letters - Gérard Diffloth 2. Division of Rare and Manuscript Collections, Cornell University.

8 Couillard Afendras, Marie-Andree (July 30): Marie-Andree Couillard Afendras to Gérard Diffloth. Box 1, Folder Letters - Gérard Diffloth 2. Division of Rare and Manuscript Collections, Cornell University.

greetings and expressions of joy to you, an able co-worker in the same (long-neglected) field!"[9] Other letters also speak to this sense of the smallness of the community of Southeast Asian linguistics. "Thank you very, very much for your letter," begins a colleague in the field. "Every line from the outside world fills us with unspeakable joy, for we are very isolated here, and in the darkest moments we sometimes think that perhaps we are the only three people in the world who care."[10] In a particularly moving letter, a friend describes their uncertainty about whether to seek out a grant or start a family instead. "Being a mother is as good as being a linguist," she quips.[11] Not all personal letters are so weighty. In one a correspondent pledges to seek out stamps for Diffloth's son's collection if only he will send some back in return. Together, the letters, articles, and conference agendas sketch the cross-institutional academic network of Southeast Asianists among which Diffloth was a prominent node.

An article saved by Diffloth in his papers, *Suprasegmentals in Southeast Asia* by Paul K. Benedict, opens with a quote from J. A. Matisoff (who was himself a frequent correspondent of Diffloth), "In this field you have to be a little crazy."[12] It could be argued that Diffloth's archives reflect this craziness–boxes full of folders full of scraps of paper upon which research is scrawled in competing colors and directions. Yet within this impression of craziness lies a picture of a dynamic and curious mind, one constantly in motion. Though these records are an imperfect representation of Diffloth's esteemed career, they remain a valuable part of his legacy and, of course, they are available for consultation. I invite you to come to Cornell, explore these documents yourself and build upon that legacy. Visit rare.library.cornell.edu to schedule an appointment.

9 Suen, Jackson T. S. (1983): Jackson T. S. Suen to Gérard Diffloth. Box 3, Folder P'uman. Division of Rare and Manuscript Collections, Cornell University.

10 Lindell, Kristina Kristina Lindell to Gérard Diffloth. Box 3, Folder Letters - Diffloth - Lindell. Division of Rare and Manuscript Collections, Cornell University.

11 L., T. (1979): T.L. to Gérard Diffloth. Box 1, Folder Letters - Gérard Diffloth 2. Division of Rare and Manuscript Collections, Cornell University.

12 Benedict, Paul K (1992): Suprasegmentals in Southeast Asia. In M. Ratliff & E. Schiller (Hrsg.), *Papers from the First Annual Meeting of the Southeast Asian Linguistics Society*, 15–33. Arizona State University, Program for Southeast Asian Studies.

4

Proto-Vietic Glottal Features in Kri

Gérard Diffloth[1]

A casual glance at the Kri lexicon, as we presently know it,[2] shows that some words have a puzzling resemblance with those of Vietnamese, even when that language is culturally remote and not mutually intelligible with Kri. Some examples:

	Kri	**Việt**
'leaf'	sula̲:ʔ	lá
'to have pain'	katʊ̲:	đau
'inside'	kulɒ̲:ŋ	trong
'tongue'	la̲:ɛ	lưỡi
'be warm'	təʔɪ̲̆mʔ	ấm
'body-louse'	brɪ̲̆ŋʔ	rận

But, having no other guideline than ill-defined resemblances, these possible cognates and others only suggest historical anecdotes. Such hints are neither better nor worse than those provided for instance by the Swadesh-Grey lexicostatistical techniques.[3]

By contrast, the discipline of comparative-historical phonetics operates in a far more reliable and fertile dimension of historical knowledge. Short of a more

[1] Editor's note: this paper was submitted to Nick Enfield in 2018 as part of a collaborative project on the Kri language. Nick posthumously offered it for inclusion in this volume, as it provides an appropriate context for publication.

[2] At the time of writing, I have a total of 2782 Kri lexical items entered in an Austroasiatic etymological database; most of these Kri items were collected in the field by Nick Enfield, and some were collected by myself in joint sessions with him in Laos. Notes on Kri and Ahlau are represented here in systematic phonetics; in this type of notation, no systematic contrast is left unmarked: for instance, if vowel duration is contrastive, long vowels are marked: [v:] as well as short vowels: [v̆]. For the final glottal feature, glottalisation is marked, e.g.: [-mʔ], as well as the absence of it, e.g.: [-m∅]. For voice-register, breathy voice is marked: [v̤], as well as clear (modal) voice: [v̲] which is otherwise nearly always left unmarked. One advantage of the principle: 'leave nothing unmarked', is to facilitate electronic searches; current programs are usually not very smart at finding unmarkings, even when these have all been purposefully entered as empty spaces: they don't distinguish easily one kind of empty space from another. In the case of Vietnamese Quốc Ngữ orthography, the ngang tone has been left unmarked, as in spelling tradition.

[3] Harry Hoijer's critical comments about Glottochronology, published in 1962, are sufficient even today for denying a privileged place in historical linguistics to Glottochronology and kindred techniques. And I take this occasion, more than fifty years later, to praise the memory of Harry Hoijer, an inspiring teacher who also knew how to be demanding and patient with a first-year student in Linguistics at UCLA.

comprehensive account of Kri linguistic prehistory, I will present here some crucial sound-correspondences between Kri and two other Vietic languages, Ahlau and Vietnamese.

Kri Rime-glottalisation

As described elsewhere (Enfield and Diffloth 2009) the Kri language has a well established phonological contrast in live syllables[4] between glottalised and plain finals. Specifically, Kri has words ending in:

[-mʔ], [-nʔ], [-ɲʔ], [-ŋʔ], [-rʔ], [-lʔ], [-wʔ], or [-jʔ]

as well as words ending in:

[-m], [-n], [-ɲ], [-ŋ], [-r], [-l], [-w], or [-j]

This feature of glottalisation does not appear in Kri words ending in voiceless continuants (/-s/ and /-h/), or in those with any of the Kri oral stops (/-p/, /-t/, / -c/, /-k/).

Kri words ending with a lone glottal stop also contrast with those having open final syllables. But distinct correspondence patterns are involved here; they pertain to a history of Vietic and Austroasiatic vowels, including vowel-length and diphthongs, that is beyond our scope here.

The glottalisation feature seen in the Kri live syllables mentioned above is not obviously dependent on any other feature of Kri phonology, or even on any morphological or syntactic pattern that we can detect; it is not a feature which can be explained away by some secondary conditioning.

It is also widely distributed throughout the Kri lexicon. In a total collection of 1424 Kri words having live syllables, 568 have glottalisation, (39.9%), and 856 do not.

This is a central feature of the Kri language, and it demands a historical account.

Kri rime-glottalisation compared with Ahlau and Vietnamese

Rime-glottalisation has already been mentioned in previous historical studies of Vietic languages. For example, Ferlus (2014) has provided a summary of the history of Vietic (his 'Viet-Mươ̛ng') which includes this feature; and Diffloth (1989) had proposed to reconstruct it even further back in the Austroasiatic past, as 'creaky voice', beyond the kind of glottalisation he saw in the Vietic branch.

Here, I will present Kri and Ahlau cognates with phonetic notations of glottal features, plus Vietnamese cognates in standard Quốc Ngữ spelling. The purpose is to show that the phonetic quality and lexical wealth of the new Kri data confirm a reconstruction of rime-glottalisation in Proto-Vietic. This will in turn affect future reconstruction of other parts of the Proto-Vietic language, such as phonation types

[4] In Asian tone languages, the term 'live syllables' refers to syllables that are either open or ending in a voiced continuant; in Kri these continuants are the Nasals: -m, -n, -ɲ -ŋ, the Liquids: -r, -l, and the Approximants: -w, -j. Excluded from this live class are syllables ending in voiceless Continuants, in Kri: -h and -s. By contrast, the 'dead syllables' end in Stops; in Kri these are -p, -t, -c and -k. Syllables ending in a Glottal Stop are given a distinct historical status in Kri.

(voice register) and the rich vowel system.

Ahlau is the autonym for a Vietic language originally spoken in a village located on the banks of the Vưng river, in Kamkeut district, Bolikhamxai province of Laos. The name of this village: Thavưng or Tha Vưng (in Lao 'wharf at the Vưng River') has often been used, confusingly, as the name of one of the Vietic languages spoken there. Ahlau is spoken in what is claimed by its speakers as the original part of the large village called Thavưng. There are other closely related Vietic languages in the same village area, for example Ahau, and also Aheu /ʔahɔ:/.[5] The So-Thavung language cited in Premsirat (2000) is originally from that wider community; but, unlike Ahlau, it is in the process of becoming tonal.

Kri, Ahlau, and Vietnamese cognates with final Nasals.

	Kri	**Ahlau**	**Việt**
	-mʔ	**-mʔ**	**-m**
1) 'eight'	sa̲:mʔ	sa̲:mʔ	tám
2) 'blood'	ʔasa̲:mʔ	pasa̲:mʔ[6]	
3) 'to lick'	ʔalɛ̲:mʔ	halæ̲:mʔ	liếm
4) 'thunder'	krɨ̆mʔ		sấm
5) 'to taste'	ɗe̲:mʔ		nếm
	-m∅	**-m∅**	**-m**
6) 'crab'	kata̲:m∅	kata̲:m∅	dam
7) 'bird'	cɛ̲:m∅	(ha)ci̲:m∅	chim
8) 'sickle'	lia̲m∅		liềm
9) 'five'	ɗă̲m∅	dă̲m∅	năm
	Kri	**Ahlau**	**Việt**
	-nʔ	**-nʔ**	**-n**
10) 'to dive'	lă̲nʔ	lă̲nʔ	lặn
11) 'four'	pɔ̲:nʔ	po̲:nʔ	bốn
12) 'be ripe'	cɪ̲:nʔ	ci̲:nʔ	chín
13) 'nine'	cɨ̲:nʔ	cɨ̲:nʔ	chín
	-n∅	**-n∅**	**-n**
14) 'python'	klă̲n∅	tală̲n∅	trăn
15) 'to enter'	lɒ̲:n∅	lɒ̲:n∅	luôn
16) 'child'	kɒ̲:n∅	kɒ̲:n∅	con
17) 'earthworm'	palṵ:n∅	malṵ:n∅	giun

[5] These three autonyms: Ahlau, Ahau, Aheu, are variants of a word of Tai origin meaning 'what ?'; the Central Thai cognate: /ʔaraj/ is spelled with a 'may muan' letter indicating an older *-aɨ rime. The Ahlau words cited here were collected by myself on separate occasions several decades ago, in Thai refugee camps and in Laos; the speakers all came from the same village shown as Thavưng on the maps.

[6] This word is from the Aheu language, closely related to Ahlau.

	Kri	**Ahlau**	**Việt**
	-ɲʔ	**-ɲʔ**	**-n**
18) ‘to sell’	pa̲:ɲʔ	pa̲:ɲʔ	bán
19) ‘to borrow’	ma̲:ɲʔ		mượn
20) ‘husband’	kʊ̲:ɲʔ		
21) ‘male (animal)’	kɔ̲:ɲʔ		cún[7]
22) ‘navel’	basʊ̲:ɲʔ	pʰaɲu̲:ɲʔ	rốn
	-ɲ∅	**-ɲ∅**	**--n**
23) ‘to weave’	ta̲:ɲ∅		đan
23) ‘rib’	cirạ̈:ɲ∅		sườn
25) ‘to sniff’	hɔ̲:ɲ∅	hɒ̲:ɲ∅	hôn
26) ‘be tasty’	taŋɔ̲:ɲ∅	taŋɒ̲:ɲ∅	ngon
27) ‘ashes’	ɓụ̈:ɲ∅	bụ̈:ɲ∅	mun
	Kri	**Ahlau**	**Việt**
	-ŋʔ	**-ŋʔ**	**-ng**
28) ‘sole, palm’	kapa̲:ŋʔ	kapa̲:ŋʔ	váng
29) ‘to roast’	ʔada̲:ŋʔ	hada̲:ŋʔ	nướng
30) ‘internode’	plɒ̲:ŋʔ	palɒ̲:ŋʔ	lóng
31) ‘be bitter’	tă̲ŋʔ	tă̲ŋʔ	đắng
32) ‘to stand’	tɨ̈ŋʔ	tɨ̈ŋʔ	đứng
	-ŋ∅	**-ŋ∅**	**-ng**
33) ‘lid’	kərpạ̈:ŋ∅	kapạ̈:ŋ∅	vàng
34) ‘bone’	sạ̈:ŋ∅		xương
35) ‘flower’	pɔ̲:ŋ∅	po̲:ŋ∅	bông
36) ‘tooth’	kasă̲ŋ∅	kasă̲ŋ∅	răng
37) ‘ginger’	cikə̤:ŋ∅	cakɔ̤:ŋ∅	gừng

Kri, Ahlau, and Vietnamese cognates with final -j

	Kri	**Ahlau**	**Việt**
	-jʔ	**-jʔ**	**-i**
38) ‘be far’	caŋa:jʔ	caŋạ̈:jʔ	ngái
39) ‘salt’	ɓɒ̲:jʔ	bɒ̲:jʔ	muối
40) ‘smoke’	kuhɒ̃:jʔ	kahɒ̲:jʔ	khói
41) ‘bamboo-rat’	citụ̈:jʔ		dúi
	-j∅	**-j∅**	**-i**
42) ‘ear’	sa̲:j∅	sa̲:j∅	tai
43) ‘house-fly’	murɒ̲:j∅		ruồi
44) ‘to stink’	hɔ̲:j∅	ho̲:j∅	hôi
45) ‘tail’	tɔ̤:j∅	tɔ̤:j∅	đuôi

[7] Editor’s note: cún in Vietnamese generally means ‘puppy’ (unmarked for gender).

Kri final liquids

Proto-Vietic had two final liquids: *-r and *-l.
The Proto-Vietic contrast *-r vs. *-l is now lost in most current Vietic languages. Only Kri, Maleng Bro, and some varieties of Phong have preserved it, to our knowledge. These languages apparently constitute together a historical sub-branch of Vietic. But some varieties of the Rục language on the Vietnamese side of the border have also kept a final /-ɹ/ distinct from final /-l /. The reconstructions of proto-Vietic final *-r and *-l is certain, but the cognates and the precise role of glottalisation with these finals will require more information than we have at present. But we do know that Ahlau has merged the two finals with its only final liquid, /-l /.

In Kri, the *-r vs. *- l contrast is fully preserved, either plain or glottalised. In the Kri data collected, I see 101 examples of final /-r/, vs. 74 of final /-l/, and 65 examples of final /-rʔ/ vs. 59 of final /-lʔ/. The words with final glottalised -rʔ represent 38.4 percent of the total number of words with either kind of final -r, almost exactly the percentage we found with final nasals and with final -j. The profile with final /-lʔ/'s, 44.3 percent of all -l's, seems a little out of line, but still within the same range.

In any event, the glottalisation feature is retained in both Kri and Ahlau, independently from these mergers of final liquids.

Proto-Vietic rime-glottalisation

The correspondences shown above indicate that rime-glottalisation in live syllables is a feature of the proto-language from which Kri and Ahlau historically descend. As for Vietnamese, there are also correspondences between its tones and the glottal features of Proto Kri-Ahlau. These have direct implications for the reconstruction of Proto-Vietic.

In the above cognate sets, and in many others not cited here, Vietnamese has the tones sắc (v́) or nặng (ṿ) when the cognate Kri-Ahlau words have glottalised rimes, and the tones ngang (v) or huyền (v̀) when the Kri-Ahlau cognates do not.

The rich literature on Vietnamese tonogenesis, going back to Haudricourt (1954) and even earlier, generally agreed that there were two sets of factors at work here: one was the historical devoicing of the Proto-Vietnamese initials, the other was the Stop vs. Continuant or Open feature of the Proto-Vietnamese finals.

It is the second set of Vietnamese tones: sắc and nặng, which directly concerns us here. Early on, these tones were seen as belonging to syllables ending in Stops, the 'dead' syllables of Chinese historical tradition. But it was common knowledge that Vietnamese also had numerous other words with Continuant finals, notably with the Nasal finals, that strangely had one of these two tones, sắc or nặng. These exceptions were often overlooked or seen as marginal, partly because Mon-Khmer languages were not known at the time to have anything like glottalised final nasals or other features that could explain such Vietnamese tones.

Then, phonetically precise data on Ahlau ('Thavưng') rime-glottalisation began to be known, and I showed in Diffloth (1989) that other Austroasiatic languages, namely those of the Pearic branch and some in the Katuic branch (e.g. Talan, Ong) also had these glottal features. Rime-glottalisation could then go back not only to Proto-Vietic, but possibly also to much earlier Proto-Austroasiatic times, in the guise of a reconstructed creaky voice.

To my knowledge, nothing substantial has been published since on this problem. In fact, a recent work on Katuic history simply evacuates the reports of rime-

glottalisation in the ‘Ong/Yir/Talan’ group of Katuic languages (Sidwell, 2015, pp.12-14).

Returning to Vietnamese, the correspondences shown above confirm that the unexpected sắc and nặng Vietnamese tones found in live syllables have a ready historical explanation as regular reflexes of the Kri-Ahlau glottalised live rimes. Rime-glottalisation can then easily be reconstructed to Proto-Vietic times.

Not only this, but the perfectly steady ratio of live rime glottalisation In Kri and in Ahlau (39.9 % in a total of 1424 relevant words in Kri, and 39.8% in a total of 1014 relevant words in Ahlau) which is itself quite striking, matches almost perfectly a similar count in the Vietnamese language (39.3 % of words with glottalised (sắc or nặng) live rimes in a total of 628 relevant Vietnamese items). These figures indicate that the proto-Vietic language itself had a similar ratio of glottalised live rimes in its own lexicon, and that, by sheer luck, no historical innovation has intervened later that would have altered the statistical profiles of the descendant languages, at least regarding this feature of live rime glottalisation.

Far from being bizarre or exceptional, many words resembling Kri with their glottalised rimes should be expected when we imagine hearing Proto-Vietic conversations exchanged many centuries ago.

On the question of Proto-Vietic phonation-types

The correspondences shown above reveal something else about the sounds of Proto-Vietic. Both Kri and Ahlau have a contrast between breathy voice and clear (modal) voice in the vowels of word-final syllables. In both languages, clear voice vowels are definitely more frequent lexically than breathy vowels: in Kri, in a total of 2771 words, only 888, (32,0 %), have breathy vowels, in Ahlau, in a total of 1997 words, only 727 (36,4 %), have breathy vowels. We notice a somewhat different percentage profile among the two languages. When looking only at cognates between the two, we notice 51 cases of disagreement regarding phonation type, (10.5%) in a total of 485 Kri-Ahlau cognates. Not a large figure, but we are far from the near-perfect alignments we just saw in the case of live rime glottalisation.

In many but not all, Mon-Khmer languages, this clear vs. breathy phonation-type contrast is the result of a devoicing of initials, whereby earlier voiced initial onsets gave rise to a breathy phonation in the vowel that followed. But even a rapid inspection shows that such devoicing, even when it did occur, cannot be the whole story of voice-register contrasts in Kri-Ahlau. For Example the word for ‘nine’: /cn̤:nʔ/ (Example No.13) has a breathy vowel in both languages, whereas cognates in the rest of Austroasiatic regularly show an older voiceless *c- or *t- in the onset. Parallel cases can be seen with Kri: /tn̤ŋʔ/ ‘to stand up’ (No.32), also with Kri: /tṳ:jʔ/ ‘bamboo-rat’ (No.41), both having cognates with a historical voiceless *t- onset. There are a number of similar cases among Kri-Ahlau etyma not cited here.

To make things worse, a glance at the Vietnamese cognates given above shows further disagreements in voice-register history. As is well known, the Vietnamese tone series: ‘ngang / sắc / hỏi’ is conditioned by historically *voiceless onsets, while the tone series ‘huyền / nặng / ngã’ is conditioned by historically *voiced ones, with counter-patterns due to certain complex onsets.

But the Kri-Ahlau word for ‘nine’: /cn̤:nʔ/, with its breathy voice, has a Vietnamese cognate: chín (No.13) with a sắc tone. This tone is a regular outcome for Vietnamese since MK cognates have a *voiceless onset in this word. The same thing can be said for

the Kri word /tɨ̤ŋʔ/ ‘to stand’ (No.32) having a Vietnamese cognate: đứng with a sắc tone which is also regular considering its MK cognates; the word ‘bamboo-rat’ (No.42) with a Vietnamese cognate: dúi may present the same paradox. We see here a historical dissonance between the registers of Kri-Ahlau and the apparently regular tones of Vietnamese.

Clearly, there are historical patterns here that look different from and additional to the classical Mon-Khmer schema of devoicing for registro- genesis.

A hint of a possible solution is seen in examples No.12 and 13. The Kri word /cɪ̤:nʔ/ ‘be ripe’[8] (No.12) contrasts in voice-register with the Kri word /ci̤:nʔ/ ‘nine’ (No.13) discussed above; and yet the Vietnamese cognates are identical: chín for both. Cognates outside Vietic have a *voiceless *c- in the onset for both words,[9] showing that the Vietnamese sắc tone is in fact a regular outcome for both. Therefore, the source of the occasional disagreement between Kri-Ahlau registers and Vietnamese tones most likely resides in the vowel itself. Other register dissonances are often, but not always, found in words with Vietic high vowels; therefore, I suspect that proto-Vietic vowel-height may be a decisive factor here. The Bru language (West-Katuic) has also seen similarly conditioned register-shifts in its own history (Diffloth, 1982); and Bru is influential in the Kri-Ahlau area. But a full reconstruction of the history of Vietic vowels remains a task for the future; and the very rich history of Katuic vowels has barely been scratched so far.

Register dissonances seen within Vietic should be examined in the context of the far more ancient landscape of Proto Vieto-Katuic reconstruction.

References

Diffloth, Gérard. 1982. ‘Registres, dévoisement, timbres vocaliques: leur histoire en Katouique’. *Mon-Khmer Studies* 11, pp.47-82.

Diffloth, Gérard. 1989. ‘Proto-Austroasiatic creaky voice’, *Mon-Khmer Studies* 15, pp.139-154.

Enfield, N. J., and Gérard Diffloth, 2009. ‘Phonology and Sketch Grammar of Kri, a Vietic Language of Laos.’ *Cahiers de Linguistique - Asie Orientale* 38.1, pp.3–69.

Ferlus, Michel., 2001, rev. 2014. ‘The origin of tones in Viet-Muong’. Somsonge Burusphat. *Papers from the Eleventh Annual Conference of the Southeast Asian Linguistic Society*. Tempe, Arizona, pp.297-313.

Haudricourt, André-Georges. 1954. ‘De l’origine des tons en vietnamien’, *Journal Asiatique* 242, pp.68-82.

Hoijer, Harry, 1962. ‘Linguistic subgroupings by glottochronology and by the comparative method’, *Lingua* 11, pp.192-8.

Premsrirat, Suwilai. 2000. *So (Thavung) preliminary dictionary*. Melbourne, 189 pp.

Sidwell, Paul. 2015. *The Katuic languages, Classification, Reconstruction and Comparative Lexicon*. Lincom Europa.

[8] The meaning ‘be ripe’ is a semantic shift specific to Vietic in an etymon which means ‘be cooked’ in most Austroasiatic languages, for example Modern Mon: /cin/ ‘be cooked’.

[9] Most AA languages have a *c- and *s- in this word. The complex onset of the Khmer cognate /cʔən/ is also *voiceless, and the unexpected medial /-ʔ-/ is probably a feature inherited from the Pearic substratum of Old Khmer.

5

A Reconstruction of Proto-Ta'oi Phonology and Lexicon with a Focus on the Origins of Rime Laryngealization Contrasts

Ryan Gehrmann

1 Introduction

Diffloth's (1982) *Registres, devoicement, timbres vocaliques: Leur histoire en Katouique* (Registers, devoicing and vowel qualities: Their history in Katuic) was a landmark paper. Not only did Diffloth present the first accurate reconstruction of Proto-Katuic vocalism, which has only been marginally improved on in the intervening four decades, but he did so based on just 138 lexical comparanda. However, this is not the paper's most remarkable contribution. Through his investigation of the historical development of the vowels and registers of Pacoh, Diffloth was the first to describe an alternative registrogenetic process, quite different from the more widely distributed and better-researched *Khmer Model* of registrogenesis. While the latter model involves the gradual transphonologization of onset phonation contrasts as a register contrast and then, ultimately, as new vowel quality contrasts, Diffloth's *registrogénèse hérétique* (heretical registrogenesis) described the opposite. Diffloth shows that Proto-Katuic vowel quality contrasts became transphonologized into a register contrast in Pacoh. This discovery inspired others to look for parallel registrogenetic processes in other Austroasiatic languages, and further examples have since been described (Sidwell 2015, 2019; Gehrmann 2015, 2022a).

The significant achievements of this paper notwithstanding, there was another atypical register language among the Katuic languages which the paper did not address for lack of data: Ta'oi. Only later in Diffloth's (1989) article *Proto-Austroasiatic Creaky Voice,* does he offer analysis on the thorny problem of Ta'oi registrogenesis. This article addresses the register contrast in a Ta'oi dialect spoken in Talan Village, which Diffloth simply calls *Talan*. A contrast between modal and creaky/laryngealized voice quality is found on Talan rimes, and Diffloth characterizes this as a register contrast. Diffloth confirms and improves slightly upon Ferlus's (1974) earlier analysis of register in Ong, another named Ta'oi variety very closely related to Talan. He then turns his attention to the ultimate origin of the Ta'oi register contrast. He dismisses the possibility first raised by Ferlus that Ta'oi register might have been conditioned by Proto-Katuic onset voicing, showing that there is no correlation between the two. He then attempts to establish correspondence between the registers of Pacoh and the registers of Ta'oi but is unable to do so. In the end, he resorts to reconstructing the Ta'oi modal-creaky contrast back to Proto-Katuic, stating that the latter must be reconstructed with vowels that occur in "two distinct registers: creaky and clear." (Diffloth 1989, 144).

In an earlier publication, I proposed an alternative interpretation to Diffloth's

Proto-Katuic retention hypothesis for Ta'oi register (Gehrmann 2015). Ta'oi registrogenesis was accomplished through the heretical registrogenesis model that Diffloth proposed for Pacoh, but registrogenesis in the two languages were separate, parallel events. Proto-Katuic vowel quality contrasts do correspond to Ta'oi register differences, just like Pacoh, but different Proto-Katuic vowels developed into different modern vowels with different register assignments in many cases in modern Ta'oi and Pacoh. Diffloth was thus very close to solving the problem when he compared Talan and Pacoh register, but the direct comparison of the two registers rather than their Proto-Katuic vowel antecedents led to a null result.

In this paper, I present a more comprehensive description of Proto-Ta'oi's vowel height-conditioned registrogenesis, superseding the earlier description in Gehrmann (2015). After introducing the modern Ta'oi language and presenting a phonological description of a conservative variety, *Ta'oiq*, a phonological reconstruction of Proto-Ta'oi is presented. Supplementary materials are available online[1], including lexical evidence supporting the reconstruction of Proto-Ta'oi vocalism and register presented here and a lexical reconstruction of the language comprising approximately 1,200 words with accompanying Katuic comparanda.

2 The Ta'oi Language

Ta'oi is a Katuic language of southern Laos spoken primarily in Ta'oi and Tumlan districts of Salavan province and in isolated pockets in Champasak province.[2] A variety of glossonyms are employed by different Ta'oi speech communities. Within Ta'oi, variation in segmental phonology is mostly trivial, but there are significant differences related to the development or loss of the rime laryngealization contrast that is reconstructible to Proto-Ta'oi. Based on these differences, three primary varieties of the language are discernible in the available data: (1) Ta'oiq, (2) Ta'oi of Tha Taeng and (3) Ta'uas.

Among these, the Ta'oiq variety is the most conservative with respect to rime laryngealization. Ta'oiq subsumes a number of named doculects, including Ta'oiq [tth] (Conver et al. 2014; Gehrmann 2015), Ong [oog] (Ferlus 1974, 1979), Bru of Talan village (Huffman 1979a, Diffloth 1989), Ir [irr] (Huffman 1979b) and Katang (Ferlus 1974).[3] In Ta'oiq dialects, Proto-Ta'oi rime laryngealization developed two different reflexes under conditioning from historical coda type and vowel length, namely (1) **creaky voice** realized on the main syllable vowel or (2) **glottalized codas** resulting from the lenition of historical codas (see Section 3.2)

Ta'oi of Tha Taeng, described by L-Thongkum (2001), has no creaky-modal voice

---

1 https://doi.org/10.17605/OSF.IO/6BDQ5

2 In Vietnam, there are varieties of the Pacoh language that use a variant of the glossonym Ta'oi (e.g., Taôih as described by Nguyễn et al. (1986)). Moreover, the Vietnamese government's official designation for this Pacoh-Ta'oih group of Katuic speakers is Tà Ôi. The Pacoh-related *Ta'oih* language is not directly addressed in this paper, as it is not a part of the Ta'oi language of Laos as defined here (i.e., the group of language varieties that are directly descended from the reconstructed stage called Proto-Ta'oi).

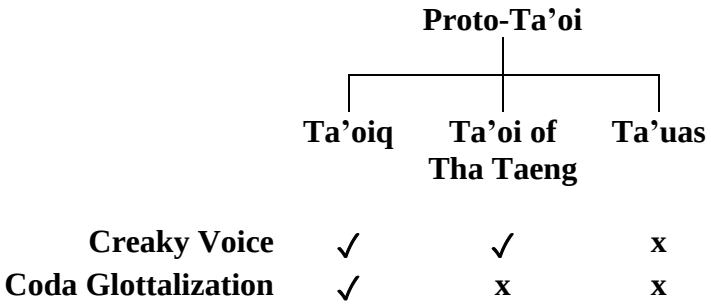
3 Note that the glossonyms *Bru* and *Katang* are usually associated with West Katuic languages of the Bru group. Bru, however, is often applied more broadly as a general designation for what the Lao call *Lao Theung* peoples; that is, Austroasiatic language speakers who live in upland areas. The name Katang is for all practical purposes shared by the southernmost Bru language speakers and the westernmost Ta'oi speakers, where they overlap in Toumlane and Ta-oi districts, Salavane province, Lao PDR.

contrast, but glottalized codas that are cognate with those found in Ta'oiq varieties are apparent. The implication is that Ta'oi of Tha Taeng and Ta'oiq both inherited Proto-Ta'oi rime laryngealization, but Ta'oi of Tha Taeng underwent a conditioned, partial neutralization of the contrast, resulting in the loss of the creaky-modal contrast.

Ta'uas, also called Ta'oih, is described briefly by Haak (1993), and unpublished lexical data is also available thanks to the efforts of Haak et al. (nd) and Ferlus (nd). Ferlus refers to the language as *Ta-oy*, but it is clearly the same variety as Haak's Ta'uas. This variety of Ta'oi is the most innovative of the three. It has neither creaky voice nor glottalized codas, but vowel correspondences that are entirely regular with the other two varieties argue that Ta'uas is another daughter of Proto-Ta'oi. Ta'uas has undergone a complete, unconditioned merger of the historical rime laryngealization contrast.

The reconstructability of Proto-Ta'oi rime laryngealization is on firm footing. There is no evidence to support the idea that creaky voice and coda glottalization in Ta'oiq and Ta'oi of Tha Taeng are conditioned, post-Proto-Ta'oi innovations. Rather, as will be demonstrated below, there is good evidence to explain their origins in a reconstructible Proto-Ta'oi rime laryngealization contrast that carries forward certain Proto-Katuic vowel quality differences (cf. Gehrmann 2015). Figure 1 summarizes the discussion so far.

Figure 1: Ta'oi classification



3 Modern Ta'oiq Phonology

In this section, a brief description of Ta'oiq phonology is presented. This variety is presented because (1) it is the most conservative variety, (2) it is the best documented variety thanks to Conver et al.'s (2014) unpublished lexicon and accompanying recordings.

3.1 Word & Syllable Structure

The phonological word may be monosyllabic or disyllabic. Disyllables are invariably iambic, comprising a prosodically prominent *main syllable* preceded by a prosodically non-prominent *presyllable*, which is comparatively restricted in terms of phonotactics and segment inventory. Monosyllables are structurally equivalent to main syllables. The maximal syllable template for Ta'oiq is presented in Figure 2.

Figure 2: Ta'oiq maximal syllable template

	C	{V/C}	C	C	V	C		
/	r	a	k	l	a	t	/	*opposite*
/	h	r	k	l	a	h	/	*to separate*

Main syllables are at minimum CV (e.g., /cạ:/ *to eat*, /nɔ:/ *this*), and an optional medial

liquid of /r/ or /l/ may appear in certain combinations (e.g., /triː/ *monitor lizard*, /klɰ̃ː/ *snail*). An optional coda may be included (e.g., /hɰ̃ːŋ/ *wasp*, /klɔːk/ *to plow)*. Vowel length contrast is neutralized in open syllables, where vowels are analyzed as long.

Presyllables are structurally deficient compared with main syllables. They must contain two segments, either /CV/ or /CC/, where the second segment constitutes the presyllable rime. /CV/ presyllables are realized as expected as [CV], and the vocalic presyllable rime /V/ is completely underspecified for vowel quality. It is typically realized as a short, mid to open central vowel and it is transcribed phonologically here as /a/.[4] Biconsonantal /CC/ presyllables can only have a liquid /r/ or /l/ as the second consonant; the only presyllables encountered so far are /hr, pr, tr, cr, kr, kl/. Presyllable /hr/ is pronounced as we would expect for a Katuic presyllable with an excrescent vocoid between the consonants (e.g., /hrblɛːʔ/ [hᵊrblɛːʔ] *to separate*). The others are realized with an excrescent vocoid after the liquid (e.g., /prlɔː/ [prᵊlɔː] *flame*) in an apparent metathesis between the liquid and the epenthetic vowel. This is unusual for a Katuic language, but it does not necessitate any change to the underlying representation of the presyllable as the syllabification of presyllables remains predictable.

Katuic languages typically allow a nasal consonant to fill the presyllable rime slot (e.g., Bru Tri /mntoːr/ [mᵊntɔor] *star*). These nasals are obligatorily homorganic to the main syllable onset and are reconstructible to Proto-Katuic (Sidwell 2005). Ta'oiq has retained prenasals, but they have been absorbed into main syllable onsets resulting in a consonant split into prenasalized and plain series. The necessity of such an analysis for modern Ta'oiq is made clear by the existence of words like /hrⁿtɰ̃ːj/ [hᵊrntɰ̃ːi] *to arrange in sequence*. In this word, a rhotic is filling the presyllable coda slot and the nasal is thus shown to be a feature of the main syllable onset. Ta'oiq speakers strongly favor this interpretation of prenasalized consonants as unitary segments (/ᴺC/), and the Ta'oiq orthography reflects this.[5] Further evidence in support of this analysis is found in the fact that voicing contrast is neutralized for prenasalized stops; this is an indication that the two formerly discrete segments are now conjoined and obliged to share a common phonation setting. Table 1 provides examples demonstrating the possible word shapes in Ta'oiq.

Table 1: Examples of possible Ta'oiq phonological word shapes

CV	/cɰ̃ː/	to eat	CVC	/moːt/	to enter
CCV	/ɟruː/	to be deep	CCVC	/klɰ̃ːŋ/	to pour
cv.CV	/kamɰ̃ː/	year	cv.CVC	/tapat/	six
cv.CCV	/raprɰ̃ː/	to hate (refl.)	cv.CCVC	/raklat/	to be opposite
cc.CV	/krnɰ̃ː/	road	cc.CVC	/prhoːt/	breath
cc.CCV	*no examples*		cc.CCVC	/hrklah/	to separate

3.2 Laryngeal Contrasts of the Ta'oiq Rime

Ta'oiq contrasts modal and non-modal phonation at two different phases of the rime. Rime-medially, there is unpredictable variation of modal voice and creaky voice, and rime-finally, there is a three-way phonation contrast: modal voicing, glottalization, and voicelessness. Table 2 provides lexical examples demonstrating the permissible combinations of rime-medial and rime-final phonation types. Notably, creaky voice

[4] The rationale for positing a phonological presyllable vowel in Ta'oiq and most other Katuic languages discussed in Gehrmann (2018) so I will not commit space to the issue here.

[5] Johanna Conver, Mackenzie Conver & Jonathan Schmutz, personal communication.

quality does not co-occur with final glottalization. The distribution of rime-medial voice quality, rime-final phonation and different coda manners of articulation is presented in Table 3, while Table 4 shows waveforms, spectrograms and F0 plots of representative example words demonstrating the five permissible combinations of rime-medial and rime-final phonation types.

Rime-final voicelessness will be considered a segmental coda /h/ here. In Ta'oiq, this voicelessness is only found directly following a vowel unlike in languages like Kri (< Vietic), which allows for voicelessness following vowels and sonorants (i.e. [V^{h} w^{h} l^{h} r^{h} j^{h}]) (Enfield & Diffloth 2009). Additionally, Ta'oiq shows unpredictable vowel duration contrast before rime-final voicelessness, strengthening the argument for a segmental coda /h/ (e.g., /lih/ *to untie* vs. /riːh/ *to choose*; /taməh/ *to ask* vs. /pəːh/ *to open*). The voiceless oral fricative /s/ is realized alternatively as a palatalized post-alveolar sibilant [ɕ] or as a debuccalized voiceless vocoid [ʝ̊] in free variation.[6] In either variant, it too behaves as a coda segment, just like /h/.

Table 2: Co-occurrence of rime-medial & rime-final phonation types with lexical examples

		Rime-Final					
		Voiced		**Glottalized**		**Voiceless**	
Rime-Medial	**Modal**	[tapaː]	*turtle*	[lampaːʔ]	*shoulder*	[tapah]	*to slap*
		[hampiː]	*vegetable*	[piːʔ]	*river mouth*	[piːh]	*poison*
		[plɔː]	*head*	[prɔːʔ]	*squirrel*	[ʔntrɔh]	*tree bark*
	Creaky	[mą̰ː]	*aunt*			[pą̰ːh]	*to cleave*
		[hʰrl̰iː]	*thorn*			[pḭːh]	*to sweep*
		[klɔ̰ː]	*snail*			[tʰmprɔ̰ːh]	*to clap*

Table 3: Distribution of laryngeal patterns and coda types

	Modal Voice *(medial)*			**Creaky Voice** *(medial)*		
	Voiced *(final)*	**Glottalized** *(final)*	**Voiceless** *(final)*	**Voiced** *(final)*	**Glottalized** *(final)*	**Voiceless** *(final)*
Open	V	Vˀ	Vʰ	V̰	-	V̰ʰ
Syllable	-	Vp Vt Vc Vk	-	-	-	-
Oral Stops	Vm Vn	Vmˀ Vnˀ Vɲˀ	-	V̰m V̰n	-	-
Nasal Stops	Vɲ Vŋ	Vŋˀ	-	V̰ɲ V̰ŋ	-	-
Approximants	Vw Vl Vj	Vwˀ Vlˀ Vjˀ	-	V̰w V̰l V̰j	-	-
Trills	Vr	-	-	V̰r	-	-
Fricatives	-	-	Vs	-	-	V̰s

[6] Reflexes of the Proto-Austroasiatic coda fricative *s behaves similarly across the Katuic languages and across the AA family (Sidwell & Rau 2015).

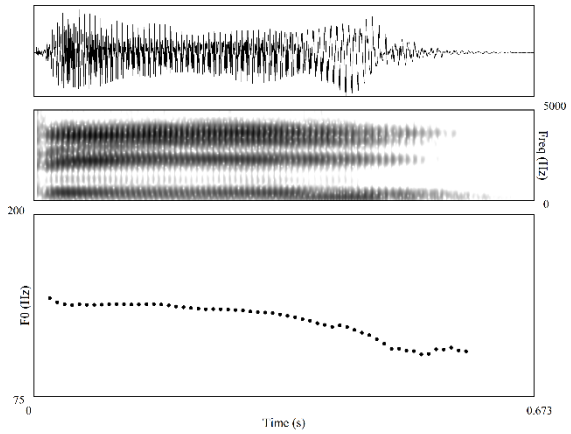
The glottalized rimes show a wider distribution vis-à-vis codas; rime-final glottalization may occur immediately post-vocalically or in combination with several different natural classes of coda, including stops, nasals and approximants. The oral stops are considered glottalized for two reasons: (1) on phonetic grounds, they are unreleased stops, co-articulated with a glottal occlusion, as is typical in East and Southeast Asia, and (2) they pattern phonologically with the other glottalized rimes in that they do not co-occur with rime-medial creaky voice. Given that final approximants may occur in modal+voiced, creaky+voiced or modal+glottalized rimes, it is tempting to simply consider these to be three syllable-level contrasts (e.g., /VC/-/V̰C/-/VCˀ/). However, vowel length contrast is expected to be neutralized in open syllables, and modal vowels with post-vocalic glottalization do show vowel length contrast (e.g., /taʔ/ *to do* vs. /taːʔ/ *iron*; /cɔʔ/ *to tie* vs. /cɔːʔ/ *head, source*). Consequently, as with coda /h/, rime-final glottalization behaves like a segmental coda /ʔ/ immediately following a vowel. We may extend this principle to glottalized coda sonorants as well. Rime-final glottalization is therefore interpreted as /ʔ/ or as a glottalized feature of sonorant codas here.

All three rime-final phonation types are accounted for through complex unitary coda segments. Voicelessness is interpreted as a final fricative /h/ or /s/. Glottalization is interpreted as a final glottal stop /ʔ/, a glottalized final sonorant /mˀ nˀ ɲˀ ŋˀ wˀ lˀ jˀ/ or an unreleased final oral stop with glottal reinforcement /p t c k/. Rime-final modal voicing is interpreted as an open syllable or as a syllable ending in a modal voiced sonorant coda /m n ɲ ŋ w l j r/.

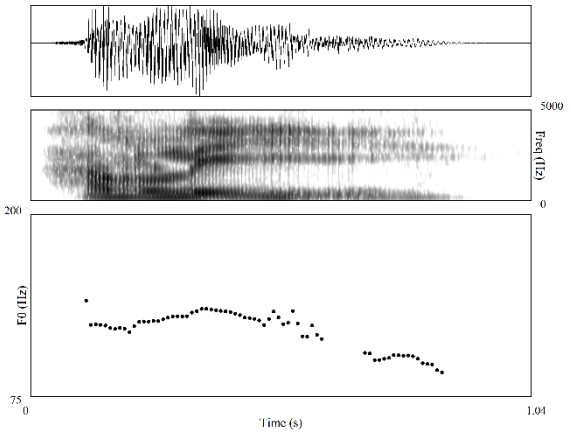
As for the two rime-medial voice qualities, these are potentially interpretable as either properties of the vowel or of the syllable. While proper instrumental acoustic investigation remains to be done, auditorily the creaky voiced rimes appear to be accompanied by a distinctly falling pitch contour (visible in Figure 3). In what follows, modal vs. creaky voice quality is marked on the main syllable vowel for Ta'oiq transcriptions, but my working hypothesis is that this is an unconventional type of register contrast, cued redundantly by voice quality and pitch contour differences (cf. Gehrmann 2022a, 2022b).

Figure 3: Examples of rime-medial and rime-final laryngeal differences

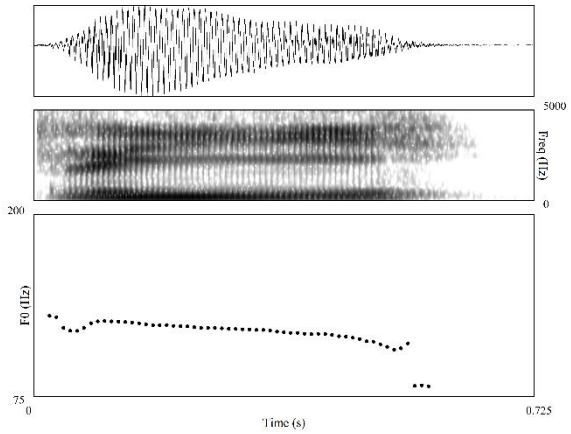
Modal Voice – Modal Voice
[tiː] hand



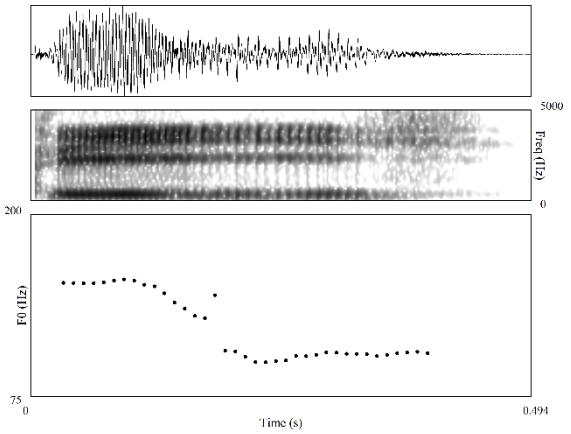
Creaky Voice – Modal Voice
[hᵊrl̰iː] *thorn*



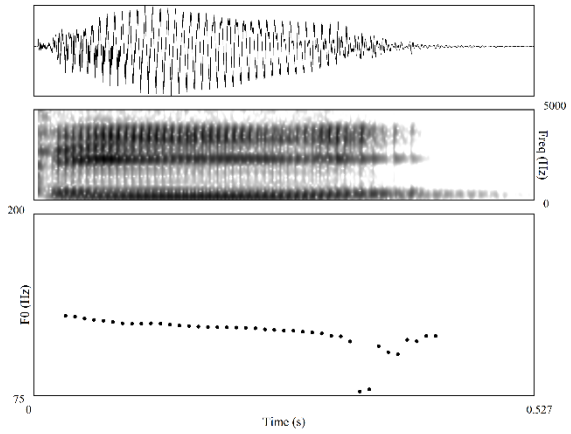
Modal Voice – Voicelessness
[riːh] to choose



Creaky Voice – Voicelessness
[t̰iːh] bland, tasteless



Modal Voice – Glottalization
[tiːʔ] old



3.3 Segment Inventory

Table 4 presents a succinct overview of Ta'oiq segmental phonology. The inventory of presyllable onset consonants is impoverished compared to the inventory of main syllable onsets. However, reduplicative morphology can produce additional presyllable onsets in morphologically complex words. The superscript nasals in parentheses on main syllable onsets indicate that these onsets are found with contrastive prenasalization, and the superscript glottal stops in parentheses on main syllable codas indicate that these consonants contrast post-glottalization. Any presyllable onset may co-occur with the presyllable vocalic rime /a/, but as noted above, only a subset may co-occur with a liquid rime (/hr, pr, tr, cr, kr, kl/). The licit tautosyllabic consonant cluster combinations are listed below simple main syllable onsets in this table. Note that voicing contrast is neutralized for prenasalized sounds; all are transcribed with voiceless stop or fricative components.

Table 4: Ta'oiq segment inventory

<table>
<tr><th colspan="2">Presyllable</th><th colspan="4">Main Syllable</th></tr>
<tr><th colspan="2">C</th><th>C/v</th><th colspan="2">C(C)</th><th colspan="2">V</th><th>C</th></tr>
<tr><td colspan="2" rowspan="4">p t c k

m ŋ
l
r
s h</td><td rowspan="4">r
a</td><td colspan="2">b d ɟ
br ɟr
bl
(m)p (n)t (ɲ)c (ŋ)k ʔ
(m)pr (n)tr (ɲ)cr (ŋ)kr
(m)pl (ŋ)kl
pʰ tʰ kʰ
m n ɲ ŋ
w l j
r
(n)s h
(n)sr</td><td colspan="2">Modal Voiced
ia ɨa ua
i: ɨ: u:
e: ə: o:
ɛ: a: ɔ:
i ɨ u
e ə o
ɛ a ɔ</td><td colspan="2">Creaky Voiced
ḭa - ṵa
ḭ: - ṵ:
ḛ: ə̰: o̰:
ɛ̰: a̰: ɔ̰:
ḭ - ṵ
- ə̰ -
ɛ̰ a̰ -</td><td rowspan="4">p t c k

m(ʔ) n(ʔ) ɲ(ʔ) ŋ(ʔ)
w(ʔ) l(ʔ) j(ʔ)
r
s h</td></tr>
</table>

4 Proto-Ta'oi Phonological Reconstruction

A reconstruction of the segmental phonology of Proto-Ta'oi is presented in this section. For the sake of brevity, the comparative data supporting the analysis below has been made available at the following DOI: (https://doi.org/10.17605/OSF.IO/6BDQ5).

This reconstruction is based on a phonological comparison of the Ta'oi doculects listed in Table 5. These data have been tabulated in my comparative lexical database of Katuic alongside numerous lexical data sets from other Katuic languages. The current number of Proto-Ta'oi lexical reconstructions is approximately 1,200. This reconstruction is part of a larger project aimed at refining the lexical and phonological reconstruction of Proto-Katuic. Undoubtedly, the Proto-Ta'oi reconstructions will undergo further revision, and the current lexicon is labeled as Version 2.

Table 5: Ta'oi lexical resources used in the Proto-Ta'oi lexical reconstruction

Ta'oiq	Ta'oiq	Conver et al. (2014)
	Ong	Ferlus (nd)
	Bru Talan	Diffloth (1989)
Ta'oi of Tha Taeng	Ta'oi Tha Taeng	L-Thongkum (2001)
Ta'uas	Ta'oi Salavan	Ferlus (nd)

4.1 Proto-Ta'oi Main Syllable Simple Onsets

Proto-Ta'oi main syllable onset consonants are straightforwardly reconstructible (see Table 6). However, onset stop voicing is marked inconsistently in the environment immediately following a nasal presyllable rime in all the Ta'oi sources available. This contrasts with other Katuic language descriptions and lexical sources, where a clear contrast of voicing is found in this environment. The Ta'oiq audio data available indicates a neutralization of voicing contrast for prenasalized stops, with all prenasalized stops being produced voiced in that variety. Taking this into account along with the inconsistent transcription of stop voicing in the other available sources, I interpret this as evidence that the neutralization of stop voicing in this environment is reconstructible to Proto-Ta'oi. I have transcribed all stops following nasals using voiceless characters in Proto-Ta'oi, recognizing that they were possibly phonetically voiced. Note however that the structural change of presyllable coda nasals to onset consonant prenasalization is only clearly demonstrated in the Ta'oiq variety, which is the focus of the description above, and so it is not reconstructed for Proto-Ta'oi.

Table 6: Correspondences supporting Proto-Ta'oi main syllable simple onset reconsruction[7]

						PT	TQ	TT	TS	PT	TQ	TT	TS	PT	TQ	TT	TS
*	b	d	ɟ			***b**	b	b	b	***p**	p	p	p	***m**	m	m	m
	p	t	c	k	ʔ	***d**	d	d	d	***t**	t	t	t	***n**	n	n	n
	m	n	ɲ	ŋ		***ɟ**	ɟ	ɟ	ɟ	***c**	c	c	c	***ɲ**	ɲ	ɲ	ɲ
	w	l	j							***k**	k	k	k	***ŋ**	ŋ	ŋ	ŋ
		r				***w**	w	w	w	***s**	s	s	s				
		s			h	***l**	l	l	l	***h**	h	h	h				
						***r**	r	r	r	***ʔ**	ʔ	ʔ	ʔ				
						***j**	j	j	j								

4.2 Proto-Ta'oi Main Syllable Complex Onsets

The complex onsets of Proto-Ta'oi presented in Table 7 parallel the simple onsets, including the issues of stop voicing after nasal presyllable rimes. There is some unexplained irregularity surrounding the reflexes of *cr in Ta'oiq.

[7] In this table and below, PT = Proto-Ta'oi, TQ = Ta'oiq, TT = Ta'oih of Tha Taeng, TS = Ta'uas.

Table 7: Correspondences supporting Proto-Ta'oi main syllable complex onset reconsruction

PT	TQ	TT	TS	PT	TQ	TT	TS	PT	TQ	TT	TS	PT	TQ	TT	TS
***pl**	pl	pl	pl	***pr**	pr	pr	pr	***bl**	bl	bl	bl	***br**	br	br	br
-	-	-	-	***tr**	tr	tr	tr	-	-	-	-	***dr**	dr	dr	dr
-	-	-	-	***cr**	cr~sr	cr	sr	-	-	-	-	***ɟr**	ɟr	ɟr	?
***kl**	kl	kl	kl	***kr**	kr	kr	kr	-	-	-	-	-	-	-	-
-	-	-	-	***sr**	sr	sr	sr	-	-	-	-	-	-	-	-

4.3 Proto-Ta'oi Main Syllable Codas

Ta'oi coda correspondences are complicated by the high register-conditioned coda lenitions in Ta'oiq and Ta'oi of Tha Taeng (see Section 4.5). The outline of the situation is shown in Table 8, where the reflexes of codas in low and high register words are shown to the left and right of the slashes, respectively.

Table 8: Correspondences supporting Proto-Ta'oi coda reconstruction[8]

						PT	TQ	TT	TS	PT	TQ	TT	TS
*	p	t	c	k	ʔ	***p**	p / mˀ	p / mˀ	p	***m**	m / mˀ	m	m
	m	n	ɲ	ŋ		***t**	t / nˀ	t / nˀ	t	***n**	n / nˀ	n	n
	w	l	j			***c**	c / jˀ	c / jˀ	c	***ɲ**	ɲ / ɲˀ	ɲ	ɲ
		r				***k**	k / ʔ	k / ʔ	k	***ŋ**	ŋ / ŋˀ	ŋ	ŋ
		s			h	***w**	w	w	w	***s**	s	s	s
						***l**	l	l	l	***h**	h	h	h
						***r**	r	r	r	***ʔ**	ʔ	ʔ	ʔ
						***j**	j	j	j				

4.4 Proto-Ta'oi Presyllables

The correspondences supporting the reconstruction of Proto-Ta'oi presyllables are presented in Table 9. The Proto-Katuic presyllable rime has tended to become simplified in many modern Katuic languages, including Ta'oi. The most conservative variety in terms of presyllable structure and inventory is Ta'oi of Tha Taeng, which retains traces of the Proto-Katuic gemination contrast for sonorant presyllable rimes (Gehrmann 2018). L-Thongkum (2001) transcribes preglottalized presyllable rimes before sonorant main syllable onset (e.g., /ˀl/) in etyma that have long, geminate sonorant consonants crossing the presyllable-main syllable boundary (e.g., /ll/) in other Katuic languages like Pacoh and Kriang. Gemination and preglottalization would appear to be cognate for these presyllable rimes (Gehrmann 2018, 132-133). Thus, Ta'oi of Tha Taeng provides important evidence for reconstructing Proto-Ta'oi presyllables when data is available. For disyllabic words lacking reflexes in Ta'oi of Tha Taeng, the Proto-Ta'oi presyllable rimes are reconstructed *a by default, but these are to be interpreted as underdetermined in those cases.

The nasal presyllable rimes transcribed as /N/ in Table 9 are underspecified for place of articulation, which is assimilated in surface forms based on the place of articulation of the main syllable onsets that follow them.

[8] The glottalization of Proto-Ta'oi coda nasals in Ta'oiq only occurs following short vowels.

Table 9: Correspondences supporting Proto-Ta'oi presyllable reconsruction

PT	TQ	TT	TS	PT	TQ	TT	TS	
***pa**	pa	pa	pa	***pr**	pr	pr	pr	
***ta**	ta	ta	ta	***tr**	ra	tr	tr	*elsewhere*
***ka**	ka	ka	ka		ta	tʔ	?	*before *r*
***ra**	ra	ra	ra	***kr**	kr	kr	kr	
***ha**	ha	ha	ha	***hr**	hr	hr	hr	
***pN**	pN	pN	pN	***tl**	ta	tʔ	?	*before *l*
***tN**	tN	tN	tN	***kl**	kl	ka	?	*elsewhere*
***kN**	kN	kN	kN		ka	kʔ	ka	*before *l*
***hN**	hN	hN	hN					
***ʔN**	ʔN	ʔN	ʔN					

4.5 Proto-Ta'oi Register

It has long been known via internal reconstruction of Ta'oiq varieties that the Ta'oiq rime-medial creak and rime-final glottalization are essentially in complementary distribution and carry forward one historical phonological property of Proto-Ta'oi (Ferlus 1974, Diffloth 1989). This property of Proto-Ta'oi is reconstructed here as contrast of *register* in keeping with my broader proposals regarding registrogenesis (Gehrmann 2015, 2022a, 2022b). Table 10 summarizes the development of Proto-Ta'oi rimes in modern Ta'oiq, building on the insights of Ferlus and Diffloth and combining them with my own subsequent work on Proto-Ta'oi phonological reconstruction. The innovations of note are:

(1) stop codas lenite to glottal(ized) consonants under the high register
(2) nasal codas become glottalized nasals after short vowels under high register
(3) short vowels lengthen under the high register before codas *-r *-s *-h.

Table 10: Modern reflexes of Proto-Ta'oi rimes in Ta'oiq (^{H} = high register, ^{L} = low register)

	***-ø**	***-ʔ**	***-p**	***-t**	***-c**	***-k**	***-m**	***-n**	***-ɲ**	***-ŋ**	***-l**	***-r**	***-j**	***-w**	***-s**	***-h**
***V:^{H}**	V:	V:ʔ	V:p	V:t	V:c	V:k	V:m	V:n	V:ɲ	V:ŋ	V:l	V:r	V:j	V:w	V:s	V:h
***V^{H}**		Vʔ	Vp	Vt	Vc	Vk	Vm	Vn	Vɲ	Vŋ	Vl	Vr	Vj	Vw	Vs	Vh
***V:^{L}**	V̰:	V:ʔ	V:mˀ	V:nˀ	V:jˀ	V:ʔ	V̰:m	V̰:n	V̰:ɲ	V̰:ŋ	V̰:l	V̰:r	V̰:j	V̰:w	V̰:s	V̰:h
***V^{L}**		Vʔ	Vmˀ	Vnˀ	Vjˀ	Vʔ	Vmˀ	Vnˀ	Vɲˀ	Vŋˀ	V̰l		V̰j	V̰w		

It must be stressed that in phonetic terms, the realization of the register contrast of Proto-Ta'oi would have differed in some important ways from the prototypical Mainland Southeast Asian register phenomenon. The more common type of register contrast is a *lax-marked* register contrast, in which the low register is characterized by comparatively marked phonetic cues, including any combination of laxer (breathy) voice quality, raised vowel quality, lowered pitch, and/or a brief voicing lag for voiceless stops. This type of register language, a Khmer Model register language (Huffman 1976, Gehrmann 2022a), is encountered frequently among Austroasiatic languages and in certain Austronesian languages (Cham, Javanese, *inter alia*). Proto-Ta'oi would have been a *tense-marked* register language, in which the more marked register is the high register,

employing a bundle of perceptual cues that include tenser (creaky) voice quality, conditioned coda mutations and/or falling pitch contour. As can be seen in Figure 3 and Table 10, modern Ta'oiq reflexes of the Proto-Ta'oi high register involve all three of these. Further examples of tense-marked register languages are found in Austroasiatic, including Sedang (< Bahnaric) (Smith 1968, 1972, 1973; Smith & Sidwell 2015), Chong (< Pearic) (Huffman 1985a, L-Thongkum 1991, Edmondson 1996, DiCanio 2009), and Pacoh (< Katuic) (Watson 1996, Gehrmann 2022b), though each has its own idiosyncrasies (Gehrmann 2022a).[9] Further discussion is provided in Section 5.2.

Each of the three Ta'oi varieties shows a different pattern of coda mutation under the high register, ranging from no mutations in Ta'uas, to mutations affecting stop codas only in Ta'oi of Tha Taeng, to mutations affecting both stop and nasal codas in Ta'oiq. Consequently, coda mutations are interpreted here as parallel, post-Proto-Ta'oi innovations and are not reconstructed for Proto-Ta'oi itself.

4.6 Proto-Ta'oi Vocalism

There is very little variation among the modern Ta'oi languages as far as vocalism is concerned. All vowels could occur in either register in Proto-Ta'oi except for the close short vowels *i, *ɨ, and *u and the close long central vowel *ɨː. Ta'oi words with these vowels do not occur in high register in my comparative lexical database, and this distribution appears to reflect the vowel system of Proto-Ta'oi. Although register contrast has been mostly neutralized in Ta'oi of Tha Taeng and completely neutralized in Ta'uas, these varieties preserve Proto-Ta'oi vowel quality categories faithfully. Because the origins of the Proto-Ta'oi register contrast lie in historical vowel quality shifts (see Section 5), and because the modern vowel quality values of Ta'oi of Tha Taeng and Ta'uas match Ta'oiq so closely, it is certain that the former two varieties formerly shared the unique register contrast of Ta'oiq and are true daughters of Proto-Ta'oi.

Table 11: Correspondences supporting Proto-Ta'oi main syllable vowel reconstruction

				PT	TQ	TT	TS	PT	TQ	TT	TS	PT	TQ	TT	TS
*	ia	ɨa	ua	***ia**	ia	ia	ia	***ɨa**	ɨa	ɨa	ɨa	***ua**	ua	ua	ua
	i(ː)	ɨ(ː)	u(ː)	***i**ː	i/iː	iː	iː	***ɨ**ː	ɨː	ɨː	ɨː	***u**ː	u/uː	uː	uː
	e(ː)	ə(ː)	o(ː)	***e**ː	eː	eː	eː	***ə**ː	əː/ɔː	əː	əː	***o**ː	oː	oː	oː
	ɛ(ː)	a(ː)	ɔ(ː)	***ɛ**ː	ɛː	ɛː	ɛː	***a**ː	aː	aː	aː	***ɔ**ː	ɔː	ɔː	ɔː
				***i**	i	i	i	***ɨ**	ə~ɨ	ɨ	ɨ	***u**	u	u	u
				***e**	e	e	e	***ə**	ə	ə	ə	***o**	o/ɔ	o	o
				***ɛ**	ɛ	ɛ	ɛ	***a**	a	a	a	***ɔ**	ɔ	ɔ	ɔ

There have been a small number of vowel innovations in the Ta'oiq variety, as indicated in Table 11.

[9] Pacoh is a close relative of Ta'oi, but the history of register formation in the two languages is separate. Pacoh is a language that appears to be losing its historical register contrast, if indeed it ever had one (see discussion in Gehrmann 2022b).

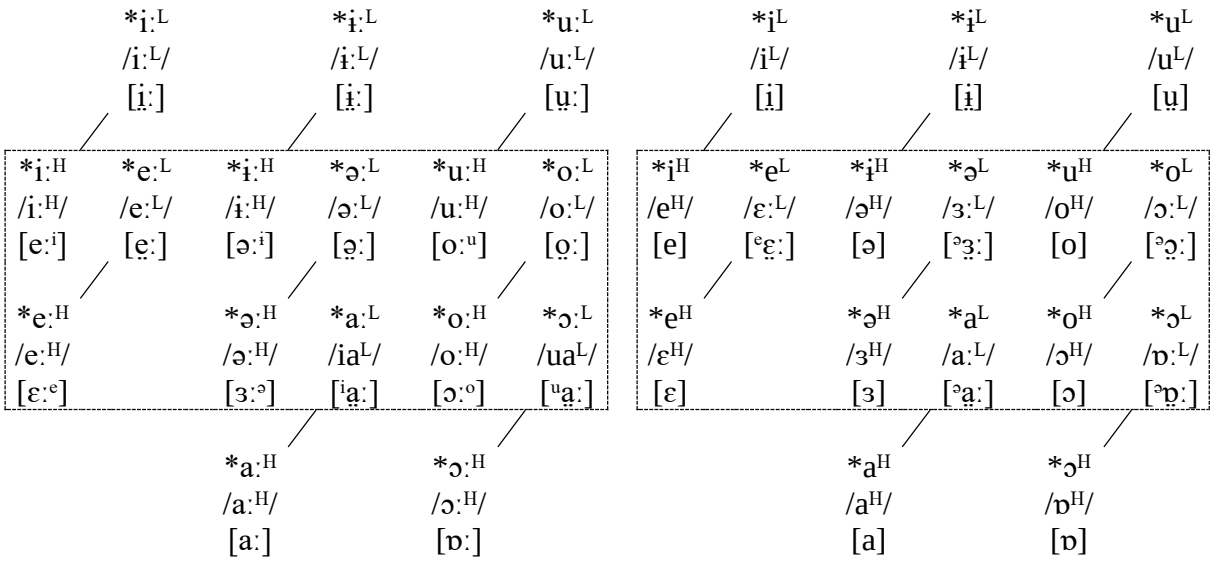
5 Proto-Ta’oi Registrogenesis

The register contrast of Proto-Ta’oi is a vital witness to the vowel quality contrasts of Proto-Katuic. It is beyond the scope of this paper to go into detail on Proto-Katuic reconstruction, but it is not possible to explain the emergence of register in Proto-Ta’oi without at least some grounding in the historical phonology of Katuic vocalism. In this section, vowel correspondences are presented which support the reconstruction of Proto-Ta’oi register and vocalism. The discussion is limited in scope to include just three representative modern languages: Ta’oi, Pacoh and Bru. These three languages were selected because in combination, they hold nearly all the evidence needed for the reconstruction of the Proto-Katuic vowel inventory. Bru and Pacoh are represented in the discussion below by exceptionally well-documented varieties: Pacoh (Watson et al. 2013) and Bru Tri (Miller & Miller 2017). Only the vowel correspondences are presented in this section but select lexical examples have been prepared and are available in the supplementary materials archived online (see Footnote 1).

5.1 Conventional Registrogenesis

In language, vowel height and voice quality covary in predictable ways. A thematic relationship between lesser vowel aperture and laxer voice quality (modal to breathy voice quality) on the one hand and greater vowel aperture and tenser voice quality (modal to creaky voice quality) on the other is well-established (Brunelle & Kirby 2016; Brunner & Żygis 2011; Esposito et al. 2019; Denning 1989; Gehrmann 2015, 2016, 2022a, 2022b; Gregerson 1976, Huffman 1985b, Lotto et al. 1997). The best documented expression of this phenomenon in the Southeast Asian context is the conditioned mutation of vowel height in Southeast Asian register languages, which is commonly referred to as *vocalic restructuring* following Huffman (1976, 1985b). Vowels under the *high register* tend to lower in vowel height over time while vowels under the *low register* will tend to raise.

Figure 4: Modern Bru Tri reflexes of Proto-Bru vowels



The chief exemplar for the vocalic restructuring of register languages is the modern Khmer language (Henderson 1952, Pinnow 1957, Jenner 1974, Huffman 1985b, Ferlus 1992), but the vocalic restructuring of register contrasts is equally evident in other

Austroasiatic languages. As an example, consider the history of the vowel height-register interaction in the Bru Tri language[10], a Katuic language of Vietnam and Laos, presented in Figure 4. The reconstructed forms represent Proto-Bru, which is reconstructed with a register contrast that is retained in modern Bru varieties (Gehrmann 2016, 2022a). We see that significant, differential vowel height mutations under the influence of register have led to a situation in which phonetically close monophthongs are found exclusively among the low register series vowels (marked with /^{L}/) while phonetically open monophthongs are conversely found only as high series vowels (marked with /^{H}/). Vowels that fall phonetically in the mid vowel range are acceptable in either register.

5.2 Registrogénèse Hérétique

Register-like contrasts involving voice quality have arisen in Ta'oi (Gehrmann 2015, 2022a), the North Bahnaric languages (Sidwell 2015), the Pearic languages (Sidwell 2019, Gehrmann 2022a) and another Katuic language, Pacoh (Diffloth 1982, Sidwell 2005, Gehrmann 2022b). These contrasts did not develop according to the classical registrogenetic model. While in Khmer Model registrogenesis, the high and low register series emerge under conditioning from historically voiceless and voiced consonant onsets, respectively, the distribution of register in this other group of languages is entirely orthogonal to historical onset voicing. As mentioned in the introduction, Diffloth (1982) was the first to successfully model the historical development of one of these unorthodox register languages in his analysis of Pacoh register, and he called this atypical registrogenetic model *heretical registrogenesis* (*registrogénèse hérétique*). In this model, earlier Proto-Katuic vowel quality contrasts converged in the vowel space to produce modern Pacoh register contrasts. Subsequent work on Pacoh and comparative Katuic only confirmed this analysis (Sidwell 2005, Gehrmann 2022b), and Diffloth's alternative registrogenetic model was eventually found to have played a role in all the register languages listed at the top of this paragraph.

In an article on Pacoh registrogenesis, I proposed to call this phenomenon *pseudoregister*, to explicitly differentiate it from register proper (Gehrmann 2022b). However, for reasons detailed in my PhD thesis, which outlines an overarching framework for modeling the emergence and development of tone and register contrasts in Austroasiatic and beyond (Gehrmann 2022a), it remains unclear just how separate *register* and *pseudoregister* really are. In lieu of this evaluative bifurcation of register into two archetypes, I now prefer a more descriptive naming system, one which references well-understood examples of different types of registrogenesis that have occurred in particular languages. Proto-Ta'oi falls under the *Sedang Model* of registrogenesis proposed there.

[10] Also known as Bru Vân Kiều – the transcription scheme used here is based on that of Vương (1999).

Table 12: Bahnar evidence for the origin of phonation contrasts in the NB language Rengao

Bahnar		**Rengao**	
Proto-North Bahnaric *-ɛːØ			
babɛː	*goat*	babiː^{H}	*goat*
r̥ɛː	*rattan*	riː^{H}	*rattan*
kanɛː	*rat*	kaniː^{H}	*rat*
ʔakɛː	*horn*	kiː^{H}	*antlers*
Proto-North Bahnaric *-ɔh			
dasɔh	*lungs*	katsuh^{H}	*lungs*
ɟɔh	*peck*	ɟuh^{H}	*peck*
kadɔh	*bark (tree)*	kaduh^{H}	*rind*
kasɔh	*spit*	cuh^{H}	*to spit*
sɔh	*light a fire*	cuh^{H}	*kindle*
Proto-North Bahnaric *-ɔːŋ			
ʔɔːŋ	*bee*	ʔoːŋ^{H}	*wasp*
ʔlɔːŋ	*tree*	loːŋ^{H}	*wood*
bɔːŋ	*casket*	boːŋ^{H}	*coffin*
gɔːŋ	*beat gong*	goːŋ^{H}	*gong*
Proto-North Bahnaric *-aC			
ɟraŋ	*house post*	ɟraŋ^{H}	*post*
maŋ	*night*	maŋ^{H}	*night*
praŋ	*clear sky*	praŋ^{H}	*end of rain*
taŋ	*bitter*	tsaŋ^{H}	*bitter*
taɓaŋ	*bamboo shoots*	tabaŋ^{H}	*sprout*
nam	*go*	nam^{H}	*go*
padam	*five*	padam^{H}	*five*
kap	*bite*	kap^{H} bar^{L}	*shut mouth*
ʔakan	*woman*	kan^{H}	*female*
panar	*wing*	manar^{H}	*wing*
mat	*eye*	mat^{H}	*eye*

Bahnar		**Rengao**	
Proto-North Bahnaric *-iːØ			
briː	*woods*	cɔː^{H} briː^{L}	*wolf (wild dog)*
ɟriː	*banyan tree*	lɔːŋ^{H} ɟriː^{L}	*banyan tree*
siː	*louse*	ciː^{L}	*louse*
tiː	*hand*	tiː^{L}	*hand*
Proto-North Bahnaric *-uh			
kuh	*salute*	kuh^{L}	*worship*
muh	*nose*	muh^{L}	*nose*
truh	*arrive*	truh^{L}	*arrive*
ʔadruh	*girl*	hadruh^{L}	*girl*
danuh	*poor*	danuh^{L}	*poor*
Proto-North Bahnaric *-uːŋ			
tuːŋ	*carry*	toːŋ^{L}	*carry*
kuːŋ	*ladder*	goːŋ^{L}	*stairs*
suːŋ	*axe*	coːŋ^{L}	*axe*
ʔjuːŋ	*stand up*	joːŋ^{L} daŋ^{L}	*sit up*
Proto-North Bahnaric *-əC			
glək	*drown*	glak^{L}	*drown*
katəŋ	*hear*	taŋ^{L}	*hear*
məŋ	*listen*	tamaŋ^{L}	*listen*
parəŋ	*strive*	raŋ^{L}	*hold*
tadəŋ	*warp*	daŋ^{L}	*approximately*
kaʔnəm	*under*	kaʔnam^{L}	*under*
hatəp	*dig hole*	tanap^{L} kajak^{H}	*burial place*
ləp	*flood*	klap^{L}	*cover*
bət	*make a dam*	bat^{L}	*dam*
kət	*to tie*	kat^{L}	*tie up*
ʔət	*hold breath*	ʔat^{L}	*stop breathing*

Sedang is a North Bahnaric language. Unlike its North Bahnaric siblings, which have all been described as having lax-marked register contrasts, Sedang has a tense-marked register contrast which parallels that of Ta'oiq typologically, complete with high register laryngealization and historical coda lenitions conditioned by the high register. Sidwell (2015) demonstrates that the historical origins of register contrast among the North Bahnaric languages are found in the transphonologization of Proto-Bahnaric vowel quality differences as differences of register, and that Sedang alone among the North Bahnaric languages has undergone *general tensing* of its register contrast. Thus,

Sedang has shifted from a lax-marked register language to a tense-marked register language. Gehrmann (2022a, 2022b) has subsequently suggested that this general tensing of an originally lax-marked, vowel height-conditioned register contrast in Sedang is not a unique event, and that parallel developments have led to the tenser-than-modal high registers in Pearic languages, Pacoh and indeed, Ta'oi. A crucial piece of evidence that supports this proposal is the fact that Sedang, Pearic, Pacoh and Ta'oi all have in common the devoicing of historically voiced stop onsets, whereas all the other North Bahnaric languages retain their historically voiced initial stops and remain lax-marked register languages. Those seeking additional details and context on this proposal should consult the references above.

Whether tense-marked or lax-marked, the vowel height-conditioned register contrasts that have been identified so far all have in common the same thematic relationship between vowel height and register that was mentioned at the beginning of Section 5.1. In conventional, onset phonation-conditioned register languages, close vowels tend to remain close in the low register, but restructure into more open vowels or diphthongs with lowered onsets in the high register (see Bru example above in Figure 4). In vowel height-conditioned register languages, close vowels develop into low register and if high register counterparts are developed for the close vowels, that is accomplished through vowel quality changes affecting a vowel that was historically non-close or diphthongal. For example, in Table 12, the North Bahnaric language Rengao has a vowel-height conditioned register contrast, whereby low register close vowels continue the historical Proto-Bahnaric close vowel series (as retained in the Central Bahnaric language Bahnar), and high register close vowels were innovated through the raising of historically non-close Proto-Bahnaric vowels.

Conversely, open vowels remain conservatively open in vowel quality under the high register in conventional register languages, while low register open vowels tend to raise and diphthongize (see Bru example above in Figure 4). Again, this thematic correlation of vowel height and register is echoed in vowel height-conditioned registrogenesis, where we find historically open vowels associated with high register, and their low register partners being innovated through vowel height changes affecting historically non-open vowels. In Table 12, we find that in Rengao, the high register open vowel /a/ continues the Proto-Bahnaric *a vowel, while the Proto-Bahnaric non-open *ə vowel has become its low register counterpart.[11]

In the next section, the vowel height-conditioned registrogenetic processes that produced the Proto-Ta'oi register contrast from earlier Proto-Katuic vowel quality contrasts will be outlined. It will be shown that the same underlying thematic relationship between close vowels and low register and between open vowels and high register found in the North Bahnaric languages drove a parallel registrogenetic process in Proto-Ta'oi.

5.3 From Proto-Katuic Vowel Height to Proto-Ta'oi Register

My own reconstruction of Proto-Katuic vocalism is presented in Table 13. While this reconstruction has not yet been published, it differs only very slightly from Sidwell's (2005) reconstruction with the reinterpretation of Sidwell's *ie, *iə, and *uo as *i̯ɜ, *iɜ, and *uɔ, respectively, and with the addition of another short, glided vowel *ᵘɔ. Table

[11] The Rengao /a/ vowel is realized phonetically as [ə̯] in low register, and so it is not actually a phonetically open vowel even today. Nevertheless, in phonological terms, [ə̯] is clearly the low register realization of /a/ (Gregerson 1976).

13 also indicates the register affiliation of the Proto-Ta'oi reflexes of each Proto-Katuic vowel, with the vowels that developed into Proto-Ta'oi low register shaded. Register outcomes are perfectly stratified in Proto-Ta'oi with respect to Proto-Katuic vowel height series, with the close and close-mid series developing into low register and the open and open-mid series developing into high register. The three opening diphthongs with [a] vowel targets pattern with the open vowels to condition Proto-Ta'oi high register.

Table 13: Gehrmann's Proto-Katuic vocalism
(shaded vowels develop low register in Proto-Ta'oi)

*	ia	ɨa	ua
	iː	ɨː	uː
	eː	əː	oː
	ɛː	ɨɜ	uɔ
		aː	ɒː

i	ɨ	u
e	ə	o
ɛ	ⁱɜ	ᵘɔ
	a	ɒ

In what remains of this section, each Proto-Ta'oi vowel's register pairs will be examined and their correspondences with Proto-Katuic vowels will be highlighted, where applicable.

5.3.1 Registrogenesis in Proto-Ta'oi Long Front Vowels

Table 14 summarizes the Proto-Ta'oi long front vowels and their wider correspondences. Two of the long front vowels, *ia^{L} and *ɛː^{L}, are Proto-Ta'oi innovations which do not carry forward any Proto-Katuic vowel categories, whereas their high register counterparts, *ia^{H} and *ɛː^{H}, are reflexes of Proto-Katuic vowels. These low register counterparts for Proto-Ta'oi *ia^{H} and *ɛː^{H} were easily innovated via borrowing, since borrowed words in Ta'oi are assigned to the unmarked low register category by default.

The register contrast in Proto-Ta'oi *iː and *eː developed with the fronting and monophthongization of two Proto-Katuic central diphthongs: *ɨa and *ɨɜ, respectively. These two Proto-Katuic diphthongs were assigned to high register according to their open and open-mid target vowels, whereas Proto-Katuic *iː and *eː developed into Proto-Ta'oi low register, being close and close-mid vowels. Even as *ɨa and *ɨɜ were transformed in terms of vowel quality to become matches with *iː and *eː, they retained their high register affinities. The historical contrast of vowel quality between Proto-Katuic *iː - *ɨa and between Proto-Katuic *eː - *ɨɜ was therefore maintained, but phonetically changed, such that it was no longer vowel quality that differentiated the pairs of vowels; register took over that role. While we cannot recapture exactly the sequence of sound changes or the relative chronology of events that led to this state of affairs, the effect of the change is apparent.

Table 14: Proto-Ta'oi long front vowel correspondences

PT	PP	Cado	Pacoh[12]	PB[13]	Bru Tri	PK
(*ia^{L})						
*ia^{H}	*ia	ia	ia	*ɛː$^{H/L}$	ɛː$^{H/L}$	*ia
*iː^{L}	*iː	iː	iː	*iː$^{(H)/L}$	iː$^{(H)/L}$	*iː
*iː^{H}	*iə	iə	iə	*ia$^{(H)/L}$	ia$^{(H)/L}$	*ɨa
*eː^{L}	*eː	ɛː	eː	*iː$^{H/L}$	iː$^{H/L}$	*eː
*eː^{H}	*ɨəː	eː	ɛː	*ɨə$^{(H)/L}$	ɨə$^{(H)/L}$	*ɨɜ
(*ɛː^{L})						
*ɛ̰ː^{H}	*ɛː	ɛː	æː	*ɛː$^{H/L}$	ɛː$^{H/L}$	*ɛː, (*ia)

As a reminder, those who wish to review lexical evidence supporting these correspondences may consult the supplementary materials online (see Footnote 1).

5.3.2 Registrogenesis in Proto-Ta'oi Long Back Vowels

Table 15 summarizes the Proto-Ta'oi long back vowels and their wider correspondences. We find parallels to the front vowel register pair formation processes just discussed here among the back vowels. Again, Proto-Ta'oi close *uː and mid *oː have innovative high registers formed via monophthongization, in this case, through a conditioned split in Proto-Katuic *uɔ, while the low registers of *uː and *oː carry forward Proto-Katuic close and close-mid vowels.[14]

Table 15: Proto-Ta'oi long back vowel correspondences

PT	PP	Pacoh	Cado	PB	Bru Tri	PK
(*ua^{L})						
*ua^{H}	*ua	ua	ua	*oː$^{H/L}$	oː$^{H/L}$	*ua
*uː^{L}	*uː	uː	uː	*uː$^{(H)/L}$	uː$^{(H)/L}$	*uː
*uː^{H}	*uə	uə	uə	*ua$^{(H)/L}$	ua$^{(H)/L}$	*uɔ[2]
*oː^{L}	*ɔː	ɔː	ɔː	*uː$^{H/L}$	uː$^{H/L}$	*oː
*oː^{H}	*oː		oː	*uə$^{(H)/L}$	uə$^{(H)/L}$	*uɔ[1]
(*ɔː^{L})						
*ɔː^{H}	*ɒː	ɒː	ɒː	*ɔː$^{H/L}$	ɔː^{H} / ua^{L}	*ɒː

[12] The status of register in Pacoh is a complicated issue, as detailed in Gehrmann's (2022b) acoustic analysis. I prefer to transcribe Pacoh with reference to vowel quality contrasts instead of register contrasts, although both transcriptions systems are legitimate.

[13] Proto-Bru registrogenesis was conditioend by a combination of onset phonation and vowel height categories, with the effect that certain Proto-Bru vowels developed register in the expected Khmer-Model patterns (transcribed $^{H/L}$), while other vowels show a pattern of skewing towards low register (transcribed $^{(H)/L}$). See discussion in Gehrmann (2022a, 76-79)

[14] A split in Proto-Katuic *uɔ led to Proto-Ta'oi *uː^{H} before codas *j, *ŋ and *k, and to Proto-Ta'oi *oː^{H} elsewhere.

5.4.3 Registrogenesis in Proto-Ta'oi Long Central Vowels

Table 16: Proto-Ta'oi long central vowel correspondences

PT	PP	Pacoh	PB	Bru Tri	PK
*ɨː^{L}	*ɨː	ɨː	*ɨː$^{(H)/L}$	ɨː$^{(H)/L}$	*ɨː
~~*ɨː^{H}~~					
*əː^{L}	*əː	oː	*əː$^{H/L}$	əː$^{H/L}$	*əː
(*əː^{H})					
(*aː^{L})					
*aː^{H}	*aː	aː	*aː$^{H/L}$	aː^{H} / ia^{L}	*aː

Table 16 summarizes the Proto-Ta'oi long central vowels and their wider correspondences. The patterns of development are quite different here, as no modern Proto-Ta'oi long central vowel register contrast is the result of the transphonologization of a Proto-Katuic vowel quality contrast. As was discussed above, register contrast was neutralized for Proto-Ta'oi *ɨː, and remains so in modern Ta'oi. Register was contrastive for Proto-Ta'oi *əː and *aː, but these were Proto-Ta'oi phonological innovations. Proto-Katuic *aː and *əː developed into high and low register, respectively, as expected based on their Proto-Katuic vowel height values. Proto-Ta'oi low register *aː was innovated through loan words, as low register is the default loanword register. The explanation for the emergence of words with high register *əː is not yet apparent, and no proposal can be given at this time, but very few words are reconstructible with *əː in high register.

5.3.4 Registrogenesis in Proto-Ta'oi Short Front Vowels

Table 17: Proto-Ta'oi short front vowel correspondences

PT	PP	Pacoh	Cado	PB	Bru Tri	PK
*i^{L}	*i	i	i	*i$^{(H)/L}$	i$^{(H)/L}$	*i ~ ɨ
~~*i^{H}~~						
(*e^{L})						
(*e^{H})						
(*ɛ^{L})						
*ɛ^{H}	*ɛ	æ	ɛ	*ɛ$^{H/L}$	ɛ$^{H/L}$	*ɛ

Table 17 summarizes the Proto-Ta'oi short front vowels and their wider correspondences. Katuic languages tend to have very few short front vowels, and Proto-Ta'oi was no exception. Proto-Ta'oi *i^{L} and *ɛ^{H} developed their expected register assignments in a regular manner from Proto-Katuic close and open vowels, respectively. No high register short close vowels are reconstructible for Proto-Ta'oi, which leaves just *e^{L}, *e^{H} and *ɛ^{L}. All the latter three represent Proto-Ta'oi innovations, mostly resulting from the fronting of central vowels in palatalizing environments.

5.3.5 Registrogenesis in Proto-Ta'oi Short Back Vowels

Table 18: Proto-Ta'oi short back vowel correspondences

PT	PP	Pacoh	Cado	PB	Bru Tri	PK
*u^{L}	*u	u	u	*u$^{(H)/L}$	u$^{(H)/L}$	*u
~~*u^{H}~~						
*o^{L}	*ɔ	ɔ	ɔ	*o$^{H/L}$	ɔ$^{H/L}$	*o
*o^{H}	*o	o		*o$^{(H)/L}$	ɔ$^{(H)/L}$	*^{u}ɔ
(*ɔ^{L})						
*ɔ^{H}	*ɒ	ɒ	ɒ	*ɔ$^{H/L}$	ɒ$^{H/L}$	*ɒ

Table 18 summarizes the Proto-Ta'oi short back vowels and their wider correspondences. Again, no high register short close vowels are reconstructible. Proto-Katuic close *u and open *ɒ developed as expected to Proto-Ta'oi *u^{L} and *ɔ^{H}, respectively, and the low register counterpart for *ɔ^{H} is not cognate with any Proto-Katuic vowels, making this yet another low register open vowel innovation, accomplished primarily through loans in the unmarked low register. Mirroring the back long vowels, Proto-Katuic *o developed low register as expected, and Proto-Ta'oi *o high register was innovated via the monophthongization of Proto-Katuic *^{u}ɔ (cf. Section 5.4.2).

5.3.6 Registrogenesis in Proto-Ta'oi Short Central Vowels

Table 19: Proto-Ta'oi *short central vowel correspondences*

PT	PP	Pacoh	Cado	PB	Bru Tri	PK
*ɨ^{L}	*ɨ	ɨ	ɨ	*ɨ$^{(H)/L}$	ə^{H} / ɨ^{L}	*ɨ
~~*ɨ^{H}~~						
*ə^{L}	*ɨ	ɨ	ɨ	*ɨ$^{(H)/L}$	ə^{H} / ɨ^{L}	*ɨ
*ə^{H}	*ɜ	ɜ	ə	*a$^{(H)/L}$	*a$^{(H)/L}$	*^{i}ɜ
*a^{L}	*a	a	a	*ə$^{H/L}$	ɜ$^{H/L}$	*ə
*a^{H}				*a$^{H/L}$	a$^{H/L}$	*a

Table 19 summarizes the Proto-Ta'oi short central vowels and their wider correspondences. Again, no high register short close vowels are reconstructible. Proto-Katuic close *ɨ and open *a developed as expected to Proto-Ta'oi *ɨ^{L} and *a^{H} respectively; but there is a partial split in the reflexes of Proto-Katuic *ɨ, as some have lowered to Proto-Ta'oi *ə^{L}. The conditions leading to this split are not apparent at this time, but since Proto-Katuic *ə has lowered to Proto-Ta'oi *a^{L}, the motivation for the split in Proto-Katuic *ɨ to Proto-Ta'oi *ɨ^{L} and *ə^{L} would seem to be to re-fill the short mid central vowel space. The lowering of Proto-Katuic *ə to be a low register counterpart for the *a vowel parallels the pattern seen in Rengao and other North Bahnaric languages (see Table 12). Finally, a register contrast for Proto-Ta'oi *ə was set up by the split in Proto-Katuic *ɨ just mentioned, which produced low register *ə^{L}, and by the monophthongization of Proto-Katuic *^{i}ɜ to high register *ə^{H}. Note that the latter development parallels the development in Proto-Katuic *ɨɜ nicely, as this long, glided vowel also developed high register, but was subsequently fronted to Proto-Ta'oi high register *e:^{H}.

6 Summary & Outlook

An understanding of Ta'oi and its phonological history are vitally important, both for the ongoing investigation into the historical phonology of the Katuic branch of Austroasiatic and for modeling the development of register in Mainland Southeast Asian languages generally. In this paper, an introduction to the wider Ta'oi language has been presented, along with a brief phonological overview of the most conservative variety: Ta'oiq. This variety preserves the Proto-Ta'oi register contrast intact, the modern expression of which is a combination of unpredictable rime-medial laryngealization and rime-final glottalization. Proto-Ta'oi was then demonstrated to be one of several atypical register languages, which have undergone a registrogenetic process that innovated register contrasts via the transphonologization of historical vowel height contrasts. The development from Proto-Katuic close and close-mid vowels to Proto-Ta'oi low register and from Proto-Katuic open and open-mid vowels to Proto-Ta'oi high register was then presented through tables of vowel correspondences between Proto-Katuic, Proto-Ta'oi, Proto-Pacoh, Pacoh, Cado, Proto-Bru and Bru Tri. These developments were discussed in overview. Finally, lexical reconstructions of Proto-Ta'oi and tables of lexical evidence supporting the reconstruction of Proto-Ta'oi vowel-heigh conditioned registrogenesis presented here were made available online as supplementary materials (see link in Footnote 1).

References

Brunelle, Marc & James Kirby. 2016. Tone and phonation in Southeast Asian languages. *Language & Linguistics Compass* 10:4. 157-207.

Brunner, Jana & Marzena Żygis. 2011. Why do glottal stops and low vowels like each other? *Proceedings of the 17th International Congress of Phonetic Sciences* (ICPhS XVII), Hong Kong, 376–379.

Conver, Johanna, Mackenzie Conver & Jonathan Schmutz. 2014. *Lexicon of Ta'oiq*. Unpublished.

Denning, Keith. 1989. The diachronic development of phonological voice quality, with special reference to Dinka and the other Nilotic languages. PhD dissertation, Stanford University.

DiCanio, Christian T. 2009. The phonetics of register in Takhian Thong Chong. *Journal of the International Phonetic Association* 39/2, 162-188.

Diffloth, Gérard. 1982. Registres, dévoisement, timbres vocaliques: leur histoire en Katouique. [Registers, devoicing and voice quality: their history in Katuic]. *Mon-Khmer Studies* 11. 47-82.

Diffloth, Gérard. 1989. Proto-Austroasiatic creaky voice. *Mon-Khmer Studies* 15. 139-154.

Edmondson, Jerold A. 1996. Voice qualities and inverse filtering in Chong. *Mon-Khmer Studies* 26, 107–116.

Enfield, N.J. & Gérard Diffloth. 2009. Phonology and sketch grammar of Kri, a Vietic language of Laos. *Cashiers de Linguistique Asie Orientale* 38(1). 3-69.

Esposito, Christina M., Morgan Sleeper & Kevin Schäfer. 2019. Examining the relationship between vowel quality and voice quality. *Journal of the International Phonetic Association* 1-32. doi:10.1017/S0025100319000094

Ferlus, Michel. 1974. La langue Ong, mutations consonantiques Proto-Ta'oi transphonologisations. [The Ong language, consonant mutations and trasphonologizations]. *Asie du Sud-Est Proto-Ta'oi Monde Insulindien* 5.113-21.

Ferlus, Michel. 1979. Formation des registres et mutations consonantiques dans les langues Mon-Khmer. [The formation of registers and consonant mutations in the Mon-Khmer languages]. *Mon-Khmer Studies* 8. 1-76.

Ferlus, Michel. 1992. Essai de phonétique historique du khmer (du milieu du premier millénaire de notre ère à l'époque actuelle). [Essay on the phonological history of Khmer (from the middle of the first millennium AD to the present day)]. *Mon-Khmer Studies* 21. 57-89.

Ferlus, Michel. nd. *Lexique comparatif: Ta-oy, Ong, Katang, Krieng*. Unpublished lexicon.

Gehrmann, Ryan. 2015. Vowel height and register assignment in Katuic. *Journal of the Southeast Asian Linguistic Society* 8. 56-70.

Gehrmann, Ryan. 2016. The West Katuic Languages: Comparative Phonology and Diagnostic Tools. Chiang Mai: Payap University MA thesis.

Gehrmann, Ryan. 2018. Katuic presyllables and derivational morphology in diachronic perspective. In Ring, Hiram & Felix Rau (eds.), *Papers from the 7th International Conference on Austroasiatic Linguistics* (Journal of the Southeast Asian Linguistics Society Special Publication No. 3), 132-156. Honolulu: University of Hawai'i Press.

Gehrmann, Ryan. 2022a. *Desegmentalization: Towards a Common Framework for the Modeling of Tonogenesis and Registrogenesis in Mainland Southeast Asia with Case Studies from Austroasiatic*. Edinburgh: University of Edinburgh PhD thesis.

Gehrmann, Ryan. 2022b. Pseudoregister in Pacoh: Preliminary acoustic analysis and implications for a general model of pseudoregister formation in Austroasiatic. In *Proceedings of the 30th Annual Meeting of the Southeast Asian Linguistics Society* (SEALS XXX).

Gregerson, Kenneth. 1976. Tongue-root and register in Mon-Khmer. In Philip N. Jenner, Laurence C. Thompson & Stanley Starosta (eds.) *Austroasiatic Studies* (Oceanic Linguistics Special Publication No. 13). Honolulu: University of Hawai'i Press.

Haak, Feikje van der. nd. *Lexicon of Ta'uas*. Unpublished lexicon.

Henderson, Eugenie. 1952. The main features of Cambodia pronunciation. *Bulletin of the School of Oriental and African Studies* 14.1. 149-174.

Huffman, Franklin. 1976. The register problem in fifteen Mon-Khmer languages. In Philip N. Jenner, Laurence C. Thompson & Stanley Starosta (eds.), *Austroasiatic Studies*, 575-590. Honolulu: The University Press of Hawaii.

Huffman, Franklin. 1979. *Analysis of Bru of Saravane, Laos*. In Paul Sidwell (ed.), *Huffman Papers*. Online: www.sealang.net/archives/huffman (Accessed 12 July, 2024).

Huffman, Franklin. 1979. *Analysis of Ir (In) of Saravane, Laos*. In Paul Sidwell (ed.), *Huffman Papers*. Online: www.sealang.net/archives/huffman (Accessed 12 July, 2024).

Huffman, Franklin. 1985a. The phonology of Chong. In Suriya Ratanakul, David Thomas & Suwilai Premsrirat (eds.), *Southeast Asian Linguistics Studies Presented to Andre G. Haudricourt*, 355-388. Thailand: Mahidol University.

Huffman, Franklin. 1985b. Vowel permutations in Austroasiatic languages. In Graham Thurgood, James A. Matisoff and David Bradley (eds.), *Linguistics of the Sino-Tibetan Area: The State of the Art. Papers Presented to Paul K. Benedict for his 71st Birthday*, 141-45. Canberra: Department of Linguistics, Research School of Pacific Studies, Australian National University.

Jenner, Philip. 1974. The development of the registers in Standard Khmer. In N.D. Liem (ed.) *South-east Asian Linguistic Studies Vol. 1*, 47-60. Canberra: Pacific Linguistics.

Lotto, A.J., L.L. Holt & K.R. Kluender. 1997. Effect of voice quality on perceived height of English vowels. *Phonetica* 54, 76-93.

L-Thongkum, Theraphan. 1991. An instrumental study of Chong registers. In Jeremy Davidson (ed.), *Essays in Mon-Khmer Linguistics in Honour of H. L. Shorto*, 141–160. London: School of Oriental and African Studies.

L-Thongkum, Theraphan. 2001. ภาษาของนานาชนเผ่าในแขวงเซกองลาวใต้. *Phasa khong nana chon phaw nai khweng se kong lao tai*. [Languages of the tribes in Xekong province southern Laos]. Bangkok: The Thailand Research Fund.

Miller, John & Carolyn Miller. 2017. *Bru-English-Vietnamese-Lao Dictionary*. SIL International. Online: https://www.webonary.org/bru/ (Accessed: 12 July, 2024)

Nguyễn Văn Lợi, Đoàn Van Phúc & Phan Xuân Thàn. 1986. *Sách hoc Tiêng Pakôh-Taôih*. [Study on the Pacoh-Taoih Language]. Hanoi, Tinh Bình tri Thiên.

Pinnow, Heinz-Jürgen. 1957. Sprachgesichliche Erwägungen zum Phonemsystem des Khmer. [Reflections on the history of the Khmer phonemic system]. *Zeitschrift für Phonetik und Allgemeine Sprachwissenschaft* 10.4. 378-391.

Sidwell, Paul. 2005. *The Katuic Languages: Classification, Reconstruction and Comparative Lexicon*. Munich: Lincom Europa.

Sidwell, Paul. 2015. Local drift and areal convergence in the restructuring of Mainland Southeast Asian languages. In N. J. Enfield and Bernard Comrie (eds.), *The Languages of Mainland Southeast Asia: The State of the Art*, 51-81. Berlin: De Gruyter Mouton.

Sidwell, Paul. 2019. Proto-Pearic and the role of vowel height in register formation. Paper presented at the *8th International Conference on Austroasiatic Linguistics*, Chiang Mai University, Chiang Mai, Thailand, Aug. 29-31, 2019.

Smith, Kenneth & Paul Sidwell. 2015. Sedang. In Mathias Jenny & Paul Sidwell (eds.), *The Handbook of Austroasiatic Languages*, 789-836. Leiden: Brill.

Smith, Kenneth. 1968. Laryngealization and delaryngealization in Sedang phonemics. *Linguistics* 38, 52-69.

Smith, Kenneth. 1972. *A Phonological Reconstruction of Proto-North-Bahnaric*. (Language Data: Asian-Pacific Series, No. 2). Santa Ana, California: Summer Institute of Linguistics.

Smith, Kenneth. 1973. Denasolaryngealization in Sedang folk linguistics. *Mon-Khmer Studies* 4, 53-62.

Vương Hữu Lễ. 1999. A new interpretation of the Bru Vân Kiêu vowel system. *Mon-Khmer Studies* 29, 97-106.

Watson, Richard. 1996. Why three phonologies for Pacoh? *Mon-Khmer Studies* 26, 197-205.

Watson, Richard, Saundra Watson & Cubuat Canxóiq. 2013. *Pacoh-Vietnamese-English Dictionary*. SIL International. Online: https://www.webonary.org/pacoh/ (Accessed 12 July, 2024).

6

Revising Proto-Aslian[1]

Paul Sidwell

1 Introduction

Gerard Diffloth's (henceforth GD) 1968 and 1977 papers on proto-Semai phonology were the first in a series of publications over subsequent years that presented comparative-historical reconstructions of various Austroasiatic (AA) branches, building towards an anticipated consolidated proto-AA reconstruction. A great significance of that early work lay in how his proto-Semai presented a model for Aslian[2] and AA reconstructions with consequences reaching through to the present day.

The Aslian branch, while attracting some very interesting and rigorous attention from concerned scholars over many decades, has nonetheless attracted modest attention in terms of comparative-historical studies. Three other papers by GD (1975, 1976a, 1976b) also insightfully discuss aspects of Semai and Aslian language history, including syllable structure and minor-syllable vocalism, although the form of proto-Aslian is not specifically discussed. GD subsequently pivoted to comparative Waic and Monic studies, leaving Aslian on the back burner. Two decades later, comparative investigations of these languages were renewed by Timothy Phillip (henceforth TP) with his (2005/2013) *Linguistic Comparison of Semai Dialects* and (2012) PhD thesis *Proto-Aslian: towards an understanding of its historical linguistic systems, principles and processes*, and then the field of Aslian phonological-reconstruction again went quiet for another decade.

GD's work on proto-Semai phonology was conducted during something of a boomtime for AA historical studies; comparative sub-branch reconstructions were emerging from SIL-affiliated scholars who worked in Indo-China, and in the UK Harry Shorto was compiling his *Mon-Khmer Comparative Dictionary* (including a phonological reconstruction). These and other efforts were delivering divergent approaches and results which saw a mix of progress and dead ends. On balance, it is now clear that GD was largely on the right track, be it by coincidences of geographical happenstance and skillful analytical insight or by recognizing genuinely archaic features in Semai that are nowadays accepted as reconstructable to proto-AA. Still, any language or language group is a mix of old and new elements, and when we treat one as archaic for comparative purposes, there is a risk of projecting too much onto the past.

Into the mid 1970s, the general view among scholars was that the AA languages of the Malay Peninsular comprised two or three distinct branches of AA. For example,

---

[1] This chapter develops on the author's presentation 32nd meeting of the Southeast Asian Linguistics Society (Chiang Mai, May 2023) *Proto-Aslian reconstruction: classification, vocalism, homeland* (DOI 10.5281/zenodo.8397374)

[2] Geoffrey Benjamin proposed the term *Aslian*, based on the Malay expression *Orang Asli* (people.indigenous) 'indigenous people') at the first ICAAL conference (Honolulu, Jan. 1973).

GD distinguished *Jahaic*, *Senoic*, and *Semelaic*[3] branches in his 1974 *Encyclopedia Brittanica* article, rather than putting them within a single Aslian branch. As the Semai lects are a major component of the Senoic (or Central Aslian) clade, GD (1976) considered the possibility that Semaic was close to being a primary branch of AA, speculating that various characteristics of proto-Semai could thus be very archaic. This would seem to have been his preferred interpretation in the light of the final sentence of DG's (1977) paper where he wrote, 'Now, we can study Semai history and climb deeper into the past of Mon-Khmer civilization.'

Decades later, when TP renewed comparative Semai studies, GD's work on Semai and Aslian influenced his work. TP's 2005 paper on Semai dialects included a reconstruction of proto-Semai that largely recapitulates GD's historical phonology while extending the list of proto-Semai etyma with essentially GD's proto-segments. Subsequently, his 2012 proto-Aslian made a major contribution to comparative AA studies, giving the scholarly world the first branch-level proto-phonology and lexicon for these languages, filling a crucial gap in AA comparative work (see Sidwell 2021 for an overview of Mainland SEA AA studies, and Sidwell & Rau 2014 for a round-up of AA comparative-historical studies). As will be seen below, TP's proto-Aslian phonology treated Semai phonology as the historical model for the Aslian branch overall, in an apparent manifestation of Teeter's law.[4]

In these circumstances, the Semai-centric approach deserves special scrutiny. The historical understanding of consonants and syllable structure reflected in GD (1968) and the later works mentioned above has stood the test of time and anticipates well the proto-AA phonology recently outlined by Sidwell & Alves (2023). Nonetheless, extending the etymological analysis of Aslian lexicon by comparison with various other AA branches yields striking results; it is possible to isolate vocalic innovations that occurred within Semai and revise the model of proto-Aslian main-syllable vowel development into one which is more coherent vis-à-vis the rest of AA. My working hypothesis is that GD paid excessive deference to Semai and Central Aslian, as uniquely within Aslian, they retained the proto-AA vowel length distinction, which was otherwise neutralised in Northern and Southern Aslian. While the feature of length is archaic, it nonetheless does not follow that the specific vowel timbres have been retained without changes, as we will see below.

2 Diffloth's Proto-Semai

Semai is a group of closely related dialects; GD (1968:65) observed that despite the apparent lectal diversity, speakers were heard to profess that 'there is one understanding', clearly indicative of a shared social identity and intercommunication. Four documented lects were selected by GD to represent the variation within Semai (Kampong Ayer, Kampong Ulu Gruntom, Boh tea Estate, Kompong Satah), correspondences established, and proto-forms reconstructed. It begins, quite properly, specifying Semai root structure with the following template:

(C_4) (C_3) (v) C_2 V C_1

In this template, v is one of /i, a, u/, C_1 may not be a voiced oral stop, there are various

[3] Corresponding to Northern, Central, and Southern Aslian in today's terminology. Semai (~Senoi)

[4] 'The language of the family you know best always turns out to be the most archaic.' Named after Karl Teeter, without specifically appearing in his writings.

collocational restrictions between C_4, C_3, and C_2, and a schwa vowel is regularly inserted between certain CC sequences to dissociate clusters. Also, C_3 is either a liquid or nasal and thus these may be analysed as separate infixed morphemes. This is essentially the same as the proposed reconstruction of the pAustroasiatic morphological word offered by Sidwell & Alves (2023:112), except that v as /i, a, u/ in pAutroasiatic remains an active question.[5] That the morphological templates between pAustroasiatic and proto-Semai should be so close is not surprising; similar patterns are repeated across AA languages today in branches that are geographically isolated from one another. This was already being discussed by scholars in the 1960s, for example, Shorto's (1960, 1963) discussion of word structures in Northern AA.

The inventory of proto-Semai consonants is given by GD (1968:67) as follows:[6]

*/	p	t	c	k	ʔ	
	b	d	ɟ	g		
	m	n	ɲ	ŋ		
	w	r	j			
		l	s[7]		h	/

Again, this essentially anticipates Sidwell & Alves (2023) pAA consonants, excepting that the ancestral implosives ***ɓ**, ***ɗ** merged with ***b**, ***d** before proto-Aslian. The Proto-Semai main-syllable nuclei have both long and short members, and any can occur with nasalisation, yielding a total of 33 distinct monophthongs as follows (GD 1968:69-71, with notation converted to IPA):

*/	iː		uː		ĩː	ɯ̃ː	ũː		i		u		ĩ	ɯ̃	ũ	
	eː	əː	oː		ẽː	ə̃ː	õː		e	ə	o		ẽ		õ	
	ɛː	aː	ɔː		ɛ̃ː	ãː	ɔ̃ː		ɛ	a	ɔ		ɛ̃	ã	ɔ̃	/

GD (1977:465-479) subsequently revised the vowel inventory, adding a diphthong, revising some monophthongs, and removing two short vowels present due to Malay borrowing:

*/	iː		ɯː	uː		i	ɯ	u
	eː		ɤː	oː			ə	
	ɛː	ʌː		ɔː		ɛ	a	ɔ
	iə	aː						/

The *iə diphthong is the notable addition; it arises from analysis of front vowel correspondences among Semai lects. While [iə] generally does not occur in modern Semai lects, the diphthong interpretation is reasonably indicated since relevant etyma have diphthong main vowels in cognates beyond Semai, where these are apparent (e.g.

---

5 I regard it likely that there was a contrastive minor-syllable vocalism of *ɪ, *ɐ, *ʊ, yet it may be beyond our methods to resolve satisfactorily.

6 Among Semai (and other Aslian languages) there is a tendency for nasal codas to be preploded, e.g. /-m/ > [-ᵇm], /-n/ > [ᵈn] etc. This preplosion is not contrastive and is generally not notated by GD. On the other hand, TP transcribes preplosion in his works as it is observed.

7 The symbol 's' is used conventionally by GD and other Aslianists for a sound that is often [ɕ] in the main-syllable onset, hence it being listed here under the palatals.

***ɟmpiər** ‘winnowing sieve’, cf. proto-Palaungic ***piɛr** ‘winnowing tray’, ***tiəʔ** ‘earth’ cf. proto-Katuic ***ktiak** ‘earth, soil’, etc.). The other changes in the proto-Semai vowel inventory do not create new entities but reflect a reinterpretation of values. Additionally, GD (1977) discusses vowel nasalization, finding that it is a largely secondary development. On the whole, nasalised vowels are found adjacent to nasal consonants or following h- or ʔ-; while such conditioning does not clearly account for all cases, it seems rather convincing that nasalisation can be discounted as a primitive feature of proto-Aslian.

It is perhaps striking that proto-Semai ***iə** is reconstructed without any corresponding back diphthong (such as ***uə**), yet this is not such an unusual pattern in AA; Sidwell (2015) proposes proto-Palaungic ***iɛ** but no corresponding back diphthong, and Sidwell (2018) proposes proto-Khasian ***ia** and no corresponding back diphthong. Nonetheless, it seems that GD modeled proto-Aslian as having distinct ***uə** and ***uɔ** main-syllable nuclei, as the topic is briefly discussed in his (1976) Temiar sketch, and in the handout he provided at the 2009 ICAAL meeting (held at Mahidol University, Thailand) entitled ‘More on Dvaravati-Old Mon’. Relevant etyma generally reflect these nuclei with /oː/ in Semai lects, so no proto-vowel distinct from proto-Semai ***o**ː is indicated.

So, by the later part of the 1970’s GD had a well-justified proto-Semai phonology, and it seems that this was integrated into GD’s wider ideas about proto-Aslian and proto-AA. In the new millennium a bright young scholar, Timothy Phillips, would renew comparative interest in Semai and Aslian, steeped in the ideas and results already presented by GD.

3 Phillip’s Proto-Semai and Proto-Aslian

TP’s *Linguistic Comparison of Semai Dialects* was first presented in 2005 and later revised and published by SIL in 2013 in its series of electronic survey reports. Much like GD’s early work on Semai, TP surveyed Semai at numerous locations, collecting wordlists and achieving many analytical insights relevant to classification, historical development, and language vitality. It is clear that TP relied upon GD’s earlier works to provide a framework for his own analyses. TP’s Appendix E presents 760 proto-Semai lexical items against 460 English glosses, and, “Of this total, 150 words are from Diffloth’s data and congruent with the current findings, while 610 words are newly added.” (2012:33)

TP also personally collected wordlists for other Aslian languages, as well as accessing published sources, and in 2012 completed a PhD thesis with the Universiti Kebangsaan (Malaysia). This presented a reconstruction of proto-Aslian phonology, lexicon, and many aspects of morpho-syntax (the latter is not discussed further here). This PhD is a landmark in comparative AA studies for the sheer scope of its subject matter, and its status as the only published consolidated proto-Aslian reconstruction. In terms of syllable and word-structure, and consonants, the results are essentially sound, and critical analyses would not lead to major changes in these aspects. However, TP’s pAslian main-syllable vowels and their supporting correspondences have some aspects that deserve close attention. These are consolidated and presented at table 1 (with Semai and pAslian forms bolded for emphasis).

Table 1: Summary of Phillips (2012) Aslian main-syllable long vowel and diphthong correspondences

Northern								Central				Southern				
Batek	Jahai	Minriq	Mintil	Tonga	Kensiu	Chewong	Jah Hut	Semai	Temiar	Lanoh	Semnam	Semaq Beri	Semelai	Temoq	Mah Meri	Phillips-PAslian
i	i	i	i	i	i	i	i	**i**ː	iː	iː	iː	i/e	i/e	i/e	i/e	***i**ː₁*
ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	i	**i**ː	iː	iː	iː	?	?	?	?	***i**ː₂*
ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ə	ə	**e**ː	eː	eː	eː	e	e	e	e	***e**ː₁*
i	i	i	i	i	i	ɛ?	ɛ	**e**ː	eː	eː	eː	ɛ?	ɛ?	ɛ?	ɛ?	***e**ː₂*
ɛ	ɛ	ɛ	ɛ	ɛ	ɛ	ɛ/a	ɛ	**ɛ**ː	ɛː	ɛː	ɛː	ɛ	ɛ	ɛ	ɛ	***ɛ**ː₁*
o	o	o	o	o	o	o	ə	**ɛ**ː	ɛː	ɛː	ɛː	ɛ	ɛ	ɛ	ɛ	***ɛ**ː₂*
ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	**i**ː	ɨː	ɨː	ɨː	ɨ	ɨ	o	ɨ	***i**ː*
ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	u	**ɤ**ː	?	?	?	u	u	u	u	***ɯ**ː*
u	u	u	u	u	u	u	o/ɔ	**u**ː	uː	uː	uː	o	o	o	o	***u**ː₁*
o	o	o	o	o	o	o	o	**u**ː	uː	uː	uː	o?	o?	o?	o?	***u**ː₂*
ɔ	ɔ	ɔ	ɔ	ɔ	ɔ	ɔ	ɔ	**o**ː	oː	oː	oː	u	u	u	u	***o**ː₁*
ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	ɨ	o	**o**ː	oː	oː	oː	u	u	u	u	***o**ː₂*
ɔ	ɔ	ɔ	ɔ	ɔ	ɔ	ɔ	ɔ	**ɔ**ː	ɔː	ɔː	ɔː	ɔ	ɔ	ɔ	ɔ	***ɔ**ː₁*
ɔ	ɔ	ɔ	ɔ	ɔ	ɔ	ɔ	ɔ	**ɔ**ː	ɔː	ɔː	ɔː	u	u	u	u	***ɔ**ː₂*
a	a	a	a	a	a	a	a	**a**ː	aː	aː	aː	a	a	a	a	***a**ː₁*
i/e	i/e	i/e	i/e	i/e	i/e	e	a	**a**ː	aː	aː	aː	a	a	a	a	***a**ː₂*
ɛ	ɛ	?		iɛ	iɛ	ɛ	ɛ/iɛ	**ɛ**ː/**e**ː/**i**ː/**ja**ː	iɛː	?	ieː/ɛː	ɛ	ɛ	ɛ	e	***iɛ***
a	a	a	a	a	a	a	wɔ/wa/wɛ	**o**ː	ɔː	ɔː	ɔː	o	o	o	o	***ua***
a/e?	ə	a/e?	a/e?	a/e?	a/e?	ia?	wo	**o**ː	wɔː	?	ᵘoː	ɔ	ɔ	ɔ	ɔ	***uə***
ɛ	ɛ	ɛ	ɛ	ɛ	iɛ	iɛ	o?	**o**ː	wɔː	?	ᵘoː	ɔ	ɔ	ɔ	ɔ	***uəN***

The first aspect of the correspondences to discuss here is the extent to which they are organised into pairs, e.g. ***ii₁, *ii₂; *ee₁, *ee₂,*** etc. At first blush, this might be interpreted as evidence of a registers system that had split the vowel system, but I don't believe this to be the case. Rather, the organisation of the correspondences in this manner is an outcome of several intersecting factors:

1) Correspondences are compiled from etyma from diverse sources: words inherited from pAA, Aslian and lower order innovations, and borrowings. It is apparent that TP had rather lax criteria for including etymologies at the Aslian level, presumably maximizing his pool of comparative data in order to enrich the analytical opportunities.
2) TP reckoned that there being a strong general tendency among Aslian languages for a "3x3" type vowel inventory, this was also likely for proto-Aslian, and thus a way should be found to organise a proliferation of correspondences into something like short and long 3x3 matrices.
3) Central Aslian vowel timbre and length values were given priority as reflecting proto-Aslian values, and where Northern and Southern values diverged this was taken as indicating secondary developments.

In relation to point 1) it is particularly striking that many of TP's etymologies lack Southern Aslian cognates. For example, his group ***ii₁*** consists of four etyma, none

having Southern cognates. Other correspondence groups variously have only one or two Southern reflexes, and this leaves the analysis incomplete. It is clear that there are many lexical innovations at the Central-Northern level included in TP's data, and these are projected to the proto-Aslian level without justification.

On point 3), it is worth quoting TP directly:

> *In the correspondence sets that follow it is usually not clear what the full nature of the original proto-vowels were. While it is not unusual for historical linguists to be unable to pin down the exact nature of a proto-phoneme, the chaos of the vowels in Aslian virtually guarantee that we will never be certain of the Proto-Aslian vowel system. In choosing a proto-vowel for each correspondence set I have tended to favor the vowel quality found in the CAs languages for the reason that CAs languages are the only ones that preserve vowel length distinctions and the reflex of PMK *a/*aa in CAs languages also tends to be preserved. This is not to say that CAs languages are necessarily more conservative for all the vowels. However, no doubt some bias has been introduced toward CAs languages being considered the more conservative.*
> *(Phillips 2012:94)*

TP was entirely reasonable to rely upon Central Aslian vowel length values, as these are lost by mergers in both Northern and Southern Aslian, but it is a jump to reckon that as Central Aslian is conservative in vowel length, it is also conservative in vowel timbre generally. TP remarks multiple times on the apparent "chaos" of the Aslian vowels, and a lack of obvious conditioning environments to explain changes. Yet this hardly builds confidence in the reconstruction; the task of explanation could certainly benefit from comparisons with Austroasiatic data outside of Aslian, to assist with sorting out exactly what is and is not "chaotic". As I already alluded, this appears to be an example of Teeter's law; TP was steeped in the Semai language, and GD's earlier comparative Semai work with his forceful declaration that was the key to "…climb deeper into the past of Mon-Khmer civilization", and this familiarity leads to a sentiment that can coloured otherwise objective analytical thinking.

4 Aslian family structure and implications.

The current received classification of Aslian languages emerges from Benjamin (1976) and was refined more recently by the computational phylogenetic study of Dunn et al. (2011). The picture that clearly emerges is that there is basic split between Southern Aslian and a Central-Northern sub-branch. One language, Jah-Hut, has an ambiguous place within Central-Northern Aslian, apparently having an intermediate position in the tree (see figure 1).

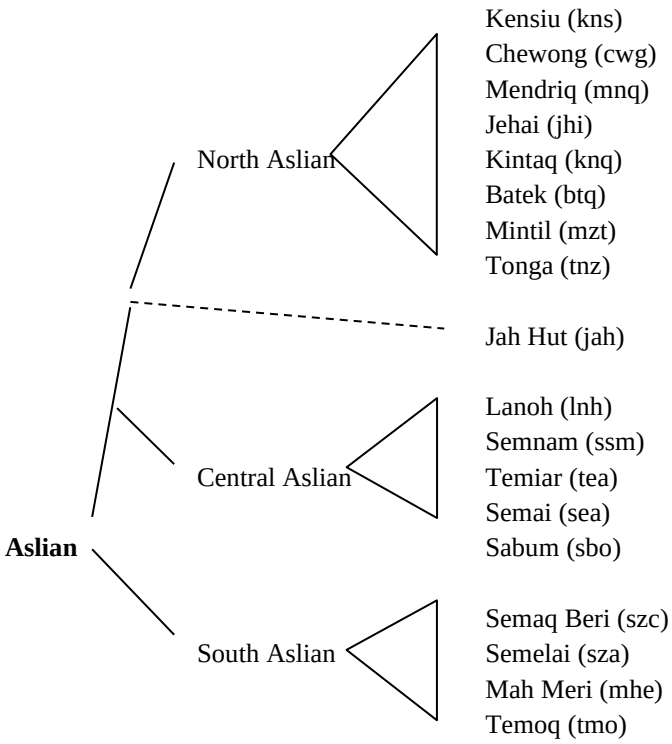


Figure 1: Aslian family tree, based on Dunn et al. (2011).

Comparing the family tree to the long vowel correspondences in Table 1, one notes that the three sub-branches form coherent blocks phonologically. It appears that the vowel length distinction was lost independently in Southern and Northern Aslian, and this has important implications for the sequencing of vocalic changes.

The basic idea that I apply in these situations is that when the length distinction is lost in AA languages, long and short vowels of the same or similar timbre merge. Naturally, this can play out is subtly different ways, especially as long and short vowels may have significant timbre differences in the first place, but in the absence of specific indications otherwise, the assumption is that long and short 3x3 inventories will merge straightforwardly. In this case, it is significant that while Central and Northern Aslian form a sub-branch, they show strong discrepancies in corresponding vowels. For example, in TP's ***ɛɛ₂** the Northern reflex is [o], for ***ee₁** the Northern reflex is [ɨ], ***ee₂** the Northern reflex is [i], ***oo₂** the Northern reflex is [ɨ], and so forth. This is the kind of thing that TP was referring to as "chaos" in the vowels.

There are several broad points to make about these Central-Northern correspondences. Firstly, we note that GD (1977) revised his proto-Semai vocalism such that he introduced significant asymmetries between the long and short vowels. In this context, it is notable that five of TP's correspondences are dominated by [ɨ] reflexes among Northern lects. The prevalence of [ɨ] indicates a centralising of vowels took place. Shorter vowels tend to be pronounced in a more lax, less peripheral manner, so I hypothesize that proto-Northern Aslian long vowels came to be pronounced with shorter duration and concomitantly less peripheral timbre, in the process of losing the contrastive length distinction. As this proceeded, we can expect that timbre mergers occurred, and also that some shifts happened to avoid mergers, perhaps on a word-by-word basis to avoid mergers and confusions in relation to specific lexical items.

Another factor is correspondences that arise from innovations and loans. The most

obvious of these is in the ***aa₁**, ***aa₂** sets; the first set is dominated by loans/innovations, while the second is predominantly AA etyma. The data, taken directly from TP's thesis, are reproduced here as tables 2 and 3.

Table 2: Phillips (2012:116) *Table 3.69 Proto-Aslian *aa₁*

	PAs	CAs	JAH	NAs	SAs
'blowgun'	*bəlaaw	bəlaaw SEA	bəlaw	bəlaw KNS	bəlaw MHE
'flower'	*bəkaaw	bəkaaw TEA	bukow (?)	bəkaw BTQ	bəkaw SZA
'flood'	*bəʔaak	bəʔaak SEA	baʔah	baʔak KNS	baʔah SZC
'stick/spear'	*ʔaat	ʔaat TEA	---	ʔat KNS ʔɒt (?) CWG	---
'tiger'	*ʔaap	ʔaap TEA	---	ʔap JHI	---
'turtle'	*pəʔaas	pəʔaas SEA	---	paʔas BTQ	paʔas SZC
'shelf'[36]	*paraaʔ	paraaʔ SEA	paraʔ	paʟaʔ BTQ	paya (?) SZC
'armpit'	*ləda(a)ʔ	---	---	ləda? KNS	ləda? SZC

TP's ***aa₁** includes 8 etyma, including 6 Southern Aslian reflexes. However, none of these items has an unambiguous AA heritage. I offer the following commentary.

- 'blowgun': this cultural item is restricted to Aslian. It can be expected to have been widely traded among tribal groups.

- 'flower': this has a parallel in Old Mon **pkāw** 'flower', but is also attested in Chamic, e.g. Röglai **bəka**:**w**, suggestive of borrowing from Austronesian.

- 'shelf': This is an Austronesian loan, cf. proto-Austronesian ***para** 'scaffolding', Malay **para** 'attic, shelf, rack'. Notwithstanding Shorto's (2006) reconstruction of PMK ***praʔ** 'loft, platform, rack', the word is likely an Austronesian loan into Bahnaric, Khmer, and Mon.

- 'flood', 'stick/spear', 'turtle', and 'armpit' have no apparent parallels in AA.

It is clear that the ***aa₁** terms are a mix of loans and innovations within Aslian and likely post-date the proto-Aslian stage, and may include Central-Northern innovations that were borrowed into Southern Aslian.

Table 3: Phillips (2012:117) *Table 3.70 Proto-Aslian *aa₂*

	PAs	CAs	JAH	NAs	SAs
'thorn'	*ɟarlaaʔ	ɟərlaaʔ TEA	ɟiʔlaʔ	ɟəliʔ MNQ ɟɨleʔ CWG	ɟarla SZA
'leaf'	*səlaaʔ	səlaaʔ SEA	ɬlaʔ	haliʔ KNS haleʔ CWG	sala SZC
'person'	*səmaaʔ	səmaaʔ SSM	---	həmiʔ KNQ 'outsider'	səmaʔ SZA
'top'	*kipaaŋ	kipaaᵍŋ SSM	---	kępɪgŋ KNS kərpiᵍŋ JHI	---
'tail'	*sə[n]taaʔ	səntaaʔ SEA	səntaʔ	hətiʔ KNS hateʔ CWG	hateʔ SZC (borrowed?)
'bone'	*ɟəʔaaŋ	ɟəʔaaᵍŋ SEA	ɟʔaŋ	ʔiyeᵍŋ KNS ɟiʔigŋ MNQ ɟɨʔɛgŋ CWG	ɟaʔaᵍŋ SZA
'tongue'	*ləntaak	ləntaak SEA	ləntak	lətik KNS latek CWG	---
'bamboo rat'	*dəka(a)n	---	---	dəken BTQ	dəkan SZA
'waist'	*gaal	gaal SSM 'hip'	---	gɛl JHI	---
'to carry in hand'	*taɟaak	taɟaak SEA	---	taɟek KNS	---

We see in table 3 some ten etyma, at least seven of these have solid AA etymologies ('person', waist', 'carry in hand' may be Aslian innovations). These data indicate a raising of ***a**: to [e] and [i] in northern lects, remaining distinct from reflexes of short ***a**, which are [a] and [ɛ] in the north. A clear implication is that length was still contrastive in proto-Northern Aslian when such shifts occurred. TP contrasts this correspondence with his ***aa₁**, reproduced below (table 3).

The main point I hope to have made by now is that the TP data is ripe for reanalysis, and in the next section, I offer my critique and proposals. I focus on the long vowels and diphthongs specifically; on the whole, my impression is that GD and TP have treated the proto-Aslian short vowels effectively and examining them here in detail would take up excessive space for little benefit.

5 Proto-Aslian long monophthongs

The Aslian family-tree model implies that in a bottom-up reconstruction, there should be Southern Aslian reflexes in any etymology that we project back to proto-Aslian. Absent a Southern Aslian cognate or clear evidence of deeper AA heritage, we should reasonably assume that such etyma belong to a more recent sub-branch. My proposal is to take a very conservative approach and compile those etymologies that satisfy the following criteria: reflexes are attested in both Southern Aslian and another Aslian language, and in AA languages that are geographically distant from Aslian (such as Khmuic and Palaungic) such that borrowing (directly or mediated) can reasonably be ruled out. We then reanalyse the phonological correspondences without assuming that any particular Aslian languages attest archaic features, but rather judge their relative archaism by their correspondence to AA features in which we have justified confidence.

For example, we know that Khmu (Khmuic) and Lamet (Palaungic) are particularly conservative vis-a-vis pAA values (see Sidwell & Alves 2023 for justification of this view), so if there is close phonetic agreement between these reference languages and one or more Aslian cognates, we can assume that the segmental values are archaic. Where values differ in other Aslian tongues, we assume that the divergent forms are innovative. The results should allow us to sketch out a basic model of proto-Aslian vocalism that can be extended and/or reanalysed on the bases of further work. This working method effectively tests the hypothesis of TP that proto-Semai provides an appropriate model for proto-Aslian, while informing any proposed revisions.

We have already examined ***aa₁** and ***aa₂** sets and found that ***aa₂** is valid for proto-Aslian while ***aa₁** belongs to a lower level, perhaps as a proto-Central-Northern Aslian innovation. Let's now work through the rest of TP's correspondence blocks.

Table 4: Phillips (2012:101) Table 3.49 Proto-Aslian *ɛɛ$_1$

	PAs	CAs	JAH	CWG	Other NAs	SAs
'sweet'	*gəhɛɛt	cəʔɛɛt (?) SEA	cəʔɛt (?)	gəhɛt	gɛhɛt KNS	gɛhɛt TMO
'moon'	*gəcɛɛʔ	gəcɛɛʔ SEA	---	kəcɛʔ	kəceʔ KNS	gadɛh SZC
'to squeeze'	*wɛ̃ɛ̃n	wɛ̃ɛ̃n TEA	---	wɛ̃t	wit (?) KNQ	wɛ̃t SZC
'cooked'	*[ʔən]cɛɛn	brcɛɛᵈn SEA	ʔencen	ncen	ʔəncɛᵈn MNQ	cin SZA
'bird'	*cɛɛm	cɛɛᵇm SEA	cɛm	---	cəpcip (?) BTQ (t.o. bird)	cim SZA
'to whisper'	*gəliseɛh	gliseɛh SSM	---	kiseh	kəliseh BTQ	bəkəraseh SZA
'astonished'	*pəlɛ(ɛ)ʔ	---	pəlɛʔ	pəlɛʔ	pəlɛʔ KNS	pəlɛʔ SZA
'I' (1S)	*yɛɛʔ	yeeʔ TEA	---	yɛ	yɛʔ MNQ	yɛ SZA

Only two of the ***ɛɛ₁** items have AA etymologies, 'cooked' and 'bird'. External cognates include Khmu **si**:**m** and Lamet **si**:**m**, supporting pAA[8] ***ci**:**m** 'bird' and ***ci:nʔ** 'cooked, ripe'. On that basis, I would revise the viable Aslian etymologies in ***ɛɛ₁** to ***i**:.

Table 5: Phillips (2012:103) Table 3.53 Proto-Aslian *ɛɛ$_2$

	PAs	CAs	JAH	CWG	NAs	SAs
'fruit'	*pəlɛɛʔ	pəlɛɛʔ SEA	pləʔ	ploʔ	ploʔ KNS	pəle SZA
'to seek'	*kɛɛʔ	kɛɛʔ SEA	---	koʔ	kɔʔ MNQ	---
'liver'	*[gə]rɛɛs	(ʔi)riis SEA	rəs	ros	ros JHI	gərɛs SZA
'to pound'	*sɛɛh	sɛh SEA	səh	---	səh BTQ	---
'shy'	*sɛɛl	sɛɛl SEA	---	---	məlsol KNQ	---

The ***ɛɛ₂** items include only two with Southern reflexes, 'fruit' and 'liver', although 'shy' can also be reconstructed to proto-Aslian as it has an apparent AA cognate in Sre

[8] All proto-Austroasiatic (pAA) reconstructions in this paper are from Sidwell (2024).

ɓasi:**l** ‘ashamed’[9], indicating that it properly belongs with ***ɛ**:**₁** and the Kensiw comparison məlsol may be unrelated. On the other hand, ‘fruit’ has cognates including Khmu **pleʔ** and Lamet **ple**:**ʔ**, indicating pAA ***pə.’leʔ**, and ‘liver’ has cognates in Munda such as Kharia **gɔ’rɛ** ‘liver’, Gta’ **grire** ‘heart’ indicating pAA ***gə.’re(**:**)ɛ**. Consequently, I would revise ***ɛɛ₂** to ***e**:.

Table 6: Phillips (2012:97) Table 3.46 Proto-Aslian *ee$_1$

	PAs	CAs	JAH	NAs	CWG	SAs
‘they two (3D)’	*weeh	wɛ(ɛ) TEA wiiy SSM	---	wih KNQ wɨh (?) BTQ	---	---
‘mushroom’	*bətees	bətees PSEA	tɛs	tis KNS	tis	pətih MHE
‘bat’	*pəleek	paleek SSM	---	pəlik KNS	plek	
‘to throw out’	*we(e)n	wees (?) SEA	---	həwit (?) KNS	---	wen SZA
‘to pull’	*ke(e)ŋ	kɛɛŋ TEA kɛk SSM	---	kiᵍŋ KNS keᵍŋ JHI	---	---

Table 7: Phillips (2012:98) Table 3.47 Proto-Aslian *ee$_2$

	PAs	CAs	JAH	NAs	CWG	SAs
‘water, river’	*[bə]teew	teew SEA	təw	bɯtẹw (?) KNS	bətəw	---
‘monitor lizard’	*pareeʔ	pareeʔ SEA	prɨʔ	---	---	pare SZA
‘mosquito’	*kəmeet	kəmeet SEA	kəmə̃t	kɨmɨ̃t BTQ	kɨmɨ̃t	---
‘bamboo’	*leew	leew SEA	---	---	ləw	lew SZC

TP’s group ***ee₁*** has only one viable Austroasiatic item: ‘mushroom’;[10] it has cognates including Khmu **tih** (< pKhmuic ***tis**) and Lamet **ti**:**h**, indicating pAA ***pə.’ti**:**ɛ** ‘mushroom’. Clearly, this etymon belongs in the same group as ***ɛɛ₁**, and the proto-Aslian form should be ***pəti**:**ɛ**.

TP’s ***e**:**₁** has **₂** also mostly lacks external support. Only ‘mosquito’ can be compared to Mon **həmit** ‘mosquito’, suggesting an original long ***i**: vowel for this etymon, therefore pAslian ***kəmi**:**t**. If we set aside the ‘mushroom’ and ‘mosquito’ etyma, I suggest that ***ee₁*** and ***ee₂*** groups reflect Central-Northern lexical innovations with an ***e**: nucleus, plus borrowing of several items (‘monitor’, ‘bamboo’, etc.) into Southern Aslian.

---

9 I have yet to find additional AA cognates for this etymon; this leave ne suspicious that it may be of Austronesian origin.

10 TP’s Aslian ***pəle**:**k** ‘bat’ may be compared to Shorto (2006) §421a ***laik** ‘flying creature’, but I doubt that the forms are related.

Table 8: Phillips (2012:95) Table 3.43 Proto-Aslian *ii$_1$

	PAs	CAs	JAH	NAs	SAs
'full, sated'	*bəhiiʔ	bəhiiʔ SSM	bahɐʔ	bəhiʔ KNS	bihi SZA
'to dry'	*tiil	tiil SEA	til	til JHI	---
'spider'	*tawiiŋ	tawiiᵍŋ TEA	---	tawiᵍŋ KNS	---
'turtle'	*kəpiil	gəlpiil SEA	---	kəpil KNQ	kəpel SZC
'sky'	*baliiŋ	baliiᵍŋ SEA	---	---	maleŋ SZA
'arm'	*bəliiŋ	bliiᵍŋ SSM	---	bəliᵍŋ KNS	blɛŋ SZA
'I' (1S)	*ʔiiɲ	ʔiiɲ SSM ʔɛʲɲ (?) SEA	---	ʔiᵍŋ CWG	ʔəɲ (?) SZA

Table 9: Phillips (2012:96) Table 3.44 Proto-Aslian *ii$_2$

	PAs	CAs	JAH	NAs	SAs
'to walk/go'	*ciip	ciip SEA	cip	cɨp JHI	---
'husband'	*kə(n)siir	gənsiir SEA	kəsir	kəsiy KNS	---
'mat'	*niis	niis TEA 'floor'	---	nɨs MNQ	---
'to swallow'	*[gə]liik	liik TEA lɨɨk (?) SSM liiᵇm (?) SEA	lik	lik KNS lit BTQ	gələt (?) SZC gələc (?) MHE

The ***ii$_1$** and ***ii$_2$** sets are discussed together, as the correspondences across Central and Northern Aslian are regular enough to confidently unite these as proto-CNAslian ***i**:. Among these etymologies only, 'sated' unambiguously indicates proto-Aslian ***i**:, based on external comparisons such as Khmu **biʔ**, Bahnar **phiʔ**, therefore proto-Aslian ***bəhi**:**ʔ** 'sated' (and by implication pAA ***bə.'hi**:ʔ 'sated'). However, proto-Aslian vowel length before glottal codas is apparently only phonetic, not contrastive, so from a structural point of view, this example is not different from short ***i** (consistent with the [e] reflex in Jah Hut) and does not belong in this grouping of correspondences.

This leaves us with only four items from the total of 10 that have viable Southern Aslian cognates, and 4 with AA heritage ('sky', 'arm', 'I', 'sated'). I propose that the other items ('dry', 'spider', walk/go', 'husband', 'mat') reflect innovated proto-CNAslian ***i**: forms; we expect such to arise after proto-Aslian ***i**: had lowered to [ɛ:], creating space for new forms.

We can compare 'sky' to Katu **plɛŋ**, Palaung **pleŋ**, and others, indicating pAA ***pə.'le**:**ŋ** 'sky'. The 'arm' etymon can be compared to Ruc **blɛ**:**ŋ**, Mlabri **blɛ**:**ŋ**, Kui **blɛ**:**ŋ** suggesting pAA ***bə.'lɛ**:**ŋ** 'arm'. These two, and the 'turtle' etymon (apparently an Aslian innovation), may be reconstructed as having proto-Aslian ***e**: nuclei, raising to ***i**: in proto-CNAslian while retaining [e] timbre in Northern Aslian as length was lost there.

The vowel of the 'I' etymon is raised and/or fronted by the palatal coda such that I reconstruct proto-Aslian ***ʔəɲ** 'I'. It is comparable to Mal (Khmuc) **ʔəɲ**, Khmer **ʔaɲ**, and others that ambiguously suggest pAA ***ʔəɲ** ~ ***ʔaɲ** 'I'. TP's long vowel Central Aslian ***ʔi**:**ɲ** is only supported by Semnan **ʔĩ**:**ɲ**, whereas Semai is variously recorded as

having forms **ʔɐʲŋ**, **ʔɛn**, **ʔeŋ**, which are more consistent with ***ʔəŋ**. The Semnam form is admittedly problematic, but it is likely that the long vowel is a local innovation, and it is clear that this etymon does not belong in any of the front vowel correspondence groups.

Table 10: Phillips (2012:110) *Table 3.61 Proto-Aslian *ɔɔ₁*

	PAs	CAs	JAH	NAs	SAs
'hair, feather'	*sɔɔk	sɔɔk SEA	sɔk	sɔk KNQ	suk SZA
'fire'	*ʔɔɔs	ʔɔɔs SEA	ʔɔs	ʔɔs KNS	ʔus SZA
'to smell (vt)'	*ʔɔ̃ɔ̃ŋ	ʔɔ̃ɔ̃ŋ SSM	ʔɔŋ	ʔɔ̃ŋ JHI	ʔuŋ SZC
'nose'	*mɔɔh	mɔɔh SEA	---	mɔh JHI	muh SZA

Table 11: Phillips (2012:111) *Table 3.62 Proto-Aslian *ɔɔ₂*

	PAs	CAs	JAH	NAs	SAs
't.o. gibbon'	*tawɔɔh	tawɔɔh SEA	---	tawɔh KNQ	tawɔ SZA
'to wake up'	*wɔɔk	wɔɔk SEA	---	wɔ̃k BTQ	wɔh (?) SZA
'pus'	*təkɔɔʔ	dəkɔɔʔ SEA 'boil'	---	təkɔʔ KNS	təkɔʔ SZC
'to suck'	*ɟɔ̃ɔ̃t	ɟɔ̃ɔ̃t SSM	---	ɟɔt MNQ	ɲɔɲɔt (?) MHE

We turn now to the back vowels and begin with consideration of ***ɔɔ₁**, ***ɔɔ₂**. The first remark is that the number of comparisons offered in support of these is concerningly scarce; on typological grounds, we would expect proto-Aslian long vowel *ɔ: to have been fairly numerically common, so there is already something odd about these.

All four examples supporting ***ɔɔ₁** have vowels with [u] timbre in Southern Aslian, as well as in external parallels (although 'hair' is ambiguous). For example: Palaung **huʔ**, Ching **suk**, Khasi **ʃŋuʔ** 'hair' (but also Bahnar **sɔk**, Thavung **sɔk** 'hair', and others); Mlabri **ʔu**:**lʰ**, Pacoh **ʔu**:**s** 'fire'; Khmu **muh**, Lamet **mu**:**s** 'nose'; Pacoh **hu**:**ŋ** 'smell, sniff', Rục **hu**:**ŋ** 'kiss', and others. It is also striking that the 'hair', 'fire', and 'nose' etymologies indicate a short vowel or a vowel unmarked for length in pAA, and it may be that they were secondarily lengthened in Central Aslian.

By contrast, the four etyma provided in support of ***ɔɔ₂** lack convincing AA etymologies. Two have Southern Aslian cognates with [ɔ] vowels, although these are also in environments in which length was not contrastive. For the moment, we put these aside and return to their analysis with the fuller context of the back vowel inventory, with particular reference to the diphthongs.

Table 12: Phillips (2012:111) *Table 3.58 Proto-Aslian *oo₁*

	PAs	CAs	JAH	NAs	SAs
'to order'	*ʔoor	ʔoor SEA	---	ʔɔr JHI	ʔur SZA
'under'	*kəroom	kəroom TEA	---	kəyɔᵇm KNS	karum SZA
'to burn'	*coom	cooᵇm SSM	---	cɔᵇm JHI, CWG	---
'to carry on head'	*tool	tool TEA	---	tol JHI	---
'pregnant'	*kayoot	kayoot TEA, SSM	---	kayot JHI, KNS	---
'nest'	*[kə]soom	ʔənsooᵇm SSM	som	sɔᵇm KNS, MNQ, BTQ, JHI, CWG	gəsụp MHE kəsim SZA
'male'	*kooɲ	ʔəŋkooʲɲ SSM	koɲ	---	kục MHE 'father'

Table 13: Phillips (2012:111) *Table 3.59 Proto-Aslian *oo₂*

	PAs	CAs	JAH	NAs	SAs
'thumb'	*taboo?	tabo(o)? TEA	---	tabə-? KNS tabo? JHI	---
'wasp'	*ʔoo[k/ŋ]	ʔook SEA	---	ʔɨᵍŋ JHI	---
'thigh'	*bəloo?	bəloo? TEA	blo?	bəlɨ? JHI	bəlu MHE

TP's *oo₁ and *oo₂ groups don't appear to form two distinct correspondences; the identification is based on the centralization or unrounding of vowels in three etyma in Northern Aslian, but this is a common change in Northern Aslian, so I treat this as one group. The five of these with Southern cognates also have AA etymologies, of which four indicate a vowel with [u] timbre (e.g., Khmu **kndru**:**m**, Lamet **ntru**:**m** 'underneath', Lamet **pəhu**:**m**, Car **ʔuhum** 'nest', Khmu **ku**:**ɲ** 'father;s sister's husband', Chong **kʰu**:**ɲ** 'father', Khmu **bluʔ**, Bahnar blu:, Mundari **bulu** 'thigh'). The 'to order' etymon appears to have a cognate in Old Mon ***ʔor** 'to command to'; this is also consistent with the reconstruction being ***ua** or ***uə** in the Monic context as well as the values in Central and Southern Aslian, so I would reconstruct a proto-Aslian diphthong, and assume borrowing into Northern Aslian. These considerations lead us to reassign the ***oo₁** and ***oo₂** group etyma mostly to proto-Aslian ***u**:.

Table 14: Phillips (2012:105) *Table 3.54 Proto-Aslian *uu₁*

	PAs	CAs	JAH	NAs	SAs
'head'	*kuuy	kuuy SEA	kɔy	kuy KNQ	kʰoy SZA
'evening'	*duuy	duuy SEA	doy 'night'	ʔamduy TNZ	doy SZA
'mortar'	*guul	guul SEA	gɔl	gul CWG	gol SZC
'snake'	*taɟuuʔ	taɟuuʔ SSM	---	taɟuʔ KNQ	tiɟɔ SZA
'winnow basket'	*cəruuʔ	cəruuʔ SEA	cəroʔ	---	cəroʔ SZA
'tree'	*ɟəhuuʔ	ɟəhuuʔ SEA	nahõʔ	ɟəhũʔ JHI	---

Table 15: Phillips (2012:105) *Table 3.55 Proto-Aslian *uu₂*

	PAs	CAs	JAH	NAs	SAs
'to wash'	*suuc	suuc SEA	soc	soc JHI	suc (?) SZC
'lung'	*suup	suup SEA	sop	sop JHI	sop SZC
'to do'	*ʔuuy	ʔuuy SEA	---	ʔoy JHI	ɟəʔɔy SZA
'to sow'	*ruuy	ruuy SEA	---	roy KNQ	---

Firstly, we note that the number of etyma offered in support of ***uu₁** and ***uu₂** are strikingly scarce, as high back vowels are rather common in AA. Additionally, rather few of them have any external AA support, and each of these requires some particular commentary.

- 'Head' has cognates in Nicobarese (Car **kuj**, Nancowry **kɔj** 'head'), and arguably Katuic (Bru **kuaj**, Kui **ku**:**j**, Kriang **kɔ**:**j** 'person') if we accept the sematic shift of 'head' > 'person'. However, the historical value of the vowel is not quite clear on the bases of this evidence.

- 'Mortar' is also problematic; it has support in Khmuic (Khmu **guəl**, Mal **ku**:**l**, Tai Hat **kɔ**:**l** 'mortar') and Vietic (Mương **ko**:**l**³, Thavung **ko**:**l**³ 'mortar'), and again the interpretation of the historical vowel is not clear.

- 'Snake' has parallels in Munda (Mundari **ti**ˈ**ʤu**, Kharia **ti**ˈ**ɟɔʔ** 'worm') and Monic (Mon **kəjaoˀ** 'maggot', Nyah Kur **nchù**:**ʔ** 'worms, maggots, caterpillars') if we accept the semantic equation of 'worm' and 'snake' (this is arguable).

- 'Tree' is supported by Old Mon **chuʔ** 'tree', but other AA comparisons suggest a front or central vowel (Lamet **khe**:**ʔ**, Sre **chi**, Chrau **chə**: 'tree', Old Khmer **ɟhɤ**: 'tree') and this ambiguity is reflected in Sidwell's (2004) pAA reconstruction ***ɟə.ɕi**:ˀ, ***ɟə.ɕe**:ˀ 'tree, wood'.

Table 16: Phillips (2012:107) *Table 3.57 Proto-Aslian *ɯɯ*

	PAs	CAs	JAH	NAs	SAs
'house'	*dɯɯŋ	dɤɤᵍŋ PSEA deeᵍŋ TEA, SSM	---	dɨᵍŋ JHI	duk MHE dəŋ SZC
'skin, bark'	*kətɯɯʔ	gtɤɤʔ PSEA katɛ̃ɛ̃k (?) SSM	kətoʔ	kətɨʔ MNQ, KNS, KNQ, TNZ, JHI, BTQ	kʰətʰo SZC kətɔk (?) MHE
'pig'	*[ɟa]lɯɯʔ	lɤɤʔ PSEA ləəʔ TEA	---	---	ɟalu SZA ɟalɨ SZC
'egg'	*pənlɯɯŋ	pənlɤɤᵍŋ PSEA	pəŋluŋ	---	---
'termite'	*[d/g]arɯɯɲ	gərɤɤʲɲ PSEA garuuʲɲ TEA	dəruɲ	darəɲ CWG darɨʲɲ MNQ, JHI	ruɲ SZA daɹoɲ SZC
'to suck'	*bɯɯʔ	bɤɤʔ PSEA	buʔbuʔ	buʔ CWG, TNZ	muʔ MHE
'drunk'	*bɯɯl	bɤɤl PSEA	---	bəl CWG	bul MHE, SZA, SZC
'to vomit'	*kɯɯʔ	kɤɤʔ PSEA koʔ TEA koʔ SSM	kuʔ	kɨʔ CWG, MNQ, KNS, KNQ, TNZ kəʔ BTQ, JHI	kuʔ MHE kʰuʔ SZA, SZC, TMO
'to sweep'	*[tam]pɯɯs	pɤɤs PSEA pu(u)s TEA pɔɔs SSM	---	pɨs JHI	tampɔs MHE
'drum'[35]	*cəntɯɯgŋ	cntɤɤᵍŋ PSEA cantooᵍŋ TEA	---	cantəŋ BTQ caŋtũŋ KNS cɔ̃ŋ (?) JHI	tuntuŋ MHE k(ə)ɾontogŋ TMO
'ripe'	*ʔəndɯɯm	ndɤɤᵇm PSEA ʔəniim TEA nãnɨ̃ĩm (?) SSM 'placenta'	num	---	ʔnom SZA n̥dụp MHE num SZC
'to winnow'	*gɯɯm	gɤɤbm PSEA gɛɛᵇm TEA	gəʔgəm	gəbm CWG	guᵇm SZA
'to sting'	*sɯɯc	sɤɤc PSEA suuc TEA	suc	sɨc JHI sɨŋit MNQ	---

The above four items have to be interpreted individually, but the overall consistency of reflexes across Aslian suggests that they may have the same nuclei in proto-Aslian. The other etyma offered in support of ***uu₁** and ***uu₂** are lacking external support but may be reconstructed to proto-Aslian as lexical innovations. As to the proto-Aslian timbre, the Southern and Northern reflexes strongly hint at ***o**:.

13 etyma are offered by TP in support of ***ɯɯ**. Although high and mid rounded back vowels dominate the Jah Hut and Southern reflexes, TP explains, 'The *ɯɯ correspondence set is largely based on the PSEA phoneme *ɤɤ,' (2012:108). Strikingly, seven of these etyma have strong AA support for which data supports a reconstruction of pAA ***u**:. Another two ('skin', bark', 'vomit') have weaker AA support, and also in

these cases pAA ***u**: is suggested. I table selected relevant comparisons below for consideration (Table 17).

Table 17: Austroasiatic cognates of Phillips (2012) Proto-Aslian *ɯɯ etyma.

gloss	pAA	Select AA cognates / Notes
'house'	*ɗuːŋ	Katu **duŋ** 'house'; Old Mon **ɗuŋ** 'city, area'
'skin,bark'	--	Sre **gəltaw** 'leather, hide', Chrau **nto**ː 'skin'
'pig'	--	(Not obviously related to Old Mon **clik** 'pig')
'egg'	*pən.'ləːŋʔ	Khasi *pylleng*, Mương **tləːŋ³/ kləŋ³** 'egg'
'termite'	*ʔəŋ.'ruːɲ	Khmu **druːɲ**, Khasi *kruin*, Bru **ntrụːn** 'termite'
'to suck'	*ɓuːʔ	Khmu **buʔ** 'suck (milk)', Gta' **buʾ** 'to suck', Rục **púː** 'suckle'
'drunk	*ɓuːl	Khmu **kmbuːl**, Mundari **bul**, Katu **bul** 'drunk'
'vomit'	--	Car **kuː-ʔəl** 'to vomit'
'to sweep'	*poːɛ, *paːɛ, *piːɛ	Car **fɔh**, Maram **paːt**, Riang **pis** 'to sweep'
'drum'	--	Malay *centong* 'scoop, ladle'
'ripe'	*Cən.'ɗuːm	Khmu **hnduːm**, Lamet **ntum**, Katu **dum** 'ripe'
'to winnow'	*guːmʔ	Khmu **guːm**, Lamet **kʰuːm**, Mundari **gum** 'to winnow'
'to sting'	*suːcʔ	Khmu **huːc**, Samre **suːc**, Ta'oi **suːc** 'to sting'

Four of the items require separate discussion:

- 'Pig' has possible cognates in Monic, Katuic, and Palaungic, but these support a long front vowel ***e**ː or ***i**ː, likely as an imitative of a pig's squeal. It is not clear that the Aslian etymon can be related to these.
- 'Egg' has cognates in Khasian and Vietic that strongly indicate a central ***ə**ː proto-vowel.
- 'To sweep' is quite problematic as although the wordshape is stable across AA, the timbre of the nucleus varies from front to centre to back in an apparently random manner across the family, and reconstruction remains underdetermined.
- 'Drum' appears to be a borrowing from Malay, allowing for a semantic shift, and should be removed from further consideration.

How should we analyze TP's ***ɯɯ** set? Harking back to ***uu₁** and ***uu₂**, we found that the latter are better interpreted as proto-Aslian ***o**ː; logically this creates an ***u**ː gap in the proto-Aslian system. At the same time, we find that nine of 13 of TP's ***ɯɯ** items have external parallels and Southern Aslian cognates with [u] timbre. TP (p.108) does acknowledge the AA cognates with [u], but merely remarks that, '…. many of these words (although certainly not all) can be tentatively linked to PMK *uu, …'. Analytically this was a missed opportunity.

A compelling counter-interpretation is that pAA ***u**ː continued unchanged into proto-Aslian and retained [u] timbre in Southern Aslian. In the rest of Aslian there were timbre changes: in Central Aslian there was general unrounding and variously some fronting and lowering; in Northern Aslian subsequent changes were more diverse as there were mergers with various short vowels as the length contrast was lost. Thus, I revise ***ɯɯ** to ***u**ː.

Not also that we already revised TP's ***oo₁** and ***oo₂** to proto-Aslian ***u**ː. This

implies a split in the reflexes of ***u**ː outside of North Aslian: I propose that generally the relevant nuclei lowered with some becoming pCNAslian ***o**ː (or ***ʊ**ː), while a larger proportion also derounded and became pCNAslian ***ɤ**ː. Conditioning of this split is not clear, but I hypothesize that in pCNAslian, there was an unconditioned general change of ***u**ː > ***ʊ**ː, the latter stage was unstable and there was a period of indeterminacy between **ʊ**ː**~ o**ː**~ɤ**ː which sorted out lexically. In this process, reflexes of the infrequent proto-Aslian ***ə**ː merged with ***u**ː**.**

Table 18: Phillips (2012:113) *Table 3.64 Proto-Aslian *~~ɨɨ~~*

	PAs	CAs	JAH	NAs	SAs
'fruit'	*kəbɨiʔ	kəbɨiʔ TEA	---	kəbɨʔ KNQ	kəbɨʔ SZC
'to extinguish'	*pɨit	pɨit TEA	---	pɨt JHI	pələt (?) MHE
'mountain'	*bənɨ(ɨ)m	---	bənɨm	bənɨm BTQ 'peak'	bənɨm SZA
'firefly'	*kəlɨit	kəʔlɨit SEA	kəʔlət	kɨtlɨt KNQ	---
'to break wind'	*pəhɨim	hɨim SEA pəheem TEA	hum	pəhəm CWG	---

TP's ***ɨɨ** is the last long monophthong he reconstructs. Two of the items have good AA etymologies:

- 'To break wind', cf. Chong **pʰu**ːˀ**m**, Pong (Vietic) **ksum**, Laven **pho**ː**m** 'to fart' point to pAA ***pə.ɕu**ː**m**, consistent with Jah Hut **hum**. Thus, we can assign this etymon to TP's ***ɯɯ** set, and therefore proto-Aslian ***pəhu**ː**m**.
- 'Mountain', cf. Khmu **pnɨm** 'termite hill', Old Khmer **bᵊnɔm** 'hill', etc. unambiguously indicate a short proto-vowel, so it does not belong here but with short ***ɨ**, partly from pAA ***i**.
- 'To extinguish' may be compared to Katu **pat**, Nyah Kur **pʰot**, Kompong Tom Pear **pet** 'to extinguish', all with short nuclei, but the diversity of vowels makes reconstruction problematic, but ***pɨt (**with lengthening in Temiar) seems to be a reasonable reconstruction. To this we can also add 'fruit' as length before glottal stop is also secondary.

Finally, we have 'firefly'; thismay be compared to Malay *kelip-kelip*, an expressive form associated with 'twinkling' appearance. I suggest that this should also be treated as having an underlyingly short but with expressive lengthening in Central Aslian.

6 Proto-Aslian diphthongs

TP reconstructs three proto-Aslian diphthongs: ***iɛ**, ***ua**, and ***uə**. Multiple sources for these nuclei can be identified, and consequently, reconstruction of specific items require specific revisions. I begin by discussing the examples that TP assembles under ***iɛ;** some 19 etyma are offered and the first thing I will say is that this is striking for the sheer number; in my experience, it is extremely odd for the front diphthong to be such a common segment in AA languages, and this immediately flags that there is something going here that needs to be better understood. I begin by isolating the examples that have AA etymologies/parallels.

Table 19: Phillips (2012:130) *Table 3.81 Proto-Aslian *iɛ*

	PAs	CAs	JAH	NAs	SAs
'saliva'	*ləhiɛŋ	ləhiəᵍŋ PSEA ləhyɛɛᵍŋ TEA lhɛɛᵍŋ SSM	huyɛŋ (?)	ləhiɛᵍŋ KNS ləhɛᵍŋ JHI ləhɛᵍŋ CWG	ləhɛᵍŋ SZC
'earth'	*[ʔa]tiɛʔ	tiəʔ PSEA tɛʔ TEA, SSM	tɛʔ	tiɛʔ KNS teʔ MNQ, BTQ, JHI tɛʔ CWG	ʔatɛ SZA, SZC, TMO teʔ MHE
'to steal'	*siɛc	siəc PSEA səyɛɛc TEA sɛɛc SSM	---	səcsiec KNS	sɛc SZA, SZC
'footprint'	*tiɛl	diəl PSEA	tɛl	tiel KNS tel JHI tɛl/tel CWG	---
'cheek'	*miɛŋ	[ka]miəŋ PSEA	---	mɛŋ CWG	mɛŋ SZA, SZC, TMO
'to sleep'	*tiɛk	tɛɛk SSM	ciyek (?)	tiek KNS tek BTQ, JHI	ɟətɛk SZA cətɛk SZC gətek MHE
'gecko'	*cəkʔiɛk	---	kəʔciyek	cəkʔę̣k KNS	cəkʔək (?) SZC
'to carry under arm'	*kamiɛk	---	kamyek	---	kamɛk SZA
'to thresh'	*rəmpiɛt	---	rəmpiyet	ʔempɛt KNS	---
'mouse'	*k[a]niɛʔ	kniəʔ PSEA	---	kiniʔ BTQ knɛ̃ʔ CWG	kanɛʔ SZA, MHE, TMO
'left (side)'	*wiɛl	wiəl PSEA yɛl TEA, LNH wɛɛl SSM	---	yal CWG	sawɛl SZA, SZC, TMO
'to forget'	*siɛm	siəᵇm PSEA	---	---	besebm SZC
'monitor lizard'	*gəriɛŋ	griəᵍŋ PSEA gɛryɛŋ TEA gəyeeᵍŋ SSM	---	gərɛgŋ JHI (sp.)	giyạk MHE geyaŋ SZC
'to plait (leaves)'	*liɛp	liəp PSEA	---	nəlɛp CWG	---
'root'	*rʔiɛs / ʔriɛs	rʔiəs PSEA yʔɛɛs SSM	---	ʔiyę̣s KNS ɟɨʔes MNQ, JHI ɟəʔes CWG	rɛs SZA, SZC
'centipede'	*k[a]ʔiɛp	kʔɛɛp PSEA kʔɛp TEA kʔeep SSM	kaʔep	kəʔiɛp KNS kʔep JHI, MNQ kɨʔɛp CWG	kaʔip SZA kiʔip MHE
'anus'	*kiɛt	kiət PSEA 'butt' kɛɛt TEA, SSM	kiyet	kit JHI, CWG, BTQ	---
'to wring, squeeze'	*riɛt	riət PSEA rəyɛ(ɛ)t TEA	rat	rit BTQ	---
'how many?'	*məriɛm	mriəᵇm PSEA marɔm (?) TEA yɔp (?) SSM	mərəm	---	marem SZC

Table 20: Austroasiatic cognates of Phillips (2012) Proto-Aslian ***iɛ** etyma.

gloss	pAA	Select AA cognates / Notes
'earth'	?	Katuic ***ktiak** 'earth, soil', otherwise AA forms indicate pAA ***te**ːʔ which may be consideed a distinct etymon.
'to steal'	?	Surin Khmer **sɪːc** 'to take (something) away stealthily'.
'footpint'	***tiːl** 'tracks'	Katuic ***tiːl** 'mark (leave a trace)', Chrau **teːl** 'footprint'.
'cheek'	***miəŋ**ʔ 'mouth, jaw'	Khasian *miaŋ 'cheek, jaw', Mal mieŋ 'mouth'; Mon həmɛŋ '(sides of) jaw, Vietic *mɛːŋʔ 'mouth'.
'to sleep'	***tiəkʔ**	Khasi *thiah*; Old Khmer *tyak ~ tyāk ~ tek* 'to sleep'
'mouse'	***kə.'neːʔ**	Kharia **kɔnɛ**, Khmu **kneʔ**, Khasi *khnai*, Rục **kəneːˢ** 'rat, mouse'
'root'	***ʔə.'riəɛ**	Mundari **reːʔd**, Khmu **riəs**, Nancowry **jiah**, Rục **liɛlʰ¹** 'root'
'centipede'	***kə.'ʔiːpʔ**	Khmu **kʔiːp**, Nancowry **kaʔeap**, Muak-Saak **kʰrʔip²**, Bahnar **kəʔɛːp** 'centipede'
'anus'	?	Mon **təket** 'anus' (consistent with ***iə**)
'to wring, squeeze'	***riətʔ**	Khmu **riat** 'to tie together', Bahnar **hərɛːt** 'to draw tight', Vietnamese *riết* 'to draw tight'

On balance, it is clear that some examples of pAA ***iə** were directly inherited into proto-Aslian as ***iɛ**, and their reflexes in Southern Aslian retain [ɛ]-like timbre; this is an understandable restructuring and merger of ***iɛ** when length was lost. In Central Aslian, reflexes are dominantly [ɛː, jɛː] although some Semai lects show [iː], and Northern lects show a mix of front vowels. Clearly many words with a front diphthong have been innovated in Aslian, including the likely adaptation of some Malay loans.

Of 19 etyma proposed by TP, 10 are found to have external AA support, and 6 of these indicate a diphthong (accepting 'earth' and 'anus'). Of the remainder, 'footprint' and 'centipede', and probably also 'to steal' indicate *iː. The 'mouse' word also lacks a diphthong, but the vowel was lower historically and the interpretation of the likely proto-Aslian form is problematic. Given the [ɛ] timbre of the nucleus in Southern Aslian reflexes of 'mouse', it can be reconstructed to proto-Aslian***kəniɛʔ**, anomalously in the AA context.

'Centipede' clearly came into Aslian with the *iː vowel, and this timbre remains unchanged in Southern Aslian, so the lowering/breaking of this vowel must have occurred in CNAslian. Perhaps the same happened with 'footprint' (now lost from Southern Aslian) and with 'to steal' with back-borrowing into Southern Aslian, hence proto-Aslian ***kəʔiːp**, ***tiːl**, ***siːc** respectively. Why this particular vowel development in these etyma? Above we saw that mostly *iː was stable in Southern Aslian but lowered to ***ɛ**ː in CNAslian, but just as there was some instability in the lowering of ***u**ː, there was also some in the lowering of ***i**ː, and I suppose it happened that some did not make it all the way to ***ɛ**ː.

We now move on to the last part, the analysis of the back diphthongs. TP offers some 22 etyma over 4 tables, all reproduced below. The basic proposition of distinguishing proto-Aslian ***ua** and ***uə** goes back to analysis in GD's (1976b) sketch of Jah Hut. According to GD, Jah Hut phonetically distinguishes five onglided nuclei: [wo] and [wə] from ***uə**, and [wɔ], [wa], [wɛ] from ***ua** (p.107). TP's correspondences further support this by aligning Semnam /ᵘoː/ with ***uə** and /ɔː/ with ***ua**. Southern Aslian lects tend to reflect ***uə** as /ɔ/ and ***ua** with /o/ although this does not always

hold.

TP's correspondence groups for these are tabled below (tables 21-24), and following that data of the wider AA etymologies is provided (table 25).

Table 21: Phillips (2012:130) *Table 3.81 Proto-Aslian *ua*

	PAs	CAs	JAH	NAs	SAs
'to dream'	*ʔəmpuaʔ	ʔm̥pooʔ SEA pɔɔʔ TEA tapɔʔ (?) SSM	ʔəmpwɔʔ	ʔm̥paʔ MNQ ʔipaʔ KNQ	po SZA m̥po SZC
'knee'	*kərual	kurool SEA karɔɔl TEA kayɔɔl SSM	kruwal	---	---
'roof'	*pəluaŋ	pəlooᵍŋ SEA pəlɔɔᵍŋ SSM	pərwɔŋ	---	pəl̥ok MHE paloᵍŋ SZC
'worm; caterpillar'	*kəmuar	kəmoor SEA kəmuur TEA	kəmuwar	kumɑy (?) CWG kəmay KNQ	kʰəmur SZA kəmuʟ SZC
'fingernail'	*cə(n)ruas	cəŋroos SEA cɛnᵈrɔɔs TEA cənyɔɔs SSM	cərwɛs	cəŋrɔs (?) MNQ cənᵈrɔs (?) JHI	cəros SZA
'one'	*[m/n]uay	niiy SSM nɛɛy TEA	niwɛy	nɑy CWG nay KNQ, KNS, MNQ, TNZ	muy SZA, SZC, TMO, MHE

Table 22: Phillips (2012:130) *Table 3.82 Proto-Aslian *uə*

	PAs	CAs	JAH	NAs	SAs
'fly (n)'	*ruəy	rəwɔɔy TEA rooy SEA	ruwey (?)	yay CWG yey KNS	rɔy SZA rɔy SZC
'to defecate'	*cuəh	co(o)h SEA	ciʔcuwoh	cicɒh CWG ʔəncah BTQ	
'dog'	*cuəʔ	cəwɔɔʔ TEA	cuwoʔ	wəʔ (?) BTQ	cɔ SZA
'navel'	*suək	sook SEA sᵘook SSM 'umbilical cord'	---	səksiak CWG sək (?) JHI 'umbilical cord'	---
'friend'	*ruəp	roop SEA ruwɔɔp TEA yᵘoop SSM	---	---	---
'to love'	*huəʔ	hooʔ SEA huwaaʔ TEA hᵘooᵈn (?) SSM	---	---	---

Table 23: Phillips (2012:132) *Table 3.83 Proto-Aslian **uəN*

	PAs	CAs	JAH	NAs	SAs
'child; children'	*kuən	k-n-oon PSEA kuwɔɔt TEA kəwɔn LNH k^{u}oo^{d}n SSM	---	kɛn BTQ kɛ^{d}n JHI kɛ^{d}n MNQ	kənɔn SZA, SZC, MHE, TMO
'inside'	*kəluəŋ	kəl^{u}oo^{g}ŋ SSM	---	kəliyɛ^{g}ŋ KNS kəlɛ^{g}ŋ JHI	---
'brain'	*lakuəm	lak^{u}oo^{b}m SSM kəlʔoo^{b}m SEA	lukom	kiɛ^{b}m CWG ləkiɛ^{b}m KNS ləkɛ^{b}m JHI kəlʔɛ^{b}m MNQ	lako^{b}m SZC
'to hold, grasp'	*kuəm	kəwɔɔm TEA k^{u}oom SSM	---	kəm (?) MNQ	kɨ^{b}m (?) SZC
'marrow'[46]	*suəm	las^{u}oo^{b}m SSM lɛmsuwaam TEA	suwəp (?)	səmsie^{b}m KNS	səmsɔ^{b}m SZA somsom MHE

Table 24: Phillips (2012:133) *Table 3.84 Proto-Aslian **uə residue*

	PAs	CAs	JAH	NAs	SAs
'to give'	*ʔuək (?)	ʔok TEA,SSM ʔɔɔk SEA	ʔɔk	ʔak CWG ʔɛk KNS, KNQ, TNZ, JHI, MNQ, BTQ	---
'to pull out'	*tuəs (?)	tɔs TEA tɔs SSM	---	təyes KNS tɨs CWG	---
'to whistle'	*[s/h]uəc (?)	həwɔɔc TEA hich^{u}ooc SSM	siʔyɔc	hiɟəhɨc JHI huchuac CWG həchɔc MNQ pichuc BTQ	sɔc SZA huwɛc (?) SZC həchɔc TMO məhɔc MHE
'to gape'	*kəhuəy (?)	kihɔɔy SEA kəhɔɔy TEA hihuəy SSM	keʔhway	hyhuay CWG hihay KNS hiyhəy JHI	hiyhɒy SZA hiyhɔy SZC ʔuhay MHE
'shoulder'	*kəlapuəh (?)	klapuəh SSM	kər(ə)paʔ (?) 'wing'	kɨlapəh JHI	kəmpoh (?) SZC

Table 25: Austroasiatic cognates of Phillips (2012) Proto-Aslian ***uə,*ua** etyma.

gloss	TP *V	pAA	Select AA cognates / Notes
‘dream’	***ua**	***ʔəm.’poːʔ**	Nancowry **ʔenfua**, Khmu **hmpoʔ**, Chong **pʰoʔ**, Arem **mpɔ**ː, Old Mon **ᵐpɔʾ** ‘dream’
‘worm, caterpillar’		**?**	Khmu **moːr** ‘worm with fuzz/hair’; proto-Waic ***kmɔr/l** ‘earthworm’
‘fingernail’		**?**	Lawa Umphai **ʔrɔs** ‘digit’
‘one’		***moːjʔ, *muəjʔ**	Kharia, **mɔiʔɟ**, Khmu **mo**ː**j**, Old Khmer **mu**ː**əj~mɔ**ː**j**, Chong **mo̤**ːʔ**j ~ mṳ**ːʔ**j,** Katu **muj**, Laven **mu**ː**j** ‘one’
‘fly (n.)’	***uə**	***roaj**	Sora **ro**ː**j**, Nancowry **juaj**, Khmu **rɔ**ː**j**, Bahnar **rɔ**ː**j**, Katu **rɔ**ː**j**, Mon **rùj** ‘fly’
‘dog’		***cɔʔ**	Khmu **sɔʔ**, Katu **ʔacɔ**ː, Laven **cɔ**ː, Arem **acɑ**ːʔ ‘dog’
‘child’	***uəN**	***koan**	Mundai **hɔn**, Nancowry **koan**, Khmu **kɔ**ː**n**, Katu **ʔakɑ**ː**n**, Bahnar **kɔ**ː**n**, Arem **kɑ**ː**n** ‘child’
‘inside’		***kə.’luːŋ, *kə.’luəŋ**	Khmu **kluəŋ**, Katu **kalɑ**ː**ŋ**, Stieng **klu**ː**ŋ**, Arem **tlɑ**ː**ŋ** ‘inside’
‘hold, grasp’		?	Riang **kuam**[1] ‘seize’, Wa **kn̤m** ‘hug’
‘whistle’	***uə residue**	***huəcʔ**	Lamet **hɑ**ː**c**, Old Khmer moːj, Khmer **hu**ː**əc**, Ngeq **kahuac**, Rục **hɔ́**ː**c** ‘whistle’

Of the 22 comparisons offered by TP in support of ***ua** and ***uə**, only 10 seem to have external support; on the face of it these are rather few examples upon which to establish regularity of correspondences; nonetheless, some insights are possible.

To begin with, it seems clear that proto-Aslian ***ua** aligns with pAA ***o**ː. The AA etymology for ‘one’ is ambiguous due to diphthonged reflexes in Katuic and Bahnaric, but otherwise AA forms strongly indicate ***mo**ː**j**ʔ; at the same time comparisons for ‘dream’, ‘worm, caterpillar’ and ‘fingernail’ are entirely consistent with this pAA ***o**ː, while no support for this value is found in the cognates of proto-Aslian ***uə**.

On the other hand, AA cognates with proto-Aslian ***uə** are more diverse. Notable are ‘fly’ and ‘child’ which I currently reconstruct with pAA ***oa**. As explained at the SEALS 32 meeting in Chiang Mai (May 2023),[11] I regard this proto-segment as a likely allophone of pAA ***ɔ**ː, but I am noting it with ***oa** because of diphthonged reflexes in Nicobarese that diverge from the dominant pattern. The ‘dog’ etymon is missing from Nicobarese, but we could reassign it to ***coaʔ** on the bases of the Aslian data. Similarly, ‘hold, grasp’ can also belong to this correspondence, but the data are ambiguous as we have only reflexes from Palaungic, which merges pAA ***ɔ**ː and ***uə** generally. At the same time, ‘whistle’ clearly indicates pAA ***uə** as reflexes are well distributed, and ‘inside’ is ambiguous, but I am inclined to give priority to the Khmu reflex and favour ***uə** in that etymon.

Consequently, we have evidence of proto-Aslian ***uə** reflecting both pAA ***ɔ**ː/***oa** and ***uə**. Harking back to our discussion around proto-Aslian ***ɔɔ$_1$**, ***ɔɔ$_2$** (see tables 10, 11), we found that ***ɔɔ$_1$** likely indexes a short proto-vowel, while ***ɔɔ$_2$** examples are

[11] See presentation archived at: https://drive.google.com/file/d/1iLJ2TTN3HuOixry8HDAnk2dx2i-UdoXC/view?usp=sharing

Aslian innovations. My suggestion is proto-Aslian actually merged *ɔː/***oa** and ***uə** into something like [wɔː], which shifted to [woː] in CNAslian, phonetically approximating the reflexes we still find in Jah Hut and Semnam. Historically ***wɔ**ː would have settled on [ɔ] in Southern Aslian, losing the ongliding as the length distinction was lost and this segment merged with short *ɔ. In Central Aslian the ongliding was mostly preserved, I suggest with an intermediate CNAslian form ***wo**ː, accounting for the shift to [o] timbre in Semai. In northern Aslian, the characteristic change was a dissimilation of the labial ongliding, combined with mergers, yielding the various central and front reflexes [a, ə, ɛ, iɛ, e] attested in TP's comparisons.

7 Phillip's proto-Aslian diphthongs

Our review of the data and reconstructions in the light of wider AA comparisons leads to a significant revision of proto-Aslian vocalism. Table 26 broadly summarises the results, excluding the roles of external and internal borrowing and idiosyncratic changes.

Table 26: Schematic summary of revised Proto-Aslian long vowels & diphthongs.

PAA (Sid2024)	**Revised Proto-Aslian**	**Revised Proto-CAslian**	**Proto-Aslian (TP)**	NAslian reflexes	Jah Hut reflexes	Semai reflexes	Other CAslian reflexes	SAslian reflexes
***e**ː	***e**ː	***i**ː	***i**ː₁	i	i	iː	iː	i/e
Ø	Ø	***i**ː	***i**ː₂	ɨ	i	iː	iː	?
Ø	***e**ː	***e**ː	***e**ː₁	ɨ/ə	ə	eː	eː	e
Ø	***e**ː	***e**ː	***e**ː₂	i/ɛ?	ɛ	eː	eː	ɛ?
***i**ː	***i**ː	***ɛ**ː	***ɛ**ː₁	ɛ/a	ɛ	ɛː	ɛː	ɛ
***e**ː	***e**ː	***ɛ**ː	***ɛ**ː₂	o	ə	ɛː	ɛː	ɛ
***ɨ**	***ɨ**	***ɨ**ː	***ɨ**ː	ɨ	ɨ	ɨː	ɨː	ɨ/o
u**ː/ə**ː	***u**ː/***ə**ː	***ʊ**ː~ ɤː	***ɯ**ː	ɨ	u	ɤː	?	u
***u**ː	***o**ː	***ʊ**ː	***u**ː₁	u	o/ɔ	uː	uː	o
Ø	***o**ː	***ʊ**ː	***u**ː₂	o	o	uː	uː	o?
***u**ː	***u**ː	**o**ː	***o**ː₁	ɔ	ɔ	oː	oː	u
***u**ː	***u**ː	**o**ː	***o**ː₂	ɨ	o	oː	oː	u
*ɔ	***ɔ**ː	***ɔ**ː	***ɔ**ː₁	ɔ	ɔ	ɔː	ɔː	ɔ
Ø	***ɔ**ː	***ɔ**ː	***ɔ**ː₂	ɔ	ɔ	ɔː	ɔː	u
Ø	***a**ː	***a**ː	***a**ː₁	a	a	aː	aː	a
***a**ː	***a**ː	***a**ː	***a**ː₂	i/e	a	aː	aː	a
***iə**	***iɛ**	***iɛ**	***iɛ**	ɛ/iɛ	ɛ/iɛ	ɛː/eː/iː/jaː	iɛː/ieː/ɛː	ɛ
*oː	***uo**	***ua**	***ua**	a	wɔ/wa/wɛ	oː	ɔː	o
*ɔː/***oa**	***wɔ**ː	***wo**ː	***uə**	a/e?/ia?	wo	oː	wɔː/ ᵘoː	ɔ
***uə**		***wo**ː	***uəN**	ɛ/iɛ	o?	oː	wɔː/ ᵘoː	ɔ

From a historiographic perspective, I propose that GD's historical analyses that focused on Semai and Central Aslian, had a skewing effect on later scholarship, which we are now redressing. While the issues of syllable and word structure, and proto-consonants, were essentially worked out in the 1970s, and laid strong bases for more recent substantial progress in pAA reconstruction, more recent work also allows us to look back and self-correct. This is the nature of scientific progress, it precedes iteratively,

stepping forward (sometimes to the side or even backwards!) to converge on the truth over time. It is the boldest among us who take the first steps into the unknown.

References

Benjamin, Geoffrey. 1976. Austroasiatic Subgroupings and Prehistory in the Malay Peninsula. In Philip N. Jenner, Laurence C. Thompson, and Stanley Starosta (eds.) *Austroasiatic Studies*. Honolulu, University of Hawaii Press (Oceanic Linguistics Special Publications No. 13). pp: 37-128.

Diffloth, Gérard. 1968. Proto-Semai Phonology. *Federation Museums Journal* (new series), 13: 65-74.

Diffloth, Gérard. 1975. Les langues Mon-Khmer de Malaisie: classification historique et innovations. *Asie du Sud-Est et Monde Insulindien* 6.4:1-19.

Diffloth, Gérard. 1976a. Minor-syllable vocalism in Senoic languages. In Jenner, Philip N., Thompson, L. C. & Starosta, S. (eds). Austroasiatic studies, Part I, pp. 229-247. Honolulu: University Press of Hawaii.

Diffloth, Gérard. 1976b. *Jah-Hut, an Austroasiatic language of Malaysia*. in N.D. Liem (ed.). *Southeast Asian Linguistic Studies* Vol.2. Canberra, Australian National University, Pacific Linguistics (vol. C-No.42.) pp. 73-118.

Diffloth, Gérard. 1977. Towards a History of Mon-Khmer: Proto-Semai Vowels. *Tônan Ajia Kenkyû (Southeast Asian Studies)* 14.4:463-95.

Dunn, Michael, Niclas Burenhult, Nicole Kruspe, Sylvia Tufvesson, and Neele Becker. 2011. Aslian linguistic prehistory: A case study in computational phylogenetics. *Diachronic*a 28.3:291-323.

Phillips, Timothy. 2005/2013. *Linguistic Comparison of Semai Dialects*. SIL Electronic Survey Report 2013-010. (The original report was filed with the Economic Planning Unit, Prime Minister's Department, Malaysia, in 2005 and later published by SIL in 2013).

Phillips, Timothy. 2012. *Proto-Aslian: towards an understanding of its historical linguistic systems, principles and processes*. PhD thesis, Institut Alam Dan Tamadun Melayu Universiti Kebangsaan Malaysia.

Shorto, Harry L. 1960. Word and syllable patterns in Palaung. *Bulletin of the School of Oriental and African Studies* 23:544-57.

Shorto, Harry L. 1963. The Structural pattern of northern Mon-Khmer languages. In H.L. Shorto, (ed.) *Linguistic Comparison in South-East Asia and the Pacific*. pp 45-61.

Sidwell, Paul and Mark Alves. 2023. Re-Evaluating Shorto's MKCD Reconstructions. In Paul Sidwell & Mark Alves (eds.) *Papers from the Ninth and Tenth International Conference on Austroasiatic Linguistics*. JSEALS Special Publication No. 12. Manoa, University of Hawaii Press. E-ISSN: 1836-6821

Sidwell, Paul. 2015. *The Palaungic Languages: Classification, Reconstruction and Comparative Lexicon*. Munich, Lincom Europa.

Sidwell, Paul. 2018. *The Khasian Languages: Classification, Reconstruction and Comparative Lexicon*. Munich, Lincom Europa.

Sidwell, Paul. 2021. Classification of MSEA Austroasiatic languages. In Paul Sidwell & Mathias Jenny (eds.) *The Languages and Linguistics of Mainland Southeast Asia: A Comprehensive Guide*. Walter de Gruyter: Berlin/Boston. pp.179-206.

Sidwell, Paul. 2024. 500 Proto Austroasiatic Etyma: Version 1.0. Journal of the Southeast Asian Linguistics Society 17.1:i-xxxii

7

Phonetic and Phonological Analysis of the Mundari Vowel System[1]

Pamir Gogoi, Luke Horo and Gregory D. S. Anderson

1 Introduction

Mundari is a Kherwarian North Munda language spoken in several regions of India like Jharkhand, Odisha, Chhattisgarh, West Bengal and Assam, and in the neighbouring country Nepal. The Registrar General and Census Commissioner of India has separate entries for Munda and Mundari as a "mother tongue" whereby Munda is reported to have 413,894 speakers and Mundari is reported to have 861,378 speakers. However, people who identify themselves with either nomenclature speak the same language. Also, there are four named 'dialects' traditionally reckoned among the community, namely, Hasadaʔ, Naguri, Keraʔ and Tamaɽia but their true linguistic categorization is yet to be established[2]. Mundari has been studied by scholars in the past but instrumental analysis of the Mundari sound system is lacking. In this paper, we present preliminary observations on the phonetic properties of Mundari vowels and substantiate the findings with acoustic analysis. Existing literature mentioning the sound system of Mundari include Whitley (1873), Nottrott (1882) and Gumperz and Biligiri (1957). Subsequently, Cook (1965), Sinha (1975) and Osada (2008) have discussed the Mundari sound inventory with more detail.

The vowel system of Mundari proposed by Cook (1965) and Osada (2008) suggest that Mundari has five phonemic vowels (see Table 1).

[1] Preliminary versions of this study were made in presentations at the SEALS conference in Chiang Mai, Thailand and HISPhonCog conference in Seoul, South Korea. Support for this research was made possible under award BCS#2041248 from National Science Foundation "Words, phrases and sentences at the interface of phonology and syntax" and award PD-281083-21 National Endowment for the Humanities/DLI-DEL "Documentation and analysis of seven Munda languages and development of the Munda Virtual Archive". This support is gratefully acknowledged. Our paper is dedicated to the late Professor Gérard Diffloth, whose interest in all Austroasiatic languages including Mundari serves as an inspiration to all present and future generations of scholars working in this area.

[2] There are minor details like some voiceless stops are aspirated initially for certain Naguri speakers in comparison with Hasadaʔ or some monosyllabic words in Hasadaʔ are disyllabic with medial -h- in Naguri–but with no consistency in a historical sense whether these /h/s are old or new–or that the default position for subject clitics in both of these varieties is attached to the word immediately preceding the verb, but in Keraʔ it typically comes after the verbal word–possibly reflecting the likely Dravidian substrate in this lect, while in Tamaɽia subject clitics often appear before and after the verb. For our present purposes, the vowel systems of the named lects do not differ in meaningful ways.

Table 1: Vowel inventory of Mundari following Osada (2008)

	Front	Central	Back
High	i		u
Mid	e		o
Low		a	

Cook (1965) also argues that Mundari vowels can vary in length, nasalization or glottalization. While vowel length and nasalization are not considered phonemic, Cook suggests that the glottalized vowels, appearing in the VʔV environment, are a separate set of vowels in the language. Additionally, it is argued that the glottalized vowels always appear in VʔV environment and never in V_1-ʔ-V_2, therefore the glottal stop is not considered a separate consonant but a vocalic feature. However, the distribution of glottal stop is more complex than previous researchers have acknowledged. We remain agnostic here as to whether checked or glottally interrupted vowels are phonemically distinct sets of vowels as this has not yet been experimentally analysed. Regardless, claims that glottal stop can never occur between dissimilar vowels in Mundari is not supported by our field data. On the other hand, the work by Sinha (1975) did not clearly establish the vowel inventory of Mundari. He presented a list of 50 sound segments that he claims are phonetic representations of the vowels occurring in Mundari. However, this clearly exceeds what experimental data suggest could be the maximal vowel inventory size of the language, and indeed exceeds the likely number of total phones encountered. Thus, the work by Sinha (1975) reveals an over-differentiation of Mundari vowels. Subsequently, the most detailed and recent literature on Mundari is found in Osada (2008). Osada's work confirms the five-vowel system in Mundari offering minimal pairs; see Table 1 for the list of these phonemes. Osada also reports the presence of vowel length in Mundari under specific contexts but does not represent vowel length as a phonemic contrast in Mundari, such as is found in the closely related language, Ho. For instance, the Mundari disyllabic forms with intervocalic retroflex *ɽ* such as *hoɽo* 'person' are realised as monosyllabic lexemes with long vowels as in *ho:* 'person' in Ho (Anderson, Osada and Harrison 2008). However, while *phonemic* vowel length is unattested in Mundari, our field data provides evidence that utterance final lengthening is a *phonetic* property of Mundari vowels.

While the five-vowel system in Mundari is adequately justified by Osada (2008), the current work offers additional data to show the contrastive nature of the Mundari vowel system, offering minimal pairs that show the vowel contrasts in different syllabic structural types; these are listed in the Appendix of the paper. Thus, we find vowel contrasts in closed monosyllables, open monosyllables, in the first syllable of disyllables and in the second syllable of disyllables. All cited data used in this work are curated from field notes and supplemented by data from Hoffmann (1930-1978). We recorded speech data from native Mundari speakers. Based on acoustic analysis of the speech data we elaborate the phonetic properties of the Mundari vowel system in the paper. Thus, in the following section, §2, we discuss the methodology of the study, explaining the data source and the analysis procedure. Then in §3 we detail the findings of the study and explain the interpretation of the result and in §4 we discuss the typological and Munda-specific implications of the findings along with a few unresolved queries related to the vowels in Mundari; finally in §5 we conclude by summarizing the key findings of the paper.

2 Method

This is a preliminary study exploring the phonetic properties of vocalism in Mundari. This forms a small part of a larger project that seeks to map phono-prosodic structures onto the complex grammatical words in the Munda languages. In this section we detail how the speech data was collected in the field and the methods we adopted for analysing the vowel system in Mundari.

2.1 Speakers

The current work is based on the speech data of four native Mundari speakers, two males and two females, with three speakers having an average age of 20 years and one male speaker being 50 years old. No age graded or gender-based differences in the vowel qualities or distributions examined here were observed. The speakers live in four different villages, namely Bari, Sarjama, Sonmer and Bongda, all located in the Khunti district of Jharkhand. Although the dialectal variation of Mundari has not yet been linguistically established, the participants of the study identified themselves as belonging to the Has(a)daʔ variety. All four speakers were bilingual in Mundari and Hindi. One male and one female speaker were pursuing postgraduate studies at the time of their recording and the other two had formal education until high school.

2.2 Procedure

The text data used for this study consists of a wordlist of 400 Mundari words having 100 monosyllables and 300 disyllables in different syllable structures (see Appendix for sample wordlist). The speech data is generated by recording the four speakers while they produce the target words once in isolation and then in three different sentence frames. The sentence frames consisted of the following intonational contexts as given in (1)-(3), with (1) a carrier phrase, (2) an out of focus phrase and (3) exclusive focus phrase.

(1)

bikram _____ *kadʒime* ‘Bikram, say _____’
bikram _____ *kadʒi-me*
Bikram _____ say-2SG

(2)

bikram _____ *kadʒime surdʒan do ka* ‘Bikram, say_____, not Surjan’
bikram _____ *kadʒi-me surdʒan do ka*
Bikram_____ say-2SG Surjan TOP NEG

(3)

bikram _____ *kadʒime dʒohar do kage* ‘Bikram, say_____, not johar’
bikram _____ *kadʒime dʒohar do ka-ge*
Bikram _____ say-2SG greetings TOP NEG-EMPH

To elicit the target words, the speakers were prompted with Hindi translation of the target words and the sentence frames, which they then responded to by producing the Mundari words in isolation and in the three sentence frames in Mundari as well. Recordings were conducted in the field outdoors using a head-worn unidirectional

microphone connected to a solid-state recorder. The data was recorded at a sampling frequency of 44100 Hz and the recorded sound files were manually annotated at the word and phoneme level in Praat (Boersma & Van Heuven, 2001) by trained phoneticians. From the annotated data, a total of 7106 vowel tokens were included in this study. Vowel samples which were identified as creaky or nasal by the annotators using spectrographic and perceptual cues were excluded from this analysis. This was done because non-modal voice quality and nasality can perturb the formant frequencies and may affect the results of our formant analysis. The distribution of the count of vowels across the four speakers are plotted in Figure 1.

2.3 Analysis

Acoustic analysis of the speech data is done by measuring the vowel formants of all tokens annotated in the dataset. The first two formant frequencies F1 and F2 of each vowel were extracted using a Praat script. The formant frequencies were extracted at the vowel midpoint between the beginning and end of the glottalic pulses. The formant ceiling for male speakers were set at 5000 Hz and for female speakers were set at 5500 Hz. Subsequently, the data was subjected to statistical testing whereby significance of the vowel categories were measured using a one-way ANOVA test. ANOVA models for both F1 and F2 were performed separately. Also, a Tukey's HSD post-hoc test was conducted to find the pair-wise difference between the F1 and F2 of each vowel. Additionally, to investigate the difference between vowels in monosyllabic and disyllabic words, independent one-way ANOVA tests were performed on F1 and F2 measures of individual vowel categories, with syllable position as the dependent variable.

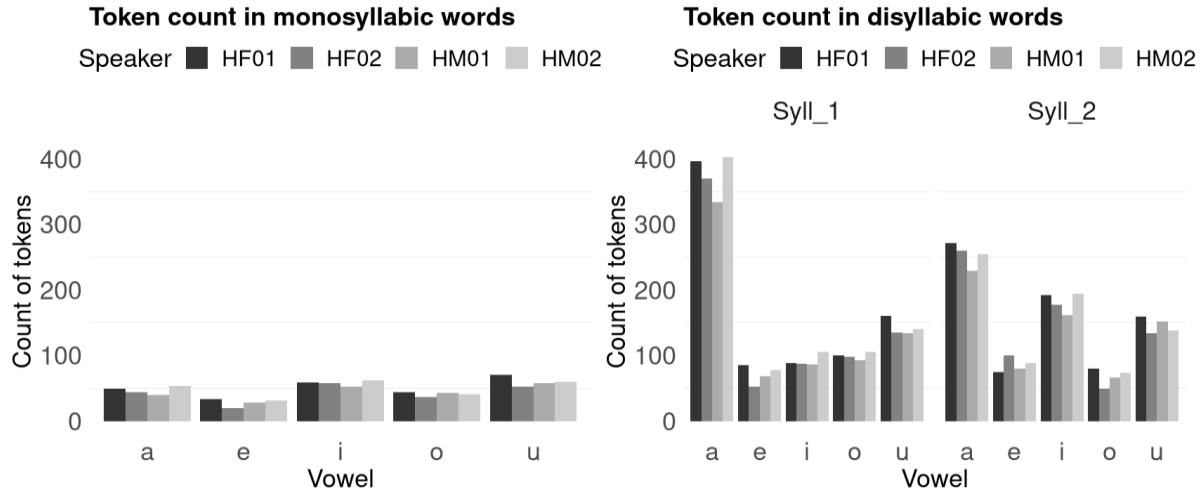


Figure 1: Distribution of vowel tokens in monosyllabic and disyllabic words

3 Results

The first and the second formant frequencies of Mundari vowels reveal that the five vowels in Mundari are categorically distinct in the vowel acoustic space. Considering the vowel system in monosyllabic words first, the mean F1 and F2 frequencies of /i, e, a, o, u/ are plotted in Figure 2. In Figure 2, formant frequencies of all the vowels are normalised across speakers and the mean is located in the white square with the ellipse showing 80% confidence level. The average formant frequencies and standard deviation of each vowel in the monosyllabic words is presented in Table 2. The mean F1 and F2 frequencies of the five vowels plotted in Figure 2 show the height and frontness features of the five vowels in Mundari. Also, it is observed that in monosyllabic words, there is a small overlap between the vowel pairs /i/ - /e/ and /u/ - /o/. This indicates the high vowels and mid vowels in Mundari may phonetically vary in quality. However, the results of a one-way ANOVA test with F1 as the dependent variable and vowels as the independent variable shows a significant effect of vowel type on F1 (F (4, 7210) =, p < .001) and the post-hoc Tukey HSD test show a significant difference between all vowels except /o/ an /e/. This is expected because F1 is a measure of vowel height and both /o/ and /e/ are mid vowels in the Mundari vowel system. Similarly, results of one-way ANOVA test with F2 as the dependent variable and vowels as the independent variable show significant effect of vowel type on F2 (F (4, 7210) =, p < .001) and the post-hoc Tukey HSD test show a significant difference between all vowels. Hence, the analysis confirms that although high and mid vowels in Mundari monosyllables may vary in their phonetic quality, they are still distinct vowel segments in the Mundari vowel system.

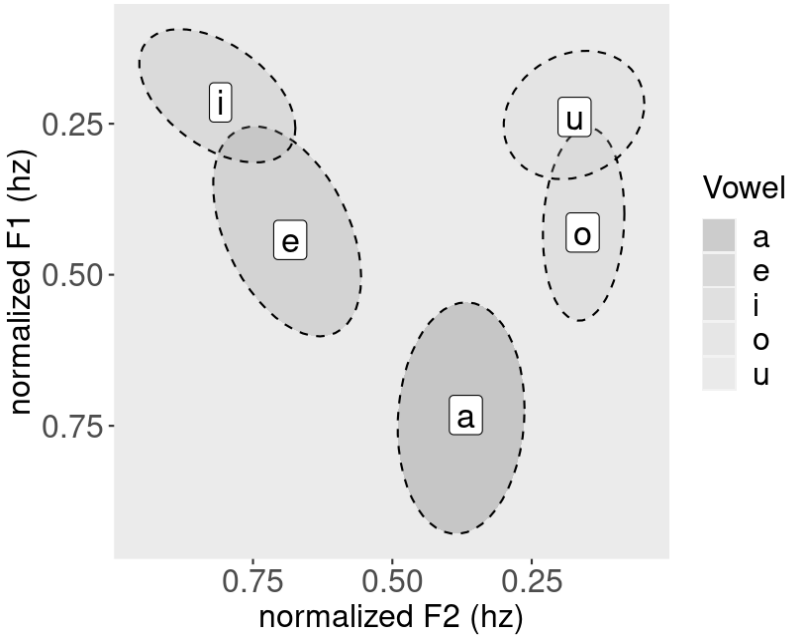


Figure 2: Vowels in monosyllabic words plotted with speaker normalized F1 and F2 values

Table 2: Mean and standard deviation of vowel formants in monosyllabic words

Vowel	Mean F1	SD F1	Mean F2	SD F2
/a/	780	95.6	1429	178
/e/	591	94.9	2064	112
/i/	441	66.4	2334	232
/o/	582	70.7	1015	171
/u/	457	57	1037	183

While the five Mundari vowels occurring in monosyllabic words were found to be significantly distinct from each other, we analysed the vowels occurring in the first and the second syllable of disyllabic words separately in order to investigate the likelihood of syllable position affecting the vowel qualities. Figure 3 shows the vowel plots drawn from the vowels occurring in the first and the second syllable of disyllabic words. The vowel formants are plotted with normalised F1 and F2 values. Also, similar to Figure 2, the white squares in Figure 3 represent the mean values and the ellipse represents an 80% confidence level. The mean and standard deviation of each vowel in the first and second syllable of disyllabic words are presented in Table 3. From the vowel plots in Figure 3 it is observed that Mundari vowels in the first and the second syllable of disyllables generally have a similar pattern with the vowels in the monosyllables. It is found that the five-vowel system is present in both the first and the second syllable of disyllabic words. This is also confirmed by the one-way ANOVA test performed on the vowels in the first syllable of disyllabic words. By considering F1 as the dependent variable and vowels as the independent variable, we find a significant effect of vowel type ($F(4, 3113) = 5306$, $p < .001$), and the post hoc tests shows a significant difference between the F1 measures of all vowel pairs occurring in the first syllable. Similar pattern was observed for vowels in the second syllable. One-way ANOVA test showed a significant effect of F1 ($F(4, 2930) = 3431$, $p < .001$) and significant difference between all vowel pairs in the second syllable as well.

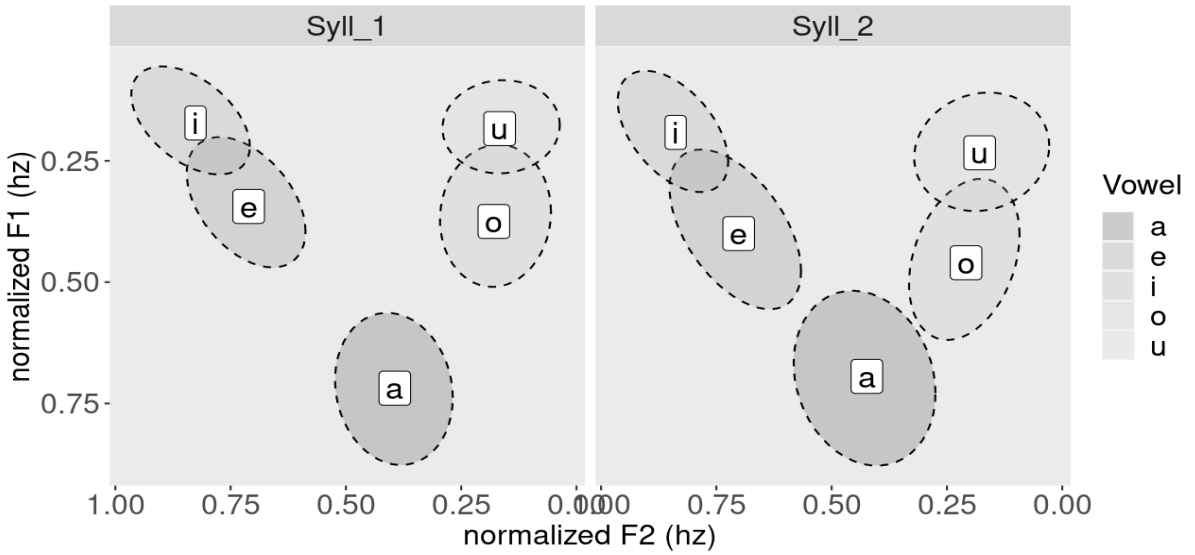


Figure 3: Vowels in disyllabic words in syllable 1 and syllable 2, plotted with speaker normalized F1 and F2 values

Table 3: Mean and standard deviation of vowel formants in disyllabic words

	Syllable 1				Syllable 2			
Vowel	Mean F1	SD F1	Mean F2	SD F2	Mean F1	SD F1	Mean F2	SD F2
/a/	775	85.1	1491	184	762	99.2	1547	196
/e/	526	74.6	2127	192	566	90.2	2119	214
/i/	411	58.5	2360	220	427	66.7	2395	227
/o/	546	69.8	1049	186	602	76.8	1101	158
/u/	421	52.8	1028	179	455	64.4	1048	180

Additionally, similar to the mid vowels in monosyllabic words, it is evident that mid vowels in the first and the second syllable also overlap with high vowels in disyllabic words. The extent of vowel overlaps between the mid and the high vowels in the vowel acoustic space in monosyllabic and disyllabic words are demonstrated with vowel density plots in Figure 4 and Figure 5. This indicates that there is a consistency in phonetic viability between the high and the mid vowels in disyllabic words as well.

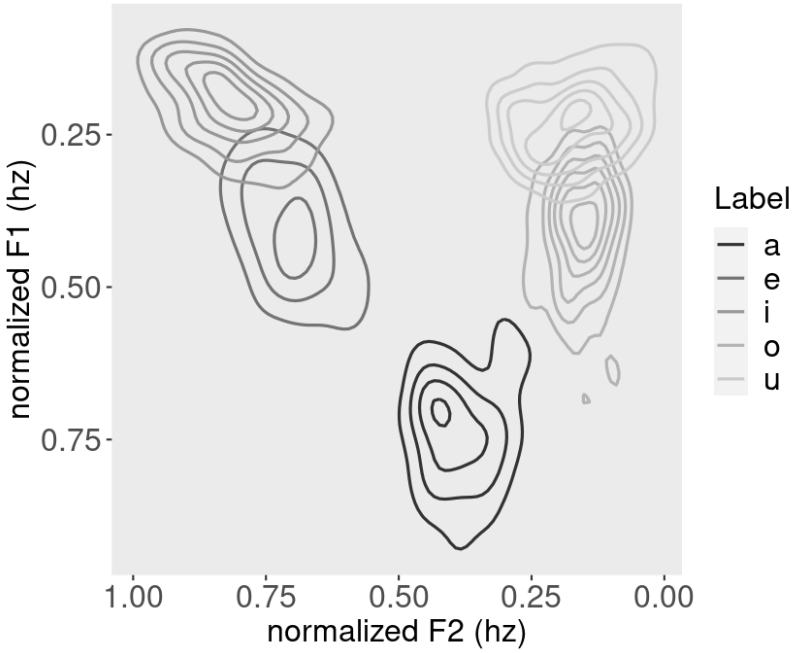


Figure 4: Vowels density in monosyllabic words, plotted with F1 and F2 values normalized across speakers

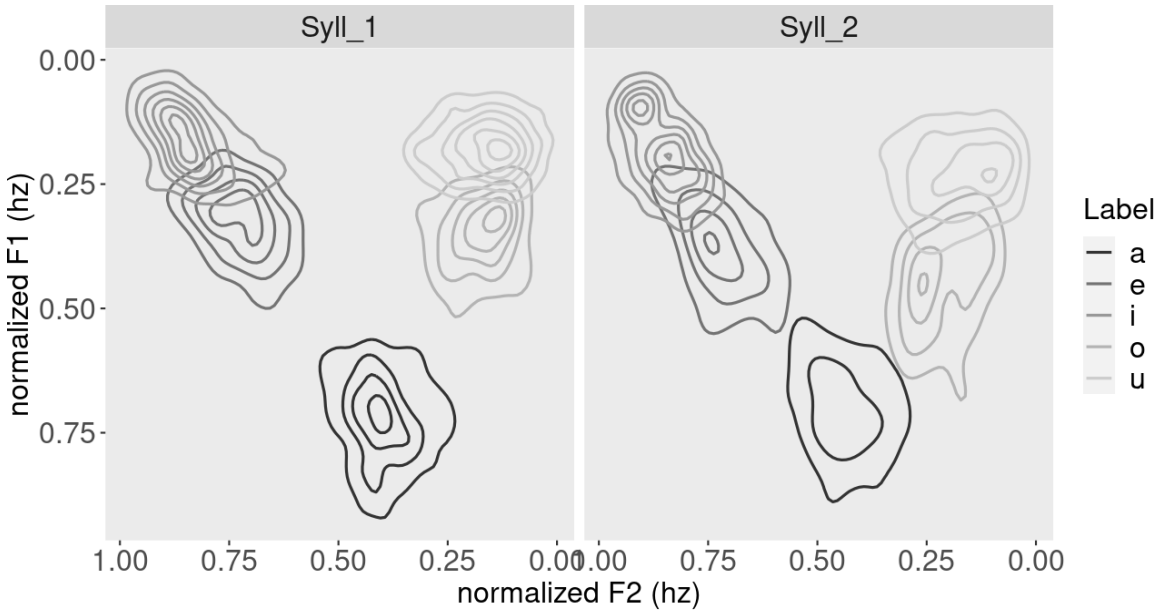


Figure 5: Vowels density in first and second syllable of disyllabic words, plotted with F1 and F2 values normalized across speakers

However, a comparison between the vowels occurring in the first and the second syllable of disyllabic words reveals that the mid vowels in the second syllable are lower than the mid vowels in the first syllable. The vowel polygon plot in Figure 6 shows the Mundari vowel system drawn from different syllable positions and plotted on the same F1 and F2 axis. It is observed that both front and back mid vowels are lower in the second syllable when compared with the same vowels in the first syllable in disyllabic words.

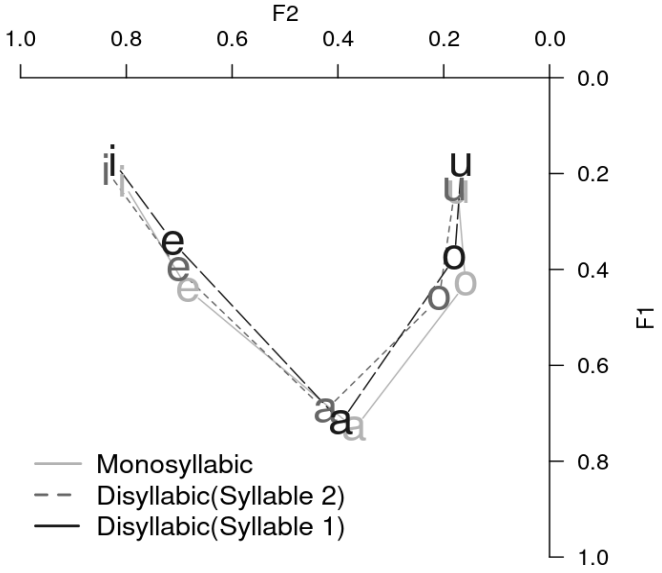


Figure 6: Mundari vowels across monosyllabic and disyllabic words

Thus, the vowel plots in Figure 6 indicate that Mundari speakers tend to produce higher (closed) and lower (open) variants of the mid vowels /e/ and /o/ in the disyllabic words. It is observed that the mid vowels are higher in the first syllable but are lower in the

second syllable.[3] In order to further examine the difference in vowel dispersion between the five vowels in Mundari, when they occur in the first and the second syllable of disyllabic words, we calculated the Euclidean distance between each vowel pair using equation (4).

(4) $$d_{xy} = \sqrt{(F1_x - F1_y)^2 + (F2_x - F2_y)^2}$$

Euclidean distance measurement is used to calculate a straight-line distance between two points in a given space. Studies have shown that perceptual difference between categorical vowels is measurable by equating the dispersion between the vowel categories in a vowel acoustic space (Liljencrants & Lindblom 1972, Lindblom 1986, 1990). In a two-dimensional vowel acoustic space, the point where a vowel segment is located represents the combined effect of its F1 and F2 frequencies. Therefore, in equation (4) d_{xy} represents the Euclidean distance between two vowels *x* and *y* based on their formant frequencies of F1 and F2. Results of the Euclidean distance measurement is presented in Table 4.

Table 4: Euclidean Distance between vowel pairs in first and second syllable

	Syllable 1				Syllable 2			
Vowel	/e/	/i/	/o/	/u/	/e/	/i/	/o/	/u/
/a/	683.01	942.16	497.8	582.83	604.65	911.77	473.83	585.88
/e/		259.83	1078.19	1104		309.03	1018.64	1076.74
/i/			1317.93	1332.04			1305.78	1347.29
/o/				126.75				156.26

From Table 4 it is evident that vowel dispersion between the front high vowel /i/ and the front mid vowel /e/ is greater in the second syllable than in the first syllable indicating that /e/ in first syllable is higher/closed and /e/ in second syllable is lowered/open. Similarly, vowel dispersion between high back vowel /u/ and the mid back vowel /o/ is found to be greater in the second syllable than in the first syllable indicating that /o/ in first syllable is higher/closed but /o/ in second syllable is lowered/open. Additionally, it is observed that, due to the higher and lower realisations of the mid vowels, vowel dispersion between the mid vowels and the low central vowel /a/ is also affected. While the low central vowel /a/ is more dispersed from the higher mid vowels in the first syllable, the dispersion is reduced with the lower mid vowels in the second syllable.

In comparison to the mid vowels in monosyllabic words, it is found that both front and back mid vowels in monosyllables are lower than their higher/closed counterparts in the first syllable (see Figure 6). Whereas the mid vowels in the second syllable appear to be closer to the mid vowels in the monosyllables. Thus, if we consider vowel realisations in monosyllables as the canonical form of Mundari vowels, it can be postulated that the higher/closed variants of the mid vowels in the first syllable are a departure from the canonical form of the Mundari vowel system.

In order to find additional evidence for this observation, we conducted a one-way ANOVA test on the F1 and F2 measurements of the vowels of the same category but

[3] Note that the West Bengal variety of Santali appears to have a phonemic contrast between higher/closed and lower/open mid vowels, at least as reported by Bodding (1922, 1929-36) and Ghosh (2008). However, the Jharkhandi lect described by Minegishi (1990) seems to lack this.

across three different groups. The first group compares F1 and F2 of all vowels in monosyllables and the first syllable of disyllabic words, the second group compares F1 and F2 of all vowels in monosyllables and the second syllable of disyllabic words and the third group compares the F1 and F2 of all vowels in the first syllable and the second syllable of disyllabic words. Results of the ANOVA test is presented in Table 5 where (*) indicates statistical difference and (!) indicates lack of statistical difference between similar vowels in different syllable positions.

Table 5: Comparison of vowels of the same category but across three groups

	F1					F2				
	i-i	e-e	a-a	o-o	u-u	i-i	e-e	a-a	o-o	u-u
Mono Vs. 1st	*	*	!	*	*	*	*	*	*	!
Mono Vs. 2nd	*	*	*	*	!	*	!	*	*	!
1st Vs. 2nd	*	*	*	*	*	!	!	*	*	*

In Table 5 the F1 measurements reveal that vowels in the first syllable of disyllabic words are always significantly higher than vowels in the second syllable. Likewise, vowels in monosyllabic words are observed to be significantly lower than both first and second syllables of disyllabic words, except for the vowels /a/ and /u/. With regards to the F2 measurements, it is observed that all vowels except the front vowels /i/ and /e/ are significantly more back in the first syllable compared to the second syllable in disyllabic words and vowels in monosyllabic words are significantly more back than both first and second syllables of disyllabic words, except for the vowels /e/ and /u/. Hence, the statistical tests confirm the initial observation that, if the Mundari vowels are best represented in monosyllables, vowels in the first syllable appear to be further away from the canonical vowel system.

Moreover, the overall vowel area space of the Mundari vowel system in the first and the second syllable of disyllables are observed to be different. Table 6 presents the polygon area of the vowel acoustic space of the Mundari vowel system based on their positions of occurrence.

Table 6: Vowel area Space of Mundari vowels in different syllable positions

	Monosyllable	Syllable 1	Syllable 2
Polygon Area	262816.7	270836.7	254833.5

The vowel area space of the Mundari vowel system in the first syllable is found to be larger than the vowel area space in the second syllable. It is likely that the vowel area space in the first syllable is expanded due to the increased dispersion between the low central vowel /a/ and the higher/closed variants of the mid vowels in the first syllable. Lessening the dispersion between the low central vowel /a/ and the lower/open variants of the mid vowels in the second syllable results in compression of the vowel area space in the second syllable. In comparison to the vowel area space of Mundari vowel system in monosyllables it is observed that the vowel area space in monosyllables is smaller than in the first syllable but larger than in the second syllable of disyllables. However, from the vowel area space measurements it cannot be generalised whether an expansion or a compression in the vowel area space is a departure from the canonical vowel area space of Mundari vowel system. Regardless, it is confirmed that syllable positions in disyllabic words can affect the vowel quality of the mid vowels as well as affect the overall vowel area space of the Mundari vowel system.

4 Discussion

Mundari has clear experimental acoustic evidence for a canonical five vowel system where the features of height, frontness and roundedness are active, yielding an inventory of /a, e, i, o, u/. Also, a predictable allophonic variation is observed. This entails a lowering of the front and back mid vowels /e/ and /o/ in the second syllable of disyllables. Consequently, the mid vowels become closer to the low vowel phoneme /a/ in the second syllables than the corresponding vowels are in the first syllable of disyllables. Besides, the vowels in monosyllables occupy a position between the first and the second syllable, although closer to the second syllable of disyllables than the first. Correspondingly, in the first syllable of disyllables, the mid vowels are closer to the high vowels /i/ and /u/ instead. Areally speaking, eastern Indo-Aryan languages typically favour lower-mid realizations of [o] as [ɔ], as does the relatively geographically close Munda language Juang, so realizations of the back rounded mid vowel as either [o] or [ɔ] in Mundari, the latter mainly in the second syllable of disyllables or in monosyllables, is not surprising. Also, several Munda languages show at least phonetic realizations of /e/ as [ɛ], so again realizations of the front unrounded mid vowel as [e] or [ɛ] in Mundari, the former most common in the initial syllable of disyllables, is in line with related languages. However, it is not yet clear if the lowering of mid-vowels is an archaic retention of an earlier Kherwarian system which likely contrasted higher/closed and lower/open mid vowels, a pattern still found in the eastern varieties of Santali spoken in West Bengal and which likely characterized earlier stages in the history of the Munda languages as a whole, since this closed/open contrast in mid-vowels is reconstructed to proto-Austroasiatic (Sidwell and Rau 2015). In other words, even in Munda languages with five or six vowel phoneme inventories synchronically like the five of Mundari, there are often traces of earlier systems, specifically of ones which likely had a closed/open contrast in mid-vowels in previous historical periods.

In addition to the phonetic properties of the Mundari vowels described above it is observed that, despite having only five phonemic vowels, there are significant co-occurrence restrictions at the foot level in Mundari. With regards to disyllables in Mundari, there appears to be a robust, foot level co-occurrence restriction (or harmonic restriction) against high vowels and mid vowels appearing in the same foot. This harmonic restriction also has reflexes in the inflection of, for example, the demonstrative stem, which is *ni* and its plural *niku*, when the productive plural marker is *-ko*, cf. the third plural pronoun *ako*. This means Mundari is like its sister Kherwarian languages Ho and Santali (of West Bengal), that exhibit foot-based vowel harmony systems (Anderson, Horo and Harrison 2024). In Ho, more than one harmony pattern is observed. According to Pucilowksi (2013), /i/ and many forms with /u/ trigger a fronting and raising of /a/ > /e/ in certain lexemes and affixes, e.g., *cilike* (~ non- harmonizing Ho *cilika* ‘how’) and *muɳeʔ* (*~ muɳa(ʔ)* ‘nose’), *tingu- eke- n- e* ‘was stood,’ *surbuɖ- te- ɖ- e* ‘he tucked it,’ *ir- ten- e* ‘harvesting,’ *huɟuʔ- ye- n- e* ‘s/he came’. This of course has the result of creating forms that are not permitted in Mundari. However, there are other patterns observed in Ho which are formally similar to Mundari, specifically raising of /o/ to [u] when adjacent to /i/, e.g. *puti* < Hindi *pothi* ‘book’ or *pulis ‘police’* < Eng. police, *dul- i- ja* ‘(they) pour it,’ < *dol*, or *bai- juʔ- wa* ‘can be made’ with the imperfective potential/ middle suffix *- (j)oʔ-* (Pucilowski 2013: 41, 43). In Santali as well there is more than a single harmonic pattern, at least in the West Bengal variety described by Bodding (1922, 1929-36) and Ghosh (2008), and moreover they are all different from the one attested in Mundari. In one type /a/ does not co-occur with /i/ or

/u/, only [ə] may instead, e.g. *əgu* 'bring', cf. Mundari *agu* 'bring' or *hipiskə* 'envy each other' (Bodding 1929-36: 159). Again, /a/ is the only non-high vowel that /i/ and /u/ can co-occur in the same foot within Mundari, so the systems are quite distinct. Another harmonic system is essentially an ATR-like harmony among mid-vowels, as in this variety of Santali there are phonemic upper-mid and lower-mid vowels in both the front and back of the vowel space. Most phonological feet permit either one set or the other *ɛkɛn* 'empty,' *ɛmɔʔ* 'giving, liberal,' *ɔnɛn* 'when,' *ɔmɔn* 'bring forth', *epel* 'raise,' *eŋgoʔt* 'bend forward,' *otʃoʔt* 'hump of bull,' *gomke* 'master' (Bodding 1929-36: 1181, 1188, 2959, 2956, 1161, 1160, 2895, 1329), but there are a number of lexical and morphemic exceptions, so the harmonic restriction is not robust *per se*. The third harmony-esque distribution of vowels observed in Santali is among nasalized and non-nasalized vowels. This is only robustly expressed as a co-occurrence restriction for / ĩ/ and / ɛ̃/ which co-occur with nasalized vowels in the same foot nearly 80% of the time (Anderson, Horo and Harrison 2024: 727). Hence, although the systems attested differ in all three of these languages, the domain of the harmony is the same, that is the foot.

Lastly, while we have the current findings at hand, there remain various outstanding questions in the analysis of the Mundari vowel system. For example, diphthongs or V1V2 sequences do occur in Mundari, the most commonly attested in our data is *ai* found for example in *lai* 'stomach' and *tai* an auxiliary stem, which of course maintain the observed harmonic restriction. However, it is still unknown if such sound sequences in Mundari function as a single phonemic unit or not. Likewise, while creaky voice is largely found in the language it is not known if a series of creaky voice vowels have arisen in speakers that do not use segmental glottal stop but rather cue its former presence in the guise of vowel creakiness (Gogoi et al. 2023-ms). Similarly, it is unknown if nasalized vowels are contrastive in any people's speech. In case of word level prosody, to date no acoustic cue seems to characterize or define the unit of phonological word[4]. There is only preliminary work showing that there is no consistent prominence pattern cued by intensity or duration in disyllabic words. Instead, duration seems to mark the right edge of utterances, not words. The only tangible evidence for right edged prominence in Mundari appears to be the rise in the fundamental frequency in the second syllable of disyllables (Horo et al. 2023). Therefore, it is yet to be determined if Mundari actually has a discernible system of prominence and whether that has any effect on the vowel quality or not. Moreover, we have not yet examined vowel quality characteristics of mid-vowels in three-, four- and five-syllable grammatical words but this is an ongoing process.

5 Conclusion

This study provides an analysis of the formant frequencies of the vowels /a, e, i, o, u/ in both monosyllabic and disyllabic words in Mundari. The vowels exhibit a consistent pattern of overlaps between specific vowel pairs /i/ - /e/ and /u/ - /o/ in both monosyllabic and disyllabic words, as evidenced by statistical significance found through ANOVA and post-hoc tests. Additionally, the results highlight significant

[4] However, morphological evidence can be used to delimit a grammatical word if it is verbal. Specifically, the placement of the subject clitic can demarcate the beginning and end of the verbal grammatical word. No such data exists for nominal forms however, so there is nothing per se that can be used to delimit nominal forms acoustically or morphotactically, or in other words nothing to define the level of word phonoprosodically nor grammatically if the word functions nominally in a syntactic sense.

patterns in vowel height (F1) and frontness (F2) across different syllabic positions in disyllabic words, specifically mid-vowels are lower in second syllables of disyllables which contrasts with their realization in first syllables of disyllabic words. Monosyllables show the same pattern as the second syllable of disyllables. A robust harmonic restriction against the co-occurrence of mid and high vowels in a single foot is observed and reported here for the first time in detail and situated within the broader typology of attested Kherwarian Munda harmony systems. Overall, this research is the first analysis of the acoustic properties of Mundari vowels, providing a foundation for further analysis and understanding of this language's phonological structure.

References

Anderson, Gregory D. S. Toshiki Osada and K. David Harrison. 2008. Ho and the other Kherwarian languages. In G. D. S. Anderson (ed.) *The Munda Languages*. Routledge Language Family Series. London: Routledge. 195-255.

Anderson, Gregory D. S., Luke Horo and K. David Harrison. 2024. Vowel Harmony in the Munda Languages. In Harry van der Hulst and Nancy Ritter (eds.) *Oxford Handbook of Vowel Harmony*. Oxford: OUP, 723-8.

Bodding, Paul Olaf. 1922. *Materials for a Santal Grammar*, Mostly Phonetic. Dumka: Santal Mission.

Bodding, Paul Olaf. 1929– 1936. *A Santal Dictionary.* Five volumes. Oslo: Norske Videnskaps Akademi.

Boersma, Paul, & Van Heuven, Vincent. 2001. Speak and unSpeak with PRAAT. *Glot International* 5 (9/10), 341-347.

Cook, Walter A. 1965. A Descriptive Analysis of Mundari: a study of the structure of the Mundari language according to the methods of linguistic science. Ph.D. Dissertation, Georgetown University.

Ghosh, Arun. Santali. In Gregory D. S. Anderson (ed.) *The Munda Languages*. Routledge Language Family Series. London: Routledge. 11-98.

Gogoi, Pamir, Luke Horo and Gregory D. S. Anderson. 2023-ms. Phonetic correlates of glottal stop in Mundari. Presented at HISPhonCog, Seoul, South Korea, May 2023. Unpublished manuscript.

Gumperz, John J. and H. S. Biligiri. 1957. Notes on the phonology of Mundari. *Indian Linguistics* 17. 6-15.

Hall-Lew, Lauren. 2010, April. Improved representation of variance in measures of vowel merger. In *Proceedings of meetings on acoustics* (Vol. 9, No. 1). AIP Publishing.

Hay, Jennifer, Paul Warren & Katie Drager. 2006. Factors influencing speech perception in the context of a merger-in-progress. *Journal of Phonetics*, *34*(4), 458-484.

Hoffman, John 1930-78. *Encyclopaedia Mundarica*. Patna: Government Press.

Horo, Luke, Pamir Gogoi, Gregory D. S. Anderson. 2023. Phonetic Correlates of Syllable Prominence in Mundari. *Proc. The Second International Conference on Tone and Intonation*, 78-82, doi: 10.21437/TAI.2023-17.

Liljencrants, Johan, & Björn Lindblom. 1972. Numerical simulation of vowel quality systems: The role of perceptual contrast. *Language*, 48(4): 839-862.

Lindblom, Björn. 1986. Phonetic universals in vowel systems. In J. J. Ohala & J. J. Jaeger (Eds.), *Experimental phonology*. Orlando: Fl.: Academic Press. 13-44.

Lindblom, Björn. 1990. On the notion of "possible speech sound". *Journal of phonetics*, 18(2): 135-152.

Minegishi, Makoto. 1990. Santali English Japanese wordlist: A preliminary report. *Journal of Asian and African Studies* 39: 69– 84.

Nottrott, Rev. Alfred. 1882. *Grammatik der Kolh-Sprache*. Gütersloh: C. Bertelsmann. Osada, Toshiki. 2008. Mundari. In G. D. S. Anderson (ed.) *The Munda Languages*. Routledge Language Family Series. London: Routledge. 99-164.

Pucilowski, Anna. 2013. Ho Morphology and Morphosyntax. Ph.D. Dissertation, University of Oregon.

Sidwell, Paul and Felix Rau. 2015. Austroasiatic Comparative-Historical Reconstruction: An Overview. In Jenny, M and Sidwell, P. (eds.), *The Handbook of Austroasiatic Languages*. Leiden: Brill, pp. 221-363.

Sinha, N. K. 1975. *Mundari Grammar.* Mysore: CIIL.

Whitley, Jabez Cornelius. 1873. *A Mundári Primer*. Bengal Secretariat Press.

Appendix A: Vowel contrast in closed monosyllables

	Mundari	**English**
1.	*ser*	‘unit of weight’; ‘to melt, smelt’
2.	*sir*	‘vein midrib and other nerves of leaves’
3.	*sen*	‘to go’
4.	*san*	‘firewood’
5.	*or*	‘pull’
6.	*ur*	‘drag’
7.	*soŋ*	‘to measure’
8.	*siŋ*	‘tree’
9.	*ol*	‘write’
10.	*il*	‘feather’
11.	*ul*	‘ripen by force’
12.	*il*	‘feather(s)’
13.	*sor*	‘to throw spear’; ‘hiss of snake’
14.	*sar*	‘arrow’
15.	*sur*	‘to choke’
16.	*om*	‘give’
17.	*am*	‘you SG’
18.	*im*	‘liver’
19.	*bar*	‘two’
20.	*bir*	‘forest’
21.	*nur*	‘solid particles coming out from hole’
22.	*nir*	‘to run’
23.	*beʔ*	‘spit’
24.	*buʔ*	‘hole’
25.	*ged*	‘disembowel’
26.	*god*	‘pluck’
27.	*rob*	‘cracking sound (of bones, branches)’
28.	*rub*	‘poisoning fish’; ‘uprooted (of tree)’

Appendix B: Vowel contrast in open monosyllables

	Mundari	**English**
1.	*le*	‘tongue’; ‘melt’
2.	*lo*	‘to burn’
3.	*lu*	‘to ladle’
4.	*la*	‘excess’; ‘dig with spade’
5.	*ro*	‘fly (n)’
6.	*ru*	‘to beat’
7.	*so*	‘to shout at fowl’
8.	*si*	‘to plough’
9.	*su*	‘to insert the hand or finger in a hole’
10.	*na*	‘now’
11.	*ni*	‘open’
12.	*ne*	‘take it! (when offering someone something by hand)’

Appendix C: Vowel contrast in first syllable of disyllables

	Mundari	**English**
1.	*mesa*	‘mix’
2.	*misa*	‘one time’
3.	*elaŋ*	‘conflagration’
4.	*alaŋ*	‘we DL INCL’
5.	*koʔo*	‘to peek’
6.	*kuʔu*	‘to cough’
7.	*hora*	‘road’
8.	*hara*	‘to grow’
9.	*duku*	‘sorrow’
10.	*diku*	‘outsider’
11.	*usar*	‘to push’
12.	*asar*	‘bow for shooting arrows’
13.	*basi*	‘stale’
14.	*bisi*	‘poison’
15.	*keɽa*	‘buffalo’
16.	*kuɽa*	‘to fold a mat’
17.	*liboʔ*	‘increase (of water) by slowly flowing in’
18.	*loboʔ*	‘fattiness, thickness of soil’
19.	*gele*	‘ear (of grain)’
20.	*gole*	‘whistling’
21.	*soŋga*	‘level path at foot of two hills facing each other’
22.	*suŋga*	‘cause irritation with stinging hairs (of caterpillar, plant)’

Appendix D: Vowel Contrast in second syllable of disyllables

	Mundari	**English**
1.	*ale*	'we PL EXCL'
2.	*alo*	'don't'
3.	*ali*	'wet something'
4.	*alu*	'potato'
5.	*bale*	'young'
6.	*bala*	'marriage'
7.	*koʔo*	'to peek'
8.	*kuʔu*	'to cough'
9.	*kaɽo*	'small'
10.	*kaɽa*	'leg'
11.	*capu*	'to touch with hand'
12.	*capi*	'to wash something'
13.	*rapud*	'to break'
14.	*rapid*	'to wink'
15.	*saru*	'variety of taro plant'
16.	*sara*	'manure'
17.	*saɽa*	'luxuriant leaves of paddy'
18.	*saɽi*	'woman's waist cloth'; 'produce a sound by beating musical instrument'
19.	*laped*	'morsel'
20.	*lapud*	'chicken pox'
21.	*ape*	'you PL'
22.	*apu*	'father'
23.	*gonḍe*	'pull someone by the leg, trip someone'
24.	*gonḍo*	'dirty of clothes'
25,	*bando*	'if not'
26.	*bandu*	'seed pods of *Spatholobus* species'

8

A Note on Khmer Historical Phonology

Ratree Wayland

1 Introduction

Despite belonging to different language families, Thai (a Tai language) and Khmer (an Austroasiatic language) have been in close contact for centuries, resulting in linguistic borrowings, first from Khmer to Thai and later from Thai to Khmer. Most Khmer loanwords likely entered the Thai language during the Ayutthaya period (1351–1767) after the Thai conquest of Angkor in 1431 (Hudak, 2018). During this period, an influx of loanwords from Pali, Sanskrit, and Khmer made the Thai language highly complex and stratified. This complexity was reflected in the expansion of titles, ranks, pronouns, royal vocabulary, and royal kin terminology, mirroring the increasing social stratification and complexity of Thai society (Hudak, 2018).

Studying lexical borrowings from Khmer to Thai is invaluable for understanding the evolution of both Thai and Khmer phonologies. The exchange of loanwords provides insight into how sounds and pronunciation patterns have been adapted and integrated into each language over time. For instance, analyzing Khmer loanwords in Thai can reveal how Thai phonology has evolved to accommodate foreign phonetic structures, highlighting shifts and phonological rules in the sound systems. This was the original topic of my doctoral dissertation planned under Diffloth's supervision at Cornell University. However, his frequent absences from Ithaca for fieldwork in Thailand and neighboring countries, along with his eventual departure from Cornell, led to the abandonment of the project and a change in my research focus. In this short paper, I evaluated a recently proposed reconstruction of two Khmer vowels, which is claimed to be supported in part by Khmer loanwords in Thai.

2 Lexical Borrowing

Borrowing is a widespread linguistic phenomenon, and no language is entirely devoid of borrowed words. The term 'borrowing' or 'loan-word' is a 'calque' or a direct translation of the German word 'Lehnwort.' This term has been considered inept or misleading since the donor language never reclaims its 'loaned' or 'borrowed' words (Haugen, 1950).

Lexical borrowing, the process of adopting words from one language into another, has traditionally been attributed to two main reasons: need and prestige (Hockett & Joseph, 2009). 'Need,' or filling a gap, is an internal cause that arises when a language lacks a term for a new concept, object, or practice introduced to the speech community. For instance, when new technologies, foods, or cultural practices are encountered, there is a need to name these new items, often leading to borrowing from the language of origin. Additionally, semantic change of a word can creates a gap in the vocabulary, which can be filled by borrowing. An example is the Old English word 'dēor,' which

originally meant ‘animal.’ When its meaning changed to ‘deer,’ the Latin word ‘animal’ was borrowed to fill the gap (McMahon, 1994).

On the other hand, ‘prestige’ is an external cause where languages borrow words from more prestigious languages or cultures. This happens when the donor language has a higher socio-cultural or economic status, leading speakers of the recipient language to adopt terms to align with that prestige. For example, languages such as Greek, Latin, German, Russian, and English have been frequent loan-givers throughout history, depending on their socio-cultural and economic power during various periods (Carling et al., 2019).

However, recent research has shifted its focus to a broader range of factors, including the need to provide labels for unique referents, associations with specific activity domains, cognitive pressure from managing presupposition domains or interaction roles, and particularly, the conflict between speaker intentions and listener expectations (Matras, 2009).

2.1 What can be borrowed?

Whitney (1881) postulated that nouns are more easily borrowed than adjectives, which in turn are more easily borrowed than verbs. On the other hand, Haugen (1950) recognized that all linguistic features could be borrowed but noted the existence of a ‘scale of adoptability’ linked to structural organization, without further elaboration. He emphasized the importance of cross-linguistic research, suggesting that borrowing patterns may vary among different languages.

Nouns are the most frequently borrowed lexical items because they typically represent tangible objects, new technologies, or cultural artifacts that lack existing equivalents in the borrowing language (Matras, 2020). Their concrete nature and the relative ease with which they can be integrated into the syntactic structures of the borrowing language contribute to their high frequency of borrowing. Elsik (2009) documented loanwords in the Romani dialect of Selice, located in southern Slovakia, using a sample list of 1,430 lexemes from the Loanword Typology project. Among these lexemes, 63% are loanwords, which entered the language at various stages. Of these loanwords, 53% (84% of all loanwords) are from Hungarian, the primary contact language. These Hungarian loanwords accounted for 63% of all nouns, 41% of verbs, 42% of adjectives, 50% of adverbs, and 23% of function words on the list.

However, the borrowability of different types of nouns is not equal, as various factors influence which nouns are more likely to be adopted into a language. For example, Brown (1999) examined factors like the frequency of use in the original language and the presence of semantic equivalents in the recipient language on the borrowing of nouns from European languages into Native American languages. The study found that the borrowability of words varied significantly, with pragmatic saliency playing a key role. Specifically, terms for living beings were more frequently borrowed than those for artifacts, and terms for animals were borrowed more often than those for plants.

In contrast, verbs are less frequently borrowed than nouns due to their syntactic and morphological complexity (Matras, 2020). Integrating a verb into a new language often requires adapting it to fit the verb conjugation patterns and syntactic rules of the borrowing language, which is more challenging than integrating nouns. According to Matras (2020), the integration of loan verbs into a recipient language can occur through several methods, each involving different degrees of modification to the original form of the verb. One method is direct insertion, where the original verb is adopted without

any changes. Another method is indirect insertion, which involves morphologically modifying the verb to fit the patterns of the recipient language. Additionally, there is the light verb construction, where the original verb is inserted into a compound construction and accompanied by an inherited verb from the recipient language. Lastly, paradigm transfer involves importing the original verb along with its original inflectional morphology, thereby transferring the entire conjugational paradigm of the verb to the recipient language. These integration strategies exist as a continuum rather than separate strategies (Matras, 2020). For example, some languages may use direct insertion, where the original verb form is not overtly modified but is assigned to a specific inherited inflection class. This class may be reserved for loans, thereby flagging the verb as a borrowed term, or it may be used for intensification of actions. In this way, the original verb, while remaining unmodified, is treated similarly to non-verbs, resembling the indirect insertion strategy (Matras, 2020).

Other parts of speech, such as adjectives and adverbs, are borrowed less frequently, followed by function words like prepositions and conjunctions, which are closely tied to the syntactic and grammatical structure of the language.

2.2 Loan Word Phonology

Haugen (1950) proposed that understanding borrowing begins with examining the behavior of bilingual speakers. He asserted that borrowing involves speakers reproducing new linguistic patterns in a different language context, which always appears as an innovation, regardless of the speaker's awareness. Haugen termed this process 'innovative reproduction'. Additionally, he differentiated 'importation,' which is the exact replication of material from another language, from 'substitution,' where some structural aspects of the borrowed item are altered. Haugen also distinguished 'loanwords' from 'hybrids' (partial word borrowings), 'loan translations' or 'calques' (replicating form-function mapping), and 'semantic loans' (replicating word meanings). Furthermore, he introduced three categories: 'loanwords' (importation without morphemic substitution), 'loan blends' (importation with partial morphemic substitution), and 'loan shifts' (complete morphemic substitution, changing the meaning of an existing word based on similarity with external words), which include 'loan translations' (words created through contact but not directly imported).

Consistent with Haugen's assumption that loanword adaptation is likely performed by advanced L2 speakers, Boersma and Hamann (2009) demonstrated that loanword adaptation can be fully explained through the phonological and phonetic comprehension and production mechanisms of the first language (L1). They offered explicit explanations for various loanword adaptation phenomena in Korean, utilizing an Optimality-Theoretic grammar model. This model operates with three levels of representation needed for describing L1 phonology: the underlying form, the phonological surface form, and the auditory-phonetic form. The model is bidirectional, meaning it applies the same constraints and rankings for both listening and speaking. These constraints and rankings are consistent for both L1 processing and loanword adaptation.

In short, the framework for understanding the borrowing process through the behavior of bilingual speakers and the distinctions between various types of borrowings provides a foundation for exploring specific cases of lexical adaptation. This theoretical background is particularly useful when examining Khmer loanwords in Thai, shedding light on the historical phonological changes in both languages.

3 Historical Pronunciations of Khmer ◌ោ◌ <au> and ◌ៃ◌ <ai>

One outstanding question in the history of Khmer phonology, recently discussed by Maspong (2024), concerns whether the graphemes ◌ោ◌ <au> and ◌ៃ◌ <ai> were pronounced as *əw and *əj, as proposed by Jenner (1974), or as *aw and *aj, as proposed by Maspong (2024). Jenner's hypothesis that *əw and *əj were the original pronunciations is based on observed rhyming patterns between ◌ូ◌ <ūv> : ◌ោ◌ <au> and ◌ី◌ <ī>: ◌ៃ◌ <ai>. Specifically, Jenner (1974) noted that <ūv> and <ī>, following voiceless onsets, could rhyme with <au> and <ai> regardless of whether the onsets were voiced or voiceless. From this, Jenner inferred that the shifts of *u: to [əw] and *i: to [əj] before voiceless onsets occurred before the vowels represented by <au> and <ai> after voiced and voiceless initials bifurcated, proposing that the original pronunciations of <au> and <ai> were *əw and *əj. If this were not the case, then words with <au> and <ai> following original voiceless onsets would not rhyme with *u: and *i: with voiceless onsets. He additionally suggested that the bifurcation of *u: and *i: predates the bifurcation of the vowels represented by <au> and <ai>.

Maspong (2024) challenged Jenner's proposal, stating that the rhyming between <ūv> and <ī> following voiceless onsets with <au> and <ai> might not accurately reflect the chronological order of vowel changes, but rather a poetic tolerance for imperfect rhymes. Additionally, Khmer loanwords in Thai are pronounced with [aw] and [aj], rather than the also possible rimes [əw] and [əj] in Thai, supporting her hypothesis that *aw and *aj were the original pronunciations.

The weakness of Jenner's proposal and Maspong's counterargument based on rhyme types lies in the inherent difficulty of determining whether we are dealing with divergent or convergent rhymes in old texts. Accurately identifying rhyme types in historical texts requires careful consideration and often remains uncertain. Maspong's argument based on Khmer loanwords in Thai is also not without issues. First, although the rimes [əw] and [əj] are phonotactically possible, Thai words with these rimes are rare. I can only think of very few Thai words with [əj] and none with [əw]. Secondly, Pittayaporn (2009) reconstructed three rimes: *aj, *auɥ, and *aw for Proto Tai (PT). The Thai or Siamese reflex of PT *aj is [aj], as in [kʰàj] 'egg,' while the Thai reflex for PT *aw is [aw], as in [kàw] 'old.' From a cross-linguistic perception point of view, a foreign sound category is more likely to be mapped to the closest (and most frequent) first language (L1) category. Therefore, Thai [aj] and [aw] could be reflexes of Khmer *aj (or *əj) and *aw (or *əw), as they were probably the closest Thai rimes to the Khmer rimes at the time of borrowing.

However, I found Maspong's proposal of an intermediate stage of *a > [ə] > [ɨ] the most intriguing. Specifically, based on rhyming patterns, she suggested that the shift of *a to [ɨ] after a voiced onset and before a glide coda might have gone through this intermediate stage. These changes were speculated to have occurred in the 17th century, coinciding with the shift of *u: to [əw] following voiceless onsets. Therefore, during this intermediate stage, the rhyming pairs <ūv> : <au> and <ī> : <ai> could be viewed as 'perfect' rhymes. However, Maspong did not clarify whether the same intermediate stage also applied to voiceless onsets, making it unclear if perfect rhymes could also be inferred in such cases. Additionally, the motivation for this intermediate stage was not addressed.

Although Maspong did not provide a motivation for this intermediate stage, it is possible that it corresponds to Huffman's (1976) and Wayland and Jongman's (2001, 2002) stage 2 in Khmer Registrogenesis. Huffman hypothesized that this stage involves 'simultaneous change in the articulation of one set of initial consonants and the

development of allophonic variation in following vowels, still in complementary distribution vis-a-vis two distinctive sets of initials' (p. 578). Wayland and Jongman (2001) further hypothesized that at this stage, voiceless and voiced onsets become tense versus lax, respectively, with accompanying changes in vowel quality (i.e., height and diphthongization) and phonation (i.e., breathy versus modal or clear).

It is further assumed that the change did not affect all vowels simultaneously, nor was the impact on vowel quality and phonation uniform across all cases. For example, Wayland and Jongman (2001) reported that breathy versus modal phonation remains phonemically distinct only for /eː/, /ɛː/, and /a/, but constitutes subphonemic variation in the remaining vowels in Chanthaburi Khmer, an older dialect of Khmer spoken in Chanthaburi, Thailand. In other words, besides the three vowels mentioned, as in Modern Khmer, the two registers of vowels in Chanthaburi Khmer can be differentiated solely by their quality (diphthongized versus non-diphthongized) with a redundant subphonemic breathy and clear phonation.

For Old Khmer *aw and *ai under discussion, it is possible that at this intermediate stage, a change in the articulation of voiced onsets resulted in a lower first formant (F1) and higher second formant (F2) frequencies, thus producing a perceived higher and more front vowel quality, with breathy phonation. Therefore, at this stage, the *a + glide codas could be produced as breathy [ə̤w] and [ə̤j] after voiced onsets. Crucially, the voiceless <tense/stiff versus voiced <lax onsets remained distinct, but complementary allophonic variation in phonation type had developed. Being subphonemic, this variation was allophonic and did not prevent perfect rhyming with [əw] and [əj] from *uː following voiceless initials. Similarly, when borrowed into Thai, the allophonic phonation difference was ignored, and both breathy *[ə̤w] and modal [əw], as well as breathy [ə̤j] and clear [əj], were mapped to the closest Thai rimes *aw and *ai, respectively. The fact that they don't rhyme with *uː after voiced onsets indicates that the change in pronunciation of the voiced onsets before *uː had moved to a more advanced stage, affecting not only its phonation but also its quality. However, Jenner's contention that the bifurcation of *uː and *iː predates the bifurcation of the vowels represented by <au> and <ai> needs further consideration.

In order for the *a + glide codas following voiceless onsets to perfectly rhyme with *uː after voiceless onsets, one must also assume, as Jenner (1974) did, that similar to voiced onsets, the *aj and *aw following voiceless onsets must be produced as [əw] and [əj], but with a modal instead of a clear phonation. According to Wayland and Jongman (2001)'s hypothesis, at this stage, the articulation of voiceless onsets was characterized by a stiff voice with an abrupt and likely imploded release, resulting acoustically in a lower F1 and F2 transition to the vowel. This formant transition could be perceptually reanalyzed as a higher (though not as front as [ə] proposed for the one induced by slack or voiced onsets) onglide to the vowel. Thus, it is likely that *aj and *aw after voiceless onsets were heard as [ɐj] and [ɐw] with a modal or clear phonation, constituting near-perfect rhymes with [əj] and [əw] from *uː after voiceless initials. Therefore, I propose that the pronunciations of the Khmer graphemes <ai> and <aw> were [ə̤j] and [ə̤w] after voiced onsets, but [ɐj] and [ɐw] after voiceless onsets. This suggests that the bifurcation of *aj and *aw occurred concurrently with that of *uː, but the difference between the two sets of vowels (after voiced versus voiceless onsets) was more phonetically salient for *uː than for *aj and *aw.

4 Summary

Whether this reconstruction ends up being accurate or not, I hope to demonstrate in this short paper that reconstructing historical phonology is a challenging task that requires evidence from various sources. While analyzing historical texts and rhyming patterns is valuable, it should be complemented with data from related languages, borrowings, and phonetic plausibility. Understanding bilingualism and cross-language speech learning is crucial when considering borrowings. By integrating insights from these diverse areas, researchers can uncover nuanced details that contribute to a more accurate reconstruction of phonological changes.

References

Boersma, Paul & Silke Hamann. 2009. Loanword adaptation as first-language phonological perception. *Loanword phonology*, 11-58.

Brown, Cecil. H. 1999. *Lexical acculturation in Native American languages* (Vol.20). Oxford University Press.

Carling, Gerd, Sandra Cronhamn, Robert Farren, Elnur Aliyev & Johan Frid. 2019. The causality of borrowing: Lexical loans in Eurasian languages. *PloS one*, 14(10), e0223588.

Haugen, Einar. 1950. The analysis of linguistic borrowing. *Language*, 26(2), 210-231.

Hock, Hans Heinrich & Brian D. Joseph. 2009. *Language history, language change, and language relationship: An introduction to historical and comparative linguistics*. Mouton de Gruyter.

Hudak, Thomas John. 2018. Thai. In *The world's major languages* (pp. 679-695). Routledge.

Huffman, Franklin E. 1976. The register problem in fifteen Mon-Khmer languages. *Oceanic Linguistics Special Publications*, (13), 575-589.

Maspong, Sireemas. 2024. Chronology of Registrogenesis in Khmer: Analyses of Poetry and Inscriptions. *Journal of the Southeast Asian Linguistics Society* 17.1: 46-62.

Matras, Yaron. 2009. *Language Contact*. Cambridge: Cambridge University Press.

Matras, Yaron. 2020. *Language Contact*. 2nd ed. Cambridge: Cambridge University Press.

McMahon, April M. S. 1994. *Understanding language change*. Cambridge University Press.

Pittayaporn, Pittayawat. 2009. The phonology of proto-tai. Doctoral Dissertation, Cornell University.

Wayland, Ratree & Allard Jongman. 2001. Chanthaburi Khmer vowels: phonetic and phonemic analyses. *Mon-Khmer Studies*, 31, 65–82.

Wayland, Ratree & Allard Jongman. 2002. Registrogenesis in Khmer: A phonetic account. *Mon-Khmer Studies*, 32, 101-115.

Whitney, William D. 1881. On mixture in language. *Transactions of the American Philological Association* (1869-1896), 12, 5-26.

9

The Expansion of Austroasiatic: an Extended Model

Roger Blench

1. Introduction

The dating and homeland of Austroasiatic has been the subject of much scholarly debate, in particular the opposition between models which propose a West to East direction and those which espouse the contrary view. Frankly, many of these arguments now seem rather empty, and based as they were on intricate but pointless linguistic arguments. The last decade has seen a major expansion of our understanding of the SE Asian Neolithic, as well as improved reconstructions of proto-languages for most Austroasiatic subgroups. Arguments which do not link the Austroasiatic expansion with the Neolithic seem to me perverse and to derive from non-linguistic presuppositions, similar to the division of Trans-Himalayan into Sinitic and 'the rest'. Moreover, it is difficult not to link the primary phase of Austroasiatic expansion to aquatic strategies, both the quest for river basins and maritime migrations. The hypothesis that the primary dispersal of Austroasiatic focused on river basins was originally advanced by Gerard Diffloth and the reconstruction of aquatic vocabulary points continues to strongly support his arguments.

What is yet to be fully established is the sequencing and routes of Austroasiatic expansion and in particular the social and cultural context of the formation of particular subgroups. For example, at least two subgroups, Aslian and Nicobaric, must have reverted to foraging, forsaking the agriculture which developed in their original home area. Although such a process is not unknown, globally it is quite unusual and requires explanation. The Munda, it now seems plausible, crossed the Bay of Bengal directly by sea, carrying with them only rice as a staple crop (Rau & Sidwell 2019). The Munda preserve no reconstructible maritime vocabulary, and probably did not travel in their own ships. Other branches of Austroasiatic, notably those in China, Mangic and Pakanic, have virtually no agricultural lexicon which can be traced back to Austroasiatic, so it may be they were also hunter-gatherers who adopted farming subsequent to their initial migrations. Finally, Rongic is a 'hidden' branch of Austroasiatic existing only as a substrate in Lepcha (Rong) in Sikkim (Blench 2023). No original Austroasiatic crop plants are attested in Rongic, arguing perhaps that they too reached this area as foragers.

This chapter[1] proposes a radical re-appraisal of the process of Austroasiatic

[1] Thanks to Frank Muyard and Paul Sidwell for comments on the first draft. Aude Favereau and Bérénice Bellina-Pryce also read through in what turned out to be a vain hope of linking the data with the Neolithic archaeology of the East coast of India. A preliminary version was presented at workshop: *Recent European research in later Southeast Asian later prehistory*, Paris, 8-9th July, 2024

expansion, taking into account the reconstructions of crop plants and material culture (or their conspicuous absence). It argues for the importance of aquatic and maritime routes and speculates on the ownership of shipping in the earliest period.

2 Early Austroasiatic speakers and their culture

2.1 Austroasiatic subgrouping

Austroasiatic is now usually divided into fourteen attested subgroups, since Pakanic and Mangic are seen to form two individual branches. Apart from these, there are two more controversial branches, Rongic and Borneo. Rongic can be detected in Austroasiatic roots in the Lepcha (Rong) language, which have cognates across a wide range of branches and thus cannot be aligned with any specific branch (Blench 2023). The evidence for a distinct Borneo branch is weaker, consisting of only a few items of Austroasiatic origin in the Austronesian languages of Borneo (Blench 2011). Compared with other language phyla, Austroasiatic shows quite low internal diversity. Given its likely age, this might be expected; it looks more like Bantu or Polynesian than Indo-European or Trans-Himalayan. Despite numerous competing proposals, Austroasiatic has no uncontroversial internal structure and may be a flat array as in Figure 1:

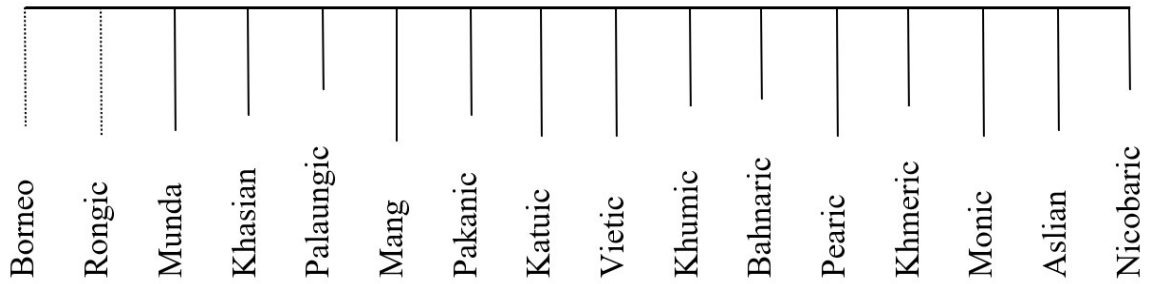


Figure 1. Flat-array structure for Austroasiatic

This is consistent with rapid dispersal at the point where the Neolithic spreads into northern mainland SE Asia from the Yangtze Valley. This can be summarised as the 'Middle Mekong' hypothesis (see discussion in Sidwell & Blench 2011). Sidwell (2022) has now revised the point of origin as the Red River Delta and the corresponding archaeological horizon, the Phùng Nguyên culture (4000–3500 BP) of northern Vietnam.

2.2 Reconstruction of subsistence

It is likely that Austroasiatic speakers were established agriculturalists at the time of initial dispersal. Cognate subsistence lexicon is common in a subset of core branches, but is conspicuous by its absence in 'outlier' branches. This has two possible interpretations. Either speakers were foragers at the point when they moved and only later adopted agriculture, or else some now unknown process caused them to lose agriculture, revert to foraging and then rebuild farming subsequently. Both of these trajectories are attested in the global archaeological record, and are considered in greater detail in §2.4.

Austroasiatic branches have a variety of common terms for crops and livestock which certainly attest to their importance in the period shortly after the initial dispersal.

Table 1 shows the attestations for crops in different branches of Austroasiatic (revised from Blench 2011):

Table 1. Proposals for crop reconstructions in Austroasiatic

Gloss	Recon.	Aslian	Bahnaric	Khasian	Khmeric	Katuic	Khmuic	Monic	Munda	Nicobaric	Palaungic	Pearic	Vietic	Shorto (2006) entry
ricefield	#sərɛː	x	x		x		x		x		x	x		§185
rice (general)	*srɔ(ː)ʔ		x		x	x		x	x					§187
husked rice	*rŋkoːʔ	x	x	x	x	x		x	x		x	x	x	§1820
paddy rice	*ɓa(ː)ʔ	x	x	x	x		x		x		x	x		§120
foxtail millet	*skɔːj	x			x[2]		x	x			x	x	x	§1447
taro	*sroʔ		x	x	x	x	x	x	x		x	x	x	§1850
sesame	#lŋa		x		x	x	x	x			x	x	x	§34
banana	#tVlVy	x				x	x			x	x		x	§1523
betel pepper	#mpluw				x	x	x	x			x	x	x	§1860

Note: Forms in Sidwell's (2024) 500- list are marked *, whereas quasi-reconstructions are marked #. Lexical reflexes supporting these etymologies and reconstructions can be found at the numbered Shorto (2006) entries (these, and other relevant data can be accessed online at sealang.net/monkhmer).

Evidence for reconstruction of *srɔ(ː)ʔ 'rice (general)' and *ɓa(ː)ʔ 'paddy rice' in Austroasiatic is convincing. Curiously, comparison of the forms of *srɔ(ː)ʔ with #sərɛː 'rice field' suggest that *srɔ(ː)ʔ may be a back formation.

The reconstruction of a word for 'foxtail millet' in Austroasiatic is apparent. In two families, Khmeric and Pearic the term applies to two other cereals, barnyard millet and Job's tears. Neither of these have been attested in the archaeological record in SE Asia, and it is presumed the name was transferred from foxtail millet.

The form #lŋa for sesame is one of those apparently shared with Austronesian at a deep level. Blust in the ACD reconstructs PMP **leŋa* 'sesame', but all his attestations are from Western Malayopolynesian, which suggests the crop was traded from the mainland quite early, corresponding the well-known Kalanay pottery tradition.

The spread of the banana from New Guinea, where it originated, has been the subject of much discussion (Blench 2020). There are multiple roots attested and some seem to cross language phylum boundaries. The Munda form (e.g. Korku *ʈora* 'wild bananas') is intriguing: was the original Austroasiatic form applied to a wild banana (very common in mainland SE Asia) and only preserved with this meaning in Munda, while being transferred to the domestic banana in the homeland area? Only more detailed lexical work can answer that.

Betel chewing has a lengthy and complex history in the Asia-Pacific region. Zumbroich (2007) is the most recent synthesis which incorporates linguistic, cultural and archaeological data. Because betel stains the teeth it can be detected on skeletal material, which creates a much fuller archaeological record than many other plants.

[2] The Khmer cognate, *skuəy* ស្ពួយ, is applied to Job's tears.

Chewing betel has two components, the pepper and the leaf, but these are often poorly distinguished in lexical sources and the reconstruction most likely refers to betel pepper. As a cultural item, borrowing in and between phyla cannot be excluded.

Apart from crops, the speakers of proto-Austroasiatic were also committed livestock producers. Almost all the major species found in the region today were already known to speakers at an early period, except horse, donkey and sheep. Table 2 shows a series of starred forms and quasi-reconstructions (revised from Blench 2011) based on widespread forms.

Table 2. Livestock quasi-reconstructions in Austroasiatic

Gloss	Recon.	Aslian	Bahnaric	Khasian	Khmeric	Katuic	Khmuic	Mangic	Monic	Munda	Nicobaric	Palaungic	Pearic	Vietic	Shorto (2006) entry
cow	#ləmboʔ	x	x			x	x	x						x	§119
buffalo	#krəpa:w	X[3]	x		x	x							x	x	§103
buffalo	#t.ri:k					x	x					x		x	§408
pig	#k.li:k				x	x			x			x	x		§417
goat	#bɛ[:]ʔ	x	x	x	x	x	x	x	x	x		x	x	x	§126
dog	*cɔʔ	x	x	x	x	x	x	x	(?)	x		x	x	x	§41
cat	#miaw	x		x		x	x				x		x	x	§1838
chicken	#sjiar		x	x		x	x	x				x			§1522
duck	#ʧtə[k]	x	x			x	x	x	x		x	x	x	x	§77

Note: Lexical reflexes supporting these etymologies and reconstructions can be found at the numbered Shorto (2006) entries (these, and other relevant data can be accessed online at sealang.net/monkhmer).

Mainland Southeast Asia has a wild bovid, the gaur, which must have preceded the domestic cow in the region. This was probably the original referent of the form reconstructed here as #ləmboʔ. Obviously cognate forms are found in Western Austronesian languages, but their patchy distribution argues they are relatively recent loanwords.

Austroasiatic has two widespread forms for ‘buffalo’, which are neatly intertwined. There are wild buffalos in SE Asia, but the usual referent for these terms is the domestic buffalo, used to pull the plough in the ricefields. The origin of Austroasiatic goes back beyond the introduction of metal, so the buffalo must have been introduced after the primary expansion. The root #krəpa:w is also apparently borrowed widely in Western Austronesian. The source of the water-buffalo is unknown, but must have been introduced in the early Metal Age and spread from branch to branch independently.

The domestic pig is one of the oldest household animals in this region. Evidence from China gives dates at around 8000 BP with a spread into mainland SE Asia at around 6000 BP, whence it spreads to the Western Austronesian region. Given this, it is somewhat surprising that the commonest root for ‘pig’ is not more widely distributed in Austroasiatic. Apparently related forms are found in Trans-Himalayan languages of NE India (Blench 2014), so it is possible there was interphylic borrowing.

Reflexes for ‘goat’ in Austroasiatic cover all the core branches. Unfortunately,

[3] As loanword.

there is absolutely no archaeozoological material for dating the introduction of goats in Southeast Asia, although they appear in the Chinese record around 4000 BP. So it is conceivable, is surprising that they were present in the early period of Austroasiatic expansion.

The dog is widely attested in Austroasiatic, despite a lack of archaeological evidence. The root *cɔʔ is suspiciously similar to PAN **asu*, suggesting possible early borrowing, especially as Katuic and Vietic have an a- prefix.

Another root widely attested in Austroasiatic is the word for 'cat' #miaw. There is no archaeozoological data supporting the presence of the cat in Neolithic SE Asia, but again they were domesticated in China relatively early (ca. 3500 BP). It might be thought that the names are merely ideophonic, but in Austronesian for example, names are quite different, often borrowed from English 'pussy'.

The chicken seems to have reached Southeast Asia in the Neolithic. The earliest remains come from Ban Non Wat, in Central Thailand, dated to between 3650-3250 BP. Assuming the Mangic forms are cognate, this would be the only evidence for livestock activity in which can be directly related to early Austroasiatic.

The duck has long been domesticated in China, and given its integration with ricefields in Southeast Asia, it is reasonable to assume it was part of the early Neolithic 'package'. Blust (ACD) reconstructs **itik* to Western Malayo-Polynesian, but it is more likely this is a mosaic of borrowings from Austroasiatic.

Of these reconstructions, the most surprising is the goat, which is not attested archaeologically but for which the linguistic evidence is very strong. It is also notable that aquatic-adapted poultry, such as ducks, appear to be more widely attested than chickens. The limited linguistic presence of chickens should be attributed to their later introduction.

However, it is clear that several branches went through highly restrictive bottlenecks in the earliest period of expansion, presumably losing and then rebuilding agriculture. The most striking case is Nicobaric, which shows no crops which can be attributed to Proto-Austroasiatic[4]. Although the Nicobarese farm today, words such as 'rice' are transparently borrowed in this case from Portuguese, e.g. Nancowry *aroʃ* < Portuguese *arroz*. Like Aslian speakers, the Nicobarese must either have reverted to foraging after leaving the Austroasiatic homeland or else never were farmers..

2.3 Correlation with SE Asian Neolithic

If the linguistic arguments are correct, then Austroasiatic is a flat array phylum with little or no internal structure. This points to a rapid early dispersal followed by incursions by Tibeto-Burman shortly after the primary expansion, creating the geographic fragmentation of individual branches apparent from the map. The reconstructions point to an agricultural society specialised in river basins, to judge by the crops grown and the pointers to aquatic subsistence (Sidwell 2010; Sidwell & Blench 2011; Blench 2018). Sidwell (2022) has 'extended' this hypothesis bringing together both inland a maritime movement. As will be seen later I do not entirely agree with his proposals, but clearly the general outlines are similar. Moreover, the original environment of speakers was tropical, although this covers a large region. Given this, we must seek to correlate this with a plausible archaeological horizon. The centre of the early dispersal of Austroasiatic is most plausibly situated in the Mekong basin as this would allow the various branches to reach their present locations by the shortest

[4] Thanks to Paul Sidwell for a comparative spreadsheet of Nicobaric languages

trajectories. Three maritime routes are posited, the Munda into north-central India, the Nicobars and to Borneo. There are no settlement dates for the Nicobars and the archaeology of India in the region where the Munda may have dispersed is little-known. However, we have to assume this in the era of early rapid dispersal, i.e. 4000-3500 BP, or else the links with other Austroasiatic subgroups would be more visible.

There is a single candidate, the SE Asian Neolithic, which satisfies all these criteria. The archaeological evidence points to a rapid expansion of the Neolithic in the Yunnan/Northern Vietnam borderland, some 4000 years ago (Higham 2002: 85 ff.). Higham (2004:47) notes:

> *The pattern of intrusive agriculturalists settling inland valleys in southern China, while the coast continued to be occupied by affluent foraging groups, is repeated in the Red River area and the contiguous coast of Vietnam.*

The most well-known site of this type is Phùng Nguyên, about 200 km. inland from Halong Bay. Dates remain problematic, but the adjacent site of Co Loa has been dated to 2000 BC. In summarising the situation, Higham (2002:352) says;

> *We find agricultural settlements being founded in the lower Red River valley, along the course of the Mekong and its tributaries, and in the Chrao Phraya valley…The dates for initial settlement, as far as they are known, are approximately the same with none earlier than about 2300 BC. Most intriguingly, the pottery vessels in many of the sites over a broad area have a similar mode of decoration. The sites reveal extended inhumation graves and an economy incorporating rice cultivation and the raising of domestic stock.*

Rispoli (2007:238) in a wide-ranging review of 'incised and impressed' pottery says;

> *The main peculiarity of the incised & impressed pottery style is its sudden appearance around the second half of the 3rd millennium B.C.E. in Neolithic sites distributed in the major river plains of mainland Southeast Asia .... Incised & impressed pottery style, moreover, does not appear in isolation, but it is associated recurrently with: small polished stone tools; stone or shell bracelets and necklace beads.*

The sudden expansion of this distinctive pottery style and associated toolkit and decorative elements is a marker of the Austroasiatic expansion. The dating of the Neolithic in SE Asia proper has been revised in recent years, and the most recent results (Higham & Higham 2009; Higham et al. 2011; Higham & Thosarat 2012; Higham 2021) which make use of Bayesian statistics, have tended to indicate more recent dates, perhaps as late as 3900 BP. Bellwood (2015:55) writes 'South Chinese Neolithic populations with food production based on rice, millet, pigs and dogs pressed southwards'…'in the centuries around 2500-2000 BCE'. However, the direct dating is not on agricultural plants but artifacts, such as shell, in burial sites. To support a more direct link with agriculture, a richer archaeobotanical record is required. Bellwood (2005:132) remarks on the wide distribution of 'incised and zone-impressed' pottery 'across parts of far southern China, northern Vietnam and Thailand after about 2500 BC'. In relation to the spread of this tradition, he says 'Peninsular Neolithic pottery has cord-marked decoration with rare incision and red-slipping, often with tripod feet or pedestals…Gua Cha in Kelantan also has fine incised pottery with zoned punctuation dating to about 1000 BC.' Given the timing proposed here, a correlation with the early phases of southern Austroasiatic expansion would not be impossible.

The dates for initial settlement, as far as they are known, are approximately the

same with none earlier than about 2300 BC. Most intriguingly, the pottery vessels in many of the sites over a broad area have a similar mode of decoration. The sites reveal extended inhumation graves and an economy incorporating rice cultivation and the raising of domestic stock.' It does not seem unreasonable in the light of our understanding of the internal structure of Austroasiatic and its level of diversity to correlate it with this particular subset of the transition from foraging to farming.

Taken together, these elements suggest that we can reconstruct the early history of Austroasiatic as follows;

a) Prior to 4000 years ago, the mainland Southeast Asian region is occupied by vegeculturalists, probably specialised in sago, yams and perhaps Musa spp. They are likely to have been Austro-Melanesian in physical type but speaking languages of an unknown and perhaps unrecoverable phylum. This can be loosely correlated with the Hoabinhian/Bacsonian horizons
b) ca. 4000 years ago, a new style of ceramics and related material culture spreads rapidly throughout the region, associated with beginnings of the Neolithic in the region, and assumed to be correlated with primary Austroasiatic dispersal.
c) Seed agriculture apparently diffused south from China, and must have merged with innovative technologies such as the crossbow and water transport to form the nucleus of the Austroasiatic phylum
d) Although these populations were primarily farmers, hunter-gatherer groups would have remained alongside them[5]
e) they develop improved types of boat, for river transport and maritime movement, accounting for a rapid dispersal in multiple directions
f) the centre of this dispersal might be in the middle Mekong or the Red River
g) Austronesian speaking peoples are simultaneously exploring the region with greatly improved maritime capacity and a 'raiding and trading' culture
h) they cross the Java Strait and populate islands west of Sumatra and well as dispersing adventive rodents around the region
i) they carry Austroasiatic speakers across the Bay of Bengal to the East Coast of India, to the Nicobars and Borneo under unknown circumstances. Few subsistence strategies persisted in the transplanted populations
j) the Aslian must have dispersed near this time, encountering Austronesian speakers in the isthmus, since they have conspicuous Austronesian culture traits. They also encounter and interact with ?Andaman-related populations, accounting for unidentifiable vocabulary in some populations
k) the core Austroasiatic populations adopt a subsistence revolution stimulates them to move both up and down the Mekong but also to spread westward to parallel river systems
l) the overall rapidity of this movement accounts for the difficulty in finding well-supported nested structures in the phylogenetic tree
m) they develop typical material culture for settled farming societies, including baskets, ceramics and cloth production
n) residual forager populations (or those have lost farming subsistence) move northwards into China and adopt a new agricultural repertoire form their neighbours
o) subsequent expansions, particularly of the Daic, Sino-Tibetan and Austronesian language phyla fragment the chain of Austroasiatic languages leading to their comparative geographic isolation in many outlying areas

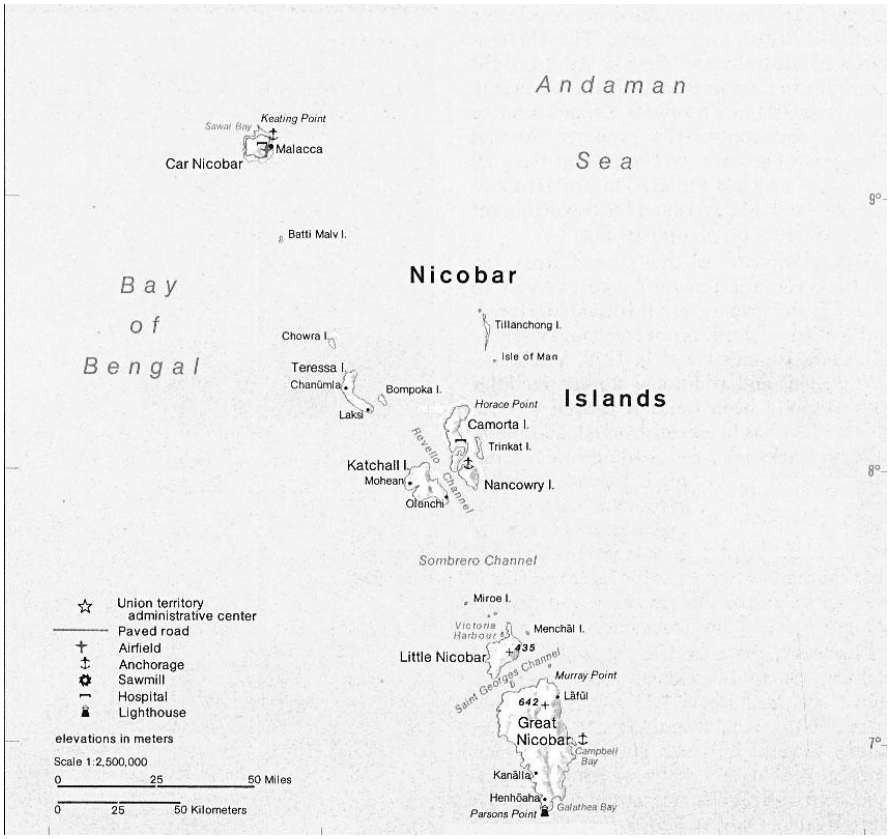
[5] Such a mixture is attested, for example, in Central Tanzania with the Southern Cushitic peoples

p) Munda languages underwent a typological shift in contact with South Asian languages, but this was limited to a single branch rather than indicative of an early two-way division in the phylum. They re-develop farming based on contact with Dravidian and later Indo-Aryan resident populations
q) It has been argued that Nicobaric languages do show significant Austronesian influence (although this was used in support of the 'Austric' argument). However, Munda languages show neither maritime vocabulary nor significant Austronesian borrowings. It is suggested this is a reflection of the conditions under which they were translocated.

4. The settlement of the Nicobars

The Nicobar Islands are an archipelago in the Indian Ocean off the northern tip of Sumatra, south of the Andamans (Map 1). They were known to the Greek geographer, Ptolemy, in manuscript versions as early as the 4th century AD. Almost all the islands are inhabited except for some small islets and all the languages spoken are branches of Nicobarese. However, a group of foragers remain in the centre of Great Nicobar, the Shom Pen, who speak a highly divergent form of the language. Blench (2013) argued that this was a language isolate, which had come under the influence of the surrounding Great Nicobar language, but Sidwell (2022b) considers it is a divergent Nicobaric language. Sidwell (2022b) classified the Nicobaric languages as in Figure 2, which also includes the various sources for Shom Pen data. One clear problem is that not all the records of Shom Pen seem to be consistent, possibly reflecting major divides between different bands. However, reviewing anew the data in Man (1889) I persist with my original argument, that the Shom Pen are an unrelated foraging group, who settled the island prior to the coming of the Nicobarese. Since a similar process evidently happened in the Andamans and Nias, this is not inherently implausible. All Nicobaric languages are fairly close to one another; Car Nicobar is the most divergent, as befits its geographical situation.

The islands have no dated archaeology, hence it is difficult to be sure when they were settled. Since the Andamanese seem to have reached their islands in deep prehistory, it is possible that the Shom Pen also arrived quite early. The Nicobarese proper plausibly arrived around 3500 BP via the same process which dispersed other Austroasiatic branches. Nicobarese languages do not have a specific demonstrable affiliation to any other branch of Austroasiatic. They differ morphologically in some ways from other subgroups, which argues they must have encountered some type of bottleneck. Significantly, Nicobarese languages show no cognates with names of crop plants elsewhere in Austroasiatic, suggesting they arrived without agriculture. The Nicobars are notable for retaining a single element of Austroasiatic hunting technology, namely the crossbow (Photo 1). This is reconstructible to proto-Austroasiatic and the Nicobarese term is presumably cognate with mainland languages.



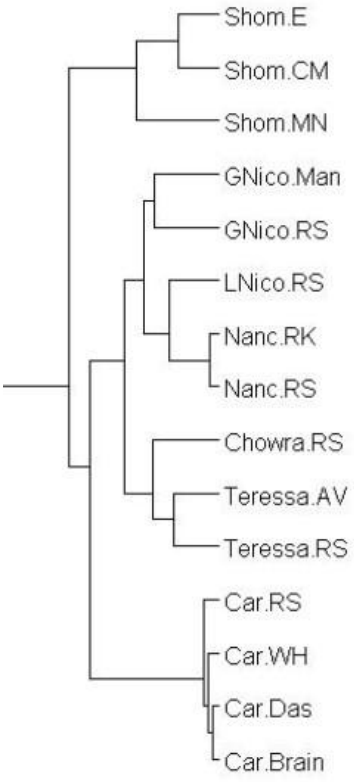
Map 1. Nicobar Islands (Source: Wikipedia[6])

Even more strikingly, they have only limited maritime lexicon; the words recorded for 'boat' today show no cognates with common Austroasiatic vocabulary. Generally, although obviously the fish they name are marine, there is no evidence for transfers from widespread river fish terms elsewhere in Austroasiatic. This makes it possible that their ancestors were carried to the Nicobars by non-Austroasiatic speakers, for unknown purposes. As will be suggested below for the Munda maritime expansion (§5), one likely candidate would be Austronesian-captained vessels. The motive for this was is unclear, but it was a period of intense maritime activity which also saw the settlement of the islands west of Sumatra (and a comparable elimination of prior Palaeolithic populations)[7].

6 Public domain: Taken from Perry-Castañeda Library (PCL), originally from CIA Indian Ocean Atlas - http://www.lib.utexas.edu/maps/islands_oceans_poles/andaman_nicobar_76.jpg

7 Paul Sidwell (p.c.) argues that small groups of SE Asian mainland Austroasiatic speakers could have migrated independently using their own maritime technology, but I do not see how this would have produced the massive lexical replacement of core Austroasiatic in Munda.

Figure 2. Nicobaric classification (Sidwell (2022b)



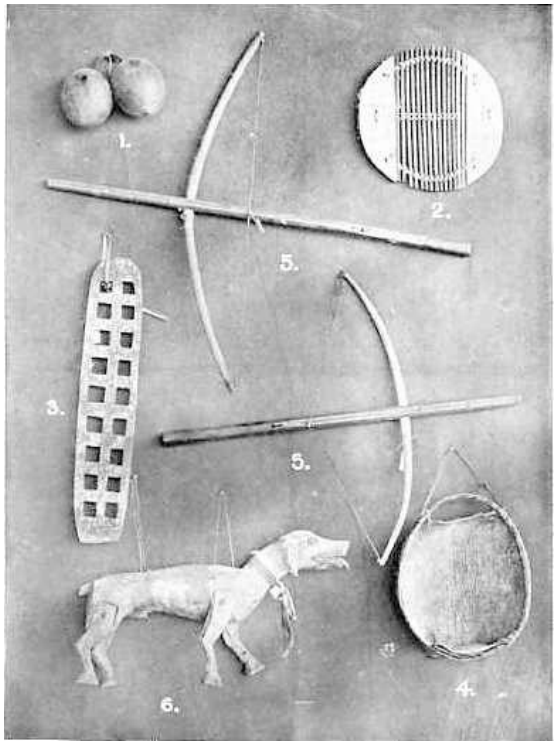


Photo 1. Nicobar crossbows from Nancowry Source: Kloss (1903)

Table 3. Austroasiatic Moklenic comparisons

Gloss	**Austroasiatic**	**Attestation**	**Moklenic**	**Attestation**
paddle	Great Nicobar	*pâūˑah*	Moklenic	*pewa, pəwaʔ*
bamboo shoot	Proto-Monic	*tɓaŋ*	Moklenic	*lubɔɔŋ*
eagle	Nicobarese	*kalâng*	Moklenic	*(ti)cum kəlaaŋ*
screwpine	Nicobarese	*larōm shakēat*[8]	Moklenic	*sakɛ[:]ʔ*
yam	Great Nicobar	*kobi(d)*	Moklen	*koboy*

If the ancestors of the Nicobarese were transhipped by Austronesians, would we expect to find linguistic evidence of this? Laurie Reid (1994, 1999) pursued the argument for Austric, a purported macrophylum which would bring together Austronesian and Austroasiatic. Although there are certainly lexical items common to both phyla, most scholars now consider this reflects early borrowing rather than genetic affiliation. However, Reid compiled evidence showing similarities in grammatical morphemes, common to Nicobarese languages and Austronesian. Sagart (2016) has argued that since these can also be traced in Trans-Himalayan they should be regarded as regional or even evidence of a deep genetic connection. A more interesting possibility

[8] Glossed as 'nine varieties having pinkish pulp'

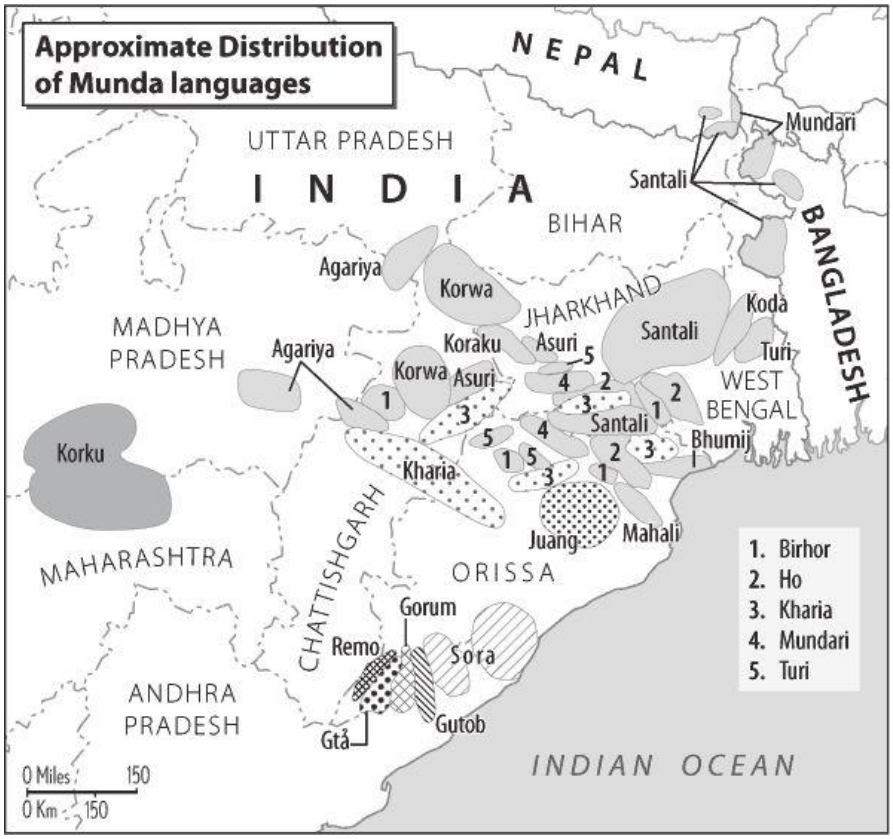
is the identification of regional loans. Larish (1999) compiled a large number of Moklenic reconstructions, as well as regional comparanda. The plausibility of these is extremely variable, but some are worth pursuing. Moklenic languages do not always reflect PMP reconstructions, hence their interest for regional vocabulary. Table 3 shows Austroasiatic Moklenic comparisons taken from his compilation. I regard this as suggestive only, but certainly worthy of more systematic investigation.

If the Nicobarese case is thought implausible, there is a strikingly strong parallel off the west coast of Africa. The Canary Islands were settled around 2000 years ago by populations from the mainland with no maritime capacity (Blench 2020a). Their fish names are all borrowed from Spanish, and at contact they had no boats at all, despite being dispersed across seven islands. Although there is evidence for early agriculture, some islands had reverted to foraging by the time European colonists arrived in the 14th century. The hypothesis is that they were landed on the Canaries by a skilled maritime population (? the Phoenicians) perhaps to collect marine resources, such as the Murex shell used to make purple dye.

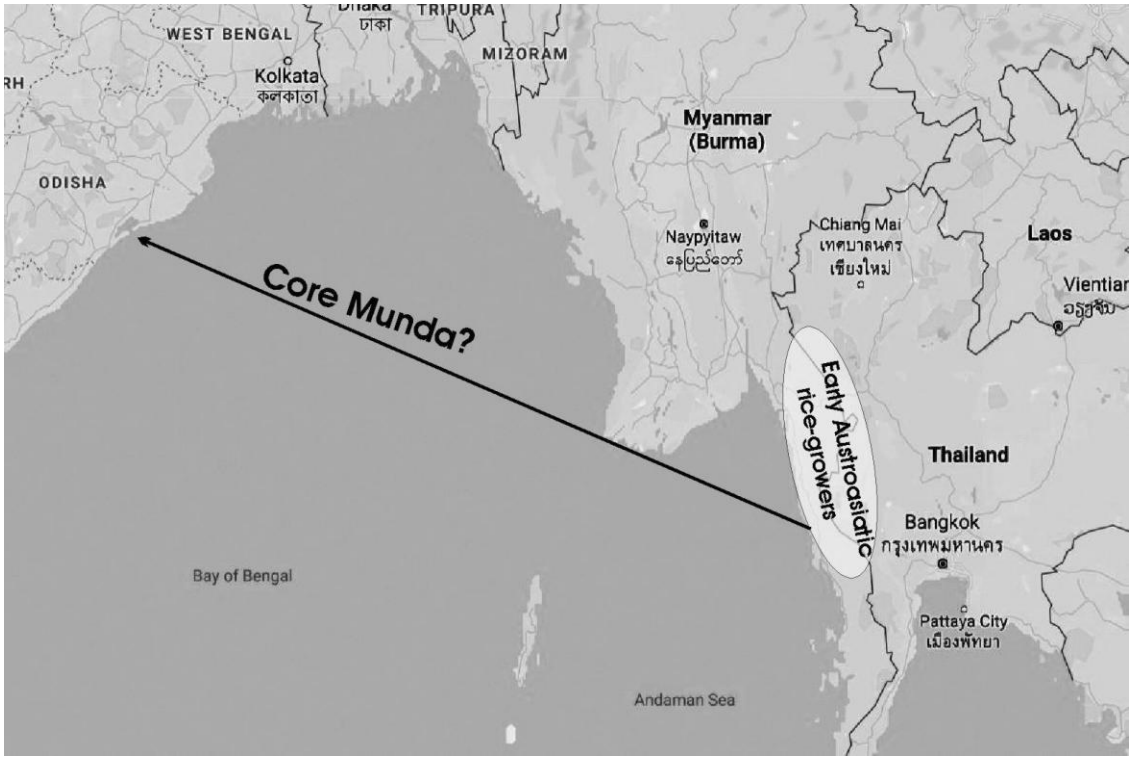
5. The Munda maritime expansion

The Munda languages are the most far-flung and geographically fragmented branch of Austroasiatic, spoken in a broad zone of Central and Northeast India (Map 2). It is usually thought that they must have spread to this region by land, given the presence of Khasian in Northeast India, although it is hard to see exactly what would have driven this dispersal. Felix Rau and Paul Sidwell (2019) have made a proposal which goes some way to resolving this problem, namely that the Munda, far from migrating by land, travelled by a maritime route across the Bay of Bengal. Sidwell (2022a) proposes a coastal route, but I argue that a direct sea crossing is more likely. This would certainly resolve the issue of the geography of Munda languages. The founding population would have been landed on the coast somewhere in modern Orissa, and dispersed north and south by the expansion of Indo-Aryans.

The Munda languages should provide a guide to the subsistence strategies of the migrants. The linguistics of cultivated plants in Munda languages have been studied in some detail in Zide & Zide (1972). This demonstrates that although rice vocabulary is attested, no other domestic plants show connections across the Bay of Bengal. Similarly, with domestic animals; apart from the chicken, these appear to have been adopted from neighbouring peoples on the Indian mainland. Munda languages have no maritime vocabulary, although river-based aquatic subsistence terms do show some cognates with mainland SE Asian lexicon (Blench 2018). Some of the existing Munda peoples are near-hunter-gatherers, all of which points to their being transported in ships owned by other peoples. Other aspects of Munda material culture, including agricultural implements and musical instruments, do attest to MSEA heritage. Blench (2022) proposes that the Munda must have travelled in Austronesian-captained shipping in the period 3500 ~ 4000 BP, leaving from the south of Myanmar on the Isthmus of Kra (Map 3).



Map 2. Munda languages (Source: Anderson 2008:2)



Map 1. Hypothetical Munda migration

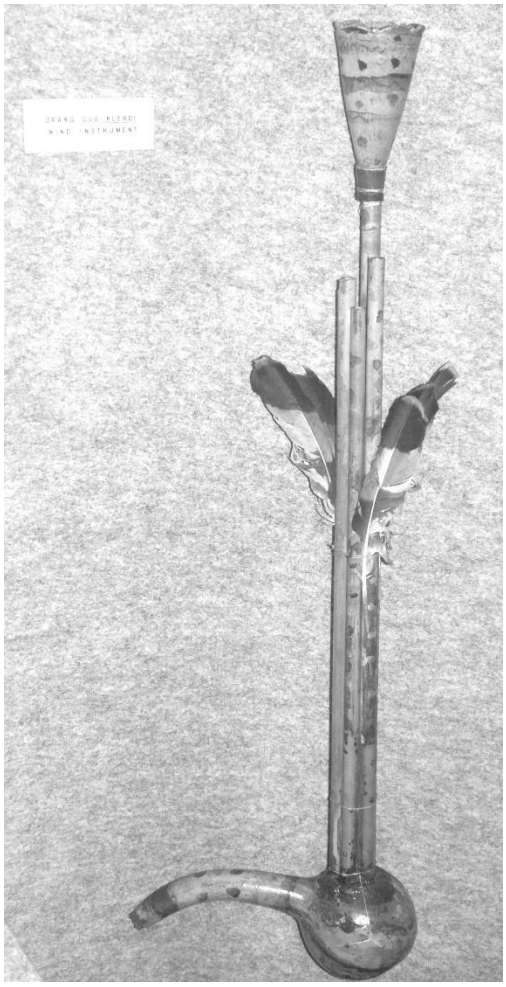


Photo 2. Dayak mouth organ (Source: Author)

If this is so, would not Munda show more evidence of contact with Austronesian? This depends strongly on the context of their historical movement. Compare this with the movement of the Barito of Borneo to Madagascar in Malay ships (Adelaar 1989). The Barito are inland people, lacking ocean skills, so it is proposed they were carried as serfs. Basic material culture and domestic animals were not transmitted, again with the sole exception of rice. Adelaar shows that Malagasy has retained words connected with maritime transfers (such as names of winds or ship parts). However, otherwise, Malagasy is conspicuously lacking in Malay loanwords. Munda would have had no reason to retain maritime lexicon and this example shows there is no necessary transfer of other types of vocabulary. A comparable example of the Guanche, the ancient inhabitants of the Canaries, is given in §4. In the New World, the Maroons of Surinam were enslaved and carried across the Atlantic. Their original languages were lost when they escaped into the forest, living along rivers deep in the interior. Their subsistence strategies are almost all rebuilt from Dutch and Amerindian techniques, with no transmission of African crops. Despite this, they maintain a strong Africa-oriented culture, which differentiates them from their neighbours. It is not at all clear why the Munda should have travelled with such a limited subsistence repertoire. However, there are plenty of comparable examples in the region and elsewhere in the world to suggest that such maritime transfers do occur.

6. Settlement in Borneo

The island of Borneo/Kalimantan is entirely populated by Austronesian speakers at present. We know that it has been settled for a very long time, since it has some of the earliest rock paintings in the world. Presumably these were created by Palaeolithic foragers, but any trace of such people is long gone. Although there are hunter-gatherers, the Punan, these seem to be Austronesians who have lost agriculture rather than a remnant population. Blust (2010) put forward the 'Macro-Borneo' hypothesis, arguing that the Austronesian languages of Borneo are too similar to be explained by initial Austronesian settlement processes and that they must have undergone levelling in the quite recent past, perhaps as late as 2000 BP.

Apart from Austronesian, there is evidence that regions of southwest Borneo show some distinctive elements of Austroasiatic culture, including the *sumpotan*, or mouth organ (Photo 2), which is characteristic of MSEA (Blench 2017, 2020b). There are a few lexical items which also appear to be cognate with Austroasiatic rather than Austronesian, for example the word for 'monkey'. Blench (2011) argued that there must have been Austroasiatic presence, probably settled from the Vietnamese coast opposite, which would have been responsible[9]. Since that paper was published, more extensive archaeological evidence has been presented, showing a strong connection between the ceramic styles of the mainland and those of this region. Blevins & Kaufman (2023) have published a far more extended discussion of possible comparisons, some of which appear to be far-fetched.

7. Northern movement to China

7.1 General

There are three isolated Austroasiatic languages spoken in South China and northern Vietnam, Mang, Bolyu and Bugan. It was originally assumed that they must form a single branch of Austroasiatic, based mainly on geography[10]. However, expanded datasets make it clear that they fall into two distinct branches, Mangic and Pakanic, i.e. Bolyu and Bugan (Sidwell 2015). All three have been heavily restructured through contact with neighbouring languages, both Tai-Kadai and Trans-Himalayan, and are now tonal and largely monosyllabic. Strikingly, however, none of these languages show cognates for subsistence crops when compared with other branches of Austroasiatic. Although their rice vocabulary is elaborated, it does not link with other branches or indeed to each other. A single lexical item, Mang *ʔaːm*^{1}, Bolyu *qam*^{53} 'husk, chaff' goes back to Proto-Austroasiatic **skaːmʔ* but this can equally well reflect the processing of wild grains by foragers.

7.2 Mang

The Mang [zng] (莽人; Vietnamese: Mảng) live primarily in Lai Châu, northwestern Vietnam. Although a genetic study has been undertaken it concluded nothing more than that the Mang were of 'southern origin', a hardly startling conclusion (Ten et al. 2007). The Vietnamese authorities have posted a brief ethnography profile on a government

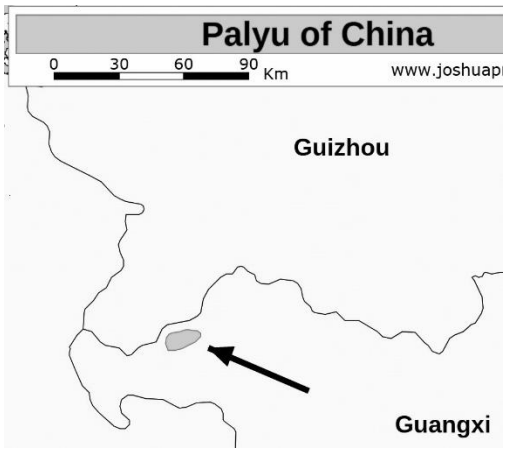
9 Sidwell (pers. comm.) argues these similarities can equally well be explained by trade

10 Unfortunately, a view still endorsed by the online Mon-Khmer Etymological Dictionary

website[11]. The Mảng are shifting cultivators and use axes to fell trees, long knives to cut branches, digging sticks, and simple ploughs. The main staple foods include dry rice and maize, and secondary food comes from cassava, sweet potatoes and pumpkins. The upland rice may be old, although the term used does not resemble other Austroasiatic languages, but all the other crops mentioned are recent, post-Columbian, introductions. The Mang term for 'winnowing basket' resembles other Austroasiatic lexemes although as it is found outside the phylum this is not absolute proof of a connection. The Mang lexicon is not known in depth, but no names for ceramics or baskets have yet been connected with other Austroasiatic branches suggesting that the Mang movement to this region was originally foraging.

7.3 Pakanic

Pakanic consists of two languages, Bolyu [ply] (巴琉语, 布流语) and Bugan [bbh], Bogan, Pakan, or Bugeng (布甘语). Documentation on both remains weak. Although the Bolyu have mostly lost their language, a survey in 2011 showed that the number of speakers had actually increased from 650 in 1980 to 1200 in 2010 (Qin & Li 2011). Their subsistence is also undescribed although they are clearly rice growers today. What subsistence terms are recorded show almost no connection with external Austroasiatic.



Map 4. Bolyu settlement (Source: Joshua Project)

[11] Mang ethnic group (nhandan.vn) & https://www.vietnamroyaltourism.com/Mang-People-in-Vietnam.html

Map 5. Bugan settlements (Source: Joshua Project)

The Bugan are slightly more numerous, with about 3000 speakers, mostly in some villages in southern Guangnan (广南) and northern Xichou (西畴), Yunnan Province, China (Map 5). What little is known about their language is summarised in Li (1996) and Li & Luo (2015). There is a part-translated wordlist with hand-written equivalents which provides basic lexical data[12]. Again the subsistence terms show hardly any correspondences with other Austroasiatic branches, although basic vocabulary clearly indicates the basic affiliation is correct.

7.4 Interpretation

The ancestors of Mang and Pakanic must have migrated north in an era when South China was highly ethnolinguistically diverse. As with Nicobaric, Mang and Pakanic, they may have been foragers or reverted to foraging, since wild animal names can be linked to other Austroasiatic branches. Although the Mang use the crossbow, *piŋ*, the term is not cognate with other Austroasiatic names. The crossbow is apparently unknown to Pakanic speakers. All three groups have then separately rebuilt agriculture based on contact with their new neighbours and adopted species appropriate to their new environment.

[12] Thanks to Paul Sidwell for the scan of this.

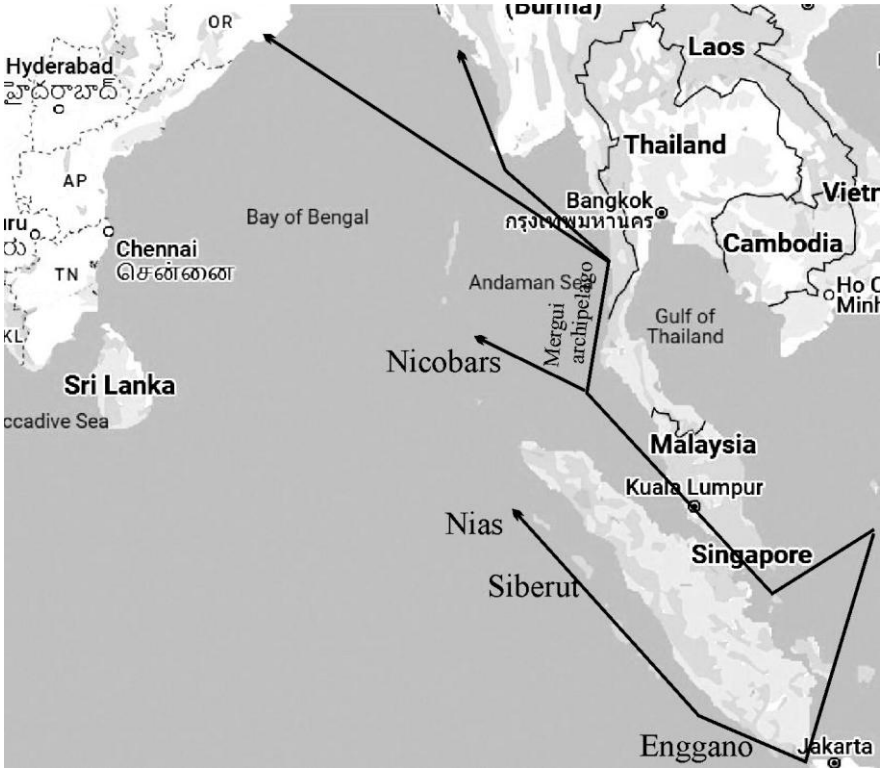
Map 6. Rong (=Lepcha) today (Source: Blench 2023)

8. To Nepal and Sikkim; Rongic a lost Austroasiatic branch?

Blench (2023) has argued that there was yet another branch of Austroasiatic, which reached the borderland between Nepal and Sikkim. In terms of common vocabulary in the Lepcha language [=Rong] with Austroasiatic, the evidence for this is quite striking. Map 6 shows the current distribution of the Lepcha people. Unfortunately, none of the shared lexicon with Austroasiatic points clearly to subsistence; Lepcha agricultural terms are all linked with their neighbours. Perhaps Lepcha was also originally a foraging population.

9. The Bay of Bengal; an early Austronesian sea?

It has been suggested in the course of this paper that ships with Austronesian owners and masters may have been active very early in the Bay of Bengal, moving goods and people. Blench (2009) pointed out that there is non-linguistic evidence for Austronesian presence in a wide area where Austronesian languages are no longer spoken, including the Arabian Gulf, the coast of East Africa and the Bay of Bengal. It was previously thought the Moklen and Moken languages spoken by the sea nomads of the Mergui archipelago were a branch of Malayic, which would date them to later than 2000 BP (e.g. Blench 2021). However, Smith (2017) has revised the classification of early Western Austronesian, arguing instead that these languages were a primary branch of Malayopolynesian, i.e. part of the dispersal out of southern Taiwan. If this is indeed the case, then the Mergui languages would have been spoken by the maritime populations exploring this area as part of the first wave of Austronesian expansion.



Map 7. Proposal for Austronesian expansion into the Bay of Bengal

This is paralleled by the problematic classification of the languages of the islands west of Sumatra, Nias, Siberut [=Mentawei] and Enggano. These languages are Austronesian, but are not closely related to the languages of adjacent Sumatra, nor do they form a group in themselves. Enggano in particular is highly divergent[13]. No in-depth historical linguistics has been undertaken which places the west Sumatran languages securely in the Austronesian tree and they may thus also be evidence for the early period of Austronesian expansion into the Bay of Bengal.

Map 7 synthesises the possible evidence for Austronesian activity in the Bay of Bengal. A migrant stream from the Taiwan/Philippines area splits into two upon reaching Sumatra. One stream passes south through the Sunda Strait between Java and Sumatra and reaches the Sumatran offshore islands, displacing Palaeolithic foragers. The northern stream passes up the west side of the isthmus, settling the Mergui archipelago. They are in intensive contact with the Austroasiatic populations resident on the mainland and capture or enslave peoples who are actual or near foragers on the mainland and carry them both across the Bay of Bengal to Orissa and to the Nicobars.

[13] The present author has put forward the hypothesis that the Enggano, who were foragers when the first Europeans arrived, and lived in remarkable 'beehive' houses similar to the Nicobars were not Austronesian. However, Edwards (2015) has argued the language is Austronesian. Archaeology would be of considerable assistance in resolving this issue.

Map 8. Distribution of the bandicoot rat

This summarises the linguistic evidence, but the presence of a highly mobile population would also have transported (unwittingly) animals and plants. The Polynesians carried the Polynesian rat, *Rattus exulans*, from its native SE Asia into the Pacific, together with a wide variety of plant species (Blench 2008). One piece of evidence for this, drawn attention to in Blench (2009), is the presence of rodent species which have been carried around the Bay of Bengal, perhaps as a pest of crops. Groves (1995) first noted these, although he did not draw the historical conclusions highlighted here. Both the fawn-coloured mouse (*Mus cervicolor*) and the lesser bandicoot rat (*Bandicota bengalensis*) are found in South Asia and Island Southeast Asia. Their 'natural' homelands are in South Asia, but they are thought to have been translocated across eastwards at some unknown point in prehistory. As an illustration of this, Map 8 shows the distribution of the bandicoot rat (Photo 3). The records for SE Asia are sporadic and discontinuous, indicating translocation. Although the rust red area is the 'natural' distribution, the bandicoot rat is regarded as a pest of rice fields in Sri Lanka and South India, as it is in Indonesia. This region is shaded yellow, suggesting it was also introduced there by the same wandering Austronesians.

Earlier authors noted some of these east-west connections without attaching dates. For example, Hornell (1920) remarked on striking similarities in boat construction to suggest 'Polynesian' influence in India, as well as pointing to coconut cultivation and toddy tapping as introduced cultural elements. Waruno Mahdi (1998, 1999) synthesised textual references and evidence for shipping types, concluding that the 'Nāgas' referred to in early texts 'typically inhabited islands, the sea coast or banks of rivers. Some of them worshipped megaliths and practised buffalo sacrifice and head-hunting' (Mahdi 1999:182). Identifying such populations with early Austronesian migrants would not be unreasonable.

Photo 3. Lesser bandicoot rat (Source: CC)

Map 9. Native range of the santol fruit (Source: Adapted from CC)

There are also domestic plants transferred across the Bay of Bengal in an early period. One example is the clove, and Mahdi (1999) also proposes relationships including words for 'lime' and 'camphor'. However, there is also phytogeographical evidence for other fruit species (Blench 2008), for example the bilimbi and carambola (*Averrhoa* spp.), the lime (*Citrus aurantifolia*), the coconut (*Cocos nucifera*), the langsat (*Lansium domesticum*), the noni (*Morinda citrifolia*) and the santol (*Sandoricum koetjape*). The santol, a fruit of limited importance today, was formerly important for its sour fruit, used in cooking. Map 9 shows the range where it is endemic, but it now occurs along the east coast of India and was presumably carried there by the early maritime trade.

Our knowledge of the archaeobotany of the region remains very limited. Even though this type of fruit should leave easily identifiable macro-remains, none have yet been reported. This is a rather obvious target for archaeological research, but also for further ethnographic work on the distribution and community use of these species.

10. A mosaic of foragers and farmers?

Archaeologists may find evidence from lacunae in the linguistic record less satisfying than evidence from excavation, but there are so many gaps in the geography of sites that linguistics has to be an important resource. Although Austroasiatic speakers are predominantly farmers today, linguistic evidence for cultivation tying them to the Austroasiatic proto-language is conspicuously absent in some subgroups. The interpretation must be that either these groups were initially foragers or that they reverted to foraging and subsequently rebuilt crop subsistence repertoires. If so, in the period when Proto-Austroasiatic was inchoate, a mosaic of early farmers and foragers, some migrated away before adopting cultivation. Table 4 is intended to capture this transition from original subsistence to agriculture.

Table 4. Austroasiatic subgroups with original subsistence strategies

Subgroup	Original subsistence	Agriculture	Core material culture
Aslian	Hunter-gatherer	Never adopted	Lost or never present
Bahnaric	Agriculture	Core strategy	Present
Borneo	Hunter-gatherer	Unclear	Unclear
Katuic	Agriculture	Core strategy	Present
Khasic	Agriculture	Core strategy	Present
Khmeric	Agriculture	Core strategy	Present
Khmuic	Agriculture	Core strategy	Present
Mangic	Hunter-gatherer	Adopted	Rebuilt
Monic	Agriculture	Core strategy	Present
Nicobaric	Hunter-gatherer	Adopted	Rebuilt
Pakanic	Hunter-gatherer	Adopted	Rebuilt
Palaungic	Agriculture	Core strategy	Present
Pearic	Agriculture	Core strategy	Present
Rongic	Hunter-gatherer	Adopted	Rebuilt
Vietic	Agriculture	Core strategy	Present

The shading for Borneo and Rongic indicates these are proposed by the author but not necessarily accepted by other Austroasiatic researchers. The concept of ‘core material culture’ represents the array of ceramics, baskets and other aspects of settled life which are widespread in the groups which were always farmers and may have been adopted in those which were originally foragers.

This involves the assumption that the core population which dispersed to become Austroasiatic of today was a mixture of farmers and foragers, rather than having a single subsistence strategy. Similar patterns occur in other parts of the world, for example, among Southern Cushitic speakers in Tanzania, where the farmers and herders of the Iraqw group co-exist with Asax foragers. Hunter-gatherers survive today among otherwise settled Austroasiatic speakers, for example, the Khmuic Mlabri in Thailand (Chazée 2001). Photo 4 shows an idealised Mlabri hunter accompanying a display in the Hill Tribes Museum, Chiang Mai.

This model still dates Austroasiatic dispersal at around 4000 BP, but assumes that it cannot be only the adoption of Neolithic technologies which was the driver. It seems possible that the introduction of new hunting and aquatic capture technologies played a key role. We know that the crossbow spread to the Nicobars (Blench 2017), and the blowpipe to the east coast of India (Blench 2009). Although we have no archaeological

evidence for river boats at this early period, terms for such boats are extremely widespread (Blench 2009). Interaction with Austronesian speakers may also have played a role as we know that common material culture elements are shared such as the distinctive back basket, or the struck bamboo zither. Blench (ms.) illustrates these type of zithers which are found on both sides of the Bay of Bengal and in the Austronesian island world. Subsequently, then the characteristic ceramics and baskets spread through the core populations who by this period have the adopted the typical attributes of sedentary culture (e.g. Alves 2022). The rapid early expansion of Austroasiatic seems to have created cultural bottlenecks, such as the loss of subsistence strategies, for example, crops and livestock. This explains the farming societies such as Mangic and Pakanic where crops terms show no correspondences with the common Austroasiatic lexicon.

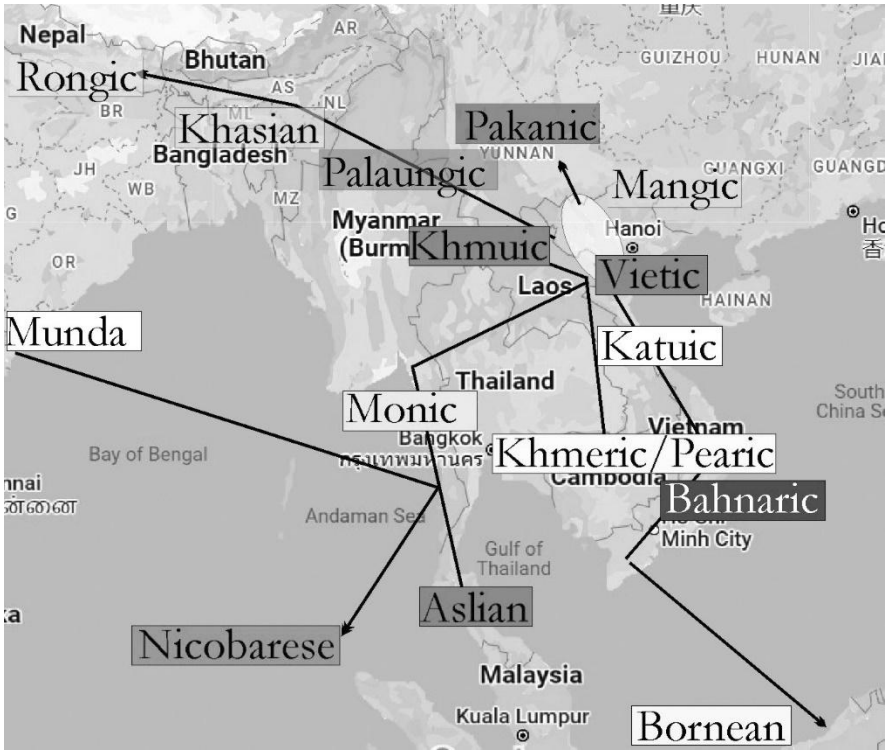
The sea remains important but its significance has not been fully explored. Open ocean voyaging played a role in the Munda, Nicobarese and Borneo expansions. If it is the case that the maritime agents were Austronesians then this might also explain the long-noticed connections between the two phyla. There is a long history of speculation that this was evidence for a genetic connection, the Austric hypothesis. Terms for 'sesame', 'dog', 'taro', 'boat type' are shared between the phyla, evidently at a deep level, but not necessarily at the level of the proto-language. The suggestion is that these similarities are the consequence of maritime interaction immediately subsequent to the first expansion of both phyla.

11. Synthesis

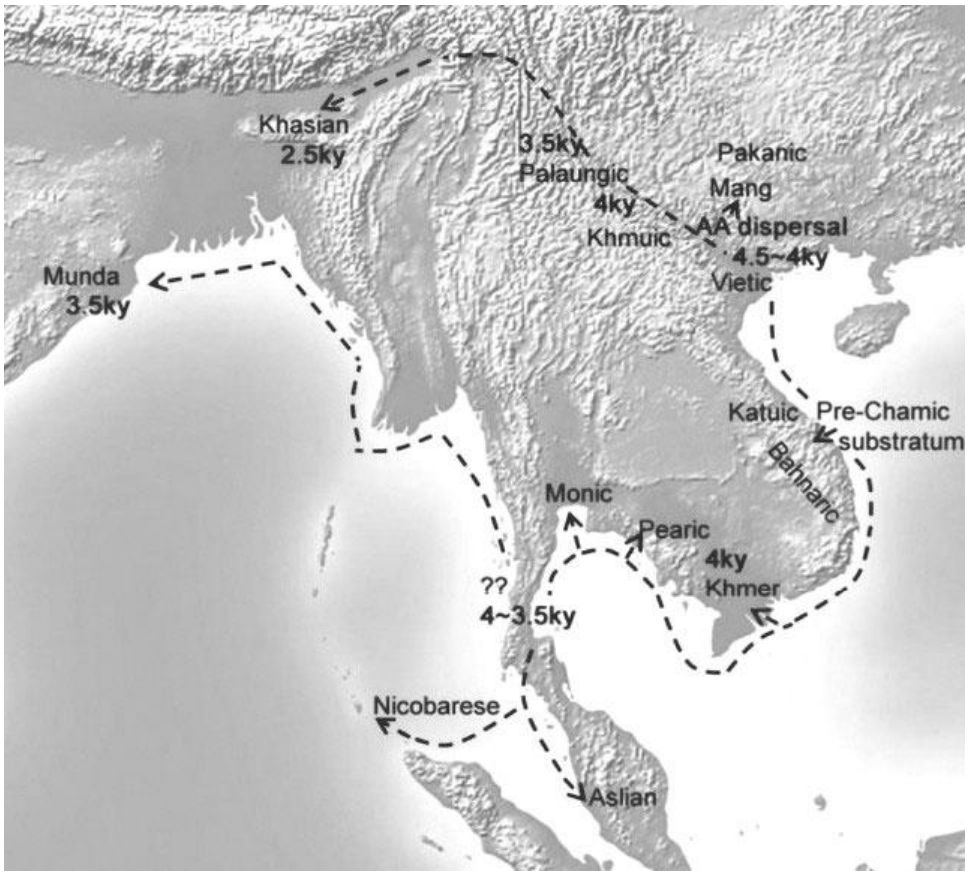
The picture presented here differs radically from the usual image of Austroasiatic expansion. In part this is because it incorporates evidence from archaeology, and supposes that even where Austroasiatic is no longer spoken, substrate lexicon can provide evidence for its former presence. It also uses negative evidence from comparative lexicon to argue that not all the groups represented in the primary expansion of Austroasiatic were farmers but were rather a mosaic with foragers. It suggests that the key innovations may have been aquatic and in hunting technology, such as the crossbow. It proposes that the common material culture of core Austroasiatic branches spread soon after the initial dispersal, explaining the numerous common terms.

Map 10 represents a synthesis of the expansion of Austroasiatic speakers, following the arguments in this paper. The geographic locations are very approximate and the arrows suggest routes, not a genetic connection. Thus, because an arrow to Eastern Nepal passes through Khasian, this does not imply I consider Rongic to have a special connection to Khasian.

Sidwell (2022a) has published a synthesis of his ideas concerning Austroasiatic dispersal (Map 11). While this re-affirms the importance of aquatic/maritime routes, the interpretation is very different. His hypothesis is that the Munda followed the coastline of the Bay of Bengal some 3kya, arriving in their present centre of gravity. I do not believe this can be integrated with the linguistic geography of Munda, given its north/south division. Most importantly, there is a plethora of evidence for the transmission of cultural and biological features *across* the Bay of Bengal. Since we have ample evidence for Austronesian shipping west of the isthmus, a direct transit across the sea is more plausible.



Map 10. Revised dispersal of Austroasiatic proposed in this paper



Map 11. Austroasiatic dispersal proposed by Sidwell (2022a)

Even so, it should be emphasised this remains speculative. We have too little archaeology in many regions to clarify the routes and dates of these migrations. The reconstruction of material culture in Austroasiatic remains at a preliminary stage and the absence of detailed lexicon some branches, such as Mangic and Pakanic, means that definitive statements cannot yet be supported. The characterisation of the Bay of Bengal as an 'Austronesian sea' goes against conventional wisdom. But as with all new paradigms, that is inevitable.

References

Adelaar, Sander A. 1989. Malay influence on Malagasy: linguistic and culture-historical implications. *Oceanic Linguistics*, 28(1):1-46.

Alves, Mark J. 2020. Historical Ethnolinguistic Notes on Proto-Austroasiatic and Proto-Vietic Vocabulary in Vietnamese. *Journal of the Southeast Asian Linguistics Society*, 13.2:xiii-xlv.

Alves, Mark J. 2022. *Preliminary Etymological Research on Words for Pottery in Mainland Southeast Asian Language History*. Presentation at the 31st Annual Meeting of the Southeast Asian Linguistic Society, the University of Hawaii, Honolulu, Hawaii. 2022.

Anderson Gregory. 2008. Inroduction to the Munda Languages. In Gregory Anderson (ed.) 2008. *The Munda Languages.* London/New York: Routledge. 1-10.

Blench, Roger M. 2007. The language of the Shom Pen: a language isolate in the Nicobar Islands. *Mother Tongue*, XII: 179-202.

Blench, Roger M. 2008. A history of fruits on the Southeast Asian mainland. *Occasional Paper 4: Linguistics, Archaeology and the Human Past*. Toshiki Osada and Akinori Uesugi eds. 115-137. Japan: Indus Project, Research Institute for Humanity and Nature, Kyoto.

Blench, Roger M. 2009. Remapping the Austronesian expansion. In: *Discovering history through language. Papers in honour of Malcolm Ross*. Bethwyn Evans (ed) 35-59. Canberra: Pacific Linguistics.

Blench, Roger M. 2011a. The role of agriculture in the evolution of Southeast Asian language phyla. In: *Dynamics of Human Diversity in Mainland SE Asia*. N.J. Enfield ed. 125-152. Canberra: Pacific Linguistics.

Blench, Roger M. 2011b. Was there an Austroasiatic presence in island SE Asia prior to the Austronesian expansion? *Bulletin of the Indo-Pacific Prehistory Association*, 30: 133-144.

Blench, Roger M. 2012. Vernacular Names for Taro in the Indo-Pacific Region: Implications for Centres of Diversification and Spread. Irrigated Taro (*Colocasia esculaenta*) in the Indo-Pacific. Matthew Spriggs, David Addison, and Peter J. Matthews eds. *Senri Ethnological Studies*, 78:21-43.

Blench, Roger M. 2014. The origins of nominal affixes in MSEA languages: convergence, contact and some African parallels. In: *Mainland Southeast Asian Languages: The State of the Art*. N.J. Enfield and Bernard Comrie eds. Canberra: Pacific Linguistics.

Blench, Roger M. 2017. Ethnographic and archaeological correlates for a mainland Southeast Asia Linguistic Area. In: *Spirits and Ships: Cultural Transfers in Early Monsoon Asia*. Acri Andrea, Roger Blench, and Alexandra Landmann eds. 207-238. Singapore: ISEAS – Yusof Ishak Institute.

Blench, Roger M. 2018. Waterworld: lexical evidence for aquatic subsistence strategies in Austroasiatic. In Hiram Ring and Felix Rau (eds.) *Papers from the Seventh International Conference on Austroasiatic Linguistics* (JSEALS Special Publication No. 3). Manoa: University of Hawaii Press. 174-193.

Blench, Roger M. 2020a. The peopling of the Canaries by the Berbers: new data and new hypotheses. *Études et documents berbères*, 45(2): 149-173.

Blench, Roger M. 2020b. The history and distribution of the free-reed mouth-organ in SE Asia. In: *EurASEAA 14, Dublin 2012, proceedings. Volume I*. Helen Lewis ed. 94-110. Oxford: Archaeopress.

Blench, Roger M. 2021. Restructuring our understanding of the South China Sea interaction sphere: the evidence from multiple disciplines. In: *Taiwan Maritime Landscapes from Neolithic to Early Modern Times: Cross-Regional Perspectives*. Paola Calanca & Frank Muyard eds. Taiwan: Academica Sinica.

Blench, Roger M. 2023. Rongic: a vanished branch of Austroasiatic. *JSEALS*, Special Issue. 46-66.

Blench, Roger M. ms. The Munda maritime dispersal: when, where and what is the evidence? ms.

Blench, Roger M. and Paul Sidwell 2011. Is Shom Pen a distinct branch of Austroasiatic? *Austroasiatic studies. ICAAL IV. Mon-Khmer Studies, Special Issue 3*: 9-18.

Blevins, Juliette and Daniel Kaufman 2023. Lexical Evidence in Austronesian for an Austroasiatic presence in Borneo. *Oceanic Linguistics* 62.2:366-413.

Blust, Robert A. 1996. Beyond the Austronesian homeland: the Austric hypothesis and its implications for archaeology. *Transactions of the American Philosophical Society, New Series* 86.5: 117-158.

Blust, Robert A. 2010. The greater north Borneo hypothesis. *Oceanic Linguistics*, 49(1):44-118.

Chazée, Laurent 2001. *The Mlabri in Laos: A World under the Canopy*. Bangkok: White Lotus Press.

Diffloth, Gérard 2005. The contribution of linguistic palaeontology and Austroasiatic. In: Laurent Sagart, Roger Blench and Alicia Sanchez-Mazas, eds. *The Peopling of East Asia: Putting Together Archaeology, Linguistics and Genetics.* 77–80. London: Routledge Curzon.

Edwards, Owen 2015. The position of Enggano within Austronesian. *Oceanic Linguistics*,:54-109.

Groves, Colin 1995. Domesticated and commensal mammals of Austronesia and their histories. In: Bellwood, Peter J., James Fox and Darrell Tryon, eds. *The Austronesians: historical and comparative perspectives*. 152–163. Canberra: Research School of Pacific and Asian Studies, Australian National University.

Higham, Charles 2002. *Early cultures of mainland Southeast Asia*. Bangkok: River Books.

Higham, Charles 2004. Mainland Southeast Asia from the Neolithic to the Iron age. In: Ian Glover & Peter Bellwood (eds.) *Southeast Asia: from prehistory to history*. 41-67. Abingdon: RoutledgeCurzon.

Higham, Charles 2013. Hunter-Gatherers in Southeast Asia: From Prehistory to the Present. *Human Biology*, 85,1-3:21-43.

Higham, Charles 2021. The Neolithic occupation of Southeast Asia. In: *The Languages and Linguistics of Mainland Southeast Asia: a comprehensive guide*. Paul Sidwell and Mathias Jenny (eds.) 19-31. Berlin/Boston; De Gruyter Mouton.

Hornell, James 1920. 67. The Common Origin of the Outrigger Canoes of Madagascar and East Africa. *Man*, 20: 134-139.

Kloss, C. Boden 1903. In the Andamans and Nicobars: The Narrative of a Cruise in the Schooner "Terrapin", with Notices of the Islands, Their Fauna, Ethnology, etc. London: John Murray, Albermare Street West.

Larish, Michael D. 1999. The position of Moken and Moklen within the Austronesian language family. PhD. University of Hawai'i at Manoa.

Li, Jinfang & Luo, Yongxian 2015. Bugan. In: *The Handbook of Austroasiatic Languages*. Paul Sidwell and Mathias Jenny (eds.) 1033–1064. Leiden: Brill.

Li, Jinfang 1996. Bugan – A New Mon–Khmer Language of Yunnan Province, China. *Mon–Khmer Studies*, 26: 135–160.

Mahdi, Waruno 1998. Linguistic data on transmission of Southeast Asian cultigens to India and Sri Lanka. In: *Archaeology and Language, vol. 2: Artefacts, languages and texts*. Blench, Roger M. and Matthew Spriggs eds. 390-415. One World Archaeology 29. London & New York: Routledge.

Mahdi, Waruno 1999. The dispersal of Austronesian boat forms in the Indian Ocean. In Blench, Roger M. and Matthew Spriggs (eds) *Archaeology and Language III*. 144-179. London & New York: Routledge.

Qin, Xiaohang & Li, Fanglan 2011. The status quo and trend of language use by Lai people. *Mon-Khmer Studies*, 40: 107–115.

Rau, Felix and Paul Sidwell 2019. The Munda maritime hypothesis. *Journal of the Southeast Asian Linguistics Society*, 12(2): 35-57.

Reid, Lawrence A. 1994. Morphological evidence for Austric. *Oceanic Linguistics*, 33(2):323-344.

Reid, Lawrence A. 1999. New linguistic evidence for the Austric hypothesis. in: E. Zeitoun and P. J.-K. Li (eds) *Selected Papers from the Eighth International Conference on Austronesian Linguistics*, Taipei: Academia Sinica.

Sagart, Laurent 2016. The Wider Connections of Austronesian: A Response to Blust (2009). *Diachronica*, 33(2): 255-281.

Sidwell, Paul 2000. *Proto South Bahnaric: a reconstruction of a Mon-Khmer language of Indo-China*. Canberra: Pacific Linguistics.

Sidwell, Paul 2005. The Katuic Languages: classification, reconstruction and comparative lexicon. Lincom Europa.

Sidwell, Paul 2010. The Austroasiatic central riverine hypothesis. *Вопросы языкового родства/Journal of Language Relationship*, 4:117-134.

Sidwell, Paul 2014. Khmuic classification and homeland. *Mon-Khmer Studies*, 43.1:47-56.

Sidwell, Paul 2014. Proto Khasian: an emerging reconstruction. In: *North East Indian Linguistics Volume 6*. Gwendolyn Hyslop, Linda Konnerth, Stephen Morey, Priyankoo Sarmah (eds.) 149-163. Canberra: Asia-Pacific Linguistics.

Sidwell, Paul 2015. The Palaungic Languages: Classification, Reconstruction and Comparative Lexicon. Lincom GmbH.

Sidwell, Paul 2022a. Austroasiatic dispersal: the AA "water-world" extended. Were the proto-Austroasiatics coastal migrants? *JSEALS*, 30: 56-72.

Sidwell, Paul 2022b. A Classification of the Nicobarese languages. *JSEALS*, 15(1):1-16

Sidwell, P. and Pascale Jacq 2003. *A Handbook of Comparative Bahnaric: volume 1-West Bahnaric*. Canberra: Pacific Linguistics.

Sidwell, Paul and Felix Rau 2015. Austroasiatic Comparative-Historical Reconstruction: An Overview. In: Jenny, Mathias and Paul Sidwell eds. 1:221-263. *The Handbook of Austroasiatic Languages*. Leiden: Brill.

Sidwell, Paul & Roger M. Blench 2011. The Austroasiatic *Urheimat* : the Southeastern Riverine Hypothesis. In: N. Enfield ed. *Dynamics of Human Diversity in Mainland SE Asia*. 317-345. Canberra: Pacific Linguistics.

Smith, Alexander D. 2017. The Western Malayo-Polynesian problem. *Oceanic Linguistics*, :435-490.

Tan, Sijie, Yang, Minhui, Yu, Haijing et al. 2007. Y-Chromosome Polymorphisms Define the Origin of the Mang, an Isolated Population in China. *Annals of Human Biology*. 34 (5): 573–581.

10

Refuting the Vieto-Katuic Hypothesis: Reconsidering Ethnohistorical Linguistic Scenarios

Mark Alves

1 The claim of a Vieto-Katuic branch of Austroasiatic

The Vieto-Katuic hypothesis is the claim that the Vietic and Katuic branches of Austroasiatic exclusively share a higher node within Austroasiatic. This hypothesis has raised ethnohistorical questions, especially with relevance to Vietnamese, a national language of a country with roughly 100 million people. Further, it touches on major questions in archaeology of the Neolithic agricultural transition. Thus, this historical linguistic claim deserves careful re-evaluation.

Diffloth (1991b) first proposed this grouping (what he called "Proto-Katuic-Vietic") primarily based on phonological evidence, while also positing that vocabulary was shared by Vietic and Katuic, though he did not provide supporting lexical data. Later, the current author (Alves 2005) proposed a few dozen lexical isoglosses shared by Vietic and Katuic. Other scholars accepted the Vieto-Katuic hypothesis in historical linguistics publications, as to be discussed in Section 2, and it has been tied to speculation about early migrations of ancestors of the Vietnamese. The hypothesis has even been noted in a chapter on prehistory in Vietnam in an English-language historical text (Kiernan's 2017 "Việt Nam: A History of the Earliest Times to the Present"), with the idea that this early group migrated northward from the Annamite Cordillera in north-central Vietnam and bordering parts of Laos.

However, (a) the historical phonological data seemed persuasive but is minimal, (b) the lexical data was more limited than that available today and was collated from paper texts, not digital sources via digital tools effective for sifting data, and (c) no archaeological evidence was presented or indeed available to support Diffloth's assertion. In view of current available data—much more than just 20 years ago—with an overview of the phonological, morphological, lexical, and ethnohistorical aspects and weighing the evidence, the Vieto-Katuic hypothesis can no longer be considered valid.

However, the goal of this article is not only to demonstrate how current data and methods show that the Vieto-Katuic hypothesis does not hold water. It is also aimed at pointing out that archaeological evidence strongly suggests a north-to-south movement of Austroasiatic speakers into Mainland Southeast Asia (MSEA hereafter), and that the result of this event significantly complicates claims of a south-to-north migration of early Vietic peoples, or a homeland near central Vietnam, a claim lacking archaeological support. Also, the linguistic evidence shows that Katuic and Vietic are distinct branches within Austroasiatic and that they share features likely due to language contact with each other at various times over history, and not necessarily with

substantial time depth.

The rest of this article presents previous studies addressing the Vieto-Katuic hypothesis, reviews factors by which the hypothesis can be re-evaluated (i.e., historical phonology, lexical data, morphology/word-formation, issues of language group homelands, and ethnohistorical matters), and summarizes problems with the Vieto-Katuic hypothesis and presents a new scenario suggested by additional data.

2 Previous Research on Vieto-Katuic and the Vietic Homeland

The quest to determine the internal classification of the Austroasiatic language family has spanned a century since the late colonial era in MSEA and neighboring areas (see Sidwell 2021b for a summary). The nine branches established in Thomas and Headley (1970) have expanded to 14 (Sidwell 2021a: 179). Proposals for higher-node groupings of branches have also fluctuated. Diffloth (1974) proposed that Mon-Khmer and Munda formed two major sub-branches, and he further divided the Mon-Khmer group into Northern, Southern, and Eastern branches. Later, he posited a more complex phylogenetic branching, with Vieto-Katuic under a major Khmero-Vietic node including Vietic, Katuic, Bahnaric, and Khmeric (Diffloth 2005a: 79).

After Diffloth's 1991b publication, a number of researchers considered the hypothesis, adding discussion and support for it. Nguyễn T. C. (1995: 13), in writing about Vietic historical phonology, supported a Vietic-Katuic group (what he called in Vietnamese *khối Proto Việt-Katu*) and noted a shared [ʔa-] presyllable on two words for animals. Alves (2005) presented what he considered to be some 40 lexical isoglosses connecting Vietic and Katuic, though Sidwell (2015) pointed out they were mostly either general Austroasiatic words or possible inter-branch loanwords. Nevertheless, Sidwell (2018) has presented a phylogenetic tree of Austroasiatic based on Bayesian methodology that grouped Vietic and Katuic, though by being focused solely on lexical data, this could indicate either language contact or common origin. Chamberlain (2019: 1590) supports the Vieto-Katuic hypothesis, assuming a homeland in north-central Vietnam based on the amount of linguistic diversity of Vietic groups in that region (1998: 109). He made further ethnohistorical assumptions that the Proto-Vietic group was made up of hunter-gatherers and that this group eventually migrated northward into the RRD to take over territory of Tai-speaking groups (Chamberlain 1998: 37).[1]

Indeed, this idea of south-to-north migration of Vietic is a central part of the narrative underlying the Vieto-Katuic hypothesis. In related ideas, Ferlus (1996: 21) considered Vietic (using the term "Viet-Muong", now considered a sub-branch of Vietic) to be close to Katuic, though still seeming to regard them as distinct branches. However, regarding the Vietic homeland, he too considered it to be near central Vietnam, just north of Katuic territory in the Upper Middle Mekong Basin, and he further posited that geographic expansion was connected to the introduction of bronze metallurgy and rice cultivation (Ferlus Ibid.). Unlike Chamberlain, who posited Vietic replacement of Tai speakers in the RRD, Ferlus tentatively speculated a scenario in which Vietic peoples migrated from central Vietnam to Nghe An province, and

[1] The historical interpretations in Chamberlain's work are not considered here. Readers may read Kelley (2017), who reviews Chamberlain's 1998 paper and suggests the historical claims are unsubstantiated, pointing out gaps and alternative explanations for Chamberlain's interpretation of Chinese historical texts.

eventually to the RRD to replace a former group of Austroasiatic speakers of an unknown speech variety. However, in neither scenario—replacement of a Tai-speaking community or Austroasiatic-speaking community—is there any supporting archaeological evidence, and as shown in Section 4, 21st-century archaeological studies run counter to aspects of these claims.

The Vieto-Katuic hypothesis has even become part of descriptions in historical texts. The historian Kiernan (2017: 46) refers to Diffloth (2005a) in writing, "Proto-Vietic diverged first from Khmeric languages around 2500 BCE, then from Katuic around 1000 BCE". Kiernan also describes a south-to-north migration of Proto-Vietic, from Nghe An, Quang Binh, and bordering Laos. The reference is a publication of Chamberlain (1998: 22-23), who cited an unpublished presentation of Diffloth (1991a).

Unfortunately, the historical linguistic methodology to obtain these dates has not been explained, nor has supporting archaeological evidence been presented in previous studies. The center-of-diversity hypothesis seems to be the primary justification for considering the central or north-central Vietnam origin, but as will be shown in Section 3.4, this approach is insufficient to overcome plentiful archaeological evidence of a north-to-south movement of Austroasiatic peoples.

Not all publications have been supportive of the Vieto-Katuic hypothesis. Sidwell (2015: 175) considers previous studies and labels them "ambiguous support." More recently, Sidwell (2021: 198-199) and Alves (2020, 2022) have presented further data and argumentation against it, with Sidwell questioning the naturalness of the sound changes Diffloth proposed (§ 3.1), and both presenting problems with previously proposed lexical isoglosses (§ 3.2). Section 3 furthers the re-evaluation of various linguistic aspects.

3 Factors in the Re-evaluation

Re-evaluating a decades-old hypothesis requires substantive data and careful methodology, especially when the result is refutation and replacement of such a hypothesis. The following aspects are considered in following subsections.

3.1 Historical phonology
3.2 Lexical data
3.3 Morphology and word-formation
4 Theories of language families' homelands and dispersals
5 Ethnohistorical linguistic data suggesting a northern Vietnam origin

While some of these matters overlap (e.g., lexical cognates and phonological features, "homeland" hypotheses and ethnohistory, etc.), each aspect is considered separately as much as possible.

3.1 Historical phonology

As noted, shared phonological innovations were posited as an indicator of a common Proto-Vieto-Katuic proto-level branch. However, despite past claims, a re-evaluation of relevant data shows there is not only no such evidence but also confounding evidence in the form of apparent innovations shared by Vietic and the Mangic and Pakanic languages to the north of Vietic (Gehrmann 2021).

Diffloth (1991b) posited that the reflexes of pAA onset **ʔ-* within a specific subset of words are Proto-Katuic **h-* and Proto-Vietic **s-* and that this constitutes a shared phonological innovation, and he provided several words in support of this claim.

Sidwell has been skeptical of this proposal (Sidwell 2005b: 196), and more recently (Sidwell 2015: 176), he questions this situation in terms of the naturalness and conditioning factors of the change. Also, we can consider Sidwell's recent (2024) reconstructions of Proto-Austroasiatic, based on a thorough re-assessment of comparative data (see Sidwell & Alves (2023) for a summary). These reconstructions show that what Diffloth reconstructed as a glottal stop in Proto-Austroasiatic was instead two distinct sounds, namely, *ʔ and *ɕ, as shown in Table 1. If Sidwell's reconstructions are valid, then Proto-Vietic *s and Proto-Katuic *h are from *ɕ ('blood', 'fart (v)', and 'breathe (v)'), what can be considered natural change from a typologically less common fricative sound to more common ones. Note that Proto-Bahnaric, like Proto-Katuic, has *h. In the other instances ('bone', 'centipede', and 'cough (v)') with a glottal stop, there is no shared pattern in Vietic and Katuic.

Table 1: Proto-Austroasiatic *ɕ and *ʔ and reflexes in Proto-Vietic and Proto-Katuic

Gloss	**Proto-Austroasiatic**	**Proto-Vietic**	**Proto-Katuic**	**Proto-Bahnaric**
blood	***ɕ**aːm, *mə.'**ɕ**aːm	*ʔa-**s**aːmʔ	*ʔa**h**aːm	*bhaːm
fart (v)	*pə.**ɕ**uːm	*k.**s**amʔ	NA	*phoːm
breathe (v)	***ɕ**əːmˀ, ***ɕ**iːmˀ	NA (Thavung pa**s**əːm³)	*pa**h**əːm	*ɟhəːm
bone	*cə.'**ʔ**aːŋ	***tʃ-ʔ**aːŋ > ɟaːŋ / tʃɨəŋ	*ʔŋ**h**aːŋ	NA / *ktsɨːŋ
centipede	*kə.'**ʔ**iːpˀ	*kr.**s**iːp	*ka**h**eːp	*kʔeːp
cough (v)	*kə.'**ʔ**ɔkˀ	NA	*ʔŋhɔk, *khɔk	*-ʔɔk

In contrast to this lack of shared phonological innovation, in recent exploration of Proto-Austroasiatic rime glottalization,[2] Gehrmann (2021) shows that Vietic rime glottalization patterns with the Mangic and Pakanic languages spoken in northern Vietnam and parts of southern China, as in Table 2. While Katuic exhibits what Gehrmann terms a "deglottalized pattern", along with branches such as Bahnaric, Khmeric, Munda, Khasian, and others, Vietic and the languages Mang, Bugan, and Bolyu constitute a "northeastern pattern" with a four-way distinction that developed from the original two-way distinction in Proto-Austroasiatic.

Table 2: Rime glottalization & the northeastern pattern
(Adapted from Gehrman 2021: Slide 9)

pAA	**Conservative**	**Deglottalized Pattern**	**Northeastern Pattern**
*-ʔ	*-ʔ	*-ø	*-ʔ *-ø
*-N	*-N	*-N	*-N *-Nˀ

[2] Diffloth (2005b) explored this in Vietic and Katuic, coming to the conclusion that it was an ancient feature in Vietic but later development in Katuic (Ibid.: 92). However, crucially, in 140 words, he finds agreement in Vietic and Katuic in rime glottalization in half of the items, and disagreement in the other half, with no apparent pattern of explanation.

Austroasiatic branches in the three types
Conservative: Palaungic, Khmuic, Monic, Aslian
Deglottalized: Katuic, Nicobaric, Munda, Khasian, Khmer, Pearic, Bahnaric
Northeastern: Vietic, Mangic/Pakanic (Mang, Bugan, Bolyu)

Though the claim of a common phonological innovation in Vietic and the Mangic and Pakanic branches is supported by comparative evidence,[3] it must be further evaluated and tested. Nevertheless, for now, this pattern serves as confounding evidence to the hypothesis of a Vieto-Katuic proto-language. And again, supporting phonological evidence connecting Vietic and Katuic exclusively is lacking.

3.2 Lexical data

In addition to phonological innovations, shared lexical innovations can be considered in the question of a common proto-language origin of languages. Vietic and Katuic naturally share many Proto-Austroasiatic etyma (e.g., 'fish (n)', 'root (n)', 'weave/plait (v),' etc.), some of which are found in Munda languages in India and Nicobaric languages in the Nicobarese islands. Such a wide geographic range with corresponding time depth presents strong evidence of their common Austroasiatic origin.

As mentioned, Diffloth (1991b: 136) posited that lexical evidence further supported a Vietic-Katuic connection, though he did not provide examples. Alves (2005) presented a few dozen words that he posited were shared exclusively by Vietic and Katuic. However, Sidwell (2015: 175) noted that many items were either found in other Austroasiatic languages or were possible loanwords.

Unlike Alves's 2005 study, using paper dictionaries, glossaries, and such, and only through visual scanning to assemble comparable word forms, this current study 20 years after that publication utilizes digital databases and other digitized resources that have come to be available since then. In addition, the databases have much more data from subsequent fieldwork and many new reconstructions of language groups, and digitized data is much more efficiently and effectively sifted and analyzed. The standard expectation now is that Austroasiatic etymological research involves:

- The *Mon-Khmer Etymological Dictionary* (and the related *Munda Etymological Dictionary*)
- Digitized dictionaries of both modern Southeast Asian languages and ancient textual sources (e.g., modern and Old Khmer, modern Vietnamese and early Vietnamese Nôm texts, etc.)
- Available reconstructions of all language families and many subgroups in the region
- Other databases and digitized lexical resources for other language families in the region (e.g., *Proto-Tai-o-Matic*, the *Sino-Tibetan Etymological Dictionary*, the *Xiaoxuetang* Chinese dialect database, etc.).

In addition, I have been collaborating with Paul Sidwell, who is spearheading a project of updated Proto-Austroasiatic reconstructions based on such data, in reassessing Shorto's (2006) Mon-Khmer Comparative Dictionary (see Sidwell and Alves 2023). This project includes the identification of attestations (or exclusions) of Shorto's items among Austroasiatic branches. As for Vietic, I have updated, reassessed, and added data

[3] Gehrmann (2023) provides a complete spreadsheet of the comparative data.

for a dozen lects to the existing dozen lects used in Ferlus's (2007) Proto-Vietic reconstructions.[4]

The result of the current reassessment is that there are at most 14 isoglosses in Proto-Vietic and Proto-Katuic, as in Table 3. This number is much smaller than the 40-plus items in Alves's 2005 study, and none of them belong to the original set. In Table 3, all the forms have comparable or the same Proto-Vietic and Proto-Katuic reconstructed forms, and—according to available database checking—comparable forms do not occur in other Austroasiatic languages or other language families. Also in Table 3, I provide my own Proto-Vietic reconstructions in parentheses when I find additional comparative Vietic data to warrant these. In some cases, the words are attested in a few Bahnaric languages, but such words are not widespread in Bahnaric, suggesting borrowing from Katuic and/or Vietic (noted in Table 3 with "likely >" and the Bahnaric forms).[5] Finally, while additional shared words may be found in, for example, Vietnamese and a single Katuic language, such words are likely loanwords in periods long after the proto-branch periods and thus cannot be used for this study.

Table 3: Isoglosses for Proto-Vietic and Proto-Katuic[6]

Glosses	**Proto-Vietic**	**Proto-Katuic**
float on water	*ɗoːs (*ɗoːl)	*ɗol 'float'
beat/thresh rice	*pɔh	*pɔh 'to beat (and 'thresh' in some languages)
wrap	*duːm	*duom 'wrap, cover' (possibly > Tarieng *toːm* 'to wrap (with cloth or leaves)')
breathe	*t-ŋəs	*tŋɨh
drip	*k-ʃɔh	*tʃoh
younger sibling	*ʔɛːm	*ʔaʔɛːm
feces/defecate[7]	*ʔɛh	*ʔɛh
gums	*ləɲ	*laɲ, *klaɲ (likely > Alak, Tarieng, Jeh, Sre)
village	*k-veːr (*k.weːl)	*wiːl, *weːl
joint/internode	*tuːc (*tuːt)	*toːt, *ʔatoːt
leech, water	*l-dɛh / l-tɛh	*ʔadeh, *deh
crayfish/shrimp	*soːm	*ʔasuom
nest/hive	*s-ʔuh > suh / ʔuh	*soh 'nest'

4 The spreadsheet with data is still being processed but will eventually be shared.

5 Paul Sidwell (p.c.) notes that West Bahnaric has experienced substantial lexical borrowing from Katuic, yet another instance of the complex language situation in this area.

6 I originally included 'galangal' (Proto-Vietic *b-riɛŋ (*C.riɛŋ) and Proto-Katuic *-rɨiŋ) and 'glutinous (of rice)' (Proto-Vietic *ɗeːp versus no Proto-Katuic form but widespread Katuic forms, cf. Pacoh *deep*, Katu *dep* 'glutinous rice', Bru *diip* 'sticky rice', Kui *diip* 'sticky, glutin'). However, for galangal, the spread of ginger in food is likely much later than a proto-language stage, and for 'glutinous', that may also have a more recent history. Thus, I do not consider these valid comparable proto-language forms back to the era in question.

7 The same form has been reconstructed by Thurgood (1999: 310) in Chamic, but he posits that this is a loanword into Chamic considering the vowel. For now, I regard this as a possible loanword from either Katuic or Vietic into Chamic, though the direction of borrowing of this noncultural word cannot be determined with certainty.

Glosses	Proto-Vietic	Proto-Katuic
inclined to, on the side of[8]	*s-gɛːŋ (*C.gɛːŋ)	Proto-Katuic *-gɛːŋ ‘lean to one side’

The data in Table 3 suggest that, most likely, these are isoglosses of Proto-Vietic and Proto-Katuic. However, the number of items is not large, and while the words are mostly not trade or cultural items, few of these can be considered truly basic vocabulary. Such words are typically Proto-Austroasiatic etyma in one or both branches. While the words are mostly not trade or cultural items, few of these can be considered truly basic vocabulary. Such words are typically Proto-Austroasiatic etyma in one or both branches. Sharing proto-forms for ‘gums’, ‘feces’, and ‘younger sibling’ is not insignificant. However, lacking evidence of shared phonological innovations, the significance of these comparanda is reduced. And a question is whether this small range of shared lexical items is sufficient alone to establish a common proto-language origin. In light of this data, we can consider the following possibilities to account for Vietic and Katuic lexical isoglosses.

- Chance similarity
- Retentions from Proto-Austroasiatic only in these two branches
- A common Vieto-Katuic proto-language stage
- Innovations in an ancient period of one branch shared with the other

When words in two neighboring groups share both phonological and semantic features, they are less likely to be the result of chance similarity, so the items in Table 3 are likely related words, but of what nature? Another possibility is that these are retentions from Proto-Austroasiatic but lost in all other branches except these two, but as such a claim can be neither proven nor disproven, that is not a usable hypothesis. And again, there is a notable lack of shared phonological innovations to support a Vieto-Katuic branch.

Instead, we are left with the possibility that these are lexical innovations in one or the other branch which were subsequently shared at some past time, though it is not possible to determine how early (e.g., before or in the Common Era). These branches may have been neighbors for thousands of years: there was plenty of time before the Common Era for this lexical exchange to have occurred. And as will be shown in Section 3.3, morphological data highlights language contact that has occurred. As to the direction of borrowing, that is difficult or impossible to determine, but borrowing may have occurred in both directions in bilingual situations.

[8] Also, see Thai *ta˥˩kʰɛɛŋ* ‘to tilt to one side, lean on one side’. It is unclear whether this is a related form or how it might have spread.

Table 4: Isoglosses for Proto-Vietic and Proto-Bahnaric[9]

Glosses	**Proto-Vietic**	**Bahnaric Reconstructions**
person, human	*ŋaːj	Proto-Bahnaric *bŋaːj
lump	NA (*kok)	Proto-Bahnaric *koʔ ‘lump (classifier)’
hedgehog/porcupine	*k-ɲiːmʔ	Proto-West-Bahnaric *kɲeːm ‘porcupine’[10]
give birth/lay/be born	*tɛh	Proto-South-Bahnaric *dɛh

Further confounding the situation is that, as I found in sifting the lexical data, four items are shared by only Vietic and Bahnaric (albeit also including some sub-branch reconstructions), as in Table 4. The number of words shared by Vietic and Bahnaric is smaller than by Vietic and Katuic, but assuming relative stability over time, neighboring branches are likely to share a larger number of words than non-adjacent ones. Sidwell (2021: 199) has noted high lexicostatistical numbers connecting Bahnaric and Katuic. Regardless, two branches of Austroasiatic can share isoglosses but not belong to a higher phylogenetic node, and borrowing in an early period must be considered one likely reason for the shared words.

In addition, some terms for numerals are shared by Vietic and Bahnaric, while Katuic has its own set. Proto-Austroasiatic etyma for ‘1’ to ‘4’ are shared by Vietic, Katuic, Bahnaric, and all branches of Austroasiatic (see Sidwell 2012), as shown in Table 5 (with Proto-Austroasiatic reconstructions here and other tables from Sidwell 2024). However, while Bahnaric and Vietic share etyma for terms for ‘5’ to ‘9’ (with various cognates in Monic, Aslian, Munda, and Mangic and Pakanic), Katuic has distinct etyma for these. The Katuic forms are not only in Katuic (e.g., ‘6’ is also found in Palaungic, and ‘8’ is also found in Munda), meaning these are not the result of innovations solely in Katuic, but rather have deep history in Austroasiatic.[11] Speculation about a scenario of Vietic and Katuic speciation, followed by their lexical innovations, and then subsequent adoption of different numeral terms above 4 would be difficult to support. Also, the cognates for ‘6’ to ‘9’ in Bolyu of the Pakanic branch are notable considering the shared phonological innovation noted in Section 3.1. Overall, though number systems can innovate, this data runs counter to a branch of Vietic and Katuic.

9 Originally, in this table, I included Vietic *p-naːŋ ‘arecanut/areca’ and Proto-North-Bahnaric *tnaːŋ ‘areca, betel’, but Paul Sidwell (p.c.) pointed out the comparable Malay *pinang* ‘areca,’ and Thurgood (1999: 300) reconstructs *pinaːŋ ‘betel (areca palm); betel-nut’. Thus, this may be a Malayo-Chamic word spreading into both Bahnaric and conservative Vietic languages (not Vietnamese), again highlighting language contact and exchange in this area.

10 It is found in the Katuic language Souei *kəɲɛɛm* ‘porcupine’, a likely loanword from Katuic. Vietic did not expand that far south until the 15th century Nam Tiến after the fall of the Champa empire.

11 In Austroasiatic, numeral terms above ‘4’ vary among the branches such that a historical explanation of their distribution is not yet possible. Nevertheless, the cognates of these words among branches in different regions, including even India, show that they are indeed ancient in Austroasiatic.

Table 5: Reconstructed numeral terms among Austroasiatic branches

Numbers	Proto-Austroasiatic	Proto-Vietic	Proto-Bahnaric	Proto-Katuic	Bolyu (Pakanic)
1	*muəjʔ, *moːjʔ	*moːc	*muəj	*muoj	mə[33]
2	*ɓaːr	*haːr[12]	*baːr	*ɓaːr	mbi[55]
3	*peːʔ	*paː	*pɛː	*pɛː	paːi[55]
4	*puənʔ	*poːnʔ	*puən	*puan	puːn[53]
5	MSEA *pɗamʔ	*ɗam	*pɗam	*səəŋ	(me[31])
6	*təˈruʔ, *pəˈruʔ	*pʰruːʔ kʰluːʔ	*t(n)raw	*tbat	pju[53]
7 7	*pɔh, *pəɛ *təˈpuːl, *təˈpuəl	*pəs NA	*tpəh NA	NA *tboːl	pei[55]
8	*təN.ɕaːm	*saːmʔ	*t(n)haːm	*tgɔːl	saːm[53]
9	MSEA *ciːnʔ	*ciːnʔ	*cin	*tgias	ɕən[53]
10	NA	*maːl, *ɟuːk	*ɟɨt	*ɟit	maːn[13]

Finally, a number of words are attested in Vietic, Katuic, Bahnaric, and sometimes other nearby language groups, such as Khmer and Khmuic, as in Table 6. While these could hypothetically be retentions from Proto-Austroasiatic, such a situation is not demonstrable or refutable. Instead, in cases of shared forms in neighboring communities, they more likely represent innovations in one branch that spread by contact to others in this geographically contained region of north to central Vietnam. This further highlights the results of long-term language contact among the branches, such that words for 'bear,'[13] 'thunder,' 'forest,' basic verbs, and other noncultural words can be and have been shared among the branches.

Table 6: Proto-language forms shared by Vietic, Katuic, Bahnaric, and other neighboring groups

Gloss	Proto-Vietic	Comparative data and notes
inclined to, on the side of[14]	*s-gɛːŋ (*C.gɛːŋ)	• Proto-Katuic *-gɛːŋ 'lean to one side' • Proto-Bahnaric *keːɔŋ 'edge'
thunder[15]	*k-rəmʔ (*krəmʔ)	• Proto-Katuic *grɨm • Proto-Bahnaric *grɨm ~ *krɨm • Khmer *krum, krɔ̀əm, krɨm - krɨm* 'sound of thunder'

[12] The unexpected *h onset in the Vietic form creates a challenge in relating it to the pAA etymon. Nevertheless, considering the robustness of retention of these basic numeral terms, it seems unlikely to have been a random lexical innovation that rhymes perfectly with an existing form, so I consider it to be related. The labial onset in pAA could be the result of alliteration with '3' and '4', but the question remains unanswered for now.

[13] The item meaning 'bear (n)' is particularly notable as currently, there is no Proto-Austroasiatic etymon for 'bear', making this perhaps the best possibility of a retention from Proto-Austroasiatic in a restricted geographic region. Otherwise, this would suggest the spread of this word from one of the branches and widespread replacement of earlier words.

[14] Also, see Thai *taˈkʰɛɛŋ* 'to tilt to one side, lean on one side'. If related, this disyllabic form is a possible instance of borrowing into Thai, though this word form is admittedly of uncertain origin.

[15] Thurgood (1999: 356) reconstructs this in Proto-Chamic but considers it a loanword from neighboring Austroasiatic languages.

tooth, fang, canine, tusk	*k-nɛːŋ	• Proto-Katuic *knɛɛŋ ‘tooth’ • Proto-South-Bahnaric *gniəŋ ~ *gneːŋ ‘tusk, canine tooth’; Proto-Central-Bahnaric *gniəŋ; Proto-North-Bahnaric *griaŋ ‘fang, canine, tusk (of boar)’; Proto-West-Bahnaric *kniaŋ ‘tusk, fang’
bear (n)	*c-guːʔ / c-kuːʔ	• Proto-Katuic *hŋkaw • Proto-Bahnaric *ckaw ~ *gaw
forest	*k-rəŋ (*krəŋ)	• Proto-Katuic *kruuŋ ‘forest’; *krɨŋ, *crɨŋ ‘virgin forest’ • Bahnaric (Jeh & Tarieng krɯŋ ‘jungle’)
split/cleave (wood) (v)	*bah / pah	• Proto-Katuic *paːh ‘split (v)’; • Proto-Bahnaric *pah ‘split, crack’; • Monic (Nyah Kur (several lects) *páh* ‘split, cut, hew’)
crow (cock) (v)	*t-karʔ	• Proto-Katuic *takar ‘crow (v)’ • Bahnaric NA (Jeh *təkar*, Alak *kakar*, Tarieng *kar*)
pomelo	*paːs	• Katuic (Ngeq *peːh*; Pacoh *piəs*) • Baharic (Tarieng *pleː piəs* / *piəç*; Jeh *plɤj piəh*)
spoon (n)	NA (*ɓuaŋ)	• Katuic (Bru *buəŋ*, Kui *buːŋ*) • Bahnaric (Tampuan *buəŋ*, Halang *ɓu̯əŋ*) • Khmuic (Khsing-Mul *buəŋ*)
go out	*ʔa-loh	• Proto-Katuic *lɔh • Proto-Bahnaric *lɔh (Possibly Katuic > Southwest Vietic)
dirty	NA (Vietnamese *nhớp*)	• Katuic (Katu, Bru *ɲəːp*) • Bahnaric (Bahnar *ʔɲəʔɲəp-ʔɲəʔɲap*, Mnong *ɲəʔ*)
king, lord, ruler	NA (*t.puə)	• Bahnaric (Sre *poa* ‘chief of a village’) • Katuic (Pacoh *vua*) • Mangic (Mang *pua*[2]) (Likely a Vietnamese loanword into all)

Proposed phonological correspondences for a Vieto-Katuic group are no longer valid, and the shared vocabulary is not enough to support phylogenetic status. Furthermore, the data shows significant long-term contact among the branches in this region, such that the shared proto-branch forms in Vietic and Katuic can be considered the result of ancient language contact.

3.3 Morphology and word-formation

This section presents word-formation features found in a few Katuic and conservative Vietic languages. These include (a) a proto-language *ʔa- presyllable on a handful of pAA words, (b) two case-marking presyllables occurring on pronouns, and (c) an instance of grammaticalization of plural pronouns used as generalized plural markers. However, none of these appear to have significant time depth and are probable later-stage innovations that spread from a Katuic language to conservative Vietic languages.

The shared *ʔa- presyllable was noted by Nguyễn T. C. (1995: 236). He suggested that Vieto-Katuic had presyllables added to some Proto-Austroasiatic words, offering examples of ‘fish’ and ‘dog’. In addition, available reconstructions show this presyllable on Proto-Vietic reconstructions for ‘bird’ and ‘hair’, which are also Proto-Austroasiatic etyma (but neither are reconstructed in Katuic with presyllables), and also ‘tortoise’ and ‘elephant’, which are not Austroasiatic in origin. ‘Elephant’ is a loanword,

possibly from Tai,[16] to which the presyllable has been added. Of course, the presyllables in Vietic used to reconstruct these are found solely in the conservative southern and western languages/lects, those which have retained disyllabic words. These items are presented in Table 7.

Table 7: Proto-Vietic reconstructions with an *ʔa- presyllable

Gloss	Proto-Vietic	Proto-Austroasiatic	Proto-Katuic	Vietnamese
dog	*ʔa-cɔːʔ	*cɔʔ ‘dog’	*ʔacɔː	chó
fish (n)	*ʔa-kaːʔ	*kaʔ ‘fish’	*ʔakaː	cá
tortoise/turtle	*ʔa-rɔː	NR[17]	NR (*ʔakɔːp ‘turtle’)	rùa
elephant	*ʔa-ɟaːŋ	NR	*ʔaciaŋ	NA (voi)
bird	*-ciːm	*ciːm ‘bird’	*ceːm	chim
hair	*-suk	*sukˀ, *sɔkˀ ‘hair’	*sok	tóc

However, other than this small number of items, no other Proto-Vietic reconstructions for animals have this presyllable. In contrast, in Proto-Katuic (Sidwell 2005a), 32 reconstructions for animal terms have an *ʔa- presyllable, and another 10 proto-forms have a nasal presyllable *ʔN-. If Proto-Vietic had this word-formation strategy, we should expect more than a few shared forms. Also, if the word for ‘elephant’ is a Tai loanword, it should date to the 2nd millennium CE, and the [ʔa-] presyllable must have been added then: this raises the possibility that these words with the [ʔa-] presyllable were borrowed at some time in the last several centuries. Finally, were these presyllables in Proto-Vietic, the onsets in Vietnamese should have lenited to voiced fricatives, such as intervocalic *k to modern Vietnamese ‘g’ [ɣ], but they have not (see Alves 2024 for an overview). This situation shows evidence of a few lexical borrowings, not retention of a proto-language word-formation strategy.

As for presyllables added to pronouns, a complete system of three types of presyllables is found on all pronouns in Pacoh of Katuic, including [ʔa-] marking accusative/dative case and a homorganic nasal [ʔN-] genitive case (Watson 1964). In Vietic, only some instances are described in Ruc, a conservative Vietic language. Nguyễn V. L. (1993:97) notes that a presyllable [pa-] can be added to Ruc pronouns to mark a beneficiary, as in 1. A sample of an [ʔa-] prefix on a pronoun in Pacoh is shown in 2. These examples show the pronouns have comparable dative senses, though the presyllables are not the same, and the syntactic positions in these two instances are different (i.e., before the direct object in 1 but after in 2).

16 The history is more complex considering the possible Chinese source (象 xiàng ‘elephant’, OC *s-daŋʔ, MC zjangX) of Proto-Tai *ɟaːŋ^{C}. Regardless, with a level tone in the Chut languages, this is a later loan, likely in the 2nd millennium CE, not one in the pre-tonogenesis period in the 1st millennium.

17 The Proto-Kherwarian reconstruction *hɔrɔ ‘tortoise’ is comparable, but the comparable data in only Vietic and possibly in Munda is not compelling enough to warrant a claim of Proto-Austroasiatic status, unless additional clarifying data is found.

1. Ruc (Nguyễn V. L. 1993: 97)
   ho:[1] muə[1] **paˈmi:[2]** kɛ:[4] ʔaˈɲɛ:l[1]
   1S buy to-2S CLSF knife
   'I bought a knife for you.'[18]

2. Pacoh (Alves 2006: 67)
   kɨ: ɟo̤:n pe:ʔ **ʔaˈdɔ:**
   1S give banana to-3S
   'I gave him the banana.'

A presyllable with the same shape as in Pacoh is the Ruc presyllable [ʔa-], which is added to pronouns in sentence-initial position but with a kind of possessive-existential sense, as in 3. This usage is parallel to that of the possessive-existential nasal presyllable in Pacoh, as in 4. Again, the phonological form does not match, but adding a presyllable to encode this semantic function is noteworthy as it is not in other Austroasiatic languages.

3. Ruc (Nguyễn V. L. 1993: 97)
   ʔaˈho:[1] ko:[3] suək[3]‿ɟu:p[3]
   LCV-1S have tobacco
   'I have tobacco (on my person).'

4. Pacoh (Alves 2006: 67)
   ʔŋ. ˈkɨ: vi: praʔ
   of-1S exist money
   'Of that which is mine, there is money.'

Finally, there is an instance of shared grammaticalization, one I have not found outside of Katuic and conservative Vietic languages. Nguyễn V. L. notes that, in Ruc, the third-person plural noun can be used to indicate plurality on a following noun, as in 5. This is parallel to the function of the third-person plural in Pacoh, as in 6.[19]

5. Ruc (Nguyễn V. L. 1993: 97)
   ho:[1] βaŋ[3] kɤəm[4] **ʔa.ˈpa:[1]** pu.ˈco:j[3]
   1S NEG meet 3P child
   'I did not meet children.'

[18] The original text samples were in a Vietnamese-orthographic system. I have converted these to IPA for the convenience of an international audience. The superscript numbers mark tones as described by Nguyễn V. L. (1993).

[19] Pacoh pronouns have developed secondary grammatical functions encoding four grammatical functions on neighboring nouns: plurality, dative, genitive, and conjunctive (Alves 2007). Other than the plurality-marking function in Ruc, I have seen no evidence of these other functions in descriptions of Ruc, or the closely related May language.

6. Pacoh (Alves 2007b: 8)

ʔa.ˈpɛː	ʔa.ˈʔɛːm	cɔːm	ləjʔ
3P	youngster	know	NEG

'Do you three young ones understand?'

However, none of the presyllables/prefixes in 1 through 6 in Ruc and Pacoh can be reconstructed to Proto-Katuic or Proto-Vietic as they are not widespread in either branch. Also, as noted, they are not phonologically related other than being presyllables with generic shapes. In Vietic, I have found this only in the Ruc language, while in Katuic, there is more but still limited comparable evidence.[20] Their developments likely happened long after a proto-language stage. I do not think this data indicates coincidental similarity and independent developments, considering the proximity of these languages and the lack of this pattern elsewhere in Mainland Southeast Asia in my knowledge. These may be grammatical borrowing in Ruc through contact with Katuic, though the situations and timing of this cannot be determined.

Overall, these shared features are interesting instances of regional typological convergence likely due to language contact between Katuic and primarily conservative minority Vietic groups, many of which have populations only in the hundreds.

4 Language family homelands and dispersals

We now consider historical linguistic methods that involve issues of ethnohistory. One goal in historical linguistics and studies of phylogenetics is to locate the original "homeland" of a language group. A more precise goal is to determine the geographic locus and timing of the spreading of ethnolinguistic group, after which daughter languages and language groups speciated.

Two common approaches in these queries include (a) the center-of-diversity hypothesis (i.e., the notion that areas with the most linguistic diversity of a language group are more likely to be the most ancient areas) and (b) the farming-language dispersal model (i.e., a situation in which agriculturalist communities settle, grow in size, migrate, settle again, thereby leading to wide areas of groups of related languages).[21] Both have been considered for Vietic language history in combination with the Vieto-Katuic hypothesis.

In the following subsections, these two approaches are re-evaluated within a broader historical linguistic context and additional extralinguistic data. Overall, considering the Austroasiatic origins of Vietic and Katuic, the farming-language dispersal is supported by archaeological data, while the center-of-diversity model is not.

---

[20] Spoken near Pacoh, the related Taoih language has been described as having a more complex case-marking system (Solntseva 1996), but these two are immediate neighbors, and language contact could account for these word-formation innovations. Available descriptions of Katu and Bru do not mention a system of prefixes on personal pronouns, though in Katu, there are nasal prefixes on various demonstratives (e.g., 'that', 'that (yonder)', 'a place', etc.) and interrogative pronouns (e.g., 'which', 'what', 'where,' etc.) (Nguyễn and Nguyễn 1998:89). Altogether, available data is insufficient to reconstruct such morphology in Proto-Katuic.

[21] Another more recent approach is Bayesian phylogeography (e.g., Wichmann and Rama 2021). This has not been applied to determine the Vietic homeland, though one recent study (Sidwell and Alves 2021) present a phylogenetic tree of Vietic using a Bayesian approach. That tree highlights the degree of linguistic diversity of the south and southwest areas, but as proposed in this study, that alone is insufficient to make claims of migration patterns: archaeological data suggests otherwise, and there are additional confounding linguistic factors to consider.

Other explanations can account for the diversity of Vietic languages to the south and west of northern Vietnam.

4.1 Center of Diversity

To determine a Vietic homeland—with an assumed Vieto-Katuic connection—some claims have focused on the center of diversity of Vietic languages (e.g., Chamberlain 1998: 10). Prior to the 15th century, the southern extent of the political control of the Đại Việt kingdom was central Vietnam until the overthrow of the Champa empire in 1471, after which Vietnamese speakers spread southward over a period of centuries (see Nguyễn Đ. Đ. 2009). Thus, the original territory of Vietic languages was in the northern part of modern-day Vietnam. Within that region, the highest degree of linguistic diversity in terms of numbers of distinct Vietic languages is in north-central Vietnam and bordering areas of Laos along the Annamite Cordillera, the Trường Sơn mountain range, where the center-of-diversity approach can be considered.

However, how useful is the center-of-diversity hypothesis in reconstructing language history? In one study (Wichman et al. (2010)), computational analysis of 82 language families worldwide was used to determine their homelands based on lexicostatistical data (i.e., percentages of cognates of 100-word basic vocabulary lists) and geographic distribution of languages. However, from the beginning of their article, Wichmann et al. emphasize the importance of using archaeological and historical information and data, and they state, "Moreover, we stress once again that the approach can only serve as a tool that should ideally be supplemented with other tools for reconstructing homelands" (Ibid.: 247). Of the 82 language families in that article, a few neighboring language groups provide perspective.

Wichmann et al.'s research suggests a locus for Proto-Kra-Dai of southeastern China in modern-day Guangxi province. This matches ideas by specialists working on Kra-Dai language history (e.g., Ostapirat 2005, Bellwood 2021:38). The major branches of Kra-Dai are in southern China, in contrast with the southwest Tai branch, which is known for having more linguistic homogeneity due to a more recent (e.g., about one millennium)[22] expansion into Mainland Southeast Asia, making that a reasonable claim. However, Kra-Dai is a language family consisting of several dozen languages, not just a dozen, as is the case in the Vietic of Austroasiatic.

For Austroasiatic, Wichmann et al.'s data puts the center of diversity along the southern coast of Thailand. This is not so far from Diffloth's (2005a: 78) claim of "the fertile flood plains of the Irrawaddy in Burma and the plains along the lower Brahmaputra in Assam and Bangladesh". However, this location does not match any published hypotheses of the Austroasiatic locus of dispersal (see Sidwell 2015 for an overview). Also, as shown in Section 4, this idea runs counter to archaeological evidence and shows the limit of an approach focusing solely on linguistic data. The case of Sino-Tibetan (aka. Trans-Himalayan) offers another useful scenario. The homeland of Sino-Tibetan has been a matter of debate, with some considering it to be in or near the Himalayan region (e.g., Blench and Post 2014), and indeed, Wichmann et al. (2010) put its homeland in the Himalayas. However, Sagart et al. (2019) hypothesize that the origins of Sino-Tibetan also lies in north China, considering lexical evidence connected to early millet production in that region (a matter related to discussion in Section 4.2 on

[22] There are differing ideas about the timing of the Daic expansion. Pittayaporn (2014) suggests a timing of the 8th to 10th centuries. However, the first Tai kingdom, Sukothai, was not established until the 1200s.

farmer-language dispersals). I will not assess the validity of Sagart et al.'s hypothesis, but it presents a claim based on linguistic and archaeological data that differs dramatically from the center of linguistic diversity of Sino-Tibetan, with hundreds of languages to the west of Chinese-speaking territory.

As for the Sinitic branch of Sino-Tibetan, it is a clear counterexample. The center of diversity of Sinitic is in southern China: nine of ten branches of Sinitic (all of which are as distinct as languages), all with dozens of dialects, are concentrated in the southeast portion of China, mostly south of the Yangtze River. Based on the diversity model, this should indicate a southeastern source of the Sinitic dispersal. However, any history text can show that it was only in the Han dynasty (c. 200 BCE to 200 CE) at the beginning of the Common Era that Chinese began to have a presence there. The current diversity appears to be the result of early Sinitic groups settling and developing distinct regional speech varieties, sometimes due to language contact with then extant languages.

The diversity-as-center concept is one criterion to be considered together with others, and the theory has supporting instances. But the counterexamples and problematic claims shown above, especially those without archaeological data, show this approach cannot stand alone.

The center of diversity of Vieto-Katuic can be put aside as the linguistic evidence does not support it, so a claim that its homeland was in the Annamite Cordillera is a moot point. As for Vietic, the highest concentration of language diversity is indeed south of the RRD. Sidwell and Alves (2021) provide the most up-to-date phylogenetic analysis of Vietic. In it, there are several main branches: the Viet-Muong, Thavung-Kri-Malieng, Chut, Pong-Tuom, and Cuoi-Tho groups. The latter four groups consist of about a dozen languages in total, and they are to the south and west of the RRD region. However, unlike the case of Kra-Dai, with several major branches and several dozen languages in total, Vietic constitutes a much smaller sized group of languages, as well as much smaller populations of the conservative bisyllabic Vietic languages. It is worth considering whether or how the smaller number of languages and population size might affect the application of the center of diversity approach. In addition, we can consider that socially prominent state-level languages may levelled past diversity.

4.2 The Language Farming Dispersal Hypothesis

The other model to consider is the language farming dispersal hypothesis: the idea that language families spread and grow due to developments in agricultural practices (though other related sociocultural features should also be considered together with agricultural practices). For Vietic, Ferlus (1996:21-22) envisioned a homeland in the middle Mekong with later movement to central Vietnam in Nghe An and Ha Tinh provinces. He speculated that in that area, there was development of metallurgy and rice cultivation, and subsequent northward expansion, though he did not offer supporting archaeological evidence.

In the last two decades, the "Two-Layer" hypothesis (e.g., Matsumura et al. 2008, etc.) and the concept of the "Neolithic Revolution" (e.g., Higham 2021, etc.) involving a wave of agriculturalists migrating from the north to the south about four millennia ago have been gaining increasing support in archaeological literature. In addition, there has been support for the farming dispersal hypothesis, at least in its weaker form: the idea of a general tendency for large language families to involve communities of agriculturalists (Hammarström 2010).

In Mainland Southeast Asia, an unknown number of languages—previous hunter-

gatherer groups of the Hoabinhian cultural complex—appear to have mixed with and/or shifted to Austroasiatic languages, consisting of groups who brought agricultural practices. Proto-Austroasiatic lexical reconstructions have been shown to have a set of terms for rice production and processing (e.g., Diffloth 2005a, re-evaluated in Alves 2023) , in addition to relevant archaeological support (e.g., Higham 2017). Proto-Vietic also has a set of rice-production terms (e.g., Alves 2020). Altogether, the notion that the hypothesized Vieto-Katuic group expanded due to the subsequent development of rice-production practices is problematic: these were grain-producing groups from the time of the Austroasiatic expansion throughout MSEA.

This information demonstrates that both Vietic and Katuic peoples, like other Austroasiatic groups, were agriculturalists who moved from the north to the south (though the specific paths to reach those points is another matter, as discussed in Section 5). As for Vietic specifically, while there is no clear evidence against claims of a south-to-north expansion of an ethnolinguistic group in that area, there is neither archaeological evidence supporting such a claim nor a clear sociocultural impetus for such a migration: it becomes unsupported speculation. The possible north-to-south movement appears at least partly due to technological innovations as part of a cultural package, including rice and millet production, domesticated dogs and pigs, and housing structures (e.g., Alves 2023), indicated by both archaeological data and Proto-Austroasiatic lexical reconstructions.

5 Ethnohistorical linguistic data suggesting a northern origin

Though the number of sub-branches of Vietic is larger towards the south than the north in northern Vietnam, archaeological evidence suggests a north-to-south movement of Austroasiatic groups. However, what evidence might suggest a northern locus of Vietic specifically? And how then might the current geographic distribution of Vietic have come to be?

Alves (2022) presents historical and ethnohistorical and archaeological data to determine the language situation during the Dong Son culture in northern Vietnam (c. 500 BCE to 200 CE (Kim 2015: 106)). That is presumably much later than the speciation of Proto-Vietic, but the 2022 study presents evidence of a continuous Austroasiatic presence in northern Vietnam from the Austroasiatic dispersal. Archaeological studies show (a) a presence of agriculturalists in northern Vietnam about 2000 BCE, considered by archaeologists to be early Austroasiatic peoples (e.g., Bellwood 2005, Higham 2017, etc.), (b) a series of related archaeological traditions from the Neolithic Phung Nguyen culture (c. 2000 BCE to 1500 BCE) and then others leading to the developed Metal-Age Dong Son period in the RRD region (Kim 2015: 106), and (c) a north-to-south expansion into Mainland Southeast Asia, as discussed in Section 4.2.

After the initial dispersal, migrating groups could have gone in various directions, and I can find no archaeological evidence that clarifies paths or chronology of the dispersal. However, archaeological evidence of the cultural package of the probable Austroasiatic-speaking Neolithic agriculturalists is found in various parts of MSEA by the early 2nd millennium BCE. This indicates widespread distribution of early Austroasiatic ethnolinguistic groups within several centuries, subsequently mixing with and/or incorporating previous hunter-gatherer groups who spoke now extinct languages. In reconsidering the Austroasiatic dispersal, Sidwell (2022) explicitly posits that the

RRD in northern Vietnam is a feasible location for this, and he raises the possibility that Austroasiatic groups travelled either by river to the northwest along the Red River or south along coastal areas, and then later further inland into central MSEA. The location of the Nicobarese is unquestionably due to maritime travel, and Rau and Sidwell (2019) have provided arguments in favor of maritime travel of Munda groups to India. Travel by water does account for some possible higher-level phylogenetic nodes (e.g., a northern group versus a southern group).

If this is the case, one possibility is that Vietic has been in northern Vietnam since the Austroasiatic dispersal. Some small Vietic groups migrated into the highlands, but without any concrete evidence of timing, this could have happened any time in the last few thousand years. Katuic is a group that migrated directly southward, though whether south or southwest via rivers, land, or the coastline and then to highlands cannot be determined. Still, the approximate timing of that branch's migration should be comparable to that of other branches.

Another matter to consider is early Chinese loanwords in Vietnamese as they provide both chronological and geographical information. Vietnamese has hundreds of early Chinese loanwords (i.e., those borrowed in the 1st millennium CE from Late Old Chinese to Middle Chinese before the period of Sino-Vietnamese vocabulary connected to Late Middle Chinese near the beginning of the 2nd millennium). The phonological evidence dating these to the first half of the 1st millennium CE is substantial (see Alves 2018 for discussion and lists of such words). At least several dozen of these are from the end-stage of Old Chinese (e.g., Alves 2024 showing loans with voiced fricative onsets connected to Old Chinese presyllabic material). Early Chinese loanwords were possibly borrowed by Vietic groups in lowland areas where the first Chinese arrived and settled, such as the Co Loa archaeological site in the RRD, as well as areas near the coast in Thanh Hoa one hundred miles south and where many Han tombs with many Chinese objects have been excavated. Some lexical borrowing likely occurred during the Eastern Han dynasty in the first two centuries CE or even earlier, coinciding with the Dong Son period. Such loanwords include many cultural terms (e.g., 'gold', 'well (noun)', 'cauldron', 'pavilion', 'brocade', 'roof tile,' 'chopsticks,' 'age,' 'misfortune,' etc.) with a wide range of cultural domains that seem unlikely to have been borrowed by hunter-gatherers living in isolated hill areas.

In available Vietic data, only a small number of early Chinese loanwords are found among the conservative Vietic languages (e.g., 'sword', 'bandit,' 'bed,' etc.). There is no evidence of Chinese in the western highlands, so these words could have been borrowed via other Vietic languages. But another possibility is that ancestors of these Vietic groups borrowed such words in the lowlands where the Han Chinese were and only later migrated to the highlands. Finally, were Vietic languages only in the southern extent at that time, while still borrowing so many Chinese words, one should expect some of these Chinese loanwords to have been borrowed into neighboring Katuic, but that is not the case.

Based on this lexical and archaeological data, the language-contact picture in the Han Dynasty appears to be a community of Vietic-speaking Metal-Age agriculturalists in lowland areas of northern Vietnam encountering Chinese speakers in the Han dynasty. Bilingualism facilitated borrowing of a range of cultural domains of these early loanwords. These are the locations of proto-urban dwellings (e.g., Co Loa, as per Kim 2015), Han tombs, and other evidence of Han-dynasty material culture. The alternative—large Chinese-speaking communities in isolated highland areas near the border with Laos—is not supported by archaeological data.

6 Discussion

The lexical isoglosses in Proto-Vietic and Proto-Katuic highlight why various researchers have speculated about an affiliation between Vietic and Katuic, but a shared branch can no longer be part of that speculation. In the past, I supported a Vieto-Katuic group (Alves 2005), but I no longer see evidence for it, but rather counterevidence against it (i.e., lack of shared phonological innovations, archaeohistorical evidence of north-to-south migration of Austroasiatic, evidence of influence via language contact, etc.). At this point, we must regard Vietic and Katuic as distinct branches in Austroasiatic, both with an original northern origin, though ones that had language contact in the past. Hypotheses involving a group migrating southward, staying in the highlands, and migrating northward again to replace either Daic groups or Austroasiatic groups are not supported.

What is even less clear is how and when Katuic came to be in central Vietnam and bordering areas of Laos, but regardless, after they Katuic speakers settled there, at various points, they were in contact with Vietic groups, resulting in lexical exchange. This could have h0appened at anytime in the past three millennia, with time for multiple periods of contact leading to both loanwords and, in some cases, sharing of morphological features.

As for the small groups of conservative Vietic languages, it cannot be stated with certainty when they arrived in their current locations. Positing that their origin was originally in the highlands (and then where did they come from before that?) is speculation without archaeological evidence. They have retained a small number of very ancient Chinese loanwords, but none of the later layers of Chinese loanwords found in Vietnamese (and to a lesser extent in Muong lects). We can also speculate that the ancestors of the now isolated highland Vietic languages borrowed words such as ‘sword’ and ‘bandit’ due to the nature of their contact with the Chinese (e.g., mercenaries), but for reasons now unknown (e.g., military conflict, incoming groups, etc.), they migrated westward into higher areas, not unlike the Aslian groups of Malaysia. The range of sociocultural types that Chamberlain (1998) describes, from hunter-gatherer to agriculturalists, could be the result of living in isolated areas for centuries or longer. And again, such early Chinese loanwords have not been found in Katuic languages, which should be expected if Vietic and Katuic peoples were residing in the same region 2,000 years ago. Regarding the reduced amount of linguistic diversity in the lowland areas, one possibility is the expansion of Viet-Muong languages, especially Vietnamese, throughout lowland areas, leading to widespread language shift, while the isolated highland groups retained their typologies and continued on regular paths of diversification that happen naturally among languages. This is to be expected in a state-level sociopolitical speech community.

Admittedly, the ideas in the previous paragraph are speculations that cannot be tested and disproven, any more than other researchers’ previous claims. But they are at least feasible considering the historical linguistic and ethnoarchaeological information considered above. Nonetheless, a Vieto-Katuic branch can no longer be part of historical linguistic and ethnohistorical consideration.

References

Alves, Mark. 2005. The Vieto-Katuic Hypothesis: Lexical Evidence. In Paul Sidwell (ed.) *SEALS XV Papers from the 15th Annual Meeting of the Southeast Asian Linguistics Society 2003*, 169–176. Pacific Linguistics, Research School of Pacific and Asian Studies, The Australian National University

Alves, Mark J. 2006. *A Grammar of Pacoh: A Mon–Khmer Language of the Central Highlands of Vietnam*. Pacific linguistics, 580. Canberra: Pacific Linguistics, Research School of Pacific and Asian Studies, the Australian National University.

Alves, Mark J. 2007. Pacoh pronouns and grammaticalization clines. *SEALS XIII, Papers from the 13th Annual Meeting of the Southeast Asian Linguistics Society (2003)*, ed. by Iwasaki Shoichi et al. Canberra, Australia, 1–12. Pacific Linguistics, Research School of Pacific and Asian Studies, The Australian National University.

Alves, Mark. 2018. Early Sino-Vietnamese lexical data and the relative chronology of tonogenesis in Chinese and Vietnamese. *Bulletin of Chinese Linguistics* 11(1-2):3–33.

Alves, Mark. 2020. Historical ethnolinguistic notes on Proto-Austroasiatic and Proto-Vietic vocabulary in Vietnamese. *Journal of the Southeast Asian Linguistics Society* 13.2*:* xiii–xlv. http://hdl.handle.net/10524/52472

Alves, Mark. 2022. The Đông Sơn speech community: Evidence for Vietic. *Crossroads* 19.2: 138–174.

Alves, Mark. 2023. Proto-Austroasiatic etymologies of words related to household structures. *Papers from the Ninth and Tenth International Conferences on Austroasiatic Linguistics*, ed. by Paul Sidwell and Mark Alves, 1–18. JSEALS Special Publication No. 12. University of Hawai'i Press.

Alves, Mark. 2024. From Vietic presyllables to Vietnamese simplex onsets. *Taiwan Journal of Linguistics* 22.1: 1–46.

Bellwood, Peter. 2005. *First Farmers: The Origins of Agricultural Societies*. Hoboken, NJ: Wiley-Blackwell.

Bellwood, Peter. 2021. Homelands and dispersal histories of Mainland Southeast Asian language families: a multidisciplinary perspective. In: *The Languages and Linguistics of Mainland Southeast Asia: A Comprehensive Guide*, ed. by Paul Sidwell and Mathias Jenny, 33–44. Boston: De Gruyter.

Blench, Roger & Mark Post. 2014. Rethinking Sino-Tibetan phylogeny from the perspective of North East Indian languages. In: *Trans-Himalayan Linguistics*, ed. by Nathan W. Hill; Thomas Owen-Smith, 71–104. Boston: Mouton de Gruyter.

Chamberlain, James. 1992. The Black Tai chronicle of Muang Mouay part I: Mythology. *The Mon-Khmer Studies Journal* 21: 19–55.

Chamberlain, James R. 1998. The origin of the Sek: Implications for Tai and Vietnamese history. *Journal of the Siam Society* 86.1-2: 27–48.

Chamberlain, James R. 2019. Vanishing Nomads: Languages and Peoples of Nakai, Laos, and Adjacent Areas. In: *Handbook of the Changing World Language Map*, ed. Stanley D. Brunn and Roland Kehrein, 1589–1605. Springer Cham.

Diffloth, Gérard. 1974. Austro-Asiatic Languages. *Encyclopedia Britannica, (Macropaedia) 2*, 15th ed.: 480–4. Chicago: Encyclopaedia Britannica.

Diffloth, Gérard. 1991a. Linguistic prehistory from a Mon-Khmer perspective. Paper presented at the conference "The High Bronze Age of Southeast Asia and South China", Hua Hin, 15 to 19 January 1991.

Diffloth, Gérard. 1991b. Vietnamese as a Mon-Khmer Language. In: *Papers from the First Annual Meeting of the Southeast Asian Linguistics Society*, ed. by Martha Ratliff and Eric Schiller, 125–139. Arizona State University, Program for Southeast Asian Studies.

Diffloth, Gérard. 2005a. The contribution of linguistic paleontology to the homeland of Austroasiatic. In: *The Peopling of East Asia: Putting Together Archaeology, Linguistics and Genetics*, ed. by Roger Blench, Laurent Sagart, & Alicia Sanchez-Mazas, 77–80. London: Routledge Curzon.

Diffloth, Gérard. 2005b. Glottalised rimes in Vietic and Katuic. In: *The 6th Pan-Asiatic International Symposium on Linguistics*, 82–94. Hanoi: Nhà Xuất Bản Khoa Học Xã Hội.

Ferlus, Michel. 1996. Langue et peuples Viet-Muong. *The Mon-Khmer Studies Journal* 26: 7–28.

Ferlus, Michel. 2007. Lexique de racines Proto Viet–Muong (Proto Vietic Lexicon). Unpublished manuscript, May 2007. http://sealang.net/monkhmer/database/

Gehrmann, Ryan. 2021. Evidence for the common origins of rime glottalization in Northeastern Austroasiatic (Vietic, Bolyu, Bugan & Mang). Presentation at ICAAL 9, Centre for Languages and Literature, Lund University 18-19 November 2021.

Gehrmann, Ryan. 2023. Mang comparative database, version 1. Unpublished.

Hammarström, Harald. 2010. A full-scale test of the language farming dispersal hypothesis. *Diachronica* 27.2: 197–213. https://doi.org/10.1075/dia.27.2.02ham

Higham, Charles. 1996. Archaeology and linguistics in Southeast Asia: implications of the Austric hypothesis. *Bulletin of the Indo-Pacific Prehistory* 14.1: 110–118.

Higham, Charles F. W. 2017. First farmers in Mainland Southeast Asia. *Journal of Indo-Pacific Archaeology* 41 (2017): 13–21.

Higham, Charles. 2021. The Neolithic occupation of Southeast Asia. In: *The Languages and Linguistics of Mainland Southeast Asia: A Comprehensive Guide*, ed. by Paul Sidwell and Mathias Jenny, 21–31. Boston: De Gruyter.

Kelley, Liam. 2017. Going backwards: Tai and Vietics. Le Minh Khai's SEAsian History Blog (+ More). Accessed 25 March 2024. https://leminhkhai.blog/2-going-backwards-tai-and-vietics/

Kiernan, Ben. 2019. *Việt Nam: A History from Earliest Time to the Present*. Oxford University Press.

Kim, Nam. 2015. *The Origins of Ancient Vietnam*. Oxford Studies in the Archaeology of Ancient States. Oxford University Press.

Matsumura, Hirofumi, Marc F. Oxenham, Yukio Dodo, Kate Domett, Nguyen Kim Thuy, Nguyen Lan Cuong, Nguyen Kim Dung, Damien Huffer, Mariko Yamagata. 2008. Morphometric affinity of the late Neolithic human remains from Man Bac, Ninh Binh Province, Vietnam: key skeletons with which to debate the 'two layer' hypothesis. *Anthropological Science* 116.2: 135–148.

Nguyễn Đình Đầu. 2009. *The Vietnamese Southward Expansion, as Viewed through the Histories, in Champa and the Archaeology of Mỹ Sơn (Vietnam)*, ed. by Andrew David Hardy, Mauro Cucarzi, & Patrizia Zolese, 61–77. Hawaii: University of Hawai'i Press.

Nguyễn, Tài Cẩn. 1995. *Giáo trình lịch sử ngữ âm tiếng Việt (sơ thảo)* [Textbook on Vietnamese historical phonology (Preliminary)]. Hà Nội: Nhà xuất bản Giáo dục.

Nguyễn Văn Lợi. 1993. *Tiếng Rục* [The Ruc language]. Hanoi: Nhà Xuất Bản Khoa Học Xã Hội.

Nguyễn Hữu Hoành and Nguyễn Văn Lợi. 1998. *Tiếng Katu* [The Katu language]. Hanoi: Nhà Xuất Bản Khoa Học Xã Hội.

Ostapirat, Weera. 2005. Kra-dai and Austronesian. In: *The peopling of East Asia: Putting together archaeology, linguistics and genetics*, ed. by Roger Blench, Laurent Sagart, & Alicia Sanchez-Mazas, 107–131. London: Routledge Curzon.

Pittayaporn, Pittayawat. 2014. Layers of Chinese loanwords in Protosouthwestern Tai as evidence for the dating of the spread of Southwestern Tai. *Manusya* 20: 47–68.

Rau, Felix & Paul Sidwell. 2019. The Munda maritime hypothesis. *Journal of the Southeast Asian Linguistics Society* 12.2: 35–57.

Sagart, Laurent, Guillaume Jacques, Yunfan Lai, Robin Ryder, Valentin Thouzeau, Simon J. Greenhill, Johann-Mattis List. 2019. Dated language phylogenies shed light on the history of Sino-Tibetan. *Proceedings of the National Academy of Sciences of the United States of America*, 116.21: 10317–10322. https://doi.org/10.1073%2Fpnas.1817972116

Schmidt, Wilhelm. 1906. Die Mon-Khmer-Völker, Ein Bindeglied Zwischen Völkern Zentralasiens Und Austronesiens. *Archiv fur Anthropologie* 33: 59–109.

Sidwell, Paul. 2005a. *The Katuic Languages: classification, reconstruction and comparative lexicon*. Lincom Europa.

Sidwell, Paul. 2005b. Proto-Katuic Phonology and the Sub-grouping of Mon-Khmer Languages. In: *SEALS XV Papers from the 15th Annual Meeting of the Southeast Asian Linguistics Society 2003*, ed. by Paul Sidwell. 193–204. Pacific Linguistics, Research School of Pacific and Asian Studies, The Australian National University.

Sidwell, Paul. 2012. Austroasiatic numerals. Web. Last accessed: 13 March 2024. https://sites.google.com/view/paulsidwell/austroasiatic-numerals

Sidwell, Paul. 2015. Austroasiatic classification. In: *The Handbook of Austroasiatic Languages*, Vol. 1, ed. by Mathias Jenny and Paul Sidwell, 144–220. Boston: Brill.

Sidwell, Paul. 2018. Austroasiatic deep chronology and the problem of cultural lexicon. Presentation given at the 28th Annual Meeting of the Southeast Asian Linguistics Society, Wenzao Ursuline University of Languages, Kaohsung Taiwan.

Sidwell, Paul. 2021a. Classification of MSEA Austroasiatic languages. In: *Languages and Linguistics of Mainland Southeast Asia: A Comprehensive Guide*, ed. by Paul Sidwell and Mathias Jenny, 179–206. Boston: De Gruyter Mouton.

Sidwell, Paul. 2021b. The History of MSEA Austroasiatic studies. In: *Languages and Linguistics of Mainland Southeast Asia: A Comprehensive Guide*, ed. by Paul Sidwell and Mathias Jenny, 61–92. Boston: De Gruyter Mouton.

Sidwell, Paul. 2022. Austroasiatic dispersal: the AA "water-world" extended. Were the Proto-Austroasiatics coastal migrants? In: *Papers from the 30th Meeting of the Southeast Asian Linguistics Society (2021)*, ed. by Mark Alves and Paul Sidwell, 56–72. JSEALS Special Publication No. 8. Honolulu: University of Hawai'i Press. http://hdl.handle.net/10524/52498

Sidwell, Paul. 2024. 500 Proto Austroasiatic Etyma: Version 1.0. *Journal of the Southeast Asian Linguistic Society* 17.1: i–xxiii. https://hdl.handle.net/10524/52519

Sidwell, Paul and Mark Alves. 2021. The Vietic languages: a phylogenetic analysis. *Journal of Language Relationship* 19.3 (2021): 166–194.

Sidwell, Paul and Mark Alves. 2023. Re-evaluating Shorto's MKCD reconstructions. In: *Papers from the Ninth and Tenth International Conference on Austroasiatic Linguistics*, ed. by Paul Sidwell & Mark Alves, 98–126. JSEALS Special Publication No. 12. Manoa, University of Hawaii Press. https://hdl.handle.net/10524/52517

Sidwell, Paul and Roger Blench. 2011. The Austroasiatic Urheimat: the Southeastern Riverine Hypothesis. In: *Dynamics of human diversity*, ed. by Nicholas Enfield, 315–344. Canberra: Pacific Linguistics.

Solntseva, V. Nina. 1996. Case-marked pronouns in the Taoih language. *The Mon-Khmer Studies Journal* 26: 33–36.

Thomas, David D. and Robert K., Jr. Headley. 1970. More on Mon-Khmer subgroupings. *Lingua* 25: 398–418.

Watson, Saundra K. 1964. Personal pronouns in Pacoh. *The Mon-Khmer Studies Journal* 1: 81–97.

Wichmann, Søren, André Müller, & Viveka Velupillai. 2010. Homelands of the world's language families: A quantitative approach. *Diachronica* 27.2: 247–276. DOI: https://doi.org/10.1075/dia.27.2.05wic

Wichmann Søren and Rama Taraka. 2021. Testing methods of linguistic homeland detection using synthetic data. *Philosophical Transactions of the Royal Society* 376.1824. https://doi.org/10.1098/rstb.2020.0202

11

Halliday Redux: Pitfalls in Mon Dialectology

Christian Bauer

Robert Halliday, author of the *Mon-English Dictionary*, spent over 40 years among the Mon in Burma and Thailand, first in Ye, then in Nakhorn Pathom, and finally in Moulmein, where he died in 1933. No modern scholar before or since had such a continuous exposure to Mon speaker communities, and it is therefore incumbent on us to pay special attention to what he had to say on linguistic matters. He pointed out that the dialect differences in Burma and in Thailand were identical:

> *"Though there are dialect differences in the spoken Mon, they are not so great as to prevent people from different districts readily understanding one another. Strange to say the very same dialectical differences are found amongst the exiles in Siam, even where at least two centuries separate them from the connection with Burma." (1922a:ii)*[1]

This was the very topic Gérard and I discussed immediately when we met for the first time in his flat in Bangkok in January 1984; he was on sabbatical leave from Chicago then, finishing the fair copy of his monograph on Nyah Kur and Monic (1984).

I was on a two-month leave from my postdoc at Monash to collect further Mon data. Both of us were puzzled by Halliday's remarks above as our field data did not reveal such a diversity and correspondence for the varieties in Thailand.

Gérard thus stated in his monograph, referring to Halliday: "Judging by the results of the present survey here that this is apparently no longer true in 1980: there is a good deal of uniformity among the descendants of these refugees, in Thailand today." (1984:41) Could it possibly be that within half a century those dialect differences in Thailand had, through convergence, been levelled?

Looking back today, it was presumptuous as neither of us had by 1984 conducted dialect surveys in Burma *in situ*; in fact, Gérard had gathered data on Burma Mon in Thailand in Sangkhlaburi at "Three Pagodas" and on a week-long tourist visa in Rangoon. For my part, I had spent nine months in villages of central Thailand between 1978-1980 as part of my SOAS PhD, and had gathered data on Ye Mon at "Three Pagodas" in late 1983.

For the record: Mon State in Burma was off-limits for foreigners until the armed opposition parties and the Burmese government reached a peace accord in 1995, and for visits to central Burma one was restricted to a 7-day tourist visa, which in the late 1980s was extended to a fortnight. It was only from 1995 onwards that I was able to spend several weeks almost each year in Mon State during term breaks and sabbaticals.[2]

While digitizing my audio tapes late last year, I noticed in some freely spoken narrative texts, recorded in Photharam, Rajburi, in 1978/79 with a particular speaker

1 On two other occasions he made similar comments; see the appendix.

2 See South 2003 for recent political developments in Mon State.

(*1935) born in that province, sporadic variation of /ɒ ~ ɛ/ in prelabial contexts /-p, -m/; thus the word for ‘to speak’ LM *huiṁ* DSM /hɒm/ > dial. /hɛm/. At the time of recording, it did not draw my attention; but after having been to Kamawet and Kalawthut in Mon State in 1997 where this is a common and regular feature and distinctive for its dialect, could it possibly be that parents or grandparents of this speaker from Photharam could have come from precisely those localities? And that the variation /ɒ > ɛ/ was in his speech sporadic only because his pattern had been levelled to conform to local usage /-ɒp, -ɒm/. This would confirm Halliday’s thesis after all.

This assumption is supported by another observation from the same recording: occasional pharyngealization, such as ‘to go out, come ~’ LM *tit* DSM /tɛt/ dial. /ŧɛʲt/ which correlates with the varieties in Kamawet and Kalawthut where this type of pharyngealization is common.[3]

Variant forms for ‘pillow’ LM *dnī* DSM /n̰i̤/ ~ni̤/[4], are also attested in Thailand.[5]

Narinthorn in her instrumental phonetic study (2011:69) lists for Rajburi, Samut Sakhorn and all of her Burma speakers monophthongal /-ɛ̤ŋ/ corresponding to diphthongal /-ɛ̤aŋ/ in Lopburi and Lamphun (/-e̤aŋ/) corresponding to DSM /-ɛ̤aŋ/ in ‘to wait’ LM *maṅ* DSM /mɛ̤aŋ/.

But the correspondences for the LM *-au* rhyme in breathy voice intersect differently: Lamphun, Rajburi, Samut Sakhorn have /-e̤a/ whereas Lamphun agrees with Burma /-ɛ̤a/, as in ‘woman, wife’ LM *brau* DSM /prɛ̤a/.[6]

LM *-ā* in breathy voice rhymes in her study are identical in all varieties in Burma and Thailand, namely /-e̤/, but differ from DSM usage /-ɛ̤a/; thus ‘monastery’ LM *bhā* DSM /phɛ̤a/ dial. /phe̤/.[7]

Although the commonest form of the LM *-ai* rhyme DSM /-oa/ in Thailand corresponds here to /-oy ~ -ɔe/—likely to be a MM retention—it co-occurs with some speakers’ /-oa/: ‘[locative] in, at’ LM (p)ḍai DSM /ɗoa/ dial. /(pə)ɗɔe ~ (pə)ɗoa/.[8]

There are another two instances that show MM retentions: ‘kinsman’ MM *kmin*, LM *mhin ~ smin*, DSM /mɛn/ dial. /min/, and ‘door’ LM *taraṅ* DSM /kərɛŋ/ dial. /təraŋ/.

The last case—noted incidentally in Photharam—shows that some dialect forms have retained a late 18th c., or earlier, form; MM /tər-/ had shifted to /kər-/ by the early 19th c.[9]

Diachronic considerations lead us to actual historical events: Common to all Mon varieties in Burma and Thailand is the use of voice quality (“register”) to mark lexical distinctions by either modal voice or breathy voice.[10] Shorto had shown by 1965 that

---

3 In his Berkeley 1982 paper Diffloth asks rhetorically if there are more than two registers in Mon, with pharyngealized voice being one of them. He thus defines register differently, which was originally meant to be a term referring to being lexically contrastive. The phenomenon of pharyngealization here is not being lexically contrastive but rather substitutes words in modal voice in certain phonological environments. In addition, this phenomenon is not regionally confined to Kamawet and Kalawthut but occurs also with speakers from Ye.

4 Shorto’s notation for voiceless nasals in DSM, and elsewhere, is /h-/ followed by a nasal, and for breathy vowels the grave accent ◌̀ instead of IPA ◌̤, in this case DSM /hnì/.

5 Prayat 1986:175.

6 ibid., p. 70.

7 In my samples actually /phɛ̤/.

8 With my Photharam speaker once /əɗoa/, otherwise /(pə)ɗɔe/ as common in that district. But see Prayat 1986, data from Samut Sakhorn, an area I had not visited.

9 Bauer 2009.

10 In Shorto’s terminology—derived from Henderson’s usage—”head register” and “chest register”, respectively.

the development of register took place between the second quarter of the 16th c. and the end of the 16th c.[11] In other words, varieties in Thailand today do not reflect stages of the language before the 16th c., and this correlates with the first of several subsequent Mon population displacements to Thailand. The history of movements of speaker communities, however, is more complex, as, for instance, following the British annexation of Tenasserim—and thus areas of today's Mon and Karen States—Mons that had settled in Thailand for a century or longer, moved then back to British controlled Mon areas.[12] This aspect helps explain why isoglosses are difficult to draw, if it is not futile altogether: continuous population movements across different linguistic contact zones (Burmese and Karen here, Thai there) and internal migrations are more likely to account for the bewildering variations.[13]

Other methodological issues may also account for differing dialect data: it should be obvious that gathering data by eliciting words from lists may skew variant data, and that it would be preferable to gather texts and record, or note, interviews and conversations.[14] Dorian 2010 has given sociological parameters to consider when studying variation across speaker populations.

Gathering data *ex-situ* has its own uncertainties: DOML gives for LM *-eṅ* rhymes /-eaŋ/ in Tarana village,[15] Kyaikmaraw township, yet I met a Mon from the neighbouring village Kawthat[16] on a visit in Bangkok's Paklat district, in conversation, who retained the MM ryhme /-eŋ/.

Other aspects that have not drawn the attention of linguists are subphonemic phenomena, such as pre-stopped nasals, which I recorded with a speaker from Ye, such as 'child, offspring' LM *kon* DSM /kon/ dial. /ko^{d}n/.

Another desideratum would deal with the emergence of new dialects, as Sakamoto 1985 has demonstrated.

---

11 Shorto 1967:247-248, paper presented at a conference on tone in Leipzig; a preliminary version was presented two years earlier at the Philological Society in London.

12 South 2003:92 and p. 363 n. 294.

13 Langham-Carter 1947:24, 36: "In 1825 Mon refugees poured in from Burmese territory [to Moulmein], many coming by sea … [p. 24] … Several noted Sayadaws from Upper Burma, finding that conditions deteriorated in King Thibaw's reign came and resided in Moulmein from about 1880. [36]" ["… coming by sea …"] is likely to refer to displacement of Mons from the Irrawaddy Delta, in particular Bassein, a narrative I heard on several occasions. Evidence for Mon presence in Bassein and environs dates until the late 16th c., and not thereafter. By 1796 Hiram Cox, passing by the mouth of the Bassein river, noted the areas on the west bank of the Irrawaddy was no longer populated by Mons. Some accounts on local history for Thailand are available, such as Sujit 2004 on communities along the Maeklong river and generally on Mon population movements by Suporn 1999.

14 The two rhyme charts, shown here at the end of my contribution, are simply meant to be a frame of reference to identify rhyme types cited in this article. I maintain, however, following Shorto, that if there is variety of spoken Mon that might be considered as some kind of standard among the wide range of variations. Shorto wrote: "The dialect […] is spoken […] in the area east of the Salween, round the Gyaing and the Ataran, and on parts of Bilugyun. It appears to command more acceptance as a standard among Burma Mons in general than might be expected in the absence of an autonomous political system, or, until recently, of an urban centre of Mon culture; its prestige has been greatly aided by the monastic organization, with its influence on those who studied in its schools" (DSM:x).

15 GPS (16.517362251874502, 97.77390530420892)

16 GPS (16.520118905855433, 97.755451706165)

Rhymes

modal voice

-i	-iʔ	-ih			-it	-in	-ip	-im
-ɔe	-ɔeʔ	-ɔeh						
-e	-eʔ	-eh	-eak	-eaŋ	-et	-en	-ep	-em
	-ɛʔ		-ɛk	-ɛŋ	-ɛt	-ɛn	-ɛp	-ɛm
-a	-aʔ	-ah	-ak	-aŋ	-at	-an	-ap	-am
-ɜ			-ɜk	-ɜŋ				
-ɒ	-ɒʔ	-ɒh			-ɒt	-ɒn	-ɒp	-ɒm
-ɔ	-ɔʔ	-ɔh	-ɔk	-ɔŋ	-ɔt	-ɔn	-ɔp	-ɔm
-o			-ok	-oŋ	-ot	-on	-op	-om
-u	-uʔ	-uh			-ut	-un	-up	-um
-ao	-aoʔ	-aoh						
-ai			-aik	-aiŋ				
			-ɔik	-ɔiŋ				
			-oik	-oiŋ				
-ui								
-ea								
-oa								

breathy voice

-i̤	-i̤ʔ	-i̤h			-i̤t	-i̤n	-i̤p	-i̤m
-e̤	-e̤ʔ	-e̤h			-e̤t	-e̤n	-e̤p	-e̤m
-ɛ̤a	-ɛ̤ʔ	-ɛ̤h	-ɛ̤ak	-ɛ̤aŋ				
			-a̤k	-a̤ŋ	-a̤t	-a̤n		
-ɜ̤	-ɜ̤ʔ	-ɜ̤h	-ɜ̤k	-ɜ̤ŋ	-ɜ̤t	-ɜ̤n	-ɜ̤p	-ɜ̤m
-o̤	-oʔ	-oh	-o̤k	-o̤ŋ	-o̤t	-o̤n	-o̤p	-o̤m
-ṳ	-ṳʔ	-ṳh			-ṳt	-ṳn	-ṳp	-ṳm
-a̤i			-a̤ik	-a̤iŋ				
			-o̤ik	-o̤iŋ				
-ṳi								
-o̤a								

---

cf. Shorto, 1966:401, 402-403

Appendix: Halliday's further remarks on dialects

"There are various dialectical differences all over the Mon country in Burma. Beginning with Ye in the southern part of the Amherst district and travelling up to Moulmein you find various differences and little changes all the way. Up river from Moulmein there is a marked difference. Crossing over to Martaban and going westward you find another change until towards Pegu you get what the Siamese Mons call the pure Mon. Strangely enough you find all this variety of dialect in Siam. This persistency of dialectical variations is quite remarkable. I have met old people in Moulmein whose parents had come over from Pegu at the beginning of the British occupation and whose dialect is still that of Pegu." (1913:10)

[not referred to in DOML]

"There are two chief dialects distinguished, and it is usual to call them the Pegu and the Martaban dialects respectively. There is, however, a great variety in the second of these, chiefly I suppose because it is the more generally used. All these differences occur amongst the Mons of Siam just as they are found here [sc. in Burma]. There are Mons in the Pathom district who exhibit the Pegu dialect in their speech, whilst in most other districts the Martaban dialect is shown with all its variations, To me it was an indication that those who spoke the Pegu dialect were of the old Mons whose fathers had come from Pegu, whereas most others were descendants of the more recent incomers from Martaban. It was very interesting to come across these old familiar differences." (1922b:78)

Abbreviations

dial. - dialect form
DOML - Diffloth 1984
DSM - Shorto 1962
LM - Literary Mon
MM - Middle Mon (14th – ~ 18th c.)
SM - Spoken Mon

References

Bauer, Christian. 2009. "When did Middle Mon end?" ICAAL 4, Bangkok.

Diffloth, Gérard. 1982. "Proto-Mon Registers: Two, Three, Four...?" *Berkeley Linguistics Society 8*: 148-157.

_____. 1984. The Dvaravati Old Mon language and Nyah Kur. Bangkok: Chulalongkorn UP.

Dorian, Nancy C. 2010. Investigating variation. The effects of social organization and social setting.Oxford Studies in Sociolinguistics. Oxford: Oxford University Press.

Halliday, Robert. 1913. "Immigration of the Mons into Siam." *JSS* 10 (3): 1-13.

_____. 1922 a (²1955). *A Mon-English dictionary*. Bangkok (reprint: Rangoon): Siam Society (Ministry of Union Culture, Govt. of the Union of Burma).

_____. 1922 b. "The Mons in Siam." *JBRS* (12): 69-79.

Langham-Carter, R[eginald] R[obert. 1947. *Old Moulmein (875-1880)*. Moulmein: The Moulmein Sun Press.

Narinthorn Sombatnan-Behr นรินถร สมบัตินนท์. [2012] 2555. “การเปรียบเทียบความเปลี่ยนแปลงของระบบสระและลักษณะทงกลสัทศาสตร์ของสระในภาษามอญไทยและมอญพม่า แนวโน้มการกลายเป็นภาษาต่างแบบ

[A comparison between the change of vowel systems and the acoustic characteristics of vowels in Thai Mon and Burmese Mon: a tendency towards different language types].” PhD, Linguistics, Chulalongkorn University.

Prayat Kitisarn. 1986. “The phonology of Mon at Ban Khlong Khru, Tambol Thasai, Amphoe Muang, Samut-Sakhon Province.” MA, ILCRD, Mahidol University.

Sakamoto, Yasuyuki 坂本恭章. 1985. “[The Mon dialect of Sam Ruan] モン語、サーム・ルアン(Sam Ruan)方言.” *[Journal of Asian and African Studies]* アジア・アフリカ言語文化研究 5: 33-42.

Shorto, Harry Leonard. 1962. *A Dictionary of modern spoken Mon*. London: Oxford UP.

_____. 1966. “Mon vowel systems.” In *In memory of J.R. Firth*, edited by C.E. Bazell et al., 398-408. London: Longmans.

_____. 1967. “The register distinctions in Mon-Khmer languages.” Wissenschaftliche Zeitschrift der Karl-Marx-Universität Leipzig [Gesellschafts- und Sprachwissenschaftliche Reihe] 16 (1/2): 245-248.

South, Ashley. 2003. *Mon nationalism and civil war in Burma*. London: RoutledgeCurzon.

Sujit Wongthes (ed.) สุจิตต์ วงษ์เทศ (บ.ก.), ed. [2004] 2547. *[The Maeklong River Basin]* ลุ่มน้ำแม่กลอง. Bangkok: Matichon.

Suporn Ocharoen สุภรณ์ โอเจริญ. [1999] 2541. *[The Mons in Thailand]* มอญในเมืองไทย. Bangkok: The Thailand Resarch Fund.

12

The Birth and Life of Monic

Mathias Jenny

1. Introduction

In 1984, Chulalongkorn University in Bangkok initiated a publication series labeled *Monic Studies*, dedicated to the Monic branch of Austroasiatic, which consists of only two members, namely Mon and Nyah Kur. Both languages come in a variety of dialects, and both go back to a common ancestor some 1500 years ago, the Old Mon language of Dvāravatī. The project, which certainly was planned to lead to continued interest and research in the Monic branch of Austroasiatic, ended after two publications in the same year: Theraphan L. Thongkum's Nyah Kur-Thai-English dictionary and Gérard Diffloth's comparative lexicon and reconstruction of what he called "Dvāravatī Old Mon" (DOM). The two volumes are the result of extensive fieldwork in Nyah Kur communities in three provinces of northeastern Thailand (Isan), namely Khorat, Chaiyaphum, and Phetchabun. Diffloth (1984) complemented his Nyah Kur data with wordlists collected from several spoken Mon varieties in Thailand and Myanmar, as well as Old Mon data as found in the inscriptions and published in Shorto's comprehensive dictionary of Mon inscriptions (Shorto 1971).

Nyah Kur (also Chaubun/Chaobon, or Niakuol, as it was called in the early publications) had been recognized earlier as related to Mon (often called Talaing or Peguan in older sources). Seidenfaden (1918) presents a few pages of vocabulary comparing Nyah Kur with Mon, both languages given in rather impressionistic transcription but clearly showing the close relationship. Thomas and Headley (1971:407) postulate a Monic branch which includes Mon and Nyah Kur (Niakuol). Shorto (1971) frequently gives Nyah Kur (Niakuol) cognates in the (unsystematic) etymological connections of Mon lexemes. Ferlus saw Nyah Kur as close enough to Mon to rely on it as collateral evidence in his reconstruction of "proto-Mon" (Ferlus 1984).

Although the close relationship between Nyah Kur and Mon had obviously been recognized for a long time, it is Gérard Diffloth who systematically showed the connections between Nyah Kur and Old Mon, first in a publication in Thai (Diffloth 1980), then in his book-length study (Diffloth 1984). By elaborating the earlier suggestions of Monic, implicitly (Seidenfaden 1918, Ferlus 1984) or explicitly (Thomas and Headley 1971), Diffloth in the 1980s formally gave birth to the Monic branch as a potential subject of in-depth study. Importantly, Diffloth (1984:1) asserted that the Nyah Kur are the descendants of the Mon speaking population of Dvāravatī, a (more or less) well established cultural area, if not political entity, located in present-day central Thailand with influence over most of the northeast, north, and possibly south of Thailand before the coming to dominance of Tai speakers migrating from the north. This claim of Nyah Kur as direct descendant of the language of Dvāravatī added a new dimension to the role of Nyah Kur and, one might expect, new importance to the further

study of the language and its people. Diffloth's hypothesis was received enthusiastically by the Mon scholar Nai Pan Hla in a 1986 paper published in the *Journal of the Siam Society*, but not much additional original or in-depth investigation followed. Hardly any linguistic research has been done in Nyah Kur after the seminal work by Diffloth and Theraphan in the 1980s. The rare exceptions include a typological account of causatives in Nyah Kur (Gainey 1990), a synchronic phonological comparison of Mon and Nyah Kur (Huffman 1990), and a sociolinguistic survey (Premsrirat 20022), besides a handful of MA theses written at Mahidol University in Thailand. Although Gérard Diffloth on several occasions mentioned the plan to do a revision of his Dvāravatī Old Mon and Nyah Kur book, he unfortunately never found the time and resources to complete this task. His 1984 publication, despite all its shortcomings (which Gérard himself was well aware of, Diffloth 1984:49-51), therefore remains the major resource for comparative Monic until now.[1]

This paper, based mainly on Diffloth 1984, complemented with available other publications on Nyah Kur and Mon, and my own observations in the field (Nyah Kur: Phetchabun 2008, Mon: Sangkhlaburi and Mon State, 1993 to 2024), presents relevant phonological and semantic developments in both Mon and Nyah Kur with the aim to position Nyah Kur in the broader context of Dvāravatī and Mainland Southeast Asia and point out its importance in areal studies. Although historical factors are included in the presentation, no claim whatsoever to completeness is made and the reader is referred to relevant sources for details.

2. Dvāravatī Old Mon (DOM) - language and history

The Old Mon language has survived to the present in two major groups of inscriptions from Burma and Thailand, respectively, dating to two periods separated by a gap of a couple of centuries. The most extensive documents in Old Mon are the inscriptions from 11th and 12th century Bagan, where Mon was a major literary language before Burmese took over the role of the main written idiom. Some of the Bagan inscriptions are rather long and provide ample material for lexical and grammatical studies of the language.[2] Mon presence is evident in Bagan epigraphy and architecture, but Old Mon as language of literature and administration may or may not have been used as spoken language by a large portion of the population in Burman-dominated Bagan (Moore 2023:8-15). Bagan Old Mon certainly was exposed to heavy Burmese influence and also exhibits a large number of Pali loanwords.

While early second millennium Bagan provides the largest corpus of Old Mon inscriptions, the oldest documents in Old Mon were found in the Chao Phraya plain and elsewhere in modern Thailand, a cultural area that has come to be known as Dvāravatī, dating to the 6th century (Wyatt 2003:17-21, Guy 2020, Watson 2020). These inscriptions are fewer and shorter than the ones found at Bagan and other locations in Burma, allowing only limited conclusions of Old Mon before Burmese influence. Pali loans are present from the earliest written documents in Mon, which is to be expected in a language that adopted the writing system from Indic sources together with Buddhist

[1] Besides Diffloth's collection of Mon data in often less than ideal circumstances, his lack of deep knowledge of Burmese led him to some obviously wrong etymologies, as in the case of *thiəŋ 'argue, think' (V94), where Nyah Kur has an obvious loan from Thai *tʰĭəŋ* 'argue' while Mon *thiəŋ* reflects a more recent loan from Burmese *tʰiɴ* 'think'.

[2] The Old Mon inscriptions are available in the *Epigraphia Birmanica* series in transcription and translation (Duroiselle et al. 1919 onwards).

practices and cultural features. The extent and internal make-up and cohesion of Dvāravatī, which is believed to have flourished in present-day Thailand from the 6th and 9th or 10th centuries, is not clear, and new discoveries in Thailand, especially the northeastern region, add to the picture, though much remains to be done (Bhumadhon 2020). The population of Dvāravatī certainly was mixed, with the Mon possibly being part of the ruling elites (see Watson 2020 for a detailed discussion). Unlike in the case of Bagan, it is likely that a sizeable portion of the population of Dvāravatī spoke a form of Old Mon, with the Nyah Kur representing the last remnants of this population. Present-day Mon communities in Thailand all represent later waves of migration from Burma, going back no further than the 16th century, though the new arrivals may have met with earlier Mon-speaking communities which have left no trace in the recorded histories of the region (McCormick and Jenny 2013:87, Baker and Phongphaichit 2017:204, Champaphan 2023:38-40). This is also true for locations traditionally connected with Dvāravatī, like Lopburi near Ayutthaya and Hariphunchai (modern Lamphun) in northern Thailand (Ongsakul 2005:32-39).

With the expansion of Khmer influence from Angkor to central and northeastern Thailand from the 10th century (Wongsathit et al. 2020, Wyatt 2003:21-25) and the intrusion of Tai speakers from the north around the 13th century (Wyatt 2003:30-49), the Mon speaking communities were separated and the ancestors of the Nyah Kur became isolated in the hills between central and northeastern Thailand (Diffloth 1984:26-27). Unlike their cousins in the Chao Phraya plain, they were cut off from contact with the bulk of Mon further west and were increasingly exposed to Khmer and local Thai/Lao influence. At least since the 10th or 11th century, what once was a single language, Dvāravatī Old Mon, began to split into two branches. Mon and Nyah Kur today are mutually unintelligible, though they still share numerous lexical items and some grammatical features. While Mon continued as a literary and everyday language in southern Burma (and probably parts of modern Thailand), it was increasingly influenced by Burmese in its phonological and grammatical structure and vocabulary. Re-immigration of Mon speakers to Ayutthaya assimilated any formerly present forms of Siam Mon. The urban elites of Dvāravatī were absorbed into the Khmer and later Tai/Siamese ruling classes, incorporating parts of Dvāravatī culture and customs in the formation of early Siam but losing their linguistic identity with a shift to Khmer and Thai. At the same time, peripheral communities like the Nyah Kur were able to maintain the linguistic, if not the cultural heritage of Dvāravatī, and preserved many archaic features, including final consonants that were lost or changed in Mon.

3. Nyah Kur as descendant of DOM

It is therefore the present-day Nyah Kur, rather than the Thai-Mon (Thai-Rāman) communities that are the true heirs of Dvāravatī Old Mon language in Thailand as reconstructed by Diffloth. Diffloth (1984) had several sources at hand to go about the reconstruction of Dvāravatī Old Mon, including primary documents in the form of contemporary inscriptions. His task was therefore rather a partial reconstruction, complementing the available records with data from modern varieties of the two sub-branches, namely spoken Mon and Nyah Kur. Other inscriptional material from post-Dvāravatī Bagan and Hariphunchai is adduced as additional evidence. While Diffloth himself was aware of the different levels of reliability of his data (Diffloth 1984:49-51), he was confident enough to base a viable description of the ancestor language of Mon and Nyah Kur on the data at hand. The fact that the close relationship between (Old)

Mon and Nyah Kur has never been seriously doubted or challenged speaks for his work. One important point is that present-day Nyah Kur is phonologically archaic enough to make it impossible to see it as the result of a hypothetical more recent migration from Mon speaking areas. Nyah Kur undeniably reflects a stage of Mon before the Middle Mon period, as witnessed, among others, in the retention of final liquids /r/ and /l/, which were lost in Mon by the 15th century (Ferlus 1984:10-11). This fact, together with the area of present-day Nyah Kur well within the frontiers of Dvāravatī, led Diffloth to the conclusion that Nyah Kur indeed represents the old Mon speaking population of Dvāravatī. While this claim wasn't openly challenged by other scholars, it also hasn't received the attention one might expect.

In the latest publication on Dvāravatī (Bennet and Watson eds. 2020), Nyah Kur is mentioned in only one paper (Watson 2020) in passing, without further elaboration. This can be interpreted as silent agreement, or as not seeing the importance of having access to a people directly connected to Dvāravatī. This neglect is a missed opportunity, as the presence of descendants of the Dvāravatī population in Thailand allows us to see and describe the ancient civilization as more than just an abstract cultural phenomenon, enshrined in museum pieces and partly renovated ruins of temples and pagodas. Dvāravatī was a cultural conglomerate of various societies and language groups with Mon speakers presumably being at least part of the ruling class. But they obviously also were part of the rural landscape, with Mon speaking communities settled far away from the cultural and religious centers. The Nyah Kur most likely reflect one of these peripheral, non-urban, non-elite groups, that nonetheless partook of the Dvāravatī culture and language, albeit in a less sophisticated and illiterate form.

While the evidence strongly suggests that the Nyah Kur indeed represent remnants of Dvāravatī Mon speakers, a couple of questions are open to discussion. On the positive evidence side, the main arguments for Diffloth's claim are the facts that:

1. Nyah Kur is a language obviously closely related to Mon but not a variety of spoken Mon,
2. it is spoken in an area that was part of the Dvāravatī sphere of influence, and
3. its archaic phonology is closer to the Old Mon of the inscriptions than to Middle or modern Mon.

On the other hand, the Nyah Kur people:

1. have no self-identification as Mon,
2. have no remembered history linking them to Dvāravatī, but rather see themselves as 'hill people', as their name indicates,
3. do not show a significant number of uniquely Mon cultural traits, such as material culture, and
4. they lack the core-cultural vocabulary considered diagnostic for Mon by some (e.g. Guillon 1999:65-66), especially OM <ḍūṅ, ḍoṅ> 'city, land, *mueang*' and <kyāk, kyek> 'Buddha, pagoda, holy object'.

Although the Nyah Kur are, and according to their own tradition always have been, settled in an area that exhibits numerous Dvāravatī sites, including Sithep, the Nyah Kur themselves do not consider themselves as Mon (Diffloth 1984:27), at least not until outsider academics told them of the claimed connection with Dvāravatī. This is hardly surprising in a rural community without written tradition and surrounded by other ethno-linguistic groups for centuries. Also, it is not clear whether the names "Mon" and "Dvāravatī" were actually used by the Dvāravatī Mon; the ethnonym <rmeñ> (or its variants) 'Mon' does not occur in any Dvāravatī Mon inscription[3] and the name Dvāravatī, occurring in a few inscriptions and on a coin found in the area, may or may not have been generally used by the indigenous population (Guy 2020, Skilling 2020). The fact that the Nyah Kur do not consider themselves 'Mon' or 'Dvāravatī' obviously is no proof that they are not in fact descendants of DOM. If the ancestors of the Nyah Kur were part of the rural population, as their endonym *ɲa̤h kur* 'people of the hills' suggests, they would not have partaken in the urban culture, which explains the lack of typical Mon cultural features and artifacts. Interestingly, Diffloth (1984:28) identified the name of a common game played among the Mon involving big tree seeds (Entada beans, *nlɛ̤ːʔ* in Nyah Kur, *hənɛ̤ʔ* in Mon) which is conserved in Nyah Kur with the cognate name. This fact aligns well with a rural society, sharing folklore but not highly formalized culture with the urban elites.

The emblematic Mon lexemes for 'town, land' (the political entity commonly referred to in historical literature in its Thai form as *mueang*) and 'Buddha, pagoda' are indeed not found in Nyah Kur, at least not in their common Mon meaning. The noun *ɗoːŋ is listed in Diffloth's reconstruction (N197) with the Nyah Kur meaning 'village, house', paralleling the semantic range of standard Thai *bâːn*. The Nyah Kur, not being part of the administrative elite and far away from any central state control, if such was actually present in Dvāravatī, obviously retained an earlier semantics of the word (roughly 'home(land)') or changed its meaning under the influence of nearby related languages (Shorto 1971:136). Similarly, the attested Old Mon word <kyāk, kyek>, referring to any holy (Buddhist) personality or object ('Buddha, pagoda, statue, holy object') is not found in this meaning in Nyah Kur. But Diffloth gives the compound *thəmɤ̤ŋ khəjaːk* 'rainbow' under entry number N163 *kjaːk. The first part of the compound, *thəmɤ̤ŋ*, though not found elsewhere in Nyah Kur, clearly corresponds to Mon *həmɤ̤ŋ* <damɔṅ> 'place, abode', the second is the Old Mon <kyāk, kyek>,[4] here presumably with a pre-Buddhist meaning 'spirit, ghost' (Shorto 1971:59-60). The rainbow in Nyah Kur is thus the 'abode of the spirits/ghosts'. The same compound is found in Old Mon in the form <dirmoṅ kyek> meaning 'image shrine' (Shorto 1971:194). This suggests that the compound is an indigenous Nyah Kur innovation at a time when the two parts were still in common use and understood.

The absence of a consciousness of Mon/Dvāravatī identity and the lack of core cultural and lexical features in Nyah Kur strengthens the case for their ancestors being rural communities rather than weakening the argument for them to be legitimate heirs of Dvāravatī Mon. Based on the evidence from Nyah Kur we can hypothsize that there was a two-way development of Dvāravatī heritage in Thailand: the culture was absorbed into Thai/Siamese kingdoms, mixed with Khmer and indigenous Tai elements, the language survived in peripheral communities remote from the cultural and religious

[3] It appears in the Sanskrit form *rāmanya* in contemporary Khmer inscriptions and may well have been an exonym for the population of Dvāravatī at that time (Wongsathit et al. 2020:140)

[4] For the initial compare *khəjaːl* 'wind' from N244 *kjaːl, Old Mon <kyāl> 'wind' (Diffloth 1984:111).

centers, presumably without literature to record and transmit the history and culture. The early days of Ayutthaya reflect the former, the Nyah Kur the latter.

4. Phonological developments from DOM to Nyah Kur and Mon

After a thousand years of separation from mainstream, urban, literary Dvāravatī Mon and exposure to other languages, both related Austroasiatic idioms and unrelated Thai/Lao, it is obvious that Nyah Kur today is a language very different from its ancestor. It is all the more remarkable that a number of features, long lost in Mon and never present in Thai/Lao, are preserved in Nyah Kur. In other cases, innovations seem to be shared with Mon, though they must have come about in independent developments in the two groups, as they are of more recent date than the separation after the decline of Dvāravatī around the 10th century. In the following paragraph, I will highlight a few of the more striking phonological features of DOM and its daughter languages.

4.1 Retention of liquids in coda position

Word-final <r> and <l> occur frequently in Old Mon (OM) inscriptions, both from Dvāravatī and later periods. By Middle Mon (MM, 15th century), these are lost in all documents, often appearing as <w> after short vowels. Alternative spellings for both OM <r> and <l> as MM <l>, <r>, and <w> suggest a certain variation of pronunciation, but may also be due to etymological spellings. Old Mon <ḅār> 'two' becomes <ḅā> in Middle Mon (Shorto 1971:405), OM <sar> 'be low' becomes MM <sar> or <saw>, likely pronounced as /sɔw/ (Shorto 1971:366). In Nyah Kur cognates, the codas are maintained in their original form and assumed pronunciation in the central (C) and southern (S) varieties, while /-r/ merges with /-l/ in the northern (N) varieties. The following examples illustrate the development of final liquids in Monic. The data for OM and MM are from Shorto (1971), Spoken Mon (SM) is based on Jenny's phonological overview (Jenny 2005:33-37):

DOM	**ID**	**Gloss**	**NK-N**	**NK-C, S**	**OM**	**MM**	**SM**
*kul	V236	'give'	*kúl*	*kúl*	*kul, kɨl*	*kiuw*	*kɒ*
*kjaːl	N244	'wind'	*khəjaːl*	*khəjaːl*	*kyāl*	*kyā*	*kya*
*ʔa(ː)r	V212	'go'	*ʔal*	*ʔaːr*	*ʔār*	*ʔā*	*ʔa*
*sar	V230	'be low'	*ɕal, chɛl*	*ɕɛr, chɛr*	*sar*	*sar, saw*	*sɔ*

One interesting case is Old Mon <kwel> 'cart', which is attested in a 7th century inscription from Lopburi (Shorto 1971:65). In Bagan Old Mon, it appears in the form <kwīl>, which leads to Middle Mon <kwī> and Spoken Mon *kwi*. The expected Nyah Kur form would be *kweːl or similar, but the actual attested form is *kiən* (Thongkum 1984:20) or *kwiən* (Shorto 1971:65). This is obviously not the reflex of the DOM form as suggested by Shorto (1971:65), but rather a loan from a local Thai/Lao variety. the Thai word for 'cart', *kwiən* is itself a loan from an Old Mon variety that retained final /l/, which regularly changes to /n/ in Thai (Jenny 2012:9).[5] The modern Nyah Kur forms are apparent re-borrowings from Thai/Lao, but ultimately go back to DOM lexemes.

[5] The same change is seen in words like OM <khal> 'small cup', which appears in Thai as *kʰăn* 'bowl' (*khal* in Nyah Kur, *khɔ* in spoken Mon).

4.2 Retention of palatals in coda position

Palatal codas <c> and <ñ> are attested in OM, but in MM changed to alveolar <t> and <n> after back vowels and <k> and <ṅ> after front vowels, respectively, and appear as such in literary Mon. Nyah Kur, on the other hand, retains the palatal codas, as seen in the following examples:

DOM	ID	Gloss	NK-N	NK-C, S	OM	MM	SM
*smɔːc	N36	'ant'	*hmuac*	*chəmuac*	-	-	*həmot*
*phiːc	V111	'fear, be afraid'	*phiːc*	*phiːc*	*phic*	*phek*	*phɔc*
*smaːɲ	V119	'ask'	*hmaːɲ*	*hmaːɲ, ɛəmaːɲ*	*smāñ*	*smān*	*hman*
*kreːɲ	N25	'parrot'	*kreːɲ*	*kreːɲ*	-	-	*krɔɲ*

In the case of 'ant', only the Nyah Kur evidence allows the reconstruction with final palatal. The word does not occur in the OM or MM inscriptions, which is not surprising given its semantics. Diffloth (1984:73) explains it vas a nominal derivate of the verb root *suːc 'to sting', which is also not found in the Mon inscriptions. The verb 'fear, be afraid' shows the regular development in MM /-c/ > /-k/ after front a non-back vowel, with subsequent re-palatalization in many SM dialects.

The word for 'be full' is reconstructed by Diffloth with a velar final in proto-Monic as *piŋ (V71), although it appears with palatal in Old Mon <piñ> and all Nyah Kur varieties *piɲ*. Without going into the details of Diffloth's (1984:285-290) argumentation, the evidence suggests that both Old Mon and proto-Nyah Kur had a palatal coda in this lexeme. This is retained in Nyah Kur, as expected, but changed (back) to velar by MM as <peṅ>. Many SM varieties have a pronunciation with a final palatal nasal as *pɔɲ,* instantiating another back-shift.

Palatal finals appear to be resilient in Monic, either being conserved, as in the case of Nyah Kur, or reintroduced through new developments, as in many modern spoken Mon varieties. The re-emergence of palatal codas in Mon is a rather recent phonological development from velar finals after non-back vowels and does not cover all dialects (Jenny 2005: 264-268).

4.3 Devoicing of onsets: a two-way development

Nyah Kur and Mon underwent devoicing of originally voiced stops in the onset position (Diffloth 1984:332-341). This is part of the "devoicing wave" that swept across all of Mainland Southeast Asia, affecting most languages of all language families in the region and leading in many cases to either tone splits or two distinct registers (phonation types), usually modal in syllables with originally voiceless onsets and breathy after originally voiced onsets (Brunelle and Tạ 2021:687-690). In many cases the register contrasts are accompanied by vowel distinctions, as in Mon and Khmer (Enfield 2021:169-174), but not in Nyah Kur.

If two closely related languages like Mon and Nyah Kur have the same register system with almost perfect correspondence of registers across the shared lexicon, it would be sound to assume that the common ancestor had the same registers as the daughter languages (Diffloth 1984:333). There is enough evidence, though, to show that proto-Monic had not lost the voicing distinction and had not yet developed phonemic registers. The most obvious evidence comes from the spelling of OM, including indigenous and loan vocabulary. Also, the two languages go separate ways in their path towards devoicing. OM voiced stops become voiceless non-aspirated stops with breathy

register and some vowel changes (*e.g. /aː/ > /ɛ̯ə/), while in Nyah Kur the voiced stops either merge with the original voiceless aspirated or, in some northern dialects, with the voiceless non-aspirated stops. There is no vowel quality change involved in Nyah Kur, but originally voiced onsets regularly produce a breathy phonation of the syllable (Diffloth 1984:332-333). In both Mon and Nyah Kur, the development of vowels after originally voiced stops is the same as after sonorants. Literary Mon does not indicate the devoicing directly, but rather retains orthographic voiced initials to represent breathy register syllables. In spite of some claims to the contrary, no spoken variety of Mon retains the voiced stops. Voiced stops can be observed in connected speech, probably more frequently in speakers with good command of spoken Burmese (Jenny 2005:34).

DOM	**ID**	**Gloss**	**NK-N**	**NK-C, S**	**OM**	**MM**	**SM**
*goːʔ	V2	‘get, able’	*kọːʔ*	*khọːʔ*	*goʔ*	*goʔ*	*kɤ̯ʔ*
*dəw	V206	‘run, escape’	*tə̯w*	*thə̯w*	*dow*	*dau*	*tɛ̯ə*
*ɟəl	V237	‘fight, collide’	*ɕə̯l*	*chə̯l*	*jal*	-	*cɤ̯*

It is interesting to note that the originally voiced stops merge with the plain voiceless stops in the southern and central, as well as some northern Nyah Kur dialects, but with the voiceless aspirated stops in other northern varieties. This attests to the shallowness of this development. This makes it problematic to apply this development to the classification of languages, as Chamberlain has done for Southwestern Tai (Chamberlain 1975:50).

The devoicing of initial stops and rise of a register contrast in Nyah Kur and Mon are independent, but areally grounded processes which may or may not have occurred at the same time. Monic in this case is just another instance of a general Mainland Southeast Asian sound change, though it provides important insights into the process, especially with regard to the two-way development in Nyah Kur dialects.

4.4 The fate of fricative/affricate+sonorant onset clusters

DOM had a rich inventory of onset clusters consisting of two or three consonants. Probably only some of these were pronounced as real clusters, while others were realized with an epenthetic shwa. Of special interest is the development of clusters of the type ‘fricative/affricate+sonorant’. The only fricative occurring in this position in DOM is /s/, the palatal stops <c>, <ch>, and <j> are likely to have been pronounced as affricates [tɕ] and [dʑ]. The reflex of these in the modern Monic varieties is [tɕ], [ɕ], [tɕʰ], or similar, never as pure palatal stop [c]. Diffloth (1984:304-332) lists most possible initial clusters of his reconstructed DOM and their development in Nyah Kur and Mon. He includes ‘fricative+liquid’ (p. 307-308) and ‘stop+nasal’ (p. 308-310), but the sequence ‘fricative+nasal’ is absent from the description, although examples are found in his comparative word list. This may be due to oversight, or, at least partly, be attributed to the fact that “it [is] difficult, for example, to reconstruct *cn- vs. *sn- with any confidence” (Diffloth 1984:309), as these two clusters merge in all modern Monic varieties.

In modern Mon, the historical sequences ‘fricative/affricate+sonorant’ regularly appear as voiceless or preaspirated sonorants. This change, which is not reflected in literary Mon, probably occurred between MM and modern Mon. In modern reading pronunciation, initial /s/ and /c/ are retained and the sonorant is pronounced as voiced, with a shwa usually separating the two initials (Jenny 2005:30-31). Old Mon <smiṅ>

‘king, prince, lord’ retains its archaic spelling in literary Mon but is pronounced [hmoɲ] in most modern varieties except for reading pronunciation, which has [səmoɲ]. The sequence <sr-> is reduced to /s/ in all Mon dialects, including reading pronunciation, though it is retained in literary Mon as <sr->.

In Nyah Kur, the northern and some central dialects show the same development as Mon, changing the ‘fricative/affricate+sonorant’ clusters to voiceless/preaspirated sonorants. Some central and the southern varieties retain the initial fricative or affricate, merging them into [tɕʰ], [ɕ], or [s]. The sequence *sr- is either simplified to /ɕ/ (in northern dialects) or retained as cluster /tɕʰr-/ in central and southern dialects.

The development of voiceless sonorants in Nyah Kur and modern Mon are independent processes, similar to the rise of registers after devoicing of originally voiced initials. Unlike the latter process, voiceless sonorants are not widespread in Mainland Southeast Asia east of present-day Myanmar. Languages that used to have voiceless sonorants in earlier stages, such as all Tai varieties, generally lost them, while others, like Khmer, never had them in their onset inventory. All modern Mon varieties exhibit initials like /hl/, /hm/, /hn/, and /hɲ/, but these are rather recent additions to the sound inventory, partly due to the increased influx of Burmese loanwords, partly due to the development from earlier clusters. While Burmese influence can be claimed in the development of Mon voiceless sonorants, Nyah Kur shows that these typologically rare phonemes can arise without external influence. The following examples illustrate the development in Nyah Kur and Mon of the complex onsets.

DOM	**ID**	**Gloss**	**NK-N**	**NK-C, S**	**OM**	**MM**	**SM**
*smaːɲ	V119	‘ask’	*hmaːɲ*	*hmaːɲ, ɕəmaːɲ*	*smāñ*	*smān*	*hman*
*sŋiːʔ	N187	‘house’	*hĩːʔ, ŋhiːʔ*	*chəŋiːʔ, ɕəŋiːʔ*	*sṅi*	*sṅi*	*hɒəʔ*
*cnaːm	N235	‘year’	*hnaːm*	*hnaːm, chənaːm*	*cnām*	*cnām*	*hnam*
*srit	N13	‘rhinoceros’	-	*chrut*	*srit*	-	*sɛt*
*sruŋ	N230	‘hole, cavity’	*ɕuŋ*	*chruŋ*	*sruṅ*	-	*saŋ*
*sloːŋ	V89	‘high, tall’	*hloːŋ*	*hloːŋ, chəloːŋ*	*sluṅ*	*sluṅ*	*hlɤŋ*
*slaːʔ	N63	‘leaf’	*hlaːʔ*	*hlaːʔ, chəlaːʔ*	*sla*	*sla*	*hlaʔ*

The historical phonology of Mon and Nyah Kur, based on Diffloth’s (1984) reconstruction of DOM and the available documents in Old and Middle Mon, show several parallel, but independent developments leading to similar outcomes in the two closely related languages. This sheds important light on the possibilities and likelihoods of sound changes with and without external influence. These insights are relevant to the assessment of claimed common developments which are used to classify languages based on “shared innovations” (Harrison 2003:232-238). Nyah Kur and Mon show that innovations, though appearing identical in two languages, may not actually be shared historically, undermining classifications based on such claims. At the same time, the Mon and Nyah Kur data challenge the absolute validity of the parsimony principle (“Ockham’s razor”, see Janda and Joseph 2003:25-26) in historical and typological linguistics.

5. Morphosyntactic developments

If the research on Nyah Kur phonology and lexicon has been scarce since Diffloth and Theraphan’s pioneering work in the 1980s, investigation in Nyah Kur morphosyntax is all but nonexistent. Diffloth (1984:263-271) includes a chapter on the development of

DOM morphology. There are traces of several inherited affixes, including causative, attributive, and nominalizing but none of them are productive in Nyah Kur (or modern Mon). Synchronically, these affixes are part of the lexicon, rather than of the morphosyntax. One exception to the lack of morphosyntactic studies is Gainey (1990), who gives a brief account of causativization in Nyah Kur. The patterns found in his study, apart from the lexicalized affixes, correspond largely to the patterns found in Thai/Lao, involving the verbs *paːʔ* 'do' and *ʔuər/ʔuəl* 'give' to express (indirect) causatives. Data collected by the present author in Phetchabun province in 2008 did not reveal significant differences in sentence structure from corresponding Thai/Lao patterns, including non-contiguous serial verbs ('take-water-come' for 'bring water') where modern Mon has contiguous patterns ('take-come-water'), often with transitivity harmony (Jenny 2014). This suggests that typologically Nyah Kur is closer to Thai/Lao than to modern Mon at least in these respects. It is not clear, though, whether Mon changed its structure, probably under Burmese influence, or Nyah Kur converged with its Thai/Lao neighbors.

6. Conclusions

The Monic branch in the "real world" was born about one thousand years ago in the course of the decline of Dvāravatī. The Mon speaking population of the Chao Phraya Plain came under increasing influence of Khmer and Thai hegemony, which led to widespread loss of the Mon language in most parts of present-day Thailand. Further west Mon continued to be influential in lower Burma, with Thaton becoming (or continuing as) an important center. Thaton had possibly already earlier served as gateway to the Indian Ocan for Dvāravatī, rather than being the center of an independent Mon kingdom (Aung Thwin 2005:79-103, especially p. 89). After the intrusion of Khmer and later Thai populations into the Dvāravatī heartland, Lower Burma became the refuge for Mon language and culture, which in turn was exposed to increasing Burmese influence up to the present day. In Thailand, Mon speakers retained their language in remote and peripheral areas, away from the centers of Khmer and Thai administration, with the Nyah Kur as last known remnants of these old Mon speakers. The new geography of the former Dvāravatī area had Khmer and Thai kingdoms in the center, bordered by the Mon (recent?) kingdom of Thaton (or rather Haṁsāvatī/Pegu) in the west and largely ungoverned hills and forests in the east, where the Mon language survived. This split of DOM into two groups that lost contact with each other, namely Mon proper in the west and Nyah Kur in the east, led to the two branches of Monic still alive today. The Mon varieties are well and alive in Myanmar, while Nyah Kur is in a much weaker position, losing both ground to Thai and several of its distinctive linguistic features (Premsrirat 2002).

After a thousand years of life in hiding, the discovery of Nyah Kur and its extensive description by Diffloth and Therapan in the 1980s brought Monic to life as a legitimate branch of Austroasiatic, giving a second, academic birth to Monic. In spite of its potential interest and relevance for the history of Dvāravatī (and Siam/Thailand) and areal studies in general, Monic keeps leading a life very much in hiding and neglect by the academic communities. While the future of Nyah Kur as a spoken language (and therefore Monic as a branch of AA) is uncertain due to mostly uncontrollable social factors, the future of Monic as a field of study depends on the controllable activities of involved researchers. The material available provides a good basis for continued work, and new material can still be added in extended documentation and description projects

as long as Nyah Kur survives in the few communities it remains to be at least a heritage language.

References

Aung Thwin, Michael A. 2005. *The mists of Rāmaññа. The legend that was Lower Burma.* Honolulu: University of Hawai'i Press.

Baker, Chris and Pasuk Phongphaichit. 2017. *A history of Ayutthaya. Siam in the early modern world.* Cambridge: Cambridge University Press.

Bennett, Anna and Hunter Watson (eds.) 2020. *Defining Dvāravatī.* Chiang Mai: Silkworm Books.

Bhumadhon, Phuthorn. 2020. New knowledge on Dvāravatī. In Bennet, Anna and Hunter Watson (eds.) *Defining Dvāravatī*, 35-47.

Brunelle, Marc and Tạ Thành Tấn. 2021. Register in the languages of mainland Southeast Asia: the state of the art. In SDidwell, Paul and Mathias Jenny (eds.) *The languages and Linguistics of Mainland Southeast Asia*. Berlin/Boston: De Gruyter Mouton, 683-706.

Chamberlain, James R. 1975. A new look at the history and classification of the Tai dialects. In Harris, J. G. and James R. Chamberlain (eds.) *Studies in Tai Linguistics in Honor of William J. Gedney*. Bangkok: Central Institute of English Language, Office of State Universities, 49–60.

Champaphan, Kamphon. 2023. *Downtown Ayutthaya*. Bangkok: Matichon Books.

Diffloth, Gérard. 1980. ญัฮกุร มอญโบราณ และอาณาจักรทวารวดี (Nyah Kur, Old Mon and the kingdom of Dvāravatī) [translated by T. L. Thongkum]. Aksornsart Journal 12.1:54-85.

Diffloth, Gérard. 1984. *The Dvāravatī Old Mon language and Nyah Kur.* Monic Studies Vol. I. Bangkok: Chulalongkorn University Printing House.

Duroiselle, Chas. and Taw Sein Ko. 1919 onwards. *Epigraphia Birmanica. Being lithic and other inscriptions of Burma*. Rangoon: Government Printing.

Ferlus, Michel. 1984. Essai de phonétique historique du môn. *Mon-Khmer Studies 12:1-90.*

Gainey, Jaroon. 1990. Causativization in Nyah Kur. *Mon-Khmer Studies* 16-17:25-30.

Guillon, Emmanuel. 1999. *The Mons. A civilization of Southeast Asia.* (Translated and edited by James V. Di Crocco). Bangkok: The Siam Society.

Guy, John. 2020. Making sense of Dvāravatī. In Bennet, Anna and Hunter Watson (eds.) *Defining Dvāravatī*, 48-63.

Harrison, S. P. 2003. On the limits of the comparative method. In Joseph, Brian D. and Richard D. Janda (eds.) *The handbook of historical linguistics.* Malden: Blackwell Publishing, 213-243.

Huffman, Franklin. 1971. *Unpublished vocabulary lists.* 495.9 Mon-Khmer, Austroasiatic general folder. David Thomas Library, Bangkok.

Huffman, Franklin. 1990. Burmese Mon, Thai Mon, and Nyah Kur: a synchronic comparison. *Mon-Khmer Studies*:31-84.

Janda, Richard D. and Brian D. Joseph. On language, change, and language change. In Joseph, Brian D. and Richard D. Janda (eds.) *The handbook of historical linguistics.* Malden: Blackwell Publishing, 3-180.

Jenny, Mathias. 2014. Transitivity and affectedness in Mon. *Journal of Mon-Khmer Studies* 43(1): 57-71.

Jenny, Mathias. 2012. The Mon language: recipient band donor between Burmese and Thai. *Journal of Language and Culture* vol. 31.2: 5-33.

Jenny, Mathias. 2005. *The verb system of Mon.* Zurich: ASAS.

McCormick, Patrick and Mathias Jenny. 2013. Contact and convergence: the Mon language in Burma and Thailand. *Cahiers de Linguistique - Asie Orientale* 42(2):77-116.

Moore, Elizabeth H. 2023. *Wider Bagan. Ancient and living Buddhist tradition.* Singapore: ISEAS Yusof Ishak Institute.

Nai Pan Hla. 1986. remnant of a lost nation & their cognate words to Old Mon epigraph. *Journal of the Siam Society* vol. 74:122-155.

Ongsakul, Sarassawadee. 2005. *History of Lan Na*. (Translated by Chitraporn Tanratanakul) Chiang Mai: Silkworm Books.

Premsrirat, Suwilai. 2002. *The future of Nyah Kur.* In Bauer, Robert S. (ed.) *Collected papers on Southeast Asian and Pacific languages.* Canberra: Pacific Linguistics, 155-165.

Seidenfaden, Erik. 1918. Some notes on the Chaubun. *Journal of the Siam Society* 12/3:1-11.

Shorto, H. L. 1971. A dictionary of the Mon inscriptions from the sixth to the sixteenth centuries. London: Oxford University Press.

Skilling, Peter. 2020. Dvāravatī in inscriptions and manuscripts. In Bennet, Anna and Hunter Watson (eds.) *Defining Dvāravatī*, 64-82.

Thomas, David and Robert K. Headley jr. 1970. More on Mon-Khmer subgroupings. *Lingua* 25:398-418.

Thongkum, Theraphan L. 1984. *Nyah Kur (Chao Bon) - Thai - English dictionary.* Monic Studies Vol. II. Bangkok: Chulalongkorn University Printing House.

Watson, Hunter Ian. 2020. Old Mon inscriptions and the extent of Dvāravatī. In Bennet, Anna and Hunter Watson (eds.) *Defining Dvāravatī*, 83-94.

Wongsathit, U-tain, Kangvol Kathsima, and Chatuphon Khotkanok. 2020. The fall of Dvāravatī as mentioned in the Khmer inscription K.1198. In Bennet, Anna and Hunter Watson (eds.) *Defining Dvāravatī*, 137-143.

Wyatt, David K. 2003. *Thailand. A short history.* 2nd edition. Chiang Mai: Silkworm Books.

13

The Agreement - Word Order Correlation in Khasi

Saralin A. Lyngdoh, Rymphang K. Rynjah

Abstract

The Khasi language is an Austroasiatic language spoken in the state of Meghalaya, India, and it follows a syntactic agreement-word order correlation that is a relatively rare phenomenon in languages. The syntax of agreement appears to suggest patterns of word order and changes that occurred over time, eventually resulting in the basic word order that exists today. By comparing the word orders of Khasi to its varieties, it is possible to observe these changes and how they led to the current word order. This suggests that the Khasi language has evolved and adapted over time, which is typical of languages to meet the changing needs of its speakers. This dynamic adaptation highlights the Khasi language as a living, changing system.

1 Introduction

The Khasi language, spoken in the northeastern region of India, primarily by the indigenous Khasis, has been extensively studied but still requires proper reconstruction of its origin and migration. It has been classified within the Austroasiatic language family, as demonstrated by the works of K.S. Nagaraja (1977), Gerard Diffloth (2005), Paul Sidwell (2011, 2018), Anne Daladier (2011), and others, which have significantly contributed to its classification and understanding of its historical development. These studies emphasize the importance of reconstructing the language's origin and migration patterns to better understand its current structure. Figure 1 illustrates a model of the genetic classification of Khasi varieties/Austroasiatic languages of Meghalaya, while Figure 2 shows their approximate geographical distribution. Within Austroasiatic, Khasi is most closely related to the Palaungic languages of Myanmar and neighboring regions (Sidwell 2021:181-182).

The Khasi language also needs re-standardization, a formal corpus, a grammar book, and an advanced dictionary. Ongoing projects such as compiling comprehensive Khasi dictionaries, linguistic classification of Khasi varieties based in regional dialects and preservation and resource building for various Natural Language Processing Applications of low resource North-Eastern Languages aim to address these challenges.

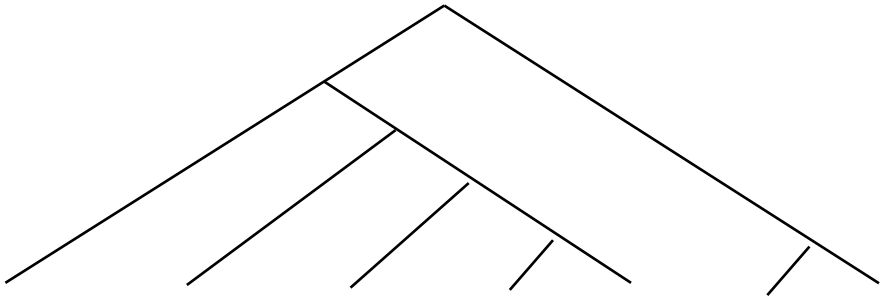


Figure 1. Relations between Khasi varieties/Austroasiatic languages of Meghalaya (based on phylogram at Sidwell 2018:30).

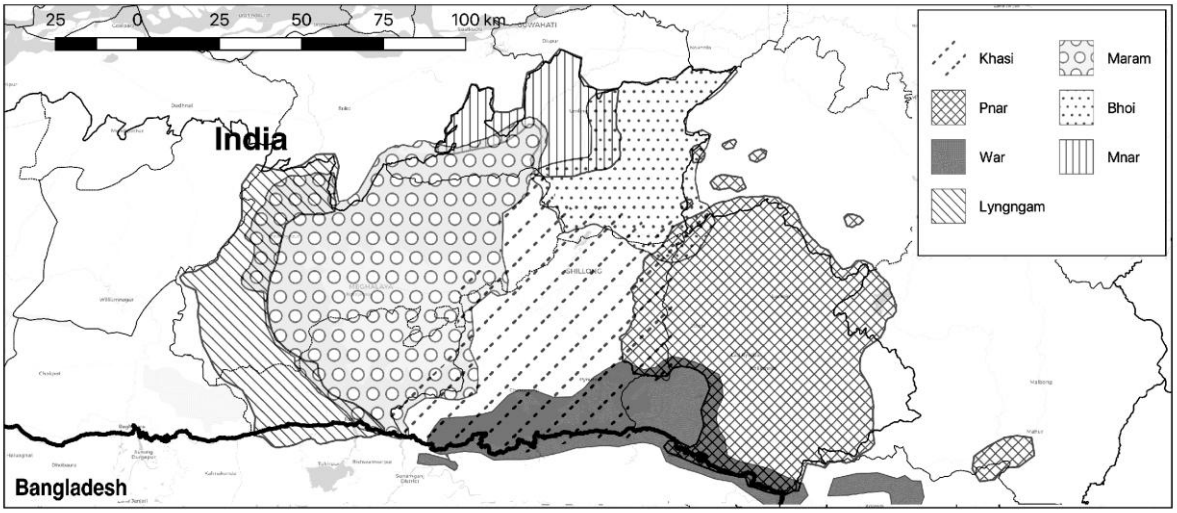


Figure 2. Map of Austroasiatic languages of Meghalaya (Hiram Ring, Creative Commons Attribution).

2 Literature Review

The examination of word order in languages greatly enriches our understanding of grammar and syntax. Steele (1978) emphasizes that basic word order and its variations are key parameters in the study of language universals. Investigating word order variation involves examining the possible boundaries within which word order can vary at the surface level. This includes looking at the conditions under which variations occur, such as case-marking on nominals and subject agreement marking on verbs. Word order variations arise from evolutionary pressures, historical trajectories, and contact-induced influences, shaping preferences for syntactic closeness or informativeness of elements (Hahn and Yu, 2022).

Language structure and communication are influenced by various perspectives. Cognitive theories, as discussed by Hawkins (2004) and Culicover (2009), emphasize the role of cognitive efficiency and ease of processing in shaping language patterns like word order and agreement. Functionalist and typological approaches, advocated by Givón (1995) and Croft (2002), argue that syntactic structures evolve to meet specific communicative needs. Sociolinguistic factors, highlighted by Labov (2001) and Milroy (1992), demonstrate that social interactions, language contact, and community norms drive language change and variation. Pragmatic motivations, explored by Levinson

(2000) and Sperber and Wilson (1995), influence syntactic choices to achieve clarity, emphasis, and efficiency in communication. These perspectives collectively provide a comprehensive understanding of how cognitive, functional, social, and pragmatic factors shape language structure and use.

According to Lyngdoh (2012), the syntactic structure of the Khasi language, particularly its agreement markers, is crucial for understanding its grammar. Khasi features concordial agreement where pronominal markers in the subject noun phrase (NP) are repeated in the verb phrase (VP), serving multiple roles such as articles and determiners. The 'Agr' marker is a separate constituent in the syntactic tree, distinct from the verb, with elements like tense and negation cliticizing onto it. Strong Agr triggers Determiner Phrase (DP) movement to satisfy the Extended Projection Principle (EPP), resulting in a pre-verbal subject position, while weak Agr allows for a post-verbal subject position and VSO word order. In passive constructions, focus on the verb can lead to head movement across the subject, resulting in VSO patterns. These insights show that Khasi permits alternative word orders like VSO due to discourse effects, with strong Agr licensing overt DP movement, highlighting the significant role of agreement markers in Khasi syntax.

Research by Rynjah and Lyngdoh (2022) explores the variations and underlying principles of word order in Khasi, influenced by both syntactic and functional factors. Their study compares word order variations in Standard Khasi and its varieties, including Pnar, War-Khasi, and War-Jaiñtia. They highlight the SVO (subject-verb-object) basic word order and the variations occurring in informal and colloquial speech. The study reveals that post-verbal subject constructions and verb-subject constructions are prevalent in these varieties but can be ungrammatical in Standard Khasi, especially with pronominal subjects. These constructions are used to focus on the topic, marking a distinct departure from the standard language.

Bedell (2011) specifically examines how verbs in Khasi show agreement with their subjects and how pronominal clitics serve as both pronouns and agreement markers. He suggests that in Khasi, subject agreement markers are strictly ordered and precede the verb, forming a set of verbal markers that are strictly ordered among themselves. The agreement clitic, if present, is the first of these markers, indicating a correlation between word order and agreement in Khasi syntax. Bedell argues against the analysis that preverbal pronominal clitics are pronoun subjects, instead presenting them as agreement markers, which is a significant aspect of the syntactic structure in Khasi.

Standard Khasi primarily exhibits an SVO word order, but variations such as VSO and VOS are also observed. Understanding these variations is crucial for effective communication in Standard Khasi, as deviations from the SVO structure can alter the intended meaning of sentences. The syntactic and morphological features of phrases in different Khasi varieties, such as Pnar, War-Khasi and War-Jaiñtia, contribute to the complexities in communication within the language (Rynjah and Lyngdoh, 2022). Recognizing and applying these word order variations accurately allows speakers to convey their messages clearly and ensure effective communication.

2 Word Order Patterning

The patterning of word order in Khasi and its varieties demonstrates an underlying structure that is not merely a matter of preference or convenience. Aside from the basic word order, as shown in example (1), Khasi exhibits a variety of syntactic features, such

as object agreement (2), verb-initial constructions without the subject (3), and post-verbal subjects (4). This structuring suggests that Khasi is organized in a way that makes it easy for users to comprehend its syntax, as well as the syntax of related varieties. Understanding these commonalities helps us gain a broader understanding of language use across different linguistic contexts.

Khasi:

(1) *ka* *meri* *ka* *ieːd* *ya* *u* *jon*
3S.FEM Meri 3S.FEM love ACC 3S.MASC Jon
'Meri loves Jon'

Pnar:

(2) *ka* *meri* *[maya* *ko* *o]* *u* *jon*
3S.FEM Meri [love 3S.FEM 3S.MASC] 3S.MASC Jon
'Meri loves Jon'

Pnar:

(3) *[dat* *u* *o]* *u* *ksew*
[hit 3S.MASC 3S.MASC] 3S.MASC dog
'He hit the dog'

Khasi:

(4) *i* *bang* *bha* *[i* *Abby]* *ya* *ka* *shriew*
3S.DIM taste INTS 3S.DIM Abby ACC 3S.FEM yam
'Abby really likes Yam'

3 Agreement-word order correlations

3.1 Pronominal Clitics

Khasi has pronominal clitics that function as agreement markers and are linked to lexical categories, such as verbs (5), adjectives (6), and functional categories like copula (7) and modals (8).

Khasi:

(5) *u* *jon* *u* *thiah*
3S.MASC Jon 3S.MASC sleep
'Jon sleeps'

(6) *u* *jon* *u* *jrong*
3S.MASC Jon 3S.MASC tall
'Jon is tall'

(7) *u* *jon* *u* *long* *u-ba* *jemnud*
3S.MASC Jon 3S.MASC COP 3S.MASC-REL gentle
'Jon is gentle'

(8) *u* *jon* *u* *lah* *ba-n* *kɨnthih*
3S.MASC Jon 3S.MASC can COMP-FUT jump
'Jon can jump'

These pronominal clitics also host other functional categories, such as negation (9), tense (10), nominalizers (11), complementizers (12), deictics (13), and question particles (14).

Khasi:

(9) *u-m* *wan* *mɨnta* *ka* *sngi*
3S.MASC-NEG come now 3S.FEM day
'He won't come today'

(10) *u-n* *wan* *lashai*
3S.MASC-FUT come tomorrow
'He won't come tomorrow'

(11) *u-ba* *jrong* (Nominalization-complementation overlap)
3S.MASC-COMP tall
'who/that is tall'

(12) *u-ba* *khot* *ya* *nga*
3S.MASC-COMP call ACC 1SG
'one who calls me'

(13) *u-ne*
3S.MASC-DEM
'This'

(14) *u-no*
3S.MASC-Q
'Which one'

These clitics also function as strong pronouns in object positions, as in (15) and (16).

Khasi:

(15) *u* *isih* *ya* *u*
3S.MASC hate ACC 3S.MASC
'He hates him'

(16) *u* *ieːd* *ya* *nga*
3S.MASC love ACC 1SG
'He loves me'

3.2 Agreement

The syntax of agreement suggests patterns of word order and historical changes that culminated in the current basic word order. Khasi, an isolating language, features consistent null subjects and stand alone words within sentences, as shown in examples (17) and (18).

Khasi:

(17) *u* *jon* *u* *long* *u-ba* *bha*
3S.MASC Jon 3S.MASC COP 3S.MASC-NMLZ good
'Jon is a good person'

(18) *u* *long* *u-ba* *bha*
3S.MASC COP 3S.MASC-NMLZ good
'He is a good person'

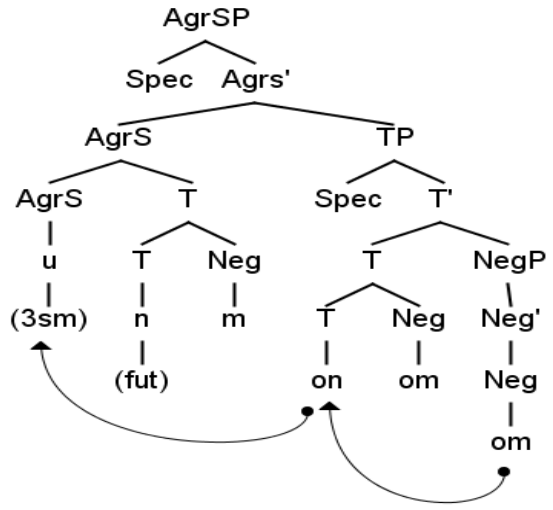
3.3 Occurrences of heads

The pattern of occurrences of all the heads is consistent:

- Agr-T(ense)-Neg(ation)
- Agr-Neg(ation)
  - Agr-T(ense)
  - Agr-Neg(ation) -T(ense)

When cliticizing onto Agr-S(ubject), both Neg head and T head drop their initial vowels, resulting in the complex Agr head being phonologically pronounced as [unim]. This complex head is syntactically integrated with the Neg head, as in (19).

(19) Example illustrating the phonological integration of Agr and Neg:



Other varieties of Khasi, however, do not exhibit this morphological complexity. In these varieties, subject agreement occurs after the verb phrase, known as post-subject agreement, as seen in examples (20) through (22). This phenomenon is absent in Standard Khasi.

Pnar:

(20) *u* *jon* *[daw* *thiah* *u]*
3S.MASC Jon [FUT sleep 3S.MASC]
'Jon will sleep'

(21) *daw [thiah u] u jon*
FUT [sleep 3S.MASC] 3S.MASC Jon
'Jon will sleep'

(22) *u jon [daw ym thiah u]*
3S.MASC Jon [FUT NEG sleep 3S.MASC]
'Jon will not sleep'

Most Khasi varieties show co-occurrences of pre-verbal and post-verbal Agr, indicating split agreement, supported by evidence including word order preferences and case marking patterns in pre-verbal and post-verbal Agr, as shown in examples (23) through (26).

Several examples are drawn from the Umñiuh and Mawlong varieties in various analyses. For clarity, Umñiuh and Mawlong are varieties of War-Khasi, as discussed in Rynjah's (2020) work "*War-Khasi and War-Jaiñtia: A Comparative Syntactic Study*". Rynjah's research emphasizes the syntactic characteristics of War-Khasi and War-Jaiñtia, identifying Umñiuh and Mawlong as two specific areas of study within the War-Khasi dialects.

Pnar:

(23) *ka thiah ko*
3S.FEM sleep 3S.FEM
'she is sleeping'

(24) *o [dat u o] u ksew*
3S.MASC [beat 3S.MASC 3S.MASC] 3S.MASC dog
'He beats the dog'

Lamin (War-Jaiñtia):

(25) *e thieh ka*
3S.FEM sleep 3S.FEM
'She is sleeping'

Umñiuh (War-Khasi):

(26) *ka thiah ka*
3S.FEM sleep 3S.FEM
'She is sleeping'

In informal contexts, split agreement in Khasi is especially evident when emphasizing the subject or topic, as shown in examples (27) and (28).

Khasi:

(27) *phi shim phi ka sopti jong nga*
2SG take 2SG 3S.FEM shirt GEN 1SG
'You took my shirt'

(28) *nga-n shim nga ka patlun*
1SG-FUT take 1SG 3S.FEM pants
'I will take the pants'

4 Historical Analyses and Theoretical Insights

4.1 Determiner Phrase (DP) Movement

Research on DP movement in Khasi, Pnar, and other varieties shows that strong agreement correlates with word order patterns. The strength of Agr in derivations motivates DP movement, supported by analyses of raising, passive, and post-verbal constructions. These types of constructions allow VSO word order as an alternative in spoken language. When V moves across subjects for focus, it loses its Agr marker. Merging Agr is critical when subjects move across VP for Case and strong D-feature checking.

4.1.1 Passive constructions

In Standard Khasi, the formal word order in passive constructions is SVO. Focus on the verb motivates V movement across the subject, resulting in VSO pattern. Example (29) shows the unmarked SVO pattern with overt Agr markers, while example (30) shows the VSO pattern with focus shift, and example (31) shows that Agr on the verb is not allowed explicitly in VSO pattern.

Khasi:

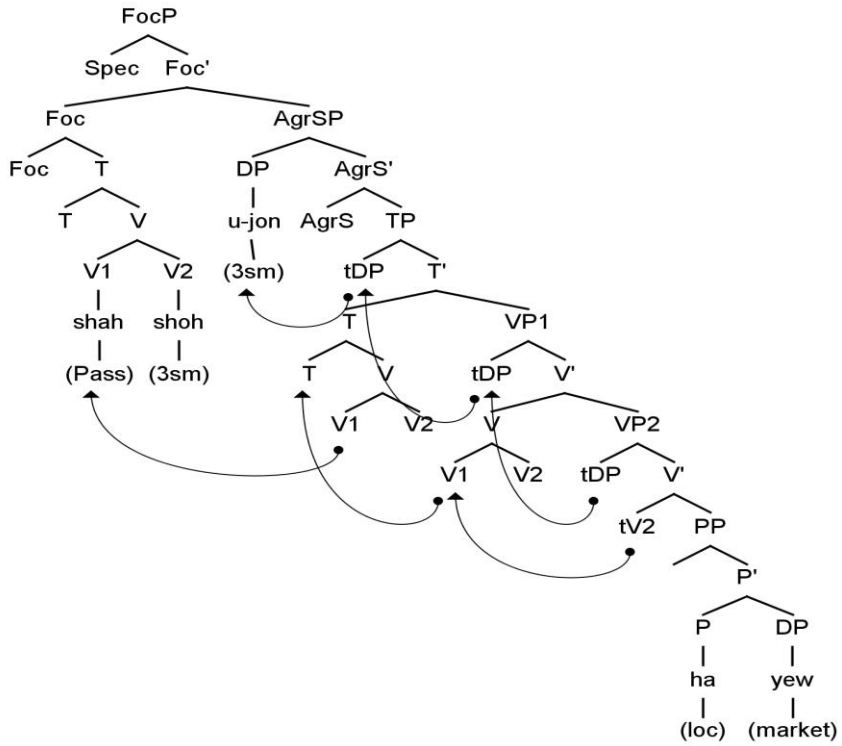
(29) *u* *jon* *u* *shah* *shoh* *ha* *yew*
3S.MASC Jon 3S.MASC PASS beat LOC market
'Jon was beaten in the market'

(30) *shah* *shoh* *u* *jon* *ha* *yew*
PASS beat 3S.MASC Jon LOC market
'Jon was beaten in the market'

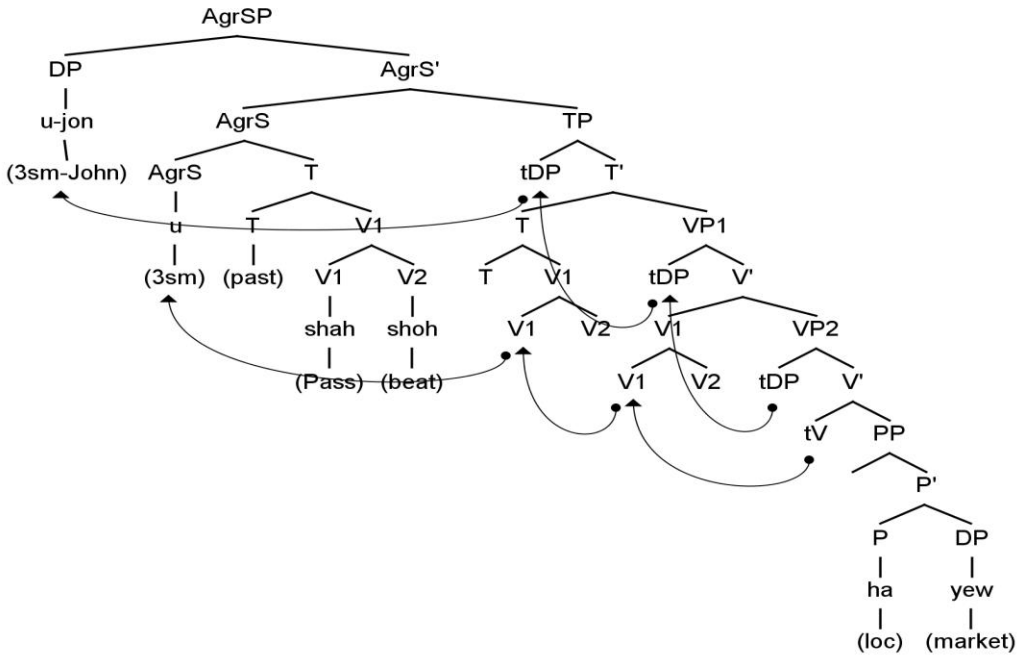
(31) *u* *jon* *u* *shah* *shoh* *ha* *yew*
3S.MASC Jon 3S.MASC PASS beat LOC market
'Jon was beaten in the market'

When Agr is merged on top of VP, as shown in examples (32) and (33), DP is forced to move out of its post-verbal position due to strong Agr.

(32) Example illustrating Focus movement in passive construction



(33) Example illustrating DP-movement in passive construction



4.1.2 Raising constructions

In Khasi, the subject is generated in one clause before moving to the subject position of a higher clause, known as "raising". This allows the subject to originate in a non-finite clause and be raised, as shown in examples (34) and (35). Example (36) is ungrammatical.

Khasi:

(34) *u* *jon* *imat* *u* *pang* *jur*
3S.MASC Jon seem 3S.MASC ill serious
'Jon seems to be seriously ill'

(35) **u* *jon* *u* *imat* *u* *pang* *jur*
3S.MASC Jon 3S.MASC seem 3S.MASC ill serious
'*Jon seems to be seriously ill'

(36) *imat* *u* *jon* *u* *pang* *jur*
seem 3S.MASC Jon 3S.MASC ill serious
'Jon seems to be seriously ill'

4.1.3 Post-verbal subject construction

Post-verbal subject constructions are common in Khasi and its varieties, especially in conversational style, allowing focus shifts to new topics, as shown in examples (37) through (39).

Khasi:

(37) *u* *wan* *u* *jon* *mynnin*
3S.MASC come 3S.MASC Jon yesterday
'Jon came yesterday'

Umñiuh (War-Khasi):

(38) *bam* *ja* *nga*
eat rice 1SG
'I am eating food'

Mawlong (War-Khasi):

(39) *sa* *jia* *nga*
eat rice 1SG
'I am eating food'

Evidence of post-verbal subjects in Khasi includes the nominative case marker with the subject pronominal clitic after the verb, emphasizing focus on the subject, as in example (40).

Khasi:

(40) *nga* *shim* ***ma-nga*** *ka* *kot* *jong* *phi* (emphasis)
1SG take NOM-1SG 3S.FEM book GEN 2SG
'(It was) I (who) took your book'

In conclusion, in all the data sets supporting a separate Agr head, one common syntactic behavior emerges: strong, interpretable Agr licenses an overt DP above it or attracts the DP to move to Spec AgrP obligatorily. Conversely, weak, uninterpretable Agr allows Spec AgrP to remain phonologically silent or morphologically empty. The necessity of strong, interpretable Agr for Spec AgrP licensing emphasizes its significant role in the syntax of the Khasi language.

4.2 Theoretical Accounts on Word Order and Agreement

Agr is a separate constituent, positioned on a separate node in the syntactic tree and detached from the verb. Evidence includes the occurrence of T(ense), Neg(ation) and Adv(erbs) between Agr and the verb.

When Agr is strong, tense and negation are suffixed to it. Morphologically and phonologically, only tense and negation particles are hosted by the Agr head in the syntactic tree. This cliticization process produces a phonological effect on the complex Agr head, as seen in the following illustrations (41)through (44).

Khasi:

(41) *u-n-m* *bam*
3S.MASC-FUT-NEG eat
'He will not eat'

(42) **u-m-n* *bam*
3S.MASC-NEG-FUT eat
'*He will not eat'

(43) *u* *ju-bam*
3S.MASC habitual-eat
'He habitually eats'

(44) *u-m* *ju-bam*
3S.MASC-NEG habitual-eat
'He habitually does not eat'

4.3 Word Order Variation

The word order of Khasi varieties differs from Standard Khasi, ranging from VSO (46, 47) to VOS (48) compared to the SVO order in Khasi (45). These variations from the standard variety do not seem to be solely attributable to language contact, but this verb-initial or predicate-initial order was likely inherited from the proto-language and retained in some languages due to linguistic and societal factors (Jenny 2015, 2020).

Khasi: SVO

(45) *nga* *ai* *u* *let* *ha* *u* *paralok*
1SG give 3S.MASC pencil DAT 3S.MASC friend
'I give the pencil to my friend'

Umñiuh (War-Khasi): VSO

(46) *ai nga u let ha u parelok*
give 1SG 3S.MASC pencil DAT 3S.MASC friend
'I give the pencil to my friend'

Lamin (War-Jaiñtia): VSO

(47) *a: nge u let he u periulok nge*
give 1SG 3S.MASC pencil DAT 3S.MASC friend 1SG.ACC
'I give the pencil to my friend'

Mawlong (War-Khasi): VOS

(48) *ai let nga yah u fralok*
give pencil 1SG DAT 3S.MASC friend
'I give the pencil to my friend'

4.4. Theoretical insights and further proposal

4.4.1 Verb Movement in Derivations

According to Bobaljik (1995), V in situ languages combine Agr and tense heads into a single inflectional phrase, while V-to-I raising languages have separate Agr and tense heads projecting two inflectional phrases. Khasi allows morphological fusion of Agr, Tense, and Neg, whereas other varieties show morphological isolation.

4.4.2 Passive Construction Uncertainty

In Khasi passive constructions, it remains uncertain whether passive markers move from their original position to satisfy syntactic requirements, or if the original structure is preserved, as shown in examples (49) through (50).

Khasi:

(49) *ki khynnah ki shah shoh*
3PL child 3PL PASS beat
'The children are beaten'

(50) *ki shah shoh ki khynnah*
3PL PASS beat 3PL child
'The children are beaten'

(51) *shah shoh ki khynnah*
PASS beat 3PL child
'The children are beaten'

Conclusion

The investigation into the agreement-word order correlation in Khasi has illuminated several key syntactic features and their implications for understanding the language's structure and evolution. The basic word order of Standard Khasi is SVO. However, it is worth questioning whether this is truly the case in all contexts. All varieties of Khasi, including Standard Khasi, commonly use VSO in actual communication but still claim SVO as the basic order in formal situations or literature, aligning with the standard norm

of Standard Khasi. This raises the question of whether the SVO word order in Khasi is an innovation.

The study reveals that pronominal clitics in Khasi serve multifunctional roles, acting as agreement markers linked to verbs, adjectives, and other functional categories. These clitics also interact with tense, negation, and other particles, creating complex Agr heads that influence word order and syntactic behavior. The presence of post-verbal subject constructions and split agreement patterns in Khasi varieties further demonstrates the dynamic nature of its syntax.

Post-VP agreement marking in almost all Khasi varieties suggests that subjects are base-generated after verbs. The V-to-I movement is blocked by the Agr head and the Tense or Neg head, causing the verb to remain in situ. Consequently, the SVO word order is derived by moving subjects across verbs. This syntactic behavior emphasizes the importance of strong, interpretable Agr in licensing overt DPs and attracting DP movement to Spec AgrP.

Historical and theoretical analyses suggest that the evolution of word order in Khasi is driven by both internal linguistic factors and external influences from neighboring languages. The strong interpretable Agr in Khasi not only licenses DP movement but also maintains syntactic integrity and coherence.

Overall, this study contributes to our understanding of the intricate relationship between agreement and word order in Khasi. By highlighting the language's unique syntactic features and their functional motivations, the research provides valuable insights into the broader principles governing language structure and change. Future research could further explore the interactions between syntactic, morphological, and phonological elements in Khasi and its varieties, offering a deeper understanding of the language's evolution and its place within the Austroasiatic family.

References

Bedell, George. 2011. *The Syntax of Agreement in Khasi*. Language in India, 11(4).

Bobaljik, Jonathan. D. 1995. *Morphosyntax: The syntax of verbal inflection.* (Doctoral dissertation, Massachusetts Institute of Technology).

Croft, William. 2002. *Typology and universals*. Cambridge University Press.

Culicover, Peter. W. 2009. *Natural language syntax*. OUP Oxford.

Daladier, Anne. 2011. *The Group Pnaric-War-Lyngngam and Khasi as a Branch of Pnaric*. Journal of the Southeast Asian Linguistics Society 4(2). 169-206.

Givón, Talmy. 1995. *Functionalism and grammar*. John Benjamins Publishing.

Hahn, Michael., & Xu, Yang. 2022. *Crosslinguistic word order variation reflects evolutionary pressures of dependency and information locality*. Proceedings of the National Academy of Sciences, 119(24), e2122604119.

Hawkins, John. A. 2004. *Efficiency and complexity in grammars*. OUP Oxford.

Jenny, Mathias. 2020. Verb-Initial Structures in Austroasiatic Languages. In *Austroasiatic syntax in areal and diachronic perspective.* Edited by Mathias Jenny, Paul Sidwell, and Mark Alves. Leiden/Boston: Brill, 21-45.

Jenny, Mathias. 2015. Syntactic diversity and change in Austroasiatic languages. In *Perspectives on Historical Syntax.* Edited by Carlotta Viti. Amsterdam: John Benjamins, 317-340.

Labov, William. 2001. *Principles of Linguistic Change,* Volume II: Social Factors. Social factors. 2.

Levinson, Stephen. C. 2000. *Presumptive meanings: The theory of generalized conversational implicature*. MIT press.

Lyngdoh, Saralin. A. 2012. *'Empty Catagories' in Khasi*. (Unpublished PhD Thesis). Delhi University.

Lyngdoh, Saralin. A. 2017. *DP Movement in Khasi* in Fabric of Indian Linguistics. Lakshi Publishers & Distributors. New Delhi

Milroy, James. 1992. *Linguistic variation and change: On the historical sociolinguistics of English*.

Nagaraja, K.S. 1977. *A descriptive analysis of Khasi*. Poona: Deccan College dissertation. (vi+321pp.)

Nagaraja, K.S. 2014. Standard Khasi. Austroasiatic Comparative-Historical Reconstruction: an overview. In Mathias Jenny & Paul Sidwell (eds.) *The handbook of Austroasiatic languages*. Leiden, Boston: Brill. pp.1145-1185.

Rynjah, Rymphang. K. 2020. War-Khasi and War-Jaiñtia: A Comparative Syntactic Study. Unpublished PhD Thesis. North-Eastern Hill University (NEHU).

Rynjah, Rymphang. K., & Lyngdoh, Saralin. A. 2022. *Word Order in Standard Khasi and its Varieties: A Comparative study of Change and Variation.* Quest Journals: Journal of Research in Humanities and Social Science Volume 10. Issue 12 (2022) pp: 279-287 ISSN(Online):2321-9467.

Rynjah, Rymphang. K., & Lyngdoh, Saralin. A. 2023. Cross-Linguistic Comparisons of Noun Phrase Constructions in Khasi Varieties. *Indian Journal of Languages and Linguistics*, doi: 10.54392/ijll2325.

Sidwell, Paul. 2011. *Proto-Khasian and Khasi-Palaungic*. Journal of the Southeast Asian Linguistics Society 4(2). 144-168.

Sidwell, Paul. 2018. *The Khasian Languages: Classification, Reconstruction and Comparative Lexicon*. Munich, Lincom Europa.

Sidwell, Paul. 2021. Classification of MSEA Austroasiatic languages. In Paul Sidwell & Mathias Jenny (eds.) *The Languages and Linguistics of Mainland Southeast Asia: A Comprehensive Guide*. Walter de Gruyter: Berlin/Boston. pp.179-206.

Sperber, Dan., & Wilson, Deirdre. 1986. *Relevance: Communication and cognition* (Vol. 142). Cambridge, MA: Harvard University Press.

Steele, Susan. 1978. *Word order variation: A typological study. In Universals of human language*, Vol.4, Syntax. ed. by Joseph H Greenberg, 585-624. California: Stanford University Press.

14

On the Semantics of Gender Assignment in Khasi

Umarani Pappuswamy

1 Introduction

The study of noun categorisation is a fundamental aspect of linguistic research, providing crucial insights into how languages systematically organise and classify nouns. Among the various mechanisms employed by languages to achieve this organisation, gender stands out as particularly insightful. Grammatical gender reflects how native speakers perceive and categorise the objects and concepts in their world. This paper focuses on the gender system of Khasi, an Austroasiatic language spoken primarily in the northeastern Indian state of Meghalaya.

The core objective of this research is to explore the semantic foundations underlying the assignment of gender in Khasi. Specifically, this study examines the gender system through the analysis and semantic classification of approximately 5000 nouns[1]. Khasi features a tripartite gender system, with *u* designating masculine nouns, *ka* marking feminine nouns, and *i* assigned to common nouns. Examples include *u dohlap* 'pancreas' (masculine), *ka pyrthei* 'earth' (feminine), and *i khyllung* 'baby' (common gender).

Although the gender assignment in Khasi demonstrates a discernible semantic core, the rules governing this system are complex and not entirely consistent, leaving many nouns unaccounted for. For instance, the gender of humans is typically assigned based on biological sex, resulting in straightforward categorisation. However, the semantic rules for other nouns are less transparent. A notable example of this complexity is found in the categorisation of fruits, where most fruits are assigned masculine gender except for the banana, which is feminine, suggesting a unique cultural perception.

This study seeks to systematically classify Khasi nouns by examining both 'natural scientific' principles, which rely on prototypical properties of the noun classes, and 'socio-cultural' elements that influence gender assignment.

This paper is organised as follows: Section 2 provides an overview of the background and theoretical framework, focusing on the noun classification system with reference to gender/noun class. Section 3 outlines gender assignment in Austroasiatic languages in general and Khasi in particular. Section 4 presents the Khasi gender system, highlighting the semantic patterns and exceptions in gender assignment. Finally, Section 5 concludes the paper by summarising the key insights of this study.

[1] The database of nouns and associated morphological features are a subset of a Khasi-English online dictionary under development by the author maintained in FieldWorks Language Explorer.

2 Background and Theoretical Framework

2.1 Why Gender/Noun Class?

'Gender' derives etymologically from Latin *genus*, via Old French *gendre*, and originally meant 'kind' or 'sort'. The terminology used to describe these systems can often be misleading. Corbett (1991, 2005) uses 'gender' as a cover term for agreement classes, while Evans (1994) prefers 'noun class.' Additionally, the term 'concordial classes' is used by many linguists. For the purpose of this study, I prefer to use the term 'gender' for small systems of two to three distinctions, which always include masculine and feminine categories.

Gender or noun class systems is one prevalent type of nominal classification system which obligatorily categorises all nouns into distinct groups (Allan, 1977; Dixon, 1986). They represent fascinating categories indeed within linguistic studies, being central in some languages while completely absent in others. Initially, the concept of gender was biologically based, distinguishing between males and females. Over time, however, it expanded to include sexless objects through associations made in myth or religion. These linguistic features serve as mechanisms for identifying and differentiating nouns, effectively creating various categories of "its."

Gender/noun class systems are particularly intriguing because they offer valuable insights into the structure of the human cognitive system and the evolution of linguistic complexity. Unlike arbitrary classifications, the categories within these systems are systematically organised based on meaningful distinctions. Craig (1986) offers additional perspectives by exploring the dynamic nature of these classifications. Craig emphasises that gender systems are not static; they evolve with changing cultural and social contexts. Her work illustrates how shifts in societal norms and values can lead to corresponding changes in the gender categories used within a language. This adaptability highlights the responsiveness of language to cultural changes, reinforcing the idea that gender systems are a living, evolving component of linguistic structure. Corbett (1991) provides a detailed typological analysis of gender systems across languages, offering insights into the universal and variable aspects of these systems. His work complements the cognitive and cultural perspectives by showing how gender systems fit into broader linguistic patterns. Senft (2000) explores noun classification in Austronesian and Papuan languages, illustrating the role of local cultural and environmental factors in shaping these systems. Similarly, Seifart (2010) provides evidence from Amazonian languages on the ecological and cultural influences on gender systems. Aikhenvald (2016) provides crucial insights into how cultural narratives and societal values shape gender systems. She highlights that these classifications are not just linguistic phenomena but are deeply intertwined with cultural identity and social structure.

These influences reflect the observation that the affiliation of nouns to gender categories is far from arbitrary. Research shows that these affiliations are systematically based on cognitive salience and cultural relevance. For instance, common distinctions include animate versus inanimate, human versus non-human, animal versus non-animal, and male versus female. These categories reflect fundamental aspects of how humans perceive and interact with the world. Additionally, gender systems often identify specific shapes and sizes, such as long versus round and big versus small. These patterns align with the neuroscientific premise that certain categories are more prominent and salient in the human cognitive system, making them more likely to be mirrored in

human communication systems (Kemmerer, 2017).

Furthermore, the categories found in gender systems are influenced by cognitive and cultural biases. For example, shapes like "long" and "round" are expected to be more common because they are significant and easily recognisable within human cognition. Cultural factors also play a role, as societies may emphasise certain categories based on their unique cultural narratives and practices (Aikhenvald, 2016). For example, studies have demonstrated that certain shape features are more likely to appear in gender systems because they are cognitively significant (Veeman et al. 2020; Basirat et al. 2021). This systematic nature of gender systems highlights their importance in understanding the interactions between language, cognition, and culture.

Understanding gender or noun class systems is crucial for several reasons. First, these systems provide insight into how different languages categorise the world, revealing underlying cognitive and cultural processes. Gender systems often reflect societal norms and values, as well as historical and mythological influences. By studying gender assignment, linguists can gain a deeper understanding of the interplay between language, thought, and culture. Additionally, gender systems play a significant role in grammatical structure and linguistic agreement. They affect how nouns interact with other parts of speech, such as adjectives, verbs, and pronouns. This interaction is essential for the coherence and cohesion of sentences, making gender a fundamental aspect of syntax and morphology. Moreover, documenting and analysing gender systems, especially in lesser-studied languages like Khasi, contributes to the broader field of linguistic typology and helps preserve linguistic diversity.

2.2 Definition and Core Characteristics

Following Hockett's (1958, p. 231) definition of gender as "classes of nouns reflected in the behavior of associated words," Corbett (1991) sees grammatical agreement as the determining criterion of gender. The assignment of gender to nouns depends on two kinds of information: the meaning of the noun and its form, which includes morphological and phonological information.

Craig (1992) argued for the existence of noun class and gender as classifier devices, primarily based on the morphosyntactic loci in which they occur. Genders are grammaticalised agreement systems that correlate, at least in part, with certain semantic characteristics, particularly in the domains of human and animate referents. They are realised through agreement with a modifier or the predicate outside the noun itself.

In many languages, there tends to be a distinction between semantic and non-semantic criteria for gender assignment. However, this principle is not strictly universal. In the context of Khasi, even animate nouns are influenced by cultural and mythological factors. For example, while one might expect biological sex to be the primary determinant for gender assignment in animals, Khasi assigns masculine gender to 'dog' (*u ksew*) and feminine gender to 'cat' (*ka miaw*), indicating a significant cultural and mythological influence beyond simple biological distinctions. This will be elaborated in §4.

2.3 Sex-based and Non-sex based Gender Systems

Gender systems in languages are frequently linked to biological sex, where the categorisation of nouns is influenced by the perceived biological distinctions between

male and female. Corbett (2013a)[2] explores how these systems can be both sex-based, where gender directly corresponds to the biological sex of the referent, and non-sex-based, where other semantic and cultural factors play a significant role in gender assignment. In sex-based systems, masculine and feminine genders typically align with male and female entities, respectively. However, non-sex-based systems incorporate a broader array of criteria, such as size, shape, and social roles, leading to more complex and culturally specific gender categorisations. This distinction is crucial in understanding the diversity of gender systems across languages and the interplay between biological, cultural, and linguistic factors in shaping them.

While there is a clear distinction between sex-based and non-sex-based systems, it is important to recognise the diversity within each of these categories. In some languages, such as French and Spanish, gender assignment strictly adheres to biological sex. Nouns denoting male beings are assigned to a masculine gender (‘le’ in French, ‘el’ in Spanish), while nouns denoting female beings are assigned to a feminine gender (‘la’ in French, ‘la’ in Spanish). This straightforward alignment with biological sex is characteristic of many Indo-European languages. Similarly, in Hindi, *ladka* ‘boy’ is masculine, and *ladki* ‘girl’ is feminine.

Conversely, other languages exhibit a more nuanced approach by integrating additional semantic features alongside biological sex. For example, in the Bantu language Swahili, gender assignment is influenced not only by biological sex but also by animacy and shape. Nouns denoting humans generally follow biological distinctions (e.g., *mwanaume* for a man, *mwanamke* for a woman), but non-human nouns are categorised based on animate versus inanimate distinctions: *mti* for a tree, *nyumba* for a house (Fidèle, 2015). This integration of animacy alongside biological sex illustrates how languages adapt gender systems to reflect broader conceptual categories beyond strict biological distinctions.

Further examples can be found in Austroasiatic languages spoken in Southeast Asia and parts of South Asia. For instance, among some Munda languages spoken in Central and Eastern India, gender assignment considers not only biological sex but also cultural associations and social roles. Nouns may be categorised based on perceived attributes such as social status, ritual significance, or even historical roles within the community. This cultural embedding of gender categories enriches linguistic expression and reflects the intricate relationship between language, culture, and environment.

These examples underscore the diversity within sex-based systems of gender assignment, demonstrating how languages employ various criteria—including biological sex, animacy, shape, and cultural associations—to categorise nouns. This variability reflects the adaptability of language to encode complex social and conceptual distinctions, highlighting the dynamic nature of linguistic diversity worldwide.

2.4 Principles of Gender Assignment

The principles by which nouns are assigned to different classes can be governed by various factors. Corbett (1991) categorises them as follows:

[2] It should be noted that in a sample of 257 languages, 145 of them have no gender system while 84 of them have sex-based and 28 of them have non-sex based gender systems.

1. Semantics: Nouns may be classified based on their meanings or semantic properties. For example, nouns representing animate objects might be assigned to one gender, while inanimate objects might be assigned to another.
2. Formal (Morphological or Phonological): The classification can also be based on the form of the noun. Morphological criteria might include specific affixes that denote gender, while phonological criteria might involve the sound patterns of the nouns.
3. A Combination of Semantics and Formal Criteria: In many languages, gender assignment is not based solely on semantics or form but rather on a combination of both. This dual approach allows for a more comprehensive classification system.

These principles, highlight the complexity and diversity of gender assignment across languages. Understanding this is crucial for gaining insights into the linguistic and cognitive processes involved in noun classification. Khasi has a combination of semantic and formal principles for assigning gender on its nouns.

3 Gender Assignment in Austroasiatic Languages

3.1 Sex-Based and Non-Sex-Based Gender Systems

Austroasiatic languages exhibit a range of gender assignment systems. While languages like Khmer and Vietnamese lack grammatical gender, Khasi features a complex tripartite system. Santali, a Munda language, has noun classes based on semantic fields rather than gender distinctions. Mon uses classifiers that convey gender distinctions, particularly for animate nouns, though less systematically than Khasi. These variations illustrate the diversity within Austroasiatic languages.

The patterns found in Austroasiatic languages seem to be mixed, as evident from Corbett (2013a) in the World Atlas of Language Structures (WALS). As shown in Table 3.1, generated on the basis of the output of the WALS Sunburst Explorer developed by Mayer et al. (2014), languages like Khmer, Semelai, and Vietnamese have no gender, while Khasi and Khmu' have sex-based gender systems. In contrast, Mundari and Nicobarese languages have non-sex-based gender systems.

Table 3.1: Sex-based and Non-sex-based Gender Systems in Austroasiatic languages

Category	Subfamily	Language
No gender	Khmer	Khmer
	Aslian	Semelia
	Viet Muong	Vietnamese
Sex-based	Palaung Khumic	Khmu'
	Khasian	Khasi
Non-sex-based	Nicobarese	Nicobarese
	Munda	Mundari

3.2 Patterns and systems of gender assignment

Understanding systems of gender assignment involves exploring how speakers categorise and assign nouns to specific genders within a language. This process can

vary significantly across different language families.

Gender assignment, which refers to the patterns and rules governing how nouns are assigned to specific genders (Corbett 2013b), involves both semantic and formal criteria. In the Austroasiatic family, gender assignment systems are particularly noteworthy for their complexity and variety. Speakers of these languages use a range of criteria to determine the gender of nouns, often relying on a mix of semantic, morphological, and phonological factors. The systems of gender assignment in Austroasiatic languages are shown in Table 3.2. Interestingly, languages that were shown as sex-based and non-sex-based in Table 3.1 show similar patterns with regard to the systems of gender assignment in that they are all primarily semantic based.

Gender systems in languages exhibit various defining properties, one of which is the presence of agreement with other linguistic elements such as adjectives and verbs (Aikhenvald, 2000, p. 21). Each noun typically belongs exclusively to one gender category. The range of genders varies significantly across languages, ranging from just two genders as seen in Portuguese, to as many as ten in Bantu languages, or even several dozen in some South American languages.

Table 3.2: Systems of Gender Assignment in Austroasiatic languages

Category	**Subfamily**	**Language**
No gender	Khmer	Khmer
	Aslian	Semelia
	Viet Muong	Vietnamese
Semantic	Nicobarese	Nicobarese
	Munda	Mundari
	Palaung Khumic	Khmu’
	Khasian	Khasi

3.3 The Case of Khasi

Turning to the case of Khasi, it has a predominant semantic assignment system with three genders:

u	masculine nouns;
ka	feminine nouns and
i	common nouns.

This contrasts with Corbett’s (2013b) observation in WALS, which states that Khasi has two genders. The assignment of gender in Khasi is largely predictable based on the semantic features or meanings associated with each noun. This structure allows for a clear delineation between masculine, feminine, and common nouns within the language, illustrating both the predictability and the semantic basis of gender assignment in Khasi.

The gender assignment system in Khasi reflects a semantic basis where all singular nouns that denote sex-differentiable entities are categorised as either masculine or feminine. Additionally, non-sex-differentiable singular nouns are assigned genders based on their perceived activity level or cultural significance within Khasi society, and various other parameters. For instance, nouns denoting active or culturally significant entities such as wood, large wooden objects, and living plants are typically categorised

as masculine. Furthermore, gender assignment in Khasi may also be influenced by mythological roles, with nouns such as the sun being categorised as feminine and the moon as masculine.

Despite the semantic core underlying Khasi noun genders, the rules governing gender assignment are highly intricate, resulting in many nouns not conforming to straightforward categorisation. These rules accommodate sets of exceptions, which although sporadic, cannot be overlooked. For example, while most fruits in Khasi are categorised as masculine, the banana stands out as an exception, possibly due to cultural or linguistic nuances where it may not be perceived strictly as a 'fruit' by native speakers as depicted in example (1). This complexity underscores the nuanced nature of gender assignment in Khasi, where both semantic principles and cultural considerations play significant roles.

(1) *u soh manir* 'litchee' (sem. domains: 1 - Universe, creation, 1.5 - Plant) M
u soh phan 'jackfruit' (sem. domains: 1 - Universe, creation, 1.5 - Plant) M
u soh pieng 'mango' (sem. domains: 5 - Daily life, 5.2 - Food, 5.2.3 - Types of food, 5.2.3.1 - Food from plants) M
ka kait 'banana' (sem. domains: 5 - Daily life, 5.2 - Food, 5.2.3.1 - Food from plants) F

As mentioned earlier, Khasi has a combination of semantic and formal principles for assigning gender on its nouns which will dealt with in the next section.

Gender in language manifests in various ways, primarily through its realisation and marking within noun phrases. One common method is through overt marking, where free morphemes are typically positioned before the noun in Khasi. This differs from Aikhenvald's assertion (2000, p. 58) that noun classes are never marked with free morphemes and that "noun class systems are typically found in languages with a fusional or agglutinating (not an isolating) profile" (Aikenvald, 2006, p.463).

Additionally, gender can be realised outside the noun itself, often within head-modifier noun phrases. This realisation frequently appears as agreement markers on modifiers like adjectives, and occasionally on modifiers from closed classes such as demonstratives and interrogatives. Furthermore, gender markers may extend beyond the noun phrase to predicates, indicating attributes intrinsic to nouns such as animacy, sex, and sometimes even shape and structure.

4 The Khasi Gender System

4.1 Previous Studies

One of the most significant contributions to the study of the Khasi gender system is Rabel (1977)'s analysis, which provides a foundational understanding of how gender is assigned in Khasi. Rabel's work is notable for its detailed categorisation of feminine and masculine nouns, offering insights into the semantic domains that influence gender assignment. Her classification system includes twenty distinct categories for feminine nouns and several for masculine nouns, reflecting both biological distinctions and cultural factors.

Rabel's categorisation of feminine nouns encompasses a wide range of semantic domains, including female human beings, female spirits and goddesses, female animals, small animals, and domestic fowls and fishes. She also includes categories such as family and clan groupings, external parts of the human body, and various illnesses,

except for skin diseases. Additionally, softwood trees and shrubs, foodstuffs, countries and cities, musical instruments, clothing, tools and implements (excluding those used for boring and digging), ring-shaped jewelry, seasons, days of the week, financial and legal terms, natural forces and landscape features, and abstract nouns are classified as feminine. This comprehensive classification highlights how gender assignment in Khasi is influenced by a combination of natural, cultural, and functional attributes.

Rabel's analysis categorises masculine nouns into several classes, including male human beings, evil spirits and male ghosts, male animals, most large animals, and most insects. Other categories encompass most internal organs of the human body, singing and talking birds, most plants (trees, shrubs, flowers), skin diseases, raw edibles (vegetables, fruits, spices), boring and digging implements, jewelry other than ring-shaped articles, long and thin (stick-like) objects, and fine particles suspended in the air. This classification highlights the combination of biological distinctions and cultural factors in the assignment of gender in Khasi.

She has also highlighted some exceptions in a few of the semantic classes in both feminine and masculine categories but does not provide much explanation as to why certain of those exceptions are present. This lack of detailed analysis leaves some questions about the underlying principles guiding these exceptions, suggesting that cultural, mythological, and religious factors might play a significant role in these anomalous cases. These influences underscore the complexity and depth of gender assignment in Khasi, demonstrating how linguistic categorisation can reflect broader cultural narratives and societal values.

In the next subsection, let us examine the key parameters for semantic differentiation and discuss additional parameters that I have come out with for gender assignment in Khasi.

4.2 Key Parameters for Semantic Differentiation of Genders

4.2.1 Common semantic parameters

Many scholars, including Corbett (1991), Craig (1992), and Aikhenvald (2006), have identified several key semantic parameters that often underlie the assignment of noun classes or gender. These parameters are also found in Khasi, as discussed below:

1. Animacy and Biological Sex: Numerous languages categorise nouns based on whether they denote animate or inanimate entities and, for animate entities, whether they are biologically male or female. This often results in the use of masculine and feminine categories, especially for nouns referring to humans and animals. However, Khasi people assign masculine and feminine genders to animals based on factors beyond biological sex, as will be discussed in §4.3 in detail. In humans, gender is assigned based on their biological sex, but in some kinship terms, like mother, an endearing term *i* is used, making it a common gender (C). This is illustrated below.

(2) *u briew* 'man' (M)
ka kynthei 'woman' (F)
u ksew 'dog' (M)
ka miaw 'cat' (F)
i kmie 'mother' (C)

2. Shape, Size, and Physical Properties: Physical characteristics such as shape, size, and other inherent properties play a significant role in determining noun classes. For example, in Khasi nouns can be classified based on their shape or size. A combination of shape, size, and physical properties is also used. Objects categorised as one-dimensional, such as long, vertical, and thin shapes, are typically assigned masculine gender. Objects that are two-dimensional such as flat shapes are classified as feminine, while three-dimensional objects, specifically those that are round, rigid and hollow, etc. also fall under the feminine category. Size-based parameters play a significant role, with larger objects generally assigned masculine gender, in contrast to smaller objects, which are typically feminine. These are exemplified below:

(3) *u dieng* 'tree' (M) [1D: tall and vertical]
ka sla 'leaf' (F) [2D: flat]
u soh pieng 'mango' (M) [3D round and rigid]
u kulai 'horse' (M) [large]
ka blang 'goat' (F) [small]

3. Cultural and Functional Attributes: The cultural significance and functional attributes of objects influence their classification. Objects that hold particular cultural importance or serve specific functions are often grouped together within the same noun class. For example, in general, a basket is referred to as masculine: *u shang* 'basket'. However, this shifts to feminine when the basket is used for specific purposes:

(4) *ka shang* 'basket' (F) [used in food preparation and storage]
u ksew 'dog' (M) [historical role as a guardian]

4. Mythological and Religious Factors: Mythological and religious beliefs significantly impact the categorisation of nouns in many languages. Entities associated with deities, spirits, or religious practices are often assigned specific genders based on their roles and attributes within cultural narratives.

(5) *ka sngi* 'sun' (F) [daughter in Khasi mythology]
u bnai 'moon' (M) [brother of sun in Khasi mythology]
ka blei-ñia 'altar' (F) [central role in connecting the spiritual and earthly realms]
ka blei 'Goddess' (F)
u blei 'God' (M)

5. Psychological and Emotional Attributes: Emotional and psychological connotations associated with certain nouns can influence their gender assignment. Words related to emotions, states of mind, or abstract qualities may be gendered based on cultural perceptions of those emotions or qualities as being more aligned with one gender.

(6) *ka jingpyrkhat* 'thought' (F)
ka jingieid 'love' (F)
ka jingsngewlem 'compassion' (F)

4.2.2 Introduction of Additional Semantic Parameters in Khasi

In addition to the common parameters for gender assignment, such as animacy, biological sex, shape, size, and cultural and functional attributes, Khasi also employs unique semantic parameters. These additional parameters—visibility, mobility, evaluative criteria, beauty, change of state, carrying, and aspects related to language

and thought—offer a more nuanced understanding of gender categorisation in Khasi.

1. Visibility: In Khasi, gender assignment often considers the aesthetic and perceptual qualities of objects. Beautiful and visible items are frequently assigned feminine gender, while ugly and less visible items tend to be masculine. This parameter underscores the cultural emphasis on aesthetics and visibility in gender categorisation.

Examples: *ka bniat* 'tooth' and *u syntiew* 'flower' are feminine, symbolising visibility and beauty, *ka ktieh* 'mud', *ka bniatktha* 'molar tooth'.

2. Mobility: The association of objects with movement or stasis plays a crucial role in gender assignment. Objects that exhibit mobility or are linked with active movement tend to be masculine, while those that are stationary or less mobile are often feminine.

Examples: *u thylliej* 'tongue' and *u ryndang* 'neck', *u snierbah* 'big intestine' which are associated with movement are masculine.

3. Evaluative Criteria: Evaluative perceptions significantly influence gender assignment in Khasi. Objects considered good or beautiful are typically assigned feminine gender, while those perceived as bad or ugly are masculine. This parameter reflects societal values and judgments encoded within the language.

Examples: ka jingkhuid 'purity', *ka jingieid* 'love' and *ka jingiarap* 'help', considered positive and good, are feminine whereas *u jingbymman* 'perversion' which is viewed as bad and generally associated with masculine qualities, is assigned masculine gender.

4. Beauty: As mentioned earlier, objects that are deemed beautiful are generally assigned a feminine gender, highlighting the cultural association of beauty with femininity. Conversely, objects that become soiled through use (become dirty), particularly those used for digging or hard work, are often assigned masculine gender, reflecting their association with labour and less aesthetic appeal.

Examples: *ka pyrda* 'curtain', and *ka pyrthei* 'earth', appreciated for their aesthetic appeal, are feminine. Tools like *u moh-khiew* 'hoe', *u tyrnem* 'hammer' which become dirty with use, are masculine.

5. Change of State: Objects associated with change or transformation are typically assigned a masculine gender. This parameter reflects the dynamic nature associated with masculinity.

Examples: *u shniuh* 'hair', which can change in length and style, and *u pyrshen* 'pimple', which appear and disappear, are masculine.

6. Carrying: Items that are used for carrying or holding objects often fall under the feminine gender, emphasising the nurturing and supportive roles associated with femininity.

Examples: *ka shang* 'basket', *ka snier* 'intestine', *ka ïing* 'house', and *ka pyrthei* 'earth', all associated with carrying or containing, are feminine.

7. Language and thought: Concepts and objects related to thought, knowledge, and language are generally assigned a feminine gender, reflecting the importance of these aspects in the Khasi culture.

Examples: *ka jingmut* 'thought', *ka ktien* 'language', and *ka jabieng* 'brain' are feminine, highlighting the cultural value placed on cognitive and communicative processes.

These additional parameters—visibility, mobility, evaluative criteria, beauty, change of state, carrying, and language and thought—demonstrate the intricate ways in which Khasi speakers categorise nouns. They reflect a sophisticated interplay between linguistic structures, cultural values, and cognitive perceptions, enriching our understanding of the Khasi gender system. There is also a notable overlap across these semantic parameters, which are generally consistent. In the next subsection, I will examine the key parameters for semantic differentiation in general and discuss the additional parameters that I have identified for gender assignment in Khasi.

4.3 Semantic Patterns, Cultural Influences, and Exceptions

In this subsection, we will explore gender assignment in Khasi across eight broad semantic domains adopted from Fieldworks Language Explorer. Our analysis will be based on the parameters outlined in §4.2. Where necessary, we will include binary features of the parameters, such as MOB + (mobility), VIS + (visibility), CARRY + (carrying), MYTH + (mythological), etc., particularly in cases of exceptions. This approach will provide a nuanced understanding of how these parameters interact within each semantic domain, highlighting patterns and deviations in gender assignment.

While there are nine semantic domains in Fieldworks Language Explorer (FLEx), the ninth domain, 'Grammar,' is excluded from this analysis. The eight domains used for our purposes are: 1. Universe and Creation, 2. Person, 3. Language and Thought, 4. Social Behaviour, 5. Daily Life, 6. Work and Occupation, 7. Physical Action, and 8. States. Though each of these domains can be expanded into several sub-domains and further sub-sub-domains at various levels (even up to seven in some cases), we will not examine examples for each specific case. Instead, we will focus on a few sub-domains to illustrate our points regarding semantic patterns and exceptions.

Domain 1. Universe and Creation

In Khasi, nouns pertaining to universe, creation, and celestial bodies demonstrate a systematic assignment predominantly adhering to feminine gender. This pattern aligns with the semantic domains of water bodies, weather phenomena, and celestial entities associated with creation and cosmic order. For instance, nouns like *ka um* 'water', *ka wah* 'river', *ka duriaw* 'sea/ocean', and *ka jingshlei um* 'flood' are classified under feminine gender due to their thematic alignment with natural elements essential for life and cosmic balance.

Conversely, exceptions to this pattern challenge the predominant categorisation. Notably, *u bnai* 'moon' and *u khlur* 'star' are assigned masculine gender despite their celestial nature and association with the sky. These exceptions suggest underlying cultural and symbolic interpretations embedded within Khasi mythology and cosmology, where specific celestial entities assume gender roles distinct from the broader semantic categories. Such exceptions highlight the nuanced interplay between linguistic structure and cultural beliefs, illustrating how gender assignment in the Khasi language integrates cosmological narratives and cultural symbolism. The following

examples illustrate this:

(7) *Celestial Bodies and Associated Phenomena:*
ka bneng ‘sky’ (sem. domains: 1 - Universe, creation, 1.1 - Sky) F
ka bam hynroh ‘eclipse’ (sem. domains: 1 - Universe, creation, 1.1 - Sky) F
ka sngi ‘sun’ (sem. domains: 1 - Universe, creation, 1.1 - Sky, 1.1.1 - Sun) F

Exceptions
u bnai ‘moon’ (sem. domains: 1.1.1.1 - Moon, 1 - Universe, creation, 1.1 - Sky, 1.1.1 - Sun) M, MYTH +.
u khlur ‘star’ (sem. domains: 1 - Universe, creation, 1.1 - Sky, 1.1.1.2 – Star, M, MYTH +.

The moon, despite its periodic changes, is recognisable as a single entity with a recognisable physiognomy, while the sun lacks such a physiognomy. In Khasi culture, the sun is personified as a woman symbolising light, life, and nurturing. Cultural narratives, such as the Creation myth, mention that *ka sngi* ‘sun’ is the daughter of *ka ramew,* ‘mother earth’. In the same legend *u bnai ‘*moon’ is personified as the son, who desires to marry his elder sister, *ka sngi*. Thus, we can see that the moon is masculine.

(8) *Land and Dwelling Places:*
ka khlaw ‘forest/jungle’ (sem. domains: 1 - Universe, creation, 1.2 - World, 1.2.1 - Land, 1.2.1.6 – Forest, grassland, desert, 1.7 - Nature, environment) F
ka khyndew ‘land’ (sem. domains: 1 - Universe, creation, 1.2 - World, 1.2.1 - Land) F
ka dewbah ‘continent’ (sem. domains: 1 - Universe, creation, 1.2 - World, 1.2.1 - Land) F

(9) *Water-bodies and Associated Phenomena:*
ka um ‘water’ (sem. domains: 1 - Universe, creation, 1.3 - Water) F
ka wah ‘river’ (sem. domains: 1 - Universe, creation, 1.3 - Water, 1.3.1 - Bodies of water, 1.3.1.3 - River) F
ka duriaw ‘sea/ocean’ (sem. domains: 1 - Universe, creation, 1.3 - Water, 1.3.1 - Bodies of water, 1.3.1.1 - Ocean, lake) F
ka jingshlei um ‘flood’ (sem. domains: 1 - Universe, creation, 1.1 - Sky, 1.1.3 - Weather, 1.1.3.7 - Flood) F

Nouns related to nature and seasons are consistently assigned feminine gender within the semantic category of universe and creation. This gender assignment reflects a systematic categorisation rooted in cultural and ecological perceptions. Objects such as *ka kyllang* ‘whirlwind’, *ka lyer* ‘wind’, *ka pyrthat* ‘thunder’, and *ka kjatsngi* ‘sunlight’ exemplify this pattern, all classified under feminine gender due to their intrinsic associations with natural elements and seasonal phenomena essential for sustenance and livelihoods.

(10) *Nature and seasons related*:
ka aiom ‘season’ (sem. domains: 1 - Universe, creation, 1.7 - Nature, environment, 1.7.1 – Natural) F
ka mariang ‘nature’ (sem. domains: 1 - Universe, creation, 1.7 - Nature, environment) F
ka synrai ‘autumn’ (sem. domains: 1 - Universe, creation) F

ka tlang ‘winter’ (sem. domains: 1 - Universe, creation) F
ka eriong ‘storm’ (sem. domains: 1 - Universe, creation, 1.1 - Sky, 1.1.3 - Weather, 1.1.3.5 - Storm) F

Plants and Associated Items

In Khasi, the gender assignment for plants and associated items demonstrates specific patterns influenced by various semantic parameters. Most plants and trees are assigned masculine gender (M). This includes general categories of plants and the trees themselves. However, a notable exception occurs with wooden objects derived from trees, which are assigned feminine gender (F).

(11) Plants and Trees:
u tyrso ‘mustard plant’ (M)
u dieng ‘tree’ (M)
ka dieng ‘wooden object’ (F)

The gender assignment of fruits and flowers generally adheres to masculine gender (M), with the exception of bananas being assigned feminine gender (F):

(12) Fruits and Flowers:
u soh ‘fruit’ (M)
u soh pieng ‘mango’ (M)
u soh ñiamtra ‘orange’ (M)
ka kait ‘banana’ (F)
u syntiew ‘flower’ (M)
*u tiewkulab ‘*rose’ (M)

Vegetables are assigned masculine gender irrespective of whether they are cooked or uncooked. However, a curry is assigned feminine gender because of the presence of liquid in it. In contrast, stimulants such as tobacco exhibit varied gender assignments. When referring to the act of smoking tobacco (*duma*), it is masculine (M). However, the tobacco leaf itself (*duma sla*) is assigned feminine gender (F), mainly due to its natural and unprocessed form and being 2 dimensional in nature.

(13) Vegetables and Stimulants:
u kubi shet ‘cabbage’ (M)
u jhur ‘vegetable’ (M)
*ka jingtah ‘*curry*’* (F)
u duma ‘tobacco (when smoked)’ (M)
ka duma sla ‘tobacco leaf’ (F)

This reflects the cultural and culinary practices associated with the preparation and consumption of food. Most cooking ingredients are assigned masculine gender, reflecting their raw and functional state. However, exceptions exist, such as bay leaf, which is feminine:

(14) Cooking Ingredients:
u rynsun ‘garlic’ (M)
ka latyrpad ‘bay leaf’ (F)

Seed is assigned masculine gender: *u symbai* ‘seed’.

Animals

Gender assignment in the Khasi language manifests a nuanced and varied pattern across different categories of animals, reflecting cultural, ecological, and linguistic factors deeply embedded within Khasi society.

Large mammals such as *u kulai* ‘horse’ and *u hati* ‘elephant’ consistently adhere to masculine gender assignment, aligning with broader cultural perceptions of these animals as symbols of strength and vitality.

(15) *Large Mammals:*
u kulai ‘horse’ (sem. domains: 1 - Universe, creation, 1.6 - Animal, 1.6.1 - Types of animals, 1.6.1.1 - Mammal) M
u hati ‘elephant’ (sem. domains: 1 - Universe, creation, 1.6 - Animal, 1.6.1 - Types of animals, 1.6.1.1 - Mammal) M

Small mammals like *ka blang* ‘goat’ exhibit feminine gender assignment, suggesting a cultural valorisation of smaller domesticated animals within Khasi society.

(16) *Small Mammals:*
ka blang ‘goat’ (sem. domains: 1 - Universe, creation, 1.6 - Animal, 1.6.1 - Types of animals, 1.6.1.1 - Mammal) F

Reptiles, typified by *u bsein* ‘snake’, are uniformly classified as masculine, resonating with their portrayal in Khasi folklore as potent and often dangerous creatures

(17) *Reptiles:*
u bsein ‘snake’ (sem. domains: 1 - Universe, creation, 1.6 - Animal, 1.6.1 - Types of animals, 1.6.1.3 - Reptile, 1.6.1.3.1 - Snake) M

Amphibians like *ka jakoit* ‘frog’ are assigned feminine gender, potentially linked to their associations with water, fertility, and ecological balance in Khasi cultural symbolism.

(18) *Amphibians:*
ka jakoit ‘frog’ (sem. domains: 1 - Universe, creation, 1.6 - Animal, 1.6.1 - Types of animals, 1.6.1.4 - Amphibian, 1.6.1.9 - Small animals) F
ka hynroh ‘toad’ (sem. domains: 1 - Universe, creation, 1.6 - Animal, 1.6.1 - Types of animals, 1.6.1.4 - Amphibian) F

Domesticated animals exhibit mixed gender such as *u ksew* ‘dog’ uphold masculine gender assignment, likely stemming from their historical roles as guardians and companions in Khasi rural life while ‘cow’ is feminine:

(19) *Domestic Animals:*
u ksew ‘dog’ (sem. domains: 1 - Universe, creation, 1.6 - Animal, 1.6.1 - Types of animals, 1.6.1.1 - Mammal) M
ka masi ‘cow’ (sem. domains: 1 - Universe, creation, 1.6 - Animal, 1.6.1 - Types of animals, 1.6.1.1 - Mammal) F

Wild carnivores including *u sing* ‘lion’ and *u khla* ‘tiger’ maintain masculine gender assignment, reflecting their roles as apex predators and symbols of power and ferocity.

(20) *Wild Animals:*
u sing ‘lion’ (sem. domains: 1 - Universe, creation, 1.6 - Animal, 1.6.1 - Types of animals, 1.6.1.1 – Mammal,1.6.1.1.2 - Carnivore) M
u myrsiang ‘fox’ (sem. domains: 1 - Universe, creation, 1.6 - Animal, 1.6.1 - Types of animals, 1.6.1.1 – Mammal, 1.6.1.1.2 - Carnivore) M

Birds
The gender assignment of birds in Khasi exhibits a diverse pattern.

(21) *ka sim* ‘bird’ (sem. domains: 1 - Universe, creation, 1.6.1 - Types of animals, 1.6.1.2 - Bird, 1.6 - Animal) F
u klew ‘peacock’(sem. domains: 1 - Universe, creation, 1.6.1 - Types of animals, 1.6.1.2 - Bird, 1.6 - Animal) M
ka han ‘duck’ (sem. domains: 1 - Universe, creation, 1.6.1 - Types of animals, 1.6.1.2 - Bird, 1.6 - Animal) F

Certain bird species such as *ka paro* ‘dove’, *ka syiar* ‘chicken’, and *u syiarryngkuh ‘rooster’* exhibit varied gender assignment, showcasing variability within avian categorisation based on specific cultural and ecological contexts. The dove and chicken are feminine, reflecting their perceived delicacy and nurturing roles, while the rooster is masculine, likely due to their striking features and significant roles in cultural narratives.

Fish and insects further exemplify the nuanced gender assignment within the Khasi language. However, the designation of cooked fish (*ka dohkha*) as feminine in general may reflect culinary traditions and domestic roles associated with food preparation in Khasi culture. Fish in general is feminine. Some fish can be masculine, for example, *u kha mukur* ‘cat fish’ (M) in both raw and cooked form.

Insects display a mixed pattern of gender assignment, with some species categorised as masculine and others as feminine. This variability underscores the complexity of linguistic categorisation influenced by ecological roles, cultural symbolism, and practical applications within Khasi daily life and belief systems.

(22) Fish and insects:
u kha mukur ‘cat fish’ (sem. domains: 1 - Universe, creation, 1.6 - Animal, 1.6.1 - Types of animals, 1.6.1.5 - Fish) M
ka dohkha ‘cooked fish’ (sem. domains: 1 - Universe, creation, 1.6 - Animal, 1.6.1 - Types of animals, 1.6.1.5 - Fish) F
ka thap-bawa ‘spider’ (F).

Domain 2: Person - Body parts

In the linguistic categorisation of human body parts within Khasi the assignment of masculine and feminine genders to specific anatomical components reflects a nuanced interplay of semantic, cultural, and pragmatic factors.

External organs such as *ka khlieh* 'head', *ka khmat* 'eye', and *ka khmut* 'nose' are classified as feminine (F). These body parts are crucial for sensory perception and communication, embodying qualities traditionally associated with femininity in Khasi culture. For instance, the head (*ka khlieh*) symbolises visibility and thought, as it houses the faculties of sight and cognition, essential for perception and intellectual pursuits. Similarly, the eye (*ka khmat*) and nose (*ka khmut*) are integral to sensory experience and are perceived through a lens that values their aesthetic and perceptive functions, aligning them with attributes of beauty and sensory awareness that are culturally feminised.

Although most external organs in Khasi are typically assigned feminine gender, there are notable exceptions. This could be due to their nature of mobility. For example:

(23) External Organs:
ka khlieh 'head' (sem. domains: 2 - Person, 2.1 - Body, 2.1.1 - Head) F
ka khmat 'eye' (sem. domains: 2 - Person, 2.1 - Body, 2.1.1 - Head, 2.1.1.1 - Eye) F
ka khmut 'nose' (sem. domains: 2 - Person, 2.1 - Body, 2.1.1 - Head, 2.1.1.3 - Nose) F

Exceptions
u thylliej 'tongue' (sem. domains: 2 - Person, 2.1 - Body, 2.1.1 - Head, 2.1.1.4 - Mouth) M; MOB +;
u ryndang 'neck' (sem. domains: 2 - Person, 2.1 - Body) M, MOB +

The tongue is highly mobile, playing a crucial role in speech and eating, which aligns with the parameter of mobility (MOB +). Similarly, the neck, being a flexible and moving part of the body that supports and facilitates head movement, exemplifies the mobility parameter (MOB +). Thus both are assigned masculine gender.

Most of the internal organs in Khasi are masculine (M), emphasising qualities such as strength, vitality, and pragmatic functionality. These organs, often not visible, are perceived as embodying endurance, movement, and internal vitality, which align with traditional masculine ideals. For instance, the pancreas (*u dohlap*) and heart (*u klongsnam*) are perhaps designated masculine due to their essential roles in digestion and circulation.

However, there are notable exceptions. The brain (*ka jabieng*) is feminine (F), highlighting its association with thought and cognitive processing, valued attributes in Khasi culture. Similarly, the intestine (*ka snier*) and large intestine (*ka snierbah*) are feminine, possibly due to their roles in carrying and processing food, aligning with the parameter of carrying.

(24) Internal Organs
u dohlap 'pancreas' (sem. domains: 2 - Person, 2.1 - Body, 2.1.8 - Internal organs) M, VIS -.
u klongsnam 'heart' (sem. domains: 2 - Person, 2.1 - Body, 2.1.8 - Internal organs, 2.1.8.1 - Heart) M, VIS -.

Exceptions

ka jabieng ‘brain’ (sem. domains: 2 - Person, 2.1 - Body, 2.1.8 - Internal organs, Brain) F, THOUGHT +.
ka snier ‘intestine’ (sem. domains: 2 - Person, 2.1 - Body, 2.1.8 - Internal organs) F, CARRY +
ka snierbah ‘big instestine’ (sem. domains: 2 - Person, 2.1 - Body, 2.1.8 - Internal organs, 2.1.8.2 - Stomach) F, CARRY +.
ka plakhun ‘womb’ F, CARRY +.

In broader terms, the categorisation of body parts in Khasi culture extends beyond mere anatomical descriptions to encompass cultural values and perceptions of bodily functions. Attributes such as carrying, visibility, and thought are feminised (F), reflecting qualities associated with nurturing, perceptiveness, and aesthetic appreciation.

In contrast, qualities related to movement, change, and pragmatic functionality are masculinised, highlighting ideals of strength, vitality, and active engagement with the external environment.

Body Functions

The categorisation of bodily fluids and functions consistently assigns feminine gender attributes to various substances and processes. Liquids such as *ka umjung* ‘urine’, *ka snam* ‘blood’, and *ka ksuit* ‘pus’ are uniformly categorised as feminine (F). This consistent gender assignment underscores their cultural perception as nurturing and vital substances within bodily processes. For instance, *ka snam* ‘blood’ is valorised for its life-giving properties and association with health, aligning it with feminine qualities of nurturing and sustenance. Similarly, urine (*umjung*) and pus (*ksuit*), while excretory in nature, are also viewed through a lens that emphasises their protective and healing functions, further reinforcing their feminine categorisation.

Body functions also reflect a consistent inclusion within domains related to bodily functions:

(25) *ka umjung* ‘urine’ (sem. domains: 2 – Person) F
ka snam ‘blood’ (sem. domains: 2 – Person), F
ka ksuit ‘pus’ (sem. domains: 2 – Person) F

The terms for various senses and body conditions are predominantly assigned feminine gender. This pattern includes words such as *ka jingiohi* ‘seeing’, *ka jingiohsngew* ‘hearing’, *ka jingmad* ‘taste’, *ka jingsma* ‘smell’, and *ka jingktah* ‘touch’. The feminine gender assignment extends to words describing body conditions as well, with terms *ka jingkhlain* ‘strong’, and *ka jingtlot* ‘weak’ all being feminine.

Diseases

In the categorisation of diseases within Khasi there is a predominantly feminine (F) classification for most illnesses, with notable exceptions. Diseases such as *ka jingkthakhlieh* ‘headache’ and *u tohjaw* ‘boil’ are categorised as feminine and masculine respectively. Headaches, for instance, are seen through a lens that emphasises their impact on overall health and well-being, reflecting their categorisation as feminine. However, boils (*tohjaw*), for example, are categorised as masculine (M), aligning with the change of state parameter.

There are also specific diseases categorised as masculine (M), diverging from the predominant feminine classification. Examples include *ñiang-lyngkut* 'leprosy' and *ñiang-blei* 'chicken pox'. These diseases are classified as masculine (M), possibly due to their severe and visibly disfiguring symptoms, which align with the parameter of ugliness (UGLY +). The masculine categorisation reflects the cultural perception of these diseases as challenging, significant health concerns that cause noticeable and often stigmatising changes to one's appearance.

Interestingly, *ñiang-khnap* 'foot and mouth disease' stands out as an exception among diseases, being categorised as feminine (F). This divergence from the predominantly masculine categorisation of diseases suggests specific cultural considerations and linguistic nuances in how this particular illness is perceived within the Khasi language framework. The categorisation of *ñiang-khnap* may be influenced by a semantic extension of the visibility parameter. As this disease affects visible parts of the body, such as the mouth and feet, it aligns with the feminine categorisation, which often includes body parts and conditions associated with visibility and care. This is illustrated below:

(26) *ka jingkthakhlieh* 'headache' (F)
u tohjaw 'boil' (M)
u ñiang-lyngkut 'leprosy' (M)
u ñiang-bley 'chicken pox' (M)

Exception:
ka ñiang-khnap 'foot and mouth disease' (F).

Domain 3: Language and Thought

In Khasi, nouns related to language, thought, and intellectual activities predominantly adhere to feminine gender. This gender assignment underscores the cultural valuation of cognitive processes, communication, and teaching, reflecting attributes traditionally associated with femininity in Khasi society. The consistent categorisation of these nouns as feminine highlights the perceived nurturing and generative qualities of thought and language. Examples of this sort include:

(27) *ka jinghikai* 'teaching' (sem. domains: 3–Language and thought) F
ka jingpyrkhat 'opinion, thought' (sem. domains: 3–Language and thought) F
ka ktien 'language' (sem. domains: Language and thought) F

Emotions and forms of communication in Khasi are similarly assigned feminine gender, reflecting their integral roles in human experience and social interaction. This is illustrated below:

(28) Emotions**:**
ka jingkmen 'happiness' (sem. domains: 3 – Language and thought) F
ka jingkyndit 'surprise' (sem. domains: 3–Language and thought) F
ka jingbitar 'anger' (sem. domains: 3–Language and thought) F
ka jinglyngngoh 'confusion' (sem. domains: 3–Language and thought) F
ka jingsngewsih 'sadness' (sem. domains: 3–Language and thought) F

(29) Communication:
ka jingkren ‘speech’ (sem. domains: 3–Language and thought) F
ka khana ‘story’ (sem. domains: 3–Language and thought) F

Domain 4: Social Behaviour

In Khasi, gender assignment in the domain of social behaviour reflects both biological distinctions and cultural values. This domain encompasses various aspects including kinship, social activities, behaviour, authority, government, law, and religion. Some of them are discussed below with examples.

Kinship:

Kinship terms are primarily based on biological sex, reflecting a clear distinction between masculine and feminine genders:

(30) Kinship terms:

Masculine:	Feminine :
u kpa ‘father’	*ka kmie* ‘mother’
u kpa tymmen ‘grandfather’	*ka kmie tymmen* ‘grandmother’
u para ‘brother’	*ka para* ‘sister’
u khun shynrang ‘son’	*ka khun kynthei* ‘daughter’

Social activity:

This sub-domain displays varied gender assignment patterns, covering a range of activities such as music, dance, weddings, and more.

Music and Dance:

Nouns associated with music and dance are predominantly feminine, emphasising the cultural importance of these activities in community gatherings and celebrations. The feminine gender assignment reflects the aesthetic and expressive qualities of music and dance, which are culturally valorised in Khasi society. Examples include:

(31) *ka jingrwai* ‘music’ (F)
ka jingshad ‘dance’ (F)

Musical instruments in Khasi are generally assigned feminine gender, reflecting their cultural significance in nurturing creativity and communal harmony through music and dance.

(32) *ka besli* ‘flute’ (F)
ka bom ‘drum’ (F)
ka maryngod ‘Khasi string instrument that resembles a violin’ *(F)*

In the analysis of the other sub-domains of ‘Social behaviour’ domain, additional parameters such as evaluative criteria and cultural significance come into play, highlighting the nuanced gender assignment in Khasi. For instance, positive attributes like purity and godliness (*ka jingkhuid* ‘purity,’ *ka jinglong riewblei* ‘Godliness’) are assigned feminine gender, reflecting the cultural perception of these traits as nurturing and life-giving, aligning with the evaluative criterion of assigning feminine gender to positive and beautiful entities. In contrast, negative traits like perversion (*u jingbymman*

'perversion') are masculine, adhering to the evaluative criterion of associating negative attributes with masculinity.

(33) Behaviour:
ka jingkhuid 'purity' (F)
ka jinglong riewblei 'Godliness' (F)
u jingbymman 'perversion' (M)

Furthermore, the cultural significance parameter is evident in the gender assignment of roles in authority and governance. Generic terms for traditional leadership and influential positions, such as *u heh* 'boss' and *u nongialam* 'leader,' are used. However, based on biological sex their female counterparts can also be used. Khasi accommodates gender-specific titles for women, like *ka myntrirangbah* 'female Prime Minister,' indicating an evolving recognition of female authority figures.

(34) Authority:
u heh 'boss' (M)
u nongialam 'leader' (M)

(35) Government:
u myntri rangbah 'Prime Minister' (M)
u myntri rangbah shnong 'Chief Minister' (M)

For female equivalents:
ka myntri rangbah 'female Prime Minister' (F)
ka myntri rangbah shnong 'female Chief Minister' (F)

In the legal domain, the use of both masculine and feminine forms for roles like judges (*u nongbishar* and *ka nongbishar*) demonstrates a formal and inclusive approach to gender representation. This dual representation aligns with the parameter of semantic and formal criteria, where gender distinctions are maintained even in professional titles.

(36) Law:
u nongbishar 'judge' (M)
ka aiñ 'law' (F)
ka ïingbishar 'court' (F)

For female equivalents:
ka nongbishar 'female judge' (F)

Religion:
Khasi religious narratives are rich with legends where divine figures exhibit both masculine and feminine traits. For instance, in the indigenous religion of Khasi, *Ka Niam Khasi*, the deity *Ka Blei Synshar* is revered as a nurturing and protective figure, embodying both maternal care and warrior strength. Similarly, *U Lei Shyllong* is celebrated for his wisdom and protective prowess, blending traits typically associated with both genders. Interestingly, both heaven and hell are feminine:

(37) *ka bneng* ‘heaven’ (F)
ka dujok ‘hell’ (F)
ka niam khasi ‘Khasi religion’ (F)
ka blei synshar ‘Goddess of wealth’ (F)
u *ryngkew u basa* ‘God of the hearth’ (M)

Domain 5: Daily Life

The everyday life and material culture of the Khasi people significantly influence gender assignment. Items frequently used in daily activities are gendered in ways that reflect their cultural significance and utility.

Cooking Utensils:

Items commonly used in food preparation and storage are typically feminine. This reflects the traditional role of women in cooking and managing household provisions, emphasising their nurturing and caretaking responsibilities.

(38) Cooking utensils:
ka khiew ‘pot’ (F)
ka shamoit ‘spoon’ (F)
ka shang ‘basket’ (F)

Furniture and household items:

Objects like *ka shuki* ‘chair’ and *ka jingthiah* ‘bed’ are feminine. These items are associated with comfort and domesticity, traditionally managed by women, highlighting their role in creating and maintaining a welcoming home environment.

(39) Furniture and household items:
ka shuki ‘chair’ (F)
ka jingthiah ‘bed’ (F)
ka shang ‘basket’ (F)
ka sharak ‘plate’ (F)

However, items like keys, which symbolise power and control, are masculine:

(40) *u shabi* ‘key’ (M)

Food and Drugs:

Items related to food and drugs are gendered based on their usage and cultural significance.

(41) *ka dawai* ‘drug’ (F)

Clothing:

Clothing items in Khasi are predominantly feminine:

(42) *ka sopti* ‘shirt’ (F)
ka jainsem ‘dress’ (F)
ka jainkpoh ‘blouse’ (F)

Jewelry:
Jewelry in Khasi is gendered in ways that reflect cultural values and aesthetic appreciation.

(43) Jewelry:
ka sati 'finger ring' (F)
ka khadu 'bracelet' (F)
ka khadu 'bangle' (F)
ka shoh-shkor 'earring' (F)

Exception
u kpieng 'necklace' (M) (Neck is masculine thus jewelry adorning it is also M).

Fire:
Items related to fire and its effects predominantly feminine. But 'spark' gets a common gender:

(44) Fire and related:
ka ding 'fire' (F)
ka shlemding 'flame' (F)
ka tdem 'smoke' (F)
i phylliah ding 'spark' (C)

Metals and Gems:
Metals and gems are gendered in Khasi, with gold typically being feminine and diamond masculine.

(45) Metals and gems:
ka ksiar 'gold' (F)
u mawlyngnai 'diamond (M)

Domain 6: Work and Occupation
In Khasi society, social roles and occupations significantly impact gender assignment. Traditionally, Khasi women are involved in weaving, farming, and domestic activities, while men often take on roles requiring physical strength or are seen as protective or authoritative. This division of labour influences the gender assignment of nouns related to these activities. A few sub-domains are considered for analysis.

Weaving Tools and Materials: Weaving tools and materials are often assigned feminine gender, reflecting the traditional role of women in weaving. We also come across instances of common gender:

(46) *ka korthain* 'loom' (F)
ka jain 'cloth' (F)
i jingpynshadksai 'spindle' (C)

Farming Equipment: Items associated with farming are typically assigned masculine gender.

(47) *u syngkai* ‘sickle’ (M)
u kba ‘hoe’ (M)

Exception:
ka lyngkor ‘plough’ (F)

Weapons and Protective Gear: Weapons and protective gear are typically feminine:

(48) *ka wait* ‘sword’ (F)
ka stieh ‘shield’ (F)
ka jamdor ‘dagger’ (F)

The categorisation of tools in Khasi reflects functional distinctions, with cooking and eating utensils predominantly classified as feminine, while tools associated with digging and heavy labour are masculine. This categorisation aligns with cultural perceptions of gender roles related to domestic and outdoor activities.

(49) *u khiew* ‘pot’ (M)
ka thain ‘spool’ (F)
ka khuri ‘bowl’ (F)

Domain 7: Physical Action
Most nouns denoting physical action are feminine:

(50) *ka ieng* ‘standing’ (F)
ka kup ‘kneeling’ (F)
ka shong ‘sitting’ (F)
ka iaid ‘walking’ (F)

Domain 8: States
Time is feminine in general. Days of the week in Khasi are feminine. They are all based on *ka sngi* ‘sun’ which is feminine. On the contrary, months of the year are masculine as they are based on *u bnai* ‘moon’ which is masculine. Location is also feminine, for example, *ka shnong* ‘city’. Other states such as emotions, shape and size have already been discussed throughout this paper.

4.4 Other Means of Gender Assignment: Morphological

Morphology plays a significant role in gender assignment in Khasi. Morphological markers such as prefixes are also used to indicate gender (also observed by Rabel 1977). For example, the prefix *jing-* is used to denote feminine gender, while *nong-* is typically used to indicate masculine gender.

However, this assignment is not always straightforward and can be influenced by semantic considerations. For example, the same root word may take on different prefixes to denote gender, thereby changing its meaning and classification. This morphological aspect of gender assignment adds another layer of complexity to the Khasi language, demonstrating how intertwined and multifaceted its gender system is.

4.4.1 Morphological Gender with *jing-*

The prefix *jing-* is consistently used to form nouns across various semantic domains, denoting feminine gender without exception. This prefix is straightforward in its gender assignment, aligning with nouns that describe abstract concepts, states, or actions (also illustrated in other sections of this paper).

(51) *ka jingbymhok* 'dishonesty' (F)
ka jingshyrkhei 'horror' (F)
ka jingshlur 'humbug' (F)
ka jingpynrit 'humiliation' (F)
ka jingmut 'idea' (F)
ka jingrai 'judgement' (F)
ka jingbha 'goodness' (F)

4.4.2 Morphological Gender with *nong-*

The prefix *nong-*, while generally indicating masculine gender, is subject to the biological sex of the noun it modifies, particularly when referring to animate beings:

(52) *u nongbishar* 'judge' (M)
u nongsynshar 'lieutenant' (M)

For nouns with animate references, semantic and biological considerations often override morphological principles. Rabel (1977) identifies *nong-* as an agentive marker that denotes masculine gender. However, in my opinion, for generic nouns, *nong-* denotes masculine gender. When specifying the biological sex, *nong-* modifies according to the sex of the referent, as seen in the examples below:

(53) *u nongtrei* 'male labourer' (M)
ka nongtrei 'female labourer' (F)

This nuanced application highlights the complexity of morphological gender assignment in Khasi, where morphology meets semantics.

4.5 Common Gender in Khasi

In Khasi, certain nouns are assigned common gender, indicating that they can refer to both masculine and feminine entities without distinction. This category is particularly evident in terms that are neutral regarding the gender of the referent. For example, *i khyllung* 'baby' (sem. domains: 2 - Person, 2.6 - Life, 2.6.4 - Stage of life, 2.6.4.1 - Baby) is a common gender noun used to refer to infants irrespective of their biological sex. Similarly, *i matbriew* 'pupil' (sem. domains: 2 - Person, 2.1 - Body, 2.1.1 - Head, 2.1.1.1 - Eye) is another example of a common gender noun, used for referring to the eye's pupil without gender specification. We also come across terms such as *i tnad* 'twig' that exemplify how common gender is applied to inanimate objects, signifying small or diminutive forms in Khasi.

Moreover, Khasi employs a unique approach to common gender through the use of *i*. This suffix is particularly prevalent in terms of endearment or diminutives. For example, when referring to a mother affectionately, *i* is used, as seen in the term for mother, *i mei*. This diminutive morpheme is not limited to people but extends to objects

as well, emphasising the small size or affection towards the item. This linguistic feature underscores the cultural nuances in Khasi, highlighting how common gender can transcend biological distinctions and apply broadly across different contexts and domains.

4.6 Gender and Loanwords in Khasi

Loanwords in Khasi are often assigned gender following the same rules as native nouns, which confirms the validity of the language's gender assignment principles. Examples:

(54) *u doktor* 'doctor' (M)
ka khop 'cup' (F)
ka skul 'school' (F)

These examples show that Khasi consistently applies its gender assignment rules to loanwords, treating them similarly to native nouns. This reinforces the robustness of Khasi gender assignment principles.

5 Conclusion

The gender assignment system in Khasi is not solely governed by semantic rules but is deeply intertwined with cultural, mythological, and societal factors. Understanding these influences provides a more comprehensive view of how Khasi speakers perceive and categorise their world. One particularly intriguing aspect of Khasi's gender system is its differentiation between humans based on biological sex, where masculine and feminine genders are assigned with relative ease. However, the application of semantic rules to other nouns defies straightforward categorisation. For instance, while the majority of fruits are classified as masculine, bananas may not be universally considered 'fruits' by native speakers.

Gender assignment in Khasi encompasses a range of parameters such as visibility, mobility, evaluative criteria, beauty, change of state, carrying, and thought and language in addition to the common semantic parameters such as animacy, shape, size, cultural and mythological. These parameters highlight the nuanced ways Khasi speakers categorise nouns, reflecting a complex interplay between linguistic, cultural, and cognitive factors. While exceptions may not represent a sizable portion of the language's noun inventory, their significance in understanding Khasi's gender assignment cannot be overstated.

This investigation reveals a systematic semantic classification of nouns in Khasi that integrates both 'natural scientific' principles and elements of 'socio-cultural nature.' The outcomes provide profound insights into the Khasi speakers' worldview as it pertains to nouns, shedding light on the cultural and linguistic intricacies woven into the fabric of the language. Additionally, the study identified specific morphological markers (*jing-* for feminine and *nong-* for masculine) that further complicate the gender assignment process, intersecting with semantic and biological criteria to add another layer of complexity to the language's gender system.

In conclusion, this research deepens our understanding of the gender assignment intricacies within the Khasi language and underscores the importance of preserving and documenting linguistic nuances in endangered languages.

COLOPHON

I extend my heartfelt gratitude to Angelina Kharkongor for not only providing invaluable data but also for introducing me to the Khasi worldview in early 2010. Thanks to Egira Shadap for working with me on the NEICOD project titled "Multilingual Interactive North-East Lexicon (MINEL)" in North Eastern Hill University, Shillong and assisting in the initial compilation of Khasi vocabulary. I also thank the many community members who have contributed data over more than a decade. Special thanks to Gamidalah War for cross-validating the data presented in this paper which have been gathered from native speakers and secondary sources such as Khasi newspapers and dictionaries. A special note of appreciation goes to Gérard Diffloth for our extensive discussions on this paper during his visit to Mysuru. Notably, this paper is dedicated to his tribute volume.

References

Allan, Keith. 1977. 'Classifiers', *Language*, 53:285–311.

Aikhenvald, Alexandra Y. 2000. *Classifiers: A typology of noun categorization devices*. Oxford: Oxford University Press.

Aikhenvald, Alexandra Y. 2006. *Language contact and language change*. Oxford: Oxford University Press.

Aikhenvald, Alexandra Y. 2016. *How gender shapes the world*. Oxford: Oxford University Press.

Basirat, Ali, Marc Allassonnière-Tang, and Aleksandrs Berdicevskis. 2021. An empirical study on the contribution of formal and semantic features to the grammatical gender of nouns. *Linguistics Vanguard*, 7(1), 20200048.

Corbett, Greville G. 1991. *Gender*. Cambridge: Cambridge University Press.

Corbett, Greville G. 2005. *Number*. Cambridge: Cambridge University Press.

Corbett, Greville G. 2013a. Sex-based and non-sex-based gender systems. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) WALS Online (v2020.3) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7385533 (Available online at http://wals.info/chapter/31

Corbett, Greville G. 2013b. Number of Genders. In: Dryer, Matthew S. & Haspelmath, Martin (eds.) WALS Online (v2020.3) [Data set]. Zenodo. https://doi.org/10.5281/zenodo.7385533 (Available online at http://wals.info/chapter/30)

Craig, Colette G. 1986. Jacaltec noun classifiers: A study in grammaticalization, *Lingua*, Volume 70 (4), pp. 241-284.

Craig, Colette G. 1992. Classifiers in a functional perspective. In: Michael Fortescue, Peter Harder and Lars Kristoffersen (eds.) *Layered Structure and Reference in a Functional Perspective*: Papers from the Functional Grammar Conference, Copenhagen, 1990. pp. 277-301.

Dixon, Robert MW. 1986. 'Noun Class and Noun Classification'. In: C. Craig (ed.) *Noun Classes and Categorization*, pp. 105–12. John Benjamins.

Evans, Nicholas. 1994. The problem of body parts and noun class membership in Australian languages, *University of Melbourne Working Papers in Linguistics* 14 (1994): 1-8.

Fidèle, Mpiranya. 2015. *Swahili grammar and workbook*. Routledge.

FieldWorks Language Explorer. https://software.sil.org/fieldworks/

Hockett, Charles F. 1958. *A course in modern linguistics*. New York: Macmillan

Kemmerer, David. 2017. ‘Categories of object concepts across Languages and brains: relevance of nominal classification systems to Cognitive Neuroscience’, *Language, Cognition and Neuroscience*, 32:401–24.

Mayer, Thomas, Bernhard Wälchli, Christian Rohrdantz and Michael Hund. 2014. From the extraction of continuous features in parallel texts to visual analytics of heterogeneous areal-typological datasets. In Nolan, Brian and Carlos Pascual-Periñán (Eds.), *Language processing and grammars: The role of functionally oriented computational models* (SLCS) (Series: Studies in Language). Amsterdam: John Benjamins, 13-38.

Rabel-Heymann, Lili. 1977. Gender in Khasi nouns. *Mon-Khmer Studies*, 6, 247-272.

Senft, Gunter. 2000. What do we really know about nominal classification systems?. In *Systems of nominal classification* (pp. 11-49). Cambridge University Press.

Seifart, Frank. 2010. Nominal classification. *Language and Linguistics Compass*, 4(8), 719-736.

Veeman, Hartger, and Ali Basira. 2020. *An exploration of the encoding of grammatical gender in word embeddings*. arXiv preprint arXiv:2008.01946.

15

Motion Serial Verb Constructions in Vietnamese: a Verbal Semantic Typology[1]

Wenjiu Du

1 Introduction

Serial verb constructions (SVCs) are prevalent in many languages around the world. In typological studies, SVCs are typically categorized into two major types: (i) asymmetric SVCs, which restrict one of the verbs to a specific class, and (ii) symmetric SVCs, which do not impose any such restrictions. These two categories can be further divided into various semantic subtypes. Among these subtypes, motion SVCs, which belong to the asymmetric category, are considered the most common across languages (Aikhenvald 2006, 2018; Durie 1997; Foley and Olson 1985). They have also been the target of extensive description and theoretical analyses over the years. While motion SVCs in Vietnamese have been subject to investigation in the literature (Clark 1978; Nguyen 1996; Srichampa 1998; Nguyen 2001; Hanske 2013; Lam 2015; Ngo 2021)[2], there is a notable lack of comparative perspectives. In this light, this paper aims to situate Vietnamese motion SVCs in a cross-linguistic context and propose a verbal semantic typology for motion SVCs in Vietnamese by refining prior classification.

The remainder of the paper is structured as follows. Section 2 provides a restrictive definition of SVCs that serves as the basis for delineating motion serialization. Section 3 reviews previous approaches to classifying motion SVCs and identifies their shortcomings. Section 4 introduces a detailed classification of Vietnamese motion SVCs, grounded in a refined typology. Finally, Section 5 concludes the paper.

2 Delimiting motion SVCs

To define motion serial verb constructions (SVCs), it is essential first to establish what constitutes SVCs, as motion SVCs represent a specific subtype within this category. The characterization of SVCs remains a contentious issue in scholarly discourse (cf. Bisang 2009; Cleary-Kemp 2015; Aikhenvald & Dixon 2019; Lovestrand 2021; Haspelmath 2016, 2022; among others). There is little consensus on their delimitation due to the lack of cross-linguistic relevance in describing SVCs across individual languages. For the purposes of the present discussion, I follow Haspelmath (2016) in treating SVCs as a comparative concept and adopt his narrow definition with slight

[1] This chapter is a write-up of a talk given at the 11th 1th International Conference on Austroasiatic Linguistics (Chiang Mai October 2023).

[2] Some previous scholarship treats directional verbs in Vietnamese as coverbs, primarily arguing based on the semantic bleaching of these verbs. In this study, I maintain that they function as lexical verbs syntactically, while their meaning is somewhat weakened.

modification as follow[3].

> *SVCs refer to productive, monoclausal constructions consisting of different independent verbs without any linking element or predicate-argument relation between the verbs.*

In other words, SVCs are characterized by the following defining properties:

(1) **Productivity**: SVCs must be non-idiomatic.
(2) **Monoclausality**: There is only one way to form the negation.
(3) **Different independent verbs**: The verbs involved must be different and capable of expressing a dynamic event without specialized coding in the predicate function, and they can stand alone outside SVCs.
(4) **No linking element**: There is an absence of connectors within the string of verbs.
(5) **No predicate-argument relation**: One verb should not be (part of) an argument of another.

This restrictive delineation fares better than previous broader ones in the sense that it enhances the feasibility of comparing SVCs cross-linguistically. In this approach, SVCs are defined independently of language-specific criteria, relying instead on universally applicable concepts such as universal conceptual-semantic and general formal principles (cf. Haspelmath 2010: 665).

The modification I introduce adjusts the term "independent verbs" from its original formulation to "*different* independent verbs", inspired by Bodomo's (1997, 2019) predicate constraint on SVCs. This adjustment aims to exclude instances involving verb reduplication within SVCs, as in (6), given the fact that verb reduplication constitutes an independent construction on its own.

(6) *Anh-ấy* ***đi*** ***đi*** ***lại*** ***lại***.
3SG.MASC go go come come
'He walked back and forth.'

Motion SVCs, therefore, specifically denote cases where one verb related to motion encodes a literal change of location (termed "translational motion" by Talmy 2000; Guillaume 2016) and occupies a restricted syntactic position or functions as a minor verb. This specification serves to exclude certain types of constructions from the category, as illustrated below.

(7) ***Về*** *nhà* ***đi****!*
return home go/PART
'Let's go back home!'

(8) *Nó* ***đáp*** *máy-bay* ***về*** *Hà Nội.*
water exit plane return Hanoi
'S/he took the plane back to Hanoi.'

[3] Note that this definition includes verb-verb compounds, which is different from Lovestrand & Ross (2021).

(9) *Cô* ***dán*** *cái* *tem* ***ở*** *phong-bì.*
2SG.FEM stick CLF stamp be.at envelope
'You have stuck the stamp on the envelop.' (Hanske 2013: 189)

Example (7) is excluded given that *đi* 'go' has has grammaticalized into a mood particle, losing its original translational meaning. Instances like (8) should also be ruled out since *về* 'return' is not in a restricted position. Example (9) is not considered because *ở* 'be at' fails to meet the criterion of being a verb in the comparative sense, which must encode a dynamic event (Haspelmath 2012).

3 Revisiting previous classification

The categorization of motion SVCs can be associated with the typology of motion events. In his seminal works, Talmy (1985, 1991, 2000) deconstructs motion events into four central components, as shown in (10), along with an external co-event that bears the relation of Manner or Cause, as in (11). The examples in (12) demonstrate the correlation between lexemes and these semantic components.

(10) Figure: the moving object
Ground: the reference object with respect to which the figure moves
Motion: the presence of motion
Path: the trajectory along which the figure moves with respect to the ground

(11) Manner: the way in which the figure moves along the path.
Cause: the reason or source from which the motion originates.

(12) a. The pencil **rolled** **off** the table.
Figure Motion + Manner + Path Ground
b. The pencil **blew** **off** the table.
Figure Motion + Cause Path Ground (Talmy 2000: 26)

Talmy distinguishes between two types of motion events: (i) "verb-framed" constructions, where the Path is integrated with the Motion as the main verb, with Manner expressed in a subordinate constituent. For example, in (13), the path of motion is represented by the main verb *entró* 'entered', wheras the manner of motion is conveyed by the non-finite verb *flotando* 'floating'. (ii) "satellite-framed" constructions, where the Path is expressed in the satellite (e.g., verb particles in English, verb affixes in German and Russian, verb complements in Chinese) adjoined to the main verb that is conflated Manner. For instance, in (14), the manner verb "ran" pairs with the path satellite "into". However, some serializing languages challenge this dichotomy. In Chinese, as seen in (15), it is unclear whether the Path component *jìn* 'enter' is encoded in the main verb or the satellite.

(13) Verb-framed: Spanish

La botella	***entró***	*a la cueva*	***(flotando)***.
the bottle	moved.in	to the cave	(floating)
Figure	Motion + Path	Ground	Manner

'The bottle floated into the cave.' (Talmy 2000: 49)

(14) Satellite-framed: English

He	***ran***		***into***	*the classroom*
Figure	Motion Manner	+	Path	Ground

(15) Equipollently-framed: Chinese

Tā	***pǎo***		***jìn***	*jiàoshì.*
3sg	run		enter	classroom
Figure	Motion Manner	+	Path?	Ground

'S/he ran into the classroom.'

In response to this issue, some linguists advocate for a third category known as "equipollently-framed" SVCs (e.g., Slobin 2004; Zlatev and Yangklang 2004), which suggests a more balanced distribution of Motion and Path components across multiple verbs within the clause.

Moreover, Talmy's framework overlooks certain types of motion SVCs (Slobin 2004; Croft et al. 2010; Vittrant 2015), such as those expressing sequential motion (16). Lovestrand (2018: 34) argues that the semantics of these constructions are better captured by the concept of "associated motion". In this vein, Lovestrand and Ross (2021) propose a typology for motion SVCs cross-linguistically. In their typology, motion SVCs are divided into two main groups: directional motion and associated motion. Associated motion SVCs are further subdivided into prior/purposive motion SVCs, concurrent motion SVCs, and subsequent motion SVCs. The distinction among the three classes of associated motion primarily hinges on the timing of the motion (cf. Koch 1984): whether the change of location occurs before the main activity instantiated by the main verb (prior), coincides with it (concurrent), or takes place afterward (subsequent).

(16) Dàgáárè (Hiraiwa & Bodomo 2008: 807; cited in Lovestrand 2018: 36)

ǹ	*dà*	***wà***	***dí***	*lá*	*kàpálà.*
1SG	PST	come	eat	FOC	fufu
Figure		Motion/Path?	?	?	

'I came and ate fufu.'

While insightful and interesting, the main issue with this typology is the potential initial confusion caused by its labels. For instance, one might argue that directional SVCs can also be interpreted as concurrent motion SVCs due to their similar timeframe. The same holds for purposive motion SVCs and subsequent motion SVCs. However, as Lovestrand and Ross (2021) assert, the differentiation can be clarified by additional semantic or syntactic criteria, that is, whether the main verb expresses a motion event (for directional vs concurrent motion SVCs) or whether the dislocation is encoded by

V1 or V2 (for purposive vs subsequent motion SVCs).

Another problem is the difficulty in categorizing directed caused accompanied motion SVCs with verbs like 'take', as in (17). Should cases like this be classified as subsequent motion SVCs (given the V2 position of 'come') or directional SVCs (based on the semantics of 'come')?

(17) Cantonese (Matthews 2006: 76)

lei^{5}	***lo*^{2}**	*di*^{1}	*saam*^{1}	***lai*^{4}**
2SG	take	PL	clothing	come

'Bring some clothes.'

The confusion arises from overlapping meanings between categories, stemming from inconsistent semantic criteria and a mixture of syntactic and semantic considerations in the classification, as shown in Table 1. Specifically, directional and purposive motion SVCs focus on verbal relationships or the lexical semantics of the verb in the construction, whereas prior, concurrent, and subsequent motion SVCs are named based on the temporal sequence that mirrors the linear verb order.

Table Error! No text of specified style in document.: Different criteria for classifying motion SVCs

Type of motion SVCs		Criteria for categorization
Directional SVCs		Constructional semantics
Associated motion SVCs	Prior/Purposive motion SVCs	Temporal iconicity/Constructional semantics
	Concurrent motion SVCs	Temporal iconicity
	subsequent motion SVCs	Temporal iconicity

4 Pursuing a verbal semantic typology

As demonstrated earlier, the existing typology of motion SVCs is problematic. Therefore, a revision of this typology is necessary to pave the way for a fine-grained classification of Vietnamese motion SVCs.

4.1 Unifying the criteria

To circumvent the problem raised in section 3, it is crucial to unify the criteria for classifying motion SVCs. Here, I adopt a verbal semantic typology that categorizes motion SVCs according to the conceptual event type expressed by the serial verbs. This approach, endorsed by Luke and Bodomo (2000: 172), facilitates a more coherent cross-linguistic comparison of SVCs by capitalizing on the consistency of situational conceptualizations across languages.

> *[I]t is very difficult to come up with cross-linguistic generalizations for serial verb constructions due to the great variety in their syntactic structure. Rather than categorizing them according to criteria like phrasal, clausal, object sharing, etc., terms which are not uniform cross-linguistically, an alternative proposal is to classify them according to the situation types they lexicalise, with the understanding that* ***situational conceptualizations across languages are more uniform than syntactic constructions across languages****. ... Once we agree on a set of serial verbs expressing various situation types, we will then be in a better position to classify them according to the situation types they express. In this way, we stand a better chance of constructing a clearer typology of serial verbs across various languages.*

Lord (1993: 2) echoes a similar sentiment: "We can relax the restrictions on surface form and instead try to characterize serial verb constructions in terms of the meanings they convey.''

Based on such a criterion, the preceding typology ought to be reformulated as follows: directional motion SVCs, purposive motion SVCs, comitant motion SVCs, and cause-effect motion SVCs. Note that this revision pertains solely to the naming system, keeping the compositional pattern unchanged as documented by Lovestrand and Ross (2021). This change seeks to improve the clarity and precision of each type. Under this verbal semantic typology, the previously ambiguous caused accompanied motion SVCs (cf. Example (17)) are unequivocally classified within the cause-effect motion SVCs category.

4.2 Towards a typology of Vietnamese motion SVCs

Before embarking on a verbal semantic typology of motion SVCs in Vietnamese, it is fundamental to ascertain the typical verb classes and their lexical semantics that characterize motion SVCs in this language.

As mentioned in section 3, the components of a motion SVC encompasses two primary types of verbal elements: the motion verb and the path verb. The motion verb denotes the manner of movement, such as 'run', 'walk', 'fly', 'float', or actions that cause displacement of an object (Figure), such as 'carry', 'throw', 'push'. The path verb specifies the direction toward a specific location (Ground), which can be categorized into deictic verbs like 'come' and 'go', and general directional (non-deictic) verbs like 'enter', 'ascend', 'descend' (cf. Chen 2023). The inventory of each verb class in Vietnamese is detailed in Table 2.

Table 2: Verb classes in Vietnamese motion SVCs

Motion verb	Manner	*đi/bước* 'walk', *chạy* 'run', *nhảy/nhẩy* 'jump', *bay* 'fly', *cuộn/lăn* 'roll', *trôi* 'float'
	Cause[4]	*chuyển* 'move, forward, transfer', *mang* "carry', *xách* "carry with the handle of the object', *đưa* 'take, bring, pass', *đem/lấy* 'take, bring', *ném/tung/vứt* 'throw', *đặt* 'put', *đẩy* 'push', *giơ* 'raise, lift'
Path verb	Decitic	*đi* 'go', *đến/lại/tới* 'come'
	Directional	*lên* 'ascend', *xuống* 'descend', *vào* 'enter', *ra* 'exit', *sang/qua* 'cross', *về* 'return', *tới* 'arrive'

Based on the refined typology in section 4.1, four types of motion SVCs can be attested in Vietnamese. In what follows, I will elucidate each pattern utilizing the compositional verbs mentioned earlier, aligned with the framework proposed by Lovestrand & Ross (2021).

Type 1: Directional motion SVCs

Directional motion SVCs are widely recognized as the most predominant subtype within motion SVCs across languages (Lovestrand & Ross 2021). They involve a verb indicating a motion event without an external cause. These constructions often feature manner-of-motion verbs combined with path-of-motion verbs, which could be deictic verbs like *đến* 'come' or *đi* 'go', as exemplified in (18) and (19), or general directional verbs like *vào* 'enter', *lên* 'ascend' or *ra* 'exit', as shown in (20)—(22). Typically, two verbs are contiguous. The path verb occupies the V2 position and the subject is the figure on the path of motion. This type captures the pure motion aspect without implying a purposive or causal relationship.

Consider the example (22) particularly. In this example, *leo* 'climb' is the motion verb encoding manner, and *lên* 'ascend' is the directional verb specifying the direction of the motion. The construction indicates a movement towards an upward location without any extra context about the purpose or cause of the motion.

(18) *Anh-ấy* ***đi*** ***đến*** *trường.*
3SG.MASC go come school
'He went to school.'

(19) *Anh-ấy* ***chạy*** ***đi*** *công-ty.*
3SG.MASC run go company
'He ran to the company.'

(20) *Con vịt* ***nhảy*** ***vào*** *hồ-nước.*
CLF duck jump enter lake
'The duck jumped into the lake.'

(21) *Cô-ấy* ***đi*** ***ra*** *vườn.*
3SG.FEM go exit garden
'She got out of the garden.'

[4] See Hanske (2013) for a detailed discussion for this type of verbs.

(22) *Người đàn-ông* ***leo*** ***lên*** *đỉnh* *núi.*
CLF man climb ascend top mountain
'The man climbed (up) to the top of the mountain.'

Type 2: Purposive motion SVCs

In purposive motion SVCs, the initial verb typically fulfills a deictic function, indicating direction or movement towards a goal, while the subsequent verb describes the purpose or intended outcome of the motion. This type has contiguous sequences of verbs in which a deictic verb is paired with another verb that may not be directly related to motion. The deictic verb normally occupies the V1 slot, with the figure always serving as the subject. This construction highlights the intention behind the movement rather than the movement itself.

Take the example (23). Here, *đến* 'come' is the deictic verb showing the direction of movement, while *làm việc* 'do work' explains the reason for coming. The purposive motion SVC clearly demonstrates the intent behind the action of coming, which is to work. The purposive interpretation will become more explicit if the purposive marker *để* is inserted between *đến* 'come' and *làm việc* 'do work'.

(23) *Cuối-cùng* *cô-ấy* *đã* ***đến*** ***làm*** *việc.*
finally 3SG.FEM PST come do work
'Finally she came to work.'

(24) *Anh-ấy* *đã* ***đến*** ***giúp*** *tôi* *chuyển* *nhà.*
3SG.MASC PST come help 1SG move house
'He came to help me moving house.'

(25) *Cô-ấy* *vừa* ***tới/đến*** ***thăm*** *bà* *của* *cô-ấy.*
3SG.FEM just come visit grandmother POSS 3SG.FEM
'She has just come to visit her grandmother.'

(26) *Hôm-qua* *tôi* ***đi*** ***mua*** *đồ.*
3SG.FEM 1SG go buy thing
'I went shopping yesterday.'

(27) *Hôm-qua* *cô-ấy* ***đi*** ***bơi.***
yesterday 3SG.FEM go swim
'She went swimming yesterday.'

Note that if the first verb is a general directional verb rather a deictic verb, the purposive marker is obligatory, as in (28).

(28) *Cô-ấy* ***vào*** ****(để)*** ***học.***
3SG.FEM enter PURP study
'She entered to study.'

Type 3: Abstract-comitant motion SVCs

Abstract-comitant motion SVCs include a non-motion verb coupled with a path-of-motion verb, where the path verb modifies the activity predicated by the non-motion verb in an abstract, metaphorical sense. The minor path verb, mostly a general directional verb like *ra* 'exit' or *lên* 'ascend', assumes the second position, as seen in (29)—(32). Deictic verbs as path verbs occur only marginally, as illustrated in (33). Similar to the two preceding types, the verbs in this category are contiguous, but the figure on the path of motion could be the subject (29) or the object argument (30)—(33). This type illustrates how the path of motion can accompany another non-motion-related action, enriching the description of the event.

Take a look at example (29) specifically. In Vietnamese, *nổ* 'burst' is not typically classified as a motion verb. It primarily denotes sudden, often violent action resulting in an eruption or explosion, rather than movement from one place to another. *Ra* 'exit' in this context does not refer to a physical event, but rather indicates the abstract outward motion of the event, adding the sense of something bursting forth or emerging suddenly. That's why I incorporate the word "abstract" into the label "comitant motion SVCs" (modifying the former "concurrent motion SVCs") to clarify this type, which sets it apart from other types based solely on its name. In this regard, this type also diverges slightly from Lovestrand & Ross (2021) in that the motion event in the sense of 'V while going' is conceptual.

(29) *Tiếng* *vỗ* *tay* ***nổ*** ***ra.***
sound clap hand burst exist
'Applause broke out.'

(30) *Cùng* ***thắp*** ***lên*** *ngọn-lửa* *tuổi* *20.*
together light ascend flame age
'Let's light up the flame of age 20.' (https://tuoitre.vn/cung-thap-len-ngon-lua-tuoi-20-20230428104210667.htm)

(31) *Anh-ấy* *đã* ***nghĩ*** ***ra*** *đáp-án.*
3SG.MASC PST think exit solution
'He camp up with the solution.'

(32) *Vấn-đề* *này* ***hóa*** ***ra*** *rất* *đơn-giản.*
problem DEM become exit very simple
'This problem turned out to be very simple.'

(33) *Cô-ấy* *không* ***quan-tâm*** ***đến/tới*** *chính-trị.*
3SG.FEM NEG care come politics
'She doesn't care about politics.'

Type 4: Cause-effect motion SVCs

Cause-effect motion SVCs involve a verb that causes motion followed by a general directional verb (34)—(37) or a deictic verb (38) that specifies the result of that motion. This type underscores the causal relationship between actions, showing how one action leads to movement in a particular direction. Consistent with the generalization of Lovestrand & Ross (2021), the figure on the path of motion largely depends on the

semantics of the event. It can be the subject, as in (34), object, as in (35) and (36), or both as in (37) and (38). Another characteristic that distinguishes this type from others is the contiguity of the serial verbs. The verbs can be contiguous as in other types (e.g., *đẩy xuống* 'push down'), or they can be separated by an element, as in *đẩy tôi xuống* 'push me down'.

As in the instance of (38), the verb *đẩy* 'push' is a caused-motion verb initiating the action, and *ra* 'out' is the general directional verb indicating the destination of the action. In this scenario, the positioning of the figure on the path of motion is undetermined. One could interpret that either the subject *anh-ấy* 'he' moves alongside the object *bàn* 'table', triggering a cumulative interpretation, or that the subject propels the object along a path of motion without moving itself. The cause-effect motion SVC provides a clear picture of the sequence of events and their causal link.

(34) *Anh-ấy* ***lăn*** ***xuống*** *tầng-dưới.*
3SG.MASC roll descend downstairs
'He rolled down stairs.'

(35) *Cô-ấy* ***cắm*** *hoa* ***vào*** *lọ.*
3SG.FEM plug flower enter vase
'She puts flowers in a vase.'

(36) *Anh-ấy* *ném* *đá* ***xuống*** *hồ.*
problem throw stone descend lake
'He threw down the stone into the lake.'

(37) *Cô-ấy* ***mang*** *sách* ***đến*** *phòng.*
3SG.FEM bring book come room
'She brought the book to the room.'

(38) *Anh-ấy* ***đẩy*** *cái* *bàn* ***ra*** *phòng* *khách.*
3SG.MASC push CLF table exit room guest
'He moved the table to the middle of the living room.'

4.3 Interim summary

The pitfalls of the previous classification of motion SVCs call for a revision of the typology. To address these issues, I adopt a verbal semantic typology, classifying motion SVCs according to the conceptual phenomenon or event type expressed in each situation. This refined typology is advantageous as it ensures that each category is distinct and non-overlapping, reducing confusion in classification. Furthermore, it amplifies the potential for meaningful cross-linguistic comparisons.

Based on such a typology, four types of motion SVCs in Vietnamese are identified. Their properties are summarized in Table 3.

Table 3: Characteristics of Vietnamese motion SVCs

Type of motion SVCs		Compositional pattern		Contiguity	Figure on path
		V1	V2		
Directional motion SVCs		manner-of-motion	directional/deictic	✓	subject
Associated motion SVCs	Purposive	deictic	free-class	✓	subject
	Abstract-comitant	non-motion	directional/deictic	✓	subject/object
	Cause-effect	cause-of-motion	directional/deictic	✓/✗	subject/object/subject+object

In terms of composition, purposive motion SVCs are unique in that the path-of-motion verb takes the V1 position and is limited to deictic verbs. Regarding contiguity, cause-effect motion SVCs are notable for their compatibility with both split and non-split configurations. When it comes to the figure on the path of motion, cause-effect motion SVCs present an intriguing scenario where the moving entity can exhibit different behaviors..

5 Conclusion

In this study, I have reappraised the classification of motion SVCs, addressing the limitations of previous typologies. I have suggested a revised typology grounded in verbal semantics. This refined classification, which includes directional, purposive, abstract-comitant, and cause-effect motion SVCs, aims to provide a clearer, more consistent framework with more distinct categories. This approach not only reduces confusion but also enhances the cross-linguistic applicability.

I have analyzed Vietnamese motion SVCs based on the new typology, demonstrating that all four types are present in Vietnamese. Each type exhibits unique characteristics in terms of compositional patterns, contiguity, and the figure on the path of motion. Basically, directional motion SVCs involve a manner-of-motion verb followed by a directional or deictic verb. Associated motion SVCs are characterized by purposive verbs in the V1 position paired with deictic verbs. Abstract-comitant motion SVCs combine non-motion verbs with path-of-motion verbs in a metaphorical sense. Cause-effect motion SVCs feature caused-motion verbs followed by directional or deictic verbs, highlighting causal relationships between actions.

Future research could further explore the nuances of these classifications, particularly in under-studied languages, to test the applicability and robustness of this typology. Additionally, examining the cognitive and pragmatic aspects of moton SVC usage in natural discourse could provide deeper insights into their functional roles in language. Comparative studies between motion SVCs in Vietnamese and other serializing languages could also advance our understanding of the universality and variability of SVC structures, contributing to broader linguistic theory.

Acknowledgement

Although I did not have the luxury of being Prof. Gérard Diffloth's student, his works on Austroasiatic linguistics provided me with a window into the study of Southeast

Asian languages back in China. I really admire his contributions to the historical linguistics of Austroasiatic languages. For this paper, I would like to extend my gratitude to the audience at the 11th International Conference on Austroasiatic Linguistics (ICAAL 11) for their helpful comments and feedback. My sincere thanks go to my language consultant, Đặng Thị Bích Phượng for kindly providing the Vietnamese data and her judgement. I am also deeply grateful to Prof. Paul Sidwell for his efforts in creating this volume.

References

Aikhenvald, Alexandra Yurievna (2018). *Serial verbs.* Oxford: Oxford University Press.

Aikhenvald, Alexandra Yurievna (2006). Serial verb constructions in typological perspective. In Alexandra Yurievna Aikhenvald & Robert Malcolm Ward Dixon (eds.), *Serial verb constructions: A cross-linguistic typology* (pp. 1-68). Oxford: Oxford University Press.

Aikhenvald, Alexandra Yurievna & Dixon, Robert Malcolm Ward (2019). Letter to the editor of *Language and Linguistics*, Serial verb constructions: A critical assessment of Haspelmath's interpretation. At https://www.academia.edu/38194874/Letter_to_the_Editor_with_a_critical_assessment_of_Haspelmath_on_serial_verbs_pdf

Bisang, Walter (2009). Serial verb constructions. *Language and Linguistics Compass, 3*(3), 792-814.

Bodomo, Adams (1997). *Paths and pathfinders: Exploring the syntax and semantics of complex verbal predicates in Dagaare and other languages.* Ph.D. dissertation, Norwegian University of Science and Technology.

Bodomo, Adams (2019). A Daagare Pandora's box: The syntax of verb serialization in an oral literature context. In James Essegbey, Dalina Kallulli & Adams Bodomo (eds.), *The grammar of verbs and their arguments: A cross-linguistic perspective* (pp. 115-149). Köln: Rüdiger Köppe Verlag.

Chen, Zhishuang (2023). Directional serial verb constructions in Mandarin – a neo-constructionist approach. *Journal of Linguistics*, *59*(4), 1-40.

Clark, Marybeth (1978). *Coverbs and case in Vietnamese.* Pacific Linguistics, Research School of Pacific and Asian Studies, The Australian National University.

Cleary-Kemp, Jessica (2015). *Serial verb constructions revisited: A case study from Koro*. Ph.D. dissertation, University of California, Berkeley.

Croft, William A., Barðdal, Jóhanna, Hollmann, Willem B., Sotirova, Violeta, & Taoka, Chiaki (2010). Revising Talmy's typological classification of complex event constructions. In Boas, Hans C. (ed.), Contrastive studies in construction grammar (pp. 201–236). Amsterdam: John Benjamins Publishing Company.

Durie, Mark (1997). Grammatical structures in verb serialization. In Alex Alsina, Joan Bresnan, & Peter Sells (eds.), *Complex predicates* (pp. 289–354). Stanford: CSLI Publications.

Foley, William Auguste and Olson, Mike. (1985). Clausehood and verb serialization. In Johanna Nicolas & Anthony C. Woodbury (eds.), *Grammar inside and outside the clause* (pp. 17–60). Cambridge: Cambridge University Press.

Guillaume, Antoine (2016). Associated motion in South America: Typological and areal perspectives. *Linguistic Typology*, *20*(1), 81-177.

Hanske, Theresa (2013). Serial verbs and change of location constructions in Vietnamese. In Daniel Hole & Elisabeth Löbel (eds.), *Linguistics of Vietnamese: An international survey* (pp. 185-214). Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110289411.185

Haspelmath, Martin (2022, November 17). Revisiting the serial verb construction concept – is it relevant for cross-linguistic comparison? Diversity Linguistics Comment. Retrieved June 26, 2024, from https://doi.org/10.58079/nsww

Haspelmath, Martin (2016). The serial verb construction: Comparative concept and cross-linguistic generalizations. *Language and Linguistics, 17*(3), 291-319.

Haspelmath, Martin (2012). How to compare major word-classes across the world's languages. In Thomas Graf, Denis Paperno, Anna Szabolcsi & Jos Tellings (eds.), Theories of everything: In honor of Edward Keenan (pp. 109–130). Los Angeles: University of California at Los Angeles.

Haspelmath, Martin (2010). Comparative concepts and descriptive categories in crosslinguistic studies. *Language, 86*(3), 663-687.

Hiraiwa, Ken & Bodomo, Adams (2008). Object-sharing as symmetric sharing: Predicate clefting and serial verbs in Dàgáárè. *Natural Language & Linguistic Theory, 26*, 795-832.

Koch, Harold (1984). The category of 'associated motion' in Kaytej. *Language in Central Australia, 1*(1), 23-34.

Lam, Quang Dong (2015). Translation of Vietnamese serial verb constructions (SVCs) and/or multi-verb constructions into English. *VNU Journal of Foreign Studies, 31*(4), 1-10.

Lovestrand, Joseph (2018). *Serial verb constructions in Barayin: Typology, description and lexical-functional grammar*. Ph.D. dissertation, University of Oxford.

Lovestrand, Joseph (2021). Serial verb constructions. *Annual Review of Linguistics, 7*(1), 109-130.

Lovestrand, Joseph & Ross, Daniel (2021). Serial verb constructions and motion semantics. In Antoine Guillaume & Harold Koch (eds.), *Associated motion* (pp. 87-128). Berlin, Boston: De Gruyter Mouton. https://doi.org/10.1515/9783110692099-003

Luke, Kang Kwang & Bodomo, Adams (2000). A comparative study of the semantics of serial verb constructions in Dagaare and Cantonese. *Languages in Contrast, 3*(2), 165-180.

Matthews, Stephen (2006). On serial verb constructions in Cantonese. In Alexandra Yurievna Aikhenvald & Robert Malcolm Ward Dixon (eds.), *Serial verb constructions: A cross-linguistic typology* (pp. 69–87). Oxford: Oxford University Press.

Ngo, Binh (2021). *Vietnamese: An essential grammar*. New York: Routledge.

Nguyen, Dinh Hoa (1996). Vietnamese verbs. *Mon-Khmer Studies, 25*, 141-160.

Nguyen, Lai (2001). *A group of directed words of movement in modern Vietnamese*. Hanoi: Social Science Publisher.

Slobin, Dan Isaac (2004). The many ways to search for a frog: Linguistic typology and the expression of motion events. In Sven Stromqvist & Ludo Verhoeven (eds.), *Relating events in narrative: Typological and contextual perspectives* (pp. 219-257). Mahwath, New Jersey: Lawrence Earlbaum Associates

Srichampa, Sophana (1998). Prepositional vs. directional coverbs in Vietnamese. *Mon-Khmer Studies, 28*, 63-83.

Talmy, Leonard (2000). *Towards a Cognitive Semantics*. Cambridge MA: MIT Press.

Talmy, Leonard (1991). Path to realization: A typology of event conflation. In *Annual Meeting of the Berkeley Linguistics Society* (pp. 480-519).

Talmy, Leonard (1985). Lexicalization patterns: Semantic structure in lexical forms. *Language typology and syntactic description, 3*(99), 36-149.

Vittrant, Alice (2015). Expressing motion: The contribution of Southeast Asian languages with reference to East Asian languages. In Nick J. Enfield & Bernard Comrie (eds.), *Languages of Southeast Asia: The state of the art* (pp. 586–632). Berlin: De Gruyter Mouton.

Zlatev, Jordan & Yangklang, Peerapat (2004). A third way to travel: The place of Thai and serial verb languages in motion event typology. In Sven Stromqvist & Ludo Verhoeven (eds.), *Relating events in narrative: Typological and contextual perspectives* (pp. 159-190). Mahwath, New Jersey: Lawrence Earlbaum Associates.

16

Vietnamese as a Heritage Language in South Korea and Japan: a Perspective from Language Policy[1]

Hong Duong Do

1 Introduction

South Korea and Japan are two countries with a large number of Vietnamese residents. In South Korea, according to the Korean Statistical Office[2], as of 2022, the number of foreigners in the country is 1.75 million, of which Vietnamese account for 209,373, making it the second-largest immigrant group in South Korea.

In Japan, since the early 2010s, Vietnamese have been the fastest-growing foreign group in the population system (according to the Japan Statistical Office)[3]. As of 2022, there are about 489,000 Vietnamese immigrants living in Japan.

With such large numbers of Vietnamese residents, the issue of cultural and language education for second-generation Vietnamese becomes an urgent requirement for the governments of both countries. Preserving language and maintaining connections with roots and origins help the second generation have a more solid development foundation in their host country. However, during the research process, we found a discrepancy in evaluating the importance of heritage language for the second generation in the two countries, and thus the language policies to develop heritage language in each country also differ.

2 Theoretical considerations

2.1 Top-down and bottom-up language policy

Top-down and bottom-up language policy are two different approaches used to manage and develop languages within communities and countries. While the top-down approach focuses on establishing language policies from the top (from the government or central management agencies down to the entire community), the bottom-up approach focuses on building language policies from the grassroots level (from the needs and desires of language users' communities) to form and promote government language policies.

The idea of the top-down and bottom-up language policy theory is proposed in

1 This chapter is a write-up of a talk given at the 11th International Conference on Austroasiatic Linguistics (Chiang Mai October 2023).

2 통계청,「인구총조사」, 2022, 2024.05.02, 성, 현재 국적 및 연령별 외국인 – 전국 https://kosis.kr/statHtml/statHtml.do?orgId=101&tblId=DT_1JA1504&conn_path=I2

3 https://www.statista.com/statistics/687809/japan-foreign-residents-total-number/

works on language and society by (Ferguson, 1959), (Hymes, 1972), (Labov, 1972),... although not specifically named. This theory is specifically presented by (Fishman, Nahirny, Hofman, & Hayden, 1966) through research on factors influencing the preservation and development of mother tongue languages in ethnic and religious communities in the United States, including the impact of language policies from both sides and how communities maintain and develop their mother tongue languages. Some important points can be observed as follows:

- Government policies strongly influence language education in the community. For example, policies to develop English or requirements to only teach English in schools have a strong impact on maintaining the mother tongue languages of minority groups.
- Some communities organize mother tongue classes (such as Jewish people organizing Yiddish language courses, Spanish people organizing Spanish language courses,...) to stimulate the use and development of mother tongue languages of minority communities in the United States.
- Inadequate support from national and grassroots policies for the preservation and development of mother tongue languages may lead to a decline and loss of mother tongue languages among younger generations in immigrant communities, posing risks of cultural and language loss for the community.

From these studies, Fishman proposes strategies or methods to maintain and develop mother tongue languages, including factors from the highest and grassroots levels. On this theoretical issue, more information can be found in the works of Joshua A. Fishman and other authors such as (Hornberger, 1998), (Spolsky, 2003) etc. Applied to the development of immigrant heritage languages for second-generation and beyond, this theory proves to be effective depending on the conditions of each country.

2.2 Vietnamese as a heritage language

Heritage language is a concept understood in both broad and narrow senses. In the broad sense, heritage language refers to possible connections between language heritage and cultural heritage. (Fishman, 2001) emphasizes specific connections within the family of language users. (Deusen-Scholl, 2003) points out that through family interactions, heritage language speakers are nurtured with a strong cultural connection to the family language. According to this definition, heritage language is considered as the second language in language proficiency (Polinsky & Kagan, 2007). However, we argue that there is a significant difference between heritage language speakers and second language speakers. Therefore, we advocate for a narrower understanding of the concept of heritage language below.

The most well-known and widely used definition of heritage language from a narrow perspective of this term is by (Valdés, 2000). According to Valdés, heritage language is defined as a minority language in society and is often learned at home during childhood. Heritage language learners grow up in an environment with a different main language from their heritage language, and learners will have more proficiency in the main language and feel more comfortable speaking it (Valdés, 2000). Thus, according to this notion, heritage language speakers are to some extent bilingual. The important criterion is that heritage language is identified as the first language acquired sequentially, but not fully acquired as individuals switch to a dominant language.

According to (Kelleher, 2010), the term heritage language is used to identify languages other than the dominant language in a specific social context. Many people

consider these as foreign languages, however, many individuals living in that society have cultural connections and roots with these languages. Therefore, these languages are not foreign to specific individuals or communities; instead, they are familiar to these individuals or communities in various ways. Some may be able to read, write, and speak the language, some may only speak or understand it when communicating, some may not understand it but are part of the family or community where the language is used. The term heritage language can be used to describe any relationship between a non-dominant language and an individual, family, or community. For further theoretical issues, please see Valdés (2000), Fishman (2001), (Polinsky M. , 1997), (Polinsky M. , 2000), (Polinsky M. , 2006), (Polinsky & Kagan, 2007), ((Montrul, 2013).

Vietnamese is considered a heritage language in cases where Vietnamese is the native language of the first generation (either parent, or both parents), used within the family when immigrating abroad. There are two identified cases:

- Vietnamese is used in families where both spouses are Vietnamese immigrants abroad.
- Vietnamese is used in families where one spouse is a Vietnamese immigrant, and the other spouse is a foreigner.

In both cases, the speakers and learners of the heritage language are second-generation and beyond. In previous studies ((Do, 2013), (Do, 2017) (Do, 2021), we have addressed the issue of preserving Vietnamese as a heritage language for mixed-heritage children (the second case). These are important sources for guiding research on teaching and learning Vietnamese as a heritage language.

3 Policies related to the maintenance of heritage languages in multicultural families in South Korea

Over the past three decades, South Korea has undergone significant societal changes and shifts in social consciousness. Previously, the concept of Korean ethnicity was based on a shared history and language within a unified national territory with a homogeneous lineage (Shin, Freda, & Yi, 1999) Therefore, many Koreans shared a strong belief and pride in a homogeneous nation in terms of ethnicity and language. However, ethnic nationalism based on a shared consciousness of ethnicity and language has also led to intolerance towards cultural diversity and ethnic diversity within Korean society. According to a survey conducted by the Ministry of Health and Welfare in 2005, 17.6% of students born from international marriages were reported to have experienced discrimination from their classmates, with the most common reason cited being “because their mother is a foreigner” (34.1%) (see (Lee, 2013).

These days that exclusivism is being challenged in the context of strong human and cultural shifts in the era of globalization. The number of immigrants entering South Korea has rapidly increased in recent years. According to statistics from the Ministry of Justice, migrant workers constitute the largest group among all foreign residents in South Korea, followed by foreigners married to Koreans (referred to as immigrant marriages). This has led to the robust formation of multicultural families in South Korea and has transformed Korean society into a multicultural society.

3.1 Language policy for multicultural families

In South Korea, language policies for minority languages (of immigrants) are closely

related to and are practically part of policies concerning "multicultural families".[4]

All children born from international marriages are guaranteed Korean citizenship and all accompanying rights at birth, according to the amended Nationality Act of 1997[5]. With a diverse population and increasing needs of children from multicultural families, in 2006, The Ministry of Education and Human Resources Development (MEHRD, now known as the Ministry of Education, Science, and Technology (MEST)) first addressed the needs of multicultural children through the announcement of the Educational Support Plan for Children from Multicultural Backgrounds (ESP). Alongside measures to support mixed-heritage children in learning Korean for integration into Korean society, there are also bilingual education support programs, encouraging children to continue using their mother tongue and requiring children to teach simple expressions in that language to classmates. However, these efforts are faint and not strongly promoted compared to measures supporting children's integration into Korean society. Therefore, multicultural education at that time (2006) was still seen as a nationalistic education imposed on multicultural children to integrate them into Korean society rather than amplifying the voices of students from different cultural backgrounds, at least within mainstream school contexts (see Lee, 2013).

The social landscape has since undergone many changes, with a stronger emphasis on supporting multicultural families and officially integrating heritage language education programs for mixed-heritage children into national language policies to support their development of unique identities. The government has addressed this issue three times through laws enacted by the Ministry of Gender Equality and Family[6], specifically:

- In the "Enforcement rule of the support for multicultural families act" (2008) of South Korea, there is a section "Provision of multilingual services" to support the language learning of mixed-heritage children.
- In Enforcement decree of the multicultural families support act (2019)
- In Multicultural families support act (2020)

The Multicultural Families Support Act has provided a legal basis for multicultural family policy. In implementing these policies, multicultural family support centers nationwide have been established to support the stable lives of immigrants and their families in South Korea (as of 2020, there are 228 centers)[7]. Language support activities include teaching Korean to immigrants (both adults and

---

4 Multicultural Family: "A general term used to refer to families composed of individuals of different ethnic and cultural origins from our own". (The Ministry of Education and Human Resources Development (MEHRD, but currently known as the Ministry of Education, Science, and Technology (MEST), 2006a, p. 1). By 2008, multicultural family was redefined as: a family consisting of a person with the nationality of the Republic of Korea and a marriage immigrant or a person with naturalization permission.

5 In which the definition of a Korean citizen includes those born to one or both parents who are Korean nationals (Government Legal Department)

6 - Ministry of Gender Equality and Family (2008), Enforcement rule of the support for multicultural families act (Ordinance of the Ministry of Health, Welfare and Family), Korean law information center
- Ministry of Gender Equality and Family (2019), Enforcement decree of the multicultural families support act (Presidential Decree), Korean law information center
- Ministry of Gender Equality and Family (2020), Multicultural families support act (Act)

7 https://www.liveinkorea.kr/portal/USA/board/boardFileDownAct.do?fileSeq=152332

children) and teaching heritage languages to mixed-heritage children.

The South Korean government is highly committed to creating a bilingual environment for multicultural children. Currently, the South Korean government has recognized eight foreign languages to be taught in secondary education (German, French, Spanish, Arabic, Russian, Japanese, Vietnamese, and Chinese) to promote two-way bilingual education in schools. At home, there is a special focus on the bilingual family environment project aimed at helping multicultural families establish an environment where children can naturally communicate in both Korean and the heritage language of the immigrant parent from an early age. This project was widely implemented nationwide in August 2023. As part of this initiative, centers are organizing counseling and training for parents on bilingualism, developing interactive programs between parents and children to create an environment where children can naturally communicate in the language of their immigrant parent at home, thereby developing their multicultural identity and supporting their development into global talents[8]. The target beneficiaries of this program are multicultural families with children under 12 years old. As of 2023, there are 210 bilingual trainers ready to implement this program in multicultural centers nationwide.

For mixed-heritage children in Korean-Vietnamese multicultural families specifically, they receive support from Korean-Vietnamese multicultural centers. At these centers, children learn about culture and heritage language (Vietnamese). The Ministry of Gender Equality in South Korea has also developed a specific curriculum for teaching Vietnamese to children in Korean-Vietnamese multicultural families, which is widely used in Korean-Vietnamese multicultural centers.

The governments of both countries have also made significant efforts to create programs to bring Vietnamese brides and mixed-heritage children back to Vietnam to visit their homeland and have these children attend short-term Vietnamese language classes. This helps mixed-heritage children understand their cultural roots, fosters a love for the Vietnamese language and culture, and encourages them to have a desire for further learning and understanding to connect with their ancestral homeland.

In addition to top-down societal implementation policies, activities supporting the teaching and learning of immigrant heritage languages are also being promoted from the grassroots level (bottom-up). Teaching and learning Vietnamese as a heritage language in South Korea receives considerable support from large corporations, enabling children in localities to learn their heritage language for free. Some private educational centers that can be mentioned include the Asia Language Institute and the Hana Center, both of which are non-profit organizations providing services to support immigrants, including heritage language courses.

Mixed-heritage children in Vietnamese-Korean families receive special support from the Hana company. The company sponsors elementary school-aged children under a program called “Hana Children of Asia”, which includes bilingual book receptions, learning Korean, Vietnamese, participating in recreational activities, and all are provided free of charge. A prominent activity in the series of educational support activities for Vietnamese-Korean mixed-heritage children is the Saturday class, where children attend the main office of Hana every Saturday to learn Korean and Vietnamese. Hana’s activities have been organized in Seoul, Gyonggi, Incheon, and Alsan.

Major corporations like Lotte Corporation have language support programs for

[8] According to the Law on Support for Multicultural Families, Article 10, Paragraph 3 regarding education and care for children and adolescents.

immigrant employees, including the development of Korean and heritage languages. Additionally, foreign communities in South Korea (for example, the Vietnamese Community Association in South Korea) also organize classes or events to support immigrants in maintaining their heritage languages.

However, it should be emphasized that despite the government's efforts in building policies and implementing language support policies for foreigners in multicultural families, the overall level of multicultural acceptance in South Korea remains low, while negative perceptions of multicultural society tend to increase. Therefore, South Korea urgently needs to enhance education for its people to increase multicultural acceptance, thereby hoping to achieve better results in language education.

3.2. Korean-Vietnamese multicultural families and the preservation, development of Vietnamese as a heritage language in South Korea.

Within the Vietnamese community in Korea, in a general context, the main immigration pathways are primarily through labor export and marriage with Koreans. Korean-Vietnamese multicultural families mainly consist of Korean men and Vietnamese women.

Regarding the educational attainment and occupations of Vietnamese individuals within Korean-Vietnamese multicultural families, with the assistance of teachers at multicultural centers in Seoul, Alsan, Gwangju, and Busan, we conducted a survey and received responses from 131 Vietnamese mothers. In terms of educational attainment, 118 individuals had a high school education (90.1%), 2 had vocational education (1.5%), and 11 had college or university degrees (8.4%). Among those with a high school education, 31.4% graduated from high school (completed 12th grade) and 68.6% had sporadic education from elementary to junior high school or had started high school but did not complete it due to marriage. In terms of occupation, 46 mothers were homemakers (35.1%), 69 worked in sales or other part-time jobs (52.7%), and 16 worked in institutions or organizations with fixed salaries (12.2%).

The role of mothers in educating their children has been proven to be paramount in the early years of a child's life and impacts their entire future. Many researchers refer to mothers as transmitters because the language a child's mother speaks at the start of his/her life is the best predictor of his/her future language use. The proficiency of youth, especially in their child-bearing years, holds the future of any language community. Put another way, young people who have strong language abilities, cultural knowledge and identify positively with the community are its insurance policy. (see (Do, 2013)

Meanwhile, most Vietnamese girls marrying Korean husbands originate from the Mekong Delta region such as Can Tho, Dong Thap, Bac Lieu, Ca Mau, Tay Ninh, or some rural areas in the North of Vietnam, characterized by difficult socioeconomic backgrounds, with educational levels mostly reaching elementary and junior high school standards in Vietnam. The educational level of the mother not only has a direct impact but also an indirect influence on child education. Mothers with higher education levels tend to have stable positions in society, enhance their role within the family, have more say in decisions concerning their children, and receive respect from their husband's family. These conditions are favorable for their children's Vietnamese language learning.

Conversely, mothers with lower education levels typically engage in domestic work at home or take low-paying jobs involving manual labor (earning less than

500,000 won per month, approximately 450 USD), and may not receive adequate respect from their husband's family, thus their influence in educating their children is diminished.

Furthermore, the influence of fathers on their children's education cannot be overlooked. Most fathers in families surveyed where the mother is Vietnamese do not speak Vietnamese. Many Korean-Vietnamese marriages are arranged, and Korean fathers often do not engage in Vietnamese-related activities or have visited Vietnam before marriage. Therefore, their understanding of Vietnamese language and culture is almost nonexistent. After marriage, Korean fathers may gain some understanding of Vietnamese culture through movies or media, but they do not learn Vietnamese. Consequently, they cannot communicate with their children in Vietnamese or understand Vietnamese culture enough to share with their wives or assist their children.

Most fathers work in physically demanding jobs, leaving little time at home to care for and educate their children, which is mainly left to the mothers. Fathers' lack of knowledge of Vietnamese leads to communication primarily in Korean within the family, making it difficult to create a bilingual environment for the children.

Korean-Vietnamese multicultural families receive support from the government through multicultural centers. These centers assist Vietnamese mothers in learning Korean to integrate into Korean society, while also helping mixed-heritage children learn Vietnamese and Vietnamese culture. In our observation, these centers play a significant role in motivating families to maintain and develop heritage languages for the second generation. However, the effectiveness of these centers in maintaining and developing heritage languages for children requires further discussion.

From previous studies (Do, 2013, 2017, 2021), we have found that the main factors influencing the maintenance of heritage languages for the second generation often include:

- Parents' awareness of the role of heritage languages
- Amount of time learners spend exposed to heritage languages
- Policies, programs supporting language learning from the government or community organizations.

Other factors also impact the above factors, such as parents' educational attainment, their social status significantly influencing their awareness of the role of heritage languages and how they support their children at home. Furthermore, these factors interact with each other, for example, effective implementation of language learning policies and programs can change parents' attitudes and awareness of heritage languages, thereby increasing the time learners spend exposed to them.

Therefore, we conducted surveys on the demand for heritage language learning among Korean parents in multicultural Vietnamese-Korean families in 2012 and 2024 to observe changes in parental attitudes toward heritage languages. Surveying over time provided us with quite interesting results.

In 100% of the Vietnamese-Korean multicultural families surveyed, the mothers were Vietnamese and the fathers were Korean.

Table 1. Information on parents' demand for children to learn Vietnamese (2012), surveying 42 multicultural families.

Mother		**Father**	
In need	No need	In need	No need
100%	0	64.3%	35.7%

Among the reasons why parents want their children to attend Vietnamese language classes, listed in order of importance:

1. To communicate with their mother and extended family.
2. To understand the culture of their homeland and preserve the heritage of Vietnamese origin.
3. To have better opportunities in the future.

Reasons why 35.7% of Korean fathers (15 people) do not feel the need for their children to learn Vietnamese, listed in order of importance:

1. Not necessary (time spent on Vietnamese classes or teaching Vietnamese could be better spent on learning Korean for better communication with the child).
2. Time-consuming (mothers should instead focus on household chores).
3. Grandparents disapprove (due to family living arrangements).
4. Causes misunderstanding within the family (father and maternal relatives do not understand Vietnamese).

However, among families where both parents are interested in their children learning Vietnamese, only 12 families send their children to cultural centers for Vietnamese classes. The rest have never sent their children to classes, with 16 mothers never speaking Vietnamese at home with their children. Reasons for this include:

1. Parents lack time.
2. Grandparents disapprove.
3. Causes misunderstanding within the family (maternal relatives do not understand Vietnamese).
4. Fear of language confusion in children.
5. Other reasons (such as centers being too far away, waiting until the child is older, etc.).

This situation is prevalent in many multicultural families with different nationalities in Korea. Therefore, cultural centers not only support parents and children at the center but also send representatives to homes to counsel multicultural families, aiming to change people's perceptions about preserving heritage languages for children. We conducted an additional survey in 2024 to observe these changes.

Table 2. Information on parents' demand for their children to learn Vietnamese (2024), survey of 131 multicultural families (answered by mothers).

Mother		**Father**	
In need	No need	In need	No need
100%	0	77,8%	22,2%

By 2024, in a survey involving 131 mothers from multicultural families, we obtained a positive result where the percentage of fathers desiring their children to learn Vietnamese has increased (77.8%). Although there are still 22.2% of fathers who do not have this desire for their children to learn Vietnamese, this number is significantly lower compared to previous surveys (especially in a larger sample size). There is potential for this number to continue to improve in the future.

The compiled reasons why mothers want their children to learn Vietnamese are ranked as follows:

1) Vietnamese people should learn Vietnamese.
2) Vietnamese is important for the future of the child.
3) Vietnamese is important for communication between mother and child.
4) Vietnamese is important for maintaining relationships with extended family (homeland, relatives), and helps the child understand their roots.

Thus, compared to 2012, mothers' perceptions of the role of Vietnamese in the lives of mixed-heritage children have changed. Firstly, mothers have a clearer understanding of the role of heritage language "Vietnamese people should learn Vietnamese." However, the biggest change is in the positions of reasons 2, 3, and 4. Previously, the role of Vietnamese in the relationship between mother and child, and between child and extended family, was prioritized, but now, the role of Vietnamese in the child's future is more emphasized. Moreover, mothers believe that they need to learn Korean to communicate with their children, so the role of Vietnamese in communication between mother and child is not as crucial. This is a concern for us in the research process because learning the mother tongue through maternal interactions in the early years of life is extremely important for a child (see Do, 2013). The fact that mothers prioritize learning Korean over their children's learning of Vietnamese will greatly affect the future interactions between mother and child, and thereby directly or indirectly affect the child's psyche. However, this is a topic that needs further discussion in a more detailed report.

There have been changes in the reasons why fathers do not have the desire for their children to attend Vietnamese classes. Fathers only mention two main reasons: firstly, they fear that their children will be discriminated against if they do not focus on learning and excelling in Korean, and secondly, they fear that it will cause family conflict. Therefore, although multicultural family policies have had an effect, there is still a need to continue to promote awareness among families, especially among fathers, in the future.

Although multicultural policies have somewhat influenced the awareness of parents regarding the role of heritage languages, families are not yet fully prepared to create a bilingual learning environment for their children. All mothers in the surveyed 131 multicultural families confirmed awareness of the government's "bilingual family environment" program. However, only 16 families (12.2%) have created a bilingual environment for their children from an early age. The remaining 115 families (87.8%) still prioritize Korean as the family language, with Vietnamese used intermittently alongside Korean in communication between mother and child, or mothers maintaining Vietnamese for listening comprehension for their children (mothers speak Vietnamese, children respond in Korean – which is the majority case). These 115 families are also not ready to implement the bilingual environment program at home.

Additionally, 89 families (67.9%) currently send their children to Vietnamese

language classes at multicultural centers, but a significant 42 families (32.1%) either do not send their children to these classes or have discontinued them. Thus, it can be observed that language policies have not yet deeply penetrated each family and have not strongly influenced the maintenance of heritage languages for mixed-heritage children in multicultural families. However, with the new bilingual program being implemented at home, there is hope for better results in the near future.

4 Policies related to the preservation and development of Vietnamese as a heritage language in Japan

Japan has also been traditionally seen as a racially homogeneous country with little cultural diversity. (Burgess, 2007) pointed out that "in practical terms, there is little concrete evidence of multiculturalism at work in contemporary Japan." However, Japanese society is currently facing the phenomenon of increasing aging and a low birth rate, leading to the need to accept a large number of immigrant workers annually to address economic and social issues. Therefore, the transformation into a multicultural society is also becoming inevitable, similar to South Korea.

4.1 Language policies related to immigrants in Japan

Observing Japan's policies towards immigrants, we find that the situation is quite similar to South Korea's initial phase before enacting laws supporting multicultural families. This is a period of forming and enhancing public awareness of a multicultural society, leading to subsequent phases: promoting measures to help immigrants not only integrate into Japanese society but also preserve their own cultural identity.

Currently, the Japanese government focuses more on supporting immigrants to integrate into Japanese society rather than on cultural exchange and preservation of minority cultures. Even though Foreign Residents Information Centres have been established in some cities to provide advice to immigrants (Immigration Bureau of Japan, 2010), these centers primarily concentrate on assisting immigrants in integrating into Japanese society rather than helping them preserve their cultural identity.

Apart from the lack of multicultural policies, there are also very few policies combating discrimination and protecting minorities and migrants ((Bradley, 2014). Moreover, there is no policy supporting children born into immigrant families (or families with foreign elements) to learn their mother tongue. In schools, the primary foreign language taught is English without emphasis on any other languages.

In 2021, a report by the Central Council for Educations highlighted the importance of heritage languages in establishing the identity of children connected to foreign countries. However, the report suggests that mother tongue education should primarily occur within the home. It also urges schools and boards of education to collaborate with civil society and other organizations to ensure that children with ties to foreign countries have opportunities to engage with their mother tongue.

Nevertheless, to date, the Japanese government still lacks policies for developing mother tongue education for immigrants. The government and some non-profit organizations have only recently begun implementing measures to support the maintenance and development of mother tongues within immigrant communities through non-profit organizations or community groups. The government may provide financial support to non-profit organizations or collaborate with immigrant community organizations to organize classes or events aimed at supporting the development of

mother tongues and preserving their cultural and linguistic heritage. Two notable organizations in Japan providing services to support the development of mother tongues for immigrants are:

- Mother Tongue Tokyo: Mother Tongue Tokyo is an organization in Tokyo aimed at supporting the preservation and development of immigrants' mother tongue and their families. They provide courses, events, and supporting materials to help immigrants maintain their mother tongue skills in their new environment.
- Mother Tongue Osaka: Similar to Mother Tongue Tokyo, the Mother Tongue Osaka organization also provides support services for immigrants and their families to help them maintain and develop their mother tongue.

In addition to these two organizations within the country, Vietnamese communities in each locality also organize classes or events to support Vietnamese language learning for immigrants and children in the community.

Osaka is the only prefecture that has implemented language development policies for immigrants and provides mother tongue education programs in elementary schools. In Osaka Prefecture, there are junior high schools where the mother tongue is taught as part of Japanese language and adaptation classes.

So, in Japan, there are currently only bottom-up language heritage support activities and they are not widespread. Japan has only recently begun to pay attention to the development of heritage language for immigrants, evidenced by the increasing number of workshops, research projects on issues related to immigrants (including language), which are receiving significant attention from the government. Therefore, in the next few years, it is possible to hope for a new situation for new research on heritage language policies in this country.

4.2 Japanese-Vietnamese Multicultural Families and the Maintenance and Development of Vietnamese as a Heritage Language.

According to statistics from the Japan Immigration Services Agency, the majority of Vietnamese in Japan are short-term laborers, trainees, or individuals with high qualifications. Meanwhile, those with the residency status "Spouse of Japanese National" constitute a relatively modest number of 4,758 individuals. Among them, Japanese-Vietnamese multicultural families include both models: Japanese husband - Vietnamese wife and Vietnamese husband - Japanese wife (not predominantly skewed towards the Korean husband - Vietnamese wife model as in South Korea).

Regarding educational background and occupations of Vietnamese individuals in Japanese-Vietnamese multicultural families, we conducted a survey with 78 Japanese-Vietnamese families in Kyushu (Okinawa) and Tokyo. These families include both men and women of Vietnamese origin. In terms of educational attainment, 49 individuals have completed secondary education (62.8%), while 29 have completed vocational college or university (37.2%). Among those with secondary education, 94% have graduated from high school (up to grade 12), and 6% are currently incomplete in their secondary education.

Regarding occupations, 43 individuals are homemakers (55.2%), 26 are involved in sales or part-time work (33.3%), and 9 work in organizations or institutions with fixed salaries (11.5%).

Thus, in terms of education, Vietnamese individuals in Japanese-Vietnamese multicultural families have relatively basic educational backgrounds (mostly completing at least high school), showing a trend towards higher education compared to Vietnamese individuals in Korean-Vietnamese multicultural families. In terms of occupation, due to the general characteristics of Japanese society, Vietnamese women are primarily engaged in homemaking, sales, or part-time work, while Vietnamese men are primarily employed outside.

Japanese-Vietnamese multicultural families are formed through both matchmaking and natural marriage processes. After marriage, these families often do not live with the Japanese parents (older generation), thereby enjoying relative independence in life and experiencing less influence from their extended families. This differs significantly from Korean-Vietnamese multicultural families, where a majority live together across multiple generations and experience substantial family influence.

The number of children in Vietnamese families (including families where both parents are Vietnamese and Japanese-Vietnamese multicultural families) in Japan is increasing. However, the number of children born into multicultural families is not high (due to the small group marrying Japanese nationals), primarily comprising children in Vietnamese immigrant families. For children with both parents being Vietnamese immigrants, both their families and society believe that maintaining and developing Vietnamese at an early age is not a significant challenge (this will be discussed further in another study). These children are focused on developing Japanese to integrate with friends and not lag behind in mainstream education. The issue of maintaining and developing Vietnamese is urgent for mixed-heritage children in Japanese-Vietnamese multicultural families. However, the number of multicultural families is not large, perhaps why the Japanese government neglects maintaining the heritage language for Japanese-Vietnamese mixed-heritage children.

Without multicultural centers like in South Korea, Japanese-Vietnamese multicultural families do not receive support or advice on teaching their children the heritage language or creating bilingual environments at home. Nevertheless, parents are actively aware of the heritage language.

We also conducted a survey to understand the attitudes of Vietnamese-Japanese multicultural families towards teaching Vietnamese to their children in 2020 (with 17 families) and 2024 (with 78 families). The situation in Japan is very different from Korea. First, Vietnamese people in Japanese multicultural families come to Japan for labor export or as experts, not primarily through marriage. They already know Japanese (basic or intermediate level), so while they have a need to improve their language skills, their existing Japanese proficiency is sufficient for their daily lives in Japan. In contrast, Vietnamese mothers in Korean-Vietnamese multicultural families mostly have poor Korean language skills, some not knowing Korean at all, thus needing to focus on learning Korean for living in Korea (thus reducing time for their children to learn Vietnamese). Second, if 100% of mothers in Korean multicultural families are mothers (in families surveyed), then in Japan, Vietnamese in multicultural families have both parents (in families surveyed). Third, regarding the proportion of parents wanting their children to attend language classes, 100% of the parents surveyed at both times agreed that teaching Vietnamese to mixed-heritage children is extremely necessary and should definitely be done. Parents believe that even though they live in Japan, their children still have Vietnamese roots, so they need to learn the language and culture of Vietnam. The reasons for this choice are mentioned as follows:

1) Vietnamese people should learn Vietnamese.
2) Children will have better opportunities in the future.
3) Bilingual education helps develop children's thinking.
4) Vietnamese is already spoken at home (as the mother tongue of the father or the mother), so it would be a waste for not learning it.

The first and second reasons are similar to the reasons why Korean parents want their children to learn heritage languages. However, the third and fourth reasons are somewhat different. Parents in Japan focus more on the child themselves, aiming at the child's internal thinking and abilities rather than external factors (such as family relationships, family conflicts, etc.). This indicates a strong demand for children to learn Vietnamese from the ground up in Japan, and with policy support, language learning can yield very positive results.

However, due to the lack of language policy support, mixed-heritage children have not been actively learning Vietnamese. In our survey of 78 families in 2024, only 13 families paid attention to bilingual education at home (16.7%), while 65 families still haven't created a bilingual environment for their children and primarily communicate with them in Japanese (83.3%). Some families have sent their children to Vietnamese language classes organized by the Vietnamese community, but these classes are small-scale and lack diversity in terms of children's proficiency levels. As children grow older, the heritage language within the family erodes, and the absence of an environment for heritage language education (from social centers, communities) will prevent them from maintaining it, leading to language loss. Therefore, 100% of the surveyed families expressed a desire for the Japanese government to have better policies for maintaining and developing heritage languages for mixed-heritage children and immigrant children.

5 Conclusion

Both South Korea and Japan have been transforming into multicultural societies, where many immigrant languages coexist alongside the official national language. Vietnamese is one of the minority languages in both countries.

Both nations have implemented policies or measures to support minority groups in integrating into their societies and have been gradually striving to help these groups maintain and develop their cultural and linguistic identities through heritage language and cultural education.

Language policies in South Korea, both top-down and bottom-up, have proven quite effective, yielding promising results nearly 20 years after implementation. There has been a noticeable positive shift in the awareness of heritage language value among multicultural Korean-Vietnamese families. In contrast, Japan currently lacks top-down language policies to develop heritage languages but has seen grassroots initiatives responding to public demand. These activities are increasingly gaining momentum and are expected to influence the government to enact top-down policies in the near future.

South Korea's policies have shown some quantitative effectiveness, but they have not yet demonstrated substantial qualitative effectiveness in family environments, as evidenced by surveys showing that language policies have not fully penetrated family settings for Korean-Vietnamese multicultural families. Despite strong grassroots support, many Korean-Vietnamese multicultural families have not created bilingual environments at home to maximize their children's linguistic potential, enabling them to become global citizens as hoped.

Survey results also indicate that regardless of government support, the family environment remains crucial for maintaining and developing children's language skills. As (Li, 2006) noted, many researchers recognize that family members and the family environment, rather than policies or laws, play a pivotal role in preserving heritage languages. Therefore, even without government support, Vietnamese-Japanese multicultural families still hold positive perceptions of heritage languages, with some families independently fostering bilingual family environments for their children. These bottom-up positive activities will contribute to pushing the Japanese government to recognize the necessity of implementing suitable policies to preserve heritage languages for second-generation immigrant children.

Language policies for multicultural families in South Korea are increasingly being improved due to South Korea's early recognition of multiculturalism and its acknowledgment of the role of preserving the cultural identity of minority groups (foreign immigrants). This is also a valuable lesson that countries transitioning to multicultural societies, such as Japan, can learn from and apply.

References

Bradley, William S. 2014. Multicultural Coexistence in Japan: Follower, Innovator, or Reluctant Late Adopter? In W. Bradley, & K. Shimizu (eds), *Multiculturalism and Conflict Reconciliation in the Asia-Pacific* (pp. 21-43). London: Palgrave Macmillan.

Burgess, Chris. 2007. Multicultural Japan? Discourse and the 'Myth' of Homogeneity. *The Asia-Pacific Journal: Japan Focus, 5*(3), 1-25.

Deusen-Scholl, Nelleke V. 2003. Toward a definition of heritage language: sociopolitical and pedagogical considerations. *Journal of Language, Identity, and Education, 2*(3), 211-230.

Do, Duong H. 2013. The role of maternal interaction in Vietnamese teaching in South Korean-Vietnamese multicultural families. In KFAS Studies, *New Asia Forum* (pp. 147-161). Seoul.

Do, Duong H. 2017. Vietnamese as a heritage language for Korean-Vietnamese mixed-race children in South Korea. *Journal of Language and Life, 12*, 22-30.

Do, Duong H. 2021. The application of Vietnamese linguistics in teaching and learning Vietnamese as a heritage language. *Language Journal, 11*, 104-115.

Ferguson, Charles D. 1959. Diglossia. *Word No 15*, 325-502.

Fishman, Joshua A. 2001. 300-plus years of heritage language education in the United States. In Peyton, Ranard & McGinnis (Eds) *Heritage languages in America: Preserving a nationalresource* (pp. 81-98). Washington DC & McHenry, IL: Center for Applied Linguistics & Delta Systems.

Fishman, Joshua A. Nahirny Vladimir C.,Hofman John E., & Hayden Robert G. 1966. *Language Loyalty in the United States: The maintenance and perpetuation of non-English mother tongues by American Ethnic and Religious groupds.* The Hague: Mouton&Co.

Hornberger, Nancy H. 1998. Language Policy, Language Education, and Language Rights: Indigenous, Immigrant, and International Perspectives. *Language in Society, 27*(4), 439-458.

Hymes, Dell. 1972. On communicative competence. In Pride & Holmes (eds), *Sociolinguistics. Selected Readings* (pp. 269-293). Harmondsworth: Penguin.

Kelleher, Ann. 2010. *Heritage.* Retrieved from https://www.cal.org/heritage/pdfs/Who-is-a-Heritage-Language-Learner.pdf [accessed 21 Nov 2023]

Labov, William. 1972. *Language in the Inner City: Studies in the Black English Vernacular.* Pennsylvania: University of Pennsylvania Press.

Lee, Siwon. 2013. Multicultural Education and Language Ideology in South Korea. *Working Papers in Educational Linguistics, 28*(1), 43-60.

Li, Guofang. 2006. The role of parents in heritage language maintenance and development: Case studies of Chinese immigrant children's home practices. In Kondo-Brown (ed), *Heritage language development: Focus on East Asian immigrants: Studies in Bilingualism Series, 32* (pp. 15-31). Amsterdam: John Benjamins.

Montrul, Silvina. 2013. Bilingualism and the Heritage Language Speaker. In T.K.Bhatia & W.C.Ritchie (eds), *The handbook of bilingualism and multilingualism* (pp. 168-189). West Sussex: Blackwell Publishing Ltd.

Polinsky, Maria. 1997. American Russian: Language loss meets language acquisition. In E. D. W.Brown (ed), *Formal approaches to Slavic linguistics* (pp. 370-407). Ann Arbor, MI: Michigan Slavic Publications.

Polinsky, Maria. 2000. The composite linguistic profile of speakers of Russian in the US. In O. Kagan & B. Rifkin (eds), *The learning and teaching of Slavic languages and cultures* (pp. 437-465). Bloomington, IN: Slavica.

Polinsky, Maria. 2006. Incomplete acquisition: American Russian. *Journal of Slavic Linguistics, 14*, 161-219.

Polinsky, Maria & Kagan, Olga. 2007. Heritage languages: In the "wild" and in the classroom. *Language and Linguistics Compass, 1(5)*, 368-395.

Shin, Giwook; Freda, James, & Yi, Gihong. 1999. The politics of ethnic nationalism in divided Korea. *Nations and nationalism, 5*, 465-484.

Spolsky, Bernard. 2003. *Language policy.* Cambridge University Press.

Valdés, Guadalupe. 2000. The teaching of heritage languages: An introduction for Slavic-teaching professionals. In O. Kagan & B. Rifkin (eds), *The Learning and Teaching of Slavic Languages and Cultures* (pp. 375-403). Bloomington, IN: Slavica.

17

Semantics of Vietnamese Rice Expressions from a Socio-Cultural Perspective[1]

Nguyễn Ngọc Bình

1 Introduction

This chapter focuses on the idioms, proverbs and aphorisms which are mentioned in the terms used to represent rice and its derived terms (seeds, plants, polished, cooked, etc.). It is also not a study of the rice expressions which involve scientific names of rice since the purpose of this kind of naming is not related to cultural elements that this study refers to. The work is based on two main approaches: The figurative language suggested by Corbett (1971: 460) and semantic domain suggested by Ottenheimer, 2006: 18).

We studied articles, journals, textbooks, and theses which are relevant to Vietnamese expressions; collecting rice expressions from dictionaries, books, theses, articles and the internet from Vietnam. Vietnamese rice expressions also were collected by a field trip and the field site is Red River Delta since it is famous throughout Vietnamese history in term of cultivation; and understanding rice expressions was also carried out by deep interviews with informants, especially the experts, scholars, and folklorists in Vietnam. Data was collected from 50 published Vietnamese books and dictionaries as well as field trips since 2015. Then expected data will be analyzed, contrasted and classified into different types and groups. Based on its contents, data will then be categorized and further sub-divided into different types. These types or domains will serve as a database for analysising the socio-cultural perspectives and further analyzed by applying related theories. In order to have validity in current usage, these data will be examined in societal usages. The total Vietnamese expressions collected are 770. The transcription based on Northern Vietnamese Dialect which adapted and edited from Đoàn Thiện Thuật (2004). A part of that data is discussed in this chapter.

2 Semantics of Vietnamese Rice Expressions

2.1 Figurative language

Corbett (1971: 460) states that figurative language is a form of speech artfully varied from common language. Corbett also divided the figurative language into two main groups: the schemes and the tropes. A scheme involves the transference of order and a trope is the transference of meaning.

The schemes are divided into 3 sub-types: schemes of words, schemes of construction, and schemes of omission. The tropes are divided into 14 sub-types:

[1] This chapter is a write-up of a talk given at the 11th 1th International Conference on Austroasiatic Linguistics (Chiang Mai October 2023).

metaphor, simile, synecdoche, metonymy, puns, anthimeria, periphrasis, personification, hyperbole, litotes, rhetorical question, irony, onomatopoeia, and oxymoron. The prominent figurative language found in Vietnamese rice expressions can be divided into metaphor, simile, hyperbole, personification, and onomatopoeia.

2.1.1 Metaphor

Metaphor is an implied comparison between two things of unlike nature that yet have something in common (Corbett, 1971: 479). Example:

(1) *Cơm* *tẻ* *mẹ* *ruột*
cooked.rice ordinary mother bowel
'Ordinary rice is the real mother' (Ordinary rice is familiar to everybody)

In (1), ordinary rice is compared with mother, the meaning marks the important role of ordinary rice.

2.1.2 Simile

Simile is an explicit comparison between two things of unlike nature that yet have something in common (Corbett, 1971: 479). The only difference between a simile and a metaphor is that in a simile the comparison is explicitly stated, usually by a word such as 'like' or 'as' while in a metaphor the comparison is just implied (Barnwell: 1980). Example:

(2) *Chắc* *như* *gạo* *bỏ* *hũ*
sure like unhusked.rice put jar
'As sure as rice put in the jar' (The certainty)

Examples (2) shows the comparison between two things of unlike nature that yet have something in common, 'chắc' (certainty) and 'gạo bỏ hũ' (rice in jar). This example uses the word 'như' (like).

2.1.3 Hyperbole

Hyperbole is the use of exaggerated terms for the purpose of emphasis or heightened effect (Corbett, 1971: 486). Example:

(3) *Con* *cá* *đánh* *ngã* *bát* *cơm*
CLF fish beat fall bowl cooked.rice
'Fish beats down the bowl of rice' (Having fish, eat much rice)

Example (3) shows the hyperbole with the exaggerated term 'đánh ngã' ('to beat down' which is normally used in martial arts) for the purpose of emphasizing the role of fish in eating.

2.1.4 Personification

Personification is the investing abstractions or inanimate objects with human qualities or abilities (Corbett, 1971: 485). Example:

(4) *Chiêm* *đùa* *mùa* *sâu*
rice.seed.in.fifth.lunar.crop joke rice.seed.in.tenth.lunar.crop deep
'Rice seed in fifth lunar month crop is fun; rice seed in tenth lunar month crop is deep'
(Shallow sowing rice seeds in fifth lunar month crop; deep sowing rice seeds in the tenth lunar month crop)

Example (4) shows the personification of investing inanimate object 'chiêm' (rice in fifth lunar crop) with human character 'đùa' (joke).

2.1.5 Onomatopoeia

Onomatopoeia is the use of words whose sound echoes the sense (Corbett, 1971: 491). Example:

(5) *Chiêm* *xấp xới* *mùa* *đợi* *nhau*
rice.tree.in.fifth.lunar.crop ONOM rice.tree.in.tenth.lunar.crop wait each.other
'Unevenly rice in fifth lunar crop; evenly rice in tenth lunar crop'
(Experience in doing the rice works)

Example (5) shows the onomatopoeia which uses the words 'xấp xới' whose sounds echo the sense of happiness.

3 Semantic domains of Social Life

Semantic domain can be defined as a 'specific area of cultural emphasis' and 'The quest was originally to see how the words that groups of humans use to describe certain things are relative to the underlying perceptions and meanings that those groups share' (Ottenheimer, 2006: 18). We suggested a major semantic domain for Vietnamese idioms and proverbs, specifically Social Life.

3.1 Social Life

The social life can be defined as a life in which there is the relationship between humans and their society. In this domain, this study will consider the relationship between rice and its society.

3.1.1 Siblings

Sibling is fundamental and important relationship in the family of Vietnamese. This relationship focuses on substance, showing the love together, and this relationship is not based on materials.

(6) *Anh em* *gạo* *đạo nghĩa* *tiền*
sibling husked.rice moral.principle money
'Brothers for the sake of rice; moral principle for the sake of money'
(A bad relationship; just for materials, not for gratitude)

In (6), the sibling relationship expresses a bad relationship, just for materiality, not for gratitude.

3.1.2 Parents and Children

The relationship between parents and children of Vietnamese is a basic and valuable relationship, and does not change through history. Parents love their children and their children show their caring and gratitude to their parents as the precious image of the Vietnamese culture.

(7) *Muốn cho gần mẹ gần cha khi vào*
want let near mother near father when into
thúng thóc khi ra quan tiền
basket paddy when Prep. CLF money
'In order to stay beside parents, come in with a basket of paddy, go out with money'
(Blame people who want to be close with parent in order to get benefit)

This example wants to criticize people who only know how to take advantage of their parents.

3.1.3 Husband and Wife

A good husband and wife relationship is the foundation of a happy family as well as the development of a sustainable society. The husband and wife relationship is a basic type of relationship in society and is found in many texts in Vietnamese culture.

(8) *Lúa tháng bảy vợ chồng rẫy nhau*
rice.plant month seven wife husband disagree each.other
'Rice in seventh lunar month made husband disagree with wife'
(Experience in doing the rice works)

By referring to the husband and wife relationship, this example expresses a fact that the crop in seventh lunar month is late with low yield and this is a reminder that husbands and wives, and families, often also have problems and arguments.

3.1.4 Husband and concubines

Polygamy was the popular situation in society before 1945, and also existed within the monarchy in Vietnam and this situation leads to the relationship of husband and concubines in the society. However, Polygamy was officially criminalized in Vietnam during the 1950s (based on Vietnamese Constitution 1992), with a lengthy prison sentence as punishment.

(9) *Ăn cơm nguội nằm nhà ngoài*
eat cooked.rice cold lie.down house outside
'Eat the cold rice; sleep outside of the house' (Disadvantage of having concubines.)

In the past, the Vietnamese men could marry many wives. This example shows the disadvantage of concubines by using the image of eating the cold rice and sleep outside of the house.

3.1.5 Son and Daughter

In ancient Vietnamese society, the role of son was often more important than the role of daughter. The son is considered as a person who will be caring for parents when their parents get old whereas daughter is to belong to another family since she have to get married with a man in that family. This situation is still permeates in the present society. The following example encourages everyone to have a fair treatment to both son and daughter.

(10) *Có* *nếp* *mừng* *nếp* *có* *tẻ*
have sticky.rice welcome sticky.rice have ordinary.rice
mừng *tẻ*
welcome ordinary.rice
'Having sticky rice, welcome it; having ordinary rice, welcome it'
(The justice between daughter and son. It is happy for having both son and daughter.)

In this above example, both sticky and ordinary rice expressed its importance, as similar as son or daughter in a family.

3.1.6 Relatives

Relatives are an indispensable element in the society of Vietnam. In an agricultural society, living together in a 'village' unit, people need to know and help each other. Relatives are people who you may share every sorrow and happiness.

(11) *Ăn* *cơm* *nhà* *dì* *uống* *nước* *nhà* *o*
eat cooked.rice house aunt.maternal drink water house aunt.paternal
'Eat rice at aunt's house (mother's side); drink water at aunt's house (father's side)' (Unstable life of the poor children).

The relationship of the maternal aunt is more emotionally strong than the relationship of the paternal aunt in Vietnamese culture. The man in the past may have many wives whereas the virtuous woman can have only one husband. Thus, normally the father's side has many grandchildren and treats their grandchildren less favourably than the mother's side.

3.1.7 Daughter-in-law and Son-in-law

Son-in-law and daughter-in-law are two important factors establishing the warmness of a family in Vietnam. The son-in-law or daughter-in-law must treat well the wife's parents or husband's parents as well as their parents. In daily life, sons-in-law and daughters-in-laws often receive assessments of their family behavior from their parents-in-law respectively. Example 12 provides tips for the son-in-law and the daughter-in-law to their parents-in-law respectively:

(12) *Làm* *rể* *chớ* *nấu* *thịt* *trâu*
do son-in-law IMP cook meat buffalo
làm *dâu* *chớ* *đồ* *xôi* *lại*
do daughter-in-law IMP steam sticky.rice again
'To be the son-in-law, should not cook the buffalo meat; to be the daughter-in-law, should not steam again the sticky rice' (Experience in cooking)

Buffalo meat is often tough if not cooked carefully and sticky rice is not delicious if cooked again.

3.1.8 King subject

Before 1945, Vietnam was a country under royal regime. The King is considered as 'son of Heaven' and had the full rights to dispose of the fate of anyone in the territory he ruled. The following example shows the 'resistance' consideration of the citizens towards the King in Vietnamese society and during the time that the royal regime gradually declined.

(13) *Cơm* *vua* *nợ* *dân*
cooked.rice King debt citizen
'Rice of the King; debt of citizens' (The King takes the rice from citizens' taxes)

This example often refers to the thoughts of men in ancient society, about being in debt to the people, being in debt to the country (eating the king's rice), but still cannot repay them.

3.1.9 Mandarins and citizens

The relationship between mandarins and citizens constitutes the governmental relations in ancient Vietnamese society. The Mandarin in this case is with money and power while the citizen is often voiceless. However, during the French colonial period, the government allowed money to buy "power". Therefore, some citizens became mandarins.

(14) *Lúa* *ông* *Láng* *bạc* *ông* *huyện* *Xanh*
rice.plant 3 PN silver 3 district PN
'Rice of Mr. Lang; silver (money) of Mr. Xanh (the head of district)'
(The power of local people)

Mr. Lang, a Lang villager, moved to live in Phuong Nhue village (Phu Tho province) and began to trade with a small shop to sell water. Then he bought land and became rich. Mr. Xanh, lived in the same village, pressed and sold the oil-tree and became rich. He then bought power to become the head of district.

3.1.10 Master and servant

The master and servant relationship played an important role in Vietnamese society. Vietnamese believed that if someone wants to be good, they must have the help of their masters. Therefore, the role of masters in the society is especially important.

(15) *Cơm* *cha* *áo* *mẹ* *công* *thầy*
cooked.rice father clothing mother merit master
'Father's rice, mother's clothes, teachers' merit'
(Show gratitude to fathers, mothers, teachers)

'Thầy' (master) is a boss in the old society. To make a living, employees always have to be patient and endure the boss's yelling and threats.

3.1.11 Individuals and society

Individuals and society is an important relationship in Vietnam society. Each individual should be responsible to society and others in society. The presentation of the individual to society showing her perceptions and educational level in society.

(16) *Bà* *tiền* *bà* *thóc* *bà* *cóc* *gì* *ai*
madam money madam paddy madam NEG what who
'Having money and rice mean everything'
(Having high position and esteem based on money, not based on talent and virtue)

(16) criticizes people who have money but look down on the poor.

3.1.12 Richness and Poorness

In Vietnamese agricultural society, rice is the measure of richness and poorness. Vietnamese people believe that if anyone has much rice, it means that he is rich and in contrast, the poor own less rice.

(17) *Khen* *nhà* *giàu* *lắm* *thóc*
praise house rich much paddy
'To praise the rich who have much paddy'
(Complimenting something of course, everyone knows it.)

(18) *Giàu* *lo* *bạc* *đói* *lo* *cơm*
rich worry silver hungry worry cooked.rice
'Richman worries in silver (money); hunger worries in rice'
(Everybody has his own stresses depending on their conditions.)

Everybody has their own stresses depending on their conditions; the rich man worries about money, and the poor man worries about how to get enough to eat.

3.1.13. Trading

Trading is an important method to earn a good living in ancient Vietnamese society. The position of trading is not considered to be as important as education or civil jobs since the Vietnamese believed that trading means deceiving others in order to gain profit.

(19) *Buôn* *trấu* *dấm* *bếp* *buôn* *tro* *trồng* *hành*
trade rice.husk brew kitchen trade ash grow onion
'Trade in rice husk for brewing on the stove, trade in ash for planting the onion'
(Praise the person who does smart business)

(19) praises the person who does smart business by taking advantage the things which they did when trading.

3.1.14 Village

In a society where 'village' culture occupies a central position, complying or encouraging the activities of the village is a requirement of every individual in society. The below example reflects the relationship of the individual to their village.

(20)	*Vỗ*	*tay*	*làng*	*cho*	*ăn*	*xôi*	*không*	*vỗ*	*tay*
	clap	hand	village	give	eat	sticky.rice	NEG	clap	hand
	làng	*lôi*	*xuống*	*hồ*					
	village	pull	downward	lake					

'Villagers treat the sticky rice if applauding; villagers pull into the lake if no applauding' (Give prominence to the living together harmoniously.)

The main structure of living of Vietnamese is village-culture. This example gives prominence to living together harmoniously in a village.

4 Discussion

In terms of figurative language, this study reveals prominent figurative language of Vietnamese rice expressions including metaphor, simile, hyperbole, personification, and onomatopoeia. These works mostly found in the researches of Vietnamese Literature such as Dương Quảng Hàm, 1943; Vũ Ngọc Phan, 1978; Đinh Gia Khánh and Chu Xuân Diên, 1973; Chu Xuân Diên, Lương Văn Đang and Phương Tri, 1975; Cao Huy Đỉnh, 1974; Bùi Văn Nguyên, Đỗ Bình Trị and others, 1978; Trần Đức Thế, 1995; Nguyễn Văn Thông, 2010... However, these studies do not specifically relate to Vietnamese rice expressions.

In regarding to semantics, Vietnamese rice expressions can be grouped into 4 main semantic domains namely nature, material life, social life and spiritual life, and many sub-domains. This classification can be found in the work of multiple authors of 'Kho tàng tục ngữ người Việt' (Vietnamese proverbs treasure, 2000). However, that collection includes only the Vietnamese proverbs without analysis and does not relate specifically to the Vietnamese rice expressions.

Further study on rice expressions should extend into other fields such as cultural studies, sociology, folklore… in order to gain all-sided information on the related issues. Future research should extend to other countries, especially in Southeast Asian countries for the sake of understanding and exchanging the related issues on rice expressions.

References

English

Duranti, Alessandro. 1997. *Linguistic Anthropology*. Cambridge: Cambridge University Press.

Walker, Anthony R. (ed.). 1994. *Rice in Southeast Asian Myth and Ritual*. Contributions to Southeast Asian Ethnography ISSN 0217-2992.

Swoyer, Chris. 2003. "The Linguistic Relativity Hypothesis". Stanford Encyclopedia of Philosophy. Metaphysics Research Lab, CSLI, Stanford University.

Greville, Corbett G. 1971. *The Expression of Gender*. De Gruyter Mouton.

Nida, Eugene A. 1975. *Componential Analysis of Meaning*. The Hague: Mouton.

Lucien, Hanks, M. 1972. Rice and Man: Agricultural Ecology in Southeast Asia. Chicago: Aldine.

Thongdee, Iam. 1994. *Rice Culture: Ritual concerning Rice and Rice Growing*. Bangkok: Saha Dhammika Limited (in Thai).

Lyons, John. 1995. *Linguistics Semantics: An Introduction*. Cambridge: Cambridge University Press.

Pisitpanporn, Naraset. 1986. *Rice cycle*. M.A. Thesis. Mahidol University.

Ottenheimer, Harriet J. 2006. *The Anthropology of Language*. Belmont, CA: Thomson Wadsworth.

Pumyoo, Watit. 2004. An Ethnosemantic study of rice terms and the conceptual system of rice in Southeast Asian Languages. M.A. Thesis. Chulalongkorn University.

Vietnamese

Chu Xuân Diên, Lương Văn Đang, Phương Tri. 1993. *Tục ngữ Việt Nam* (in lần thứ 2). Nhà xuất bản Khoa học Xã hội, Hà Nội, 420 tr (in lần thứ nhất, 1975, 390 tr).

Đỗ Bình Trị. 1999. "Những đặc điểm thi pháp của tục ngữ", Đỗ Bình Trị, *Những đặc điểm thi pháp của thể loại văn học dân gian*. Nhà xuất bản Giáo dục, Hà Nội, tr.138-163.

Đoàn Thiện Thuật. 2004. Ngữ âm tiếng Việt. Nhà xuất bản Đại học Quốc gia Hà Nội.

Hoàng Văn Hành chủ biên (1994). *Kể chuyện thành ngữ, tục ngữ,* (tái bản lần thứ nhất, có sửa chữa). Nhà xuất bản Khoa học Xã hội, Hà Nội, 384 tr (tái bản lần thứ hai,1999, 438 tr).

Nguyễn Đức Dân. 1987. "Đạo lý trong tục ngữ", *Tạp chí Văn học*. Hà Nội, (số 5), tr.57- 66.

Nguyễn Quý Thành. 1998. "Dấu ấn văn hoá trong tục ngữ", *Văn hoá dân gian*. Hà Nội, (số 4), tr.76- 79.

Nguyễn Thái Hoà. 1997. *Tục ngữ Việt Nam- Cấu trúc và thi pháp.* Nhà xuất bản Khoa học Xã hội, Hà Nội, 263 tr.

Nguyễn Trọng Lực. 1949. Tiếng nói của đồng ruộng (hay là nghề nông Việt Nam) qua ca dao, tục ngữ. Nhà sách Vĩnh Bảo, Sài Gòn.

Nguyễn Văn Mệnh. 1972. "Ranh giới giữa thành ngữ và tục ngữ", *Ngôn ngữ*. Hà Nội, (số 3), tr.12- 15.

Nguyễn Xuân Đức. 2000. "Về nghĩa của tục ngữ", *Văn hoá dân gian*. Hà Nội, (số 4), tr.48-52.

Nguyễn Xuân Kính chủ biên. 2002. *Kho tàng tục ngữ người Việt* (2 tập). Nhà xuất bản Văn hóa- Thông tin, Hà Nội, 3246 tr.

Thái Hoà. 1982. "Cơ cấu ngữ nghĩa- cú pháp của tục ngữ", *Ngôn ngữ*. Hà Nội, (số 2), tr.52-59.

18

The Role of *ruột* 'Intestine' in Vietnamese Culture and Language[1]

Hien Tran, Duong Duy Bui

1 Introduction

In Vietnamese, the word *ruột* 'intestine' denotes one of the internal organs in our body which is defined as "part of the alimentary canal, starting from the end of the stomach to the anus"[2] (Từ điển Tiếng Việt, 2000:838, my translation). For example: *Anh ấy bị viêm ruột* 'He has inflammatory bowel disease', and *các bệnh về ruột và bệnh rối loạn tiêu hóa* 'bowel diseases and digestive disorders'. In addition to referring to the internal organ, *ruột* 'intestine' is associated with different emotional and mental activities. For example: *tức lộn ruột* (lit. so angry that the intestine moves upside down: to be angry), *xót ruột* (lit. to feel a sting in the intestines: to regret), *nở từng khúc ruột* (lit. to every single piece of the intestines is expanded: to be happy), *lo cháy ruột* (lit. worries burn the intestines: to worry), and *lú ruột* (lit. in a poor intellectual state, having no memory, and no wisdom: one loses his memory or wisdom; to be absent-minded), etc. These examples suggest that the occurrence of the emotions and thoughts makes an impact on *ruột* 'intestine' and *ruột* 'intestine' provides clues for the Vietnamese speakers to understand these emotions and thoughts.

From modern medicine, we know that the intestine is the long tube-shaped organ in the abdomen that completes the process of digestion. The intestine has two parts, the small intestine and the large intestine. The small intestine absorbs nutrients and water from food for the body to use. The large intestine absorbs water and salt from the material that has not been digested as food and gets rid of any waste products left over (Medical Dictionary, The Free Dictionary online). However, the medical understanding of *ruột* 'intestine' does not help to explain "What is the function of *ruột* 'intestine' in the human body that has made it to be used to describe emotions and thoughts in Vietnamese?" and "How do Vietnamese speakers believe that the occurrences of the emotions and thoughts will affect ruột 'intestine'?".

By analyzing the uses of the word *ruột* 'intestine', this study presents metaphorical

[1] This chapter is a write-up of a talk given at the 11th 1th International Conference on Austroasiatic Linguistics (Chiang Mai October 2023). This study was funded by University of Social Sciences and Humanities (USSH), Vietnam National University under Grant number CS 2002.37.

[2] In Vietnamese, there are words such as *ruột già* (intestine-old, native Vietnamese) and *đại tràng* (big-intestine, Sino Vietnamese) referring to the large intestine and *ruột non* (intestine-young, native Vietnamese) / *tiểu tràng* (small-intestine, Sino Vietnamese) to the small intestine. My data shows that *ruột non/ tiểu tràng* or *ruột già/ đại tràng* is not used in any of the Vietnamese idioms and proverbs. This fact indicates that *ruột non/ tiểu tràng* or *ruột già/ đại tràng* are used within medical contexts due to their medical meanings. The use of the word intestine in Vietnamese suggests that language use in everyday life is not necessarily accurate as it is in medical contexts.

conceptualizations of *ruột* 'intestine' that are used to describe the emotions and thoughts and proposes conceptual metaphors that account for them. This study also investigates and presents the Vietnamese cultural model of *ruột* 'intestine' that inspires such conceptualizations. It is hoped that this study will provide insights into the conception of *ruột* 'intestine' as a Vietnamese specific cultural construct.

According to Cognitive Linguistics, we tend to understand and express emotions and thoughts which are more tangible and abstract in terms of the human body and associated bodily experiences which are more concrete. The understanding of the human body and associated bodily experiences helps us conceptualize emotions, thoughts and other mental activities (Johnson 1987, Lakoff 1987, Lakoff and Johnson 1999). This statement highlights the role of physical or physiological embodiment in structuring such abstract concepts. Physiological embodiment refers to the interaction between the body and the mind in the environment. It is "the mind emerges and takes shape from the body with which we interact with our environment. Human beings have bodies, and human embodiment shapes both what and how we know, understand, think, and reason" (Yu 2014:227). Therefore, human cognition is bodily- physiological basis, that is, embodied. It means that the way we understand and describe emotions and thoughts is motivated by physical or physiological embodiment.

The question then arises as to whether it is the physical or physiological embodiment that can be the answer for all the ways we perceive our emotions and thoughts. Note that human beings across cultures share the basic structure of the human body and many basic bodily experiences as well. This fact would lead to an assumption that the same human body part along with the same bodily experiences can be used to understand the same emotion in different cultures. However, the facts show different situations for various languages.

In English, the heart is viewed as a container of different emotions (Niemeier 2008) but it is viewed as the center of thoughts, ideas, emotions and feelings in Chinese (Yu 2008). Or different human body parts and different physiological experiences across languages can be used to describe the same emotion concept. For instance, the anger emotion in English can be understood in terms of 'hot fluid': *You make my blood boil; Let him stew* (examples from Lakoff 1987), but in Chinese, anger is described in terms of *qi* 'gas' (qi: the energy that flows in the body (Yu 1995)), *His anger qi/gas calmed down* (example in Yu 1995).

Furthermore, evidence from different languages shows that in many situations, no actual physiological responses are involved when the emotions and thoughts occur. For example, anger is described as *He reduced my flesh into crumbs* meaning he was angry in Tunisian Arabic (Maaalej 2004) and as *His chest grew weeds/ became weedy* meaning he was angry, in Akan (Ansah 2011). Thoughts in Chinese can be understood in terms of the heart as in "*However, he can only put this matter in his heart to think about (…)"* (Yu 2008:143) but thoughts are understood in terms of one's *small liver* in Indonesian (Siahaan 2008:68).These examples show that no actual physiological responses of the body parts are used to understand the emotions and thoughts. This indicates that these concepts of emotions and thoughts are not structured by physiological experiences but culture-specific experiences.

The fact that many abstract concepts are structured by cultural-specific experiences leads to cultural embodiment, a new approach to explain the historical and cultural basis of human cognition by exploring cultural models and traditions which help organize, explain viewpoints, and motivate conceptualizations of abstract concepts in the societies (Yu 1995, 2008; Kövecses 2000; Maalej 2004; Siahaan 2008; Ansah

2011). It should be noted that cultural embodiment “occurs when physiological embodiment departed from insignificant ways, thus constructing a culturally-situated form of embodiment” (Maalej 2008:396). That is, all physiological responses assumed to be associated with abstract concepts as emotions or thoughts are all ignored in a given culture. This indicates the cultural base of such abstract concepts which is the grounding of the cultural embodiment approach in cognitive linguistics.

The explanations for the fact that different languages may use different body parts and physiological or cultural experiences in descriptions of abstract concepts lie in cultural models (Yu 2003). Cultural models can be thought of as “shared, structured knowledge” (Kövecses 2004:114) which shape “what people believe, how they act, and how they speak about the world and their own experiences” (Gibbs 1999:155). Thus, the cultural model “provides the members of a cultural group with “templates” for understanding certain aspects of their lives” (Sharifian et al. 2008:12). Specifically, Yu (2003)’s study indicates that the cultural models select certain body parts of the body and certain aspects of bodily experiences which are seen as salient and meaningful to be linked to abstract concepts in order to understand them. This means that language users in different cultures interpret their body and associated bodily experiences differently, or they can attach different values to the same bodily experiences or to the same parts of the body (Maalej and Yu 2011:6) in understanding abstract concepts. Consequently, different interpretations of the body and bodily experiences lead to varied conceptual metaphors and metonymies in different languages (Maalej and Yu 2011:7).

Previous studies highlight the role of culture in shaping metaphors and explaining the preferences for certain metaphors in different cultures. For instance, Yu (2008)’s study indicates the preference for the heart metaphors as the representations of thought, ideas and emotions in Chinese is based on ancient Chinese philosophy and traditional Chinese medicine. Siaahan in her study (2008) gives evidence that the liver metaphors in Indonesian that are descriptions of emotional and mental activities are rooted in the Indonesian animistic belief.

The structure of the present paper is organized as follows: Section 2 presents the data and methodology that are used for this study. Section 3 focuses on the Vietnamese folk belief which views the intestine as the seat of human life which we argue that serves as the basis for the conceptualization of the INTESTINE [3] as representations of emotions, thoughts, human character traits and cultural values in Vietnamese. Section 4 examines linguistic evidence involving the word ruột ‘intestine’ in the Vietnamese language and compares it with those in English if relevant. At the end of this section, the Vietnamese cultural model of ruột ‘intestine’ is proposed from the perspective of cultural perception. The conclusion of the paper is presented in section 5.

2. Data and methodology

2.1. Data

The data on which this study is based come from two sources. The first is the data consists of 127 lexical units that use the word ruột ‘intestine’ and were collected from one Vietnamese dictionary and one Vietnamese dictionary of idioms and proverbs. The second is the data that represents linguistic contexts of the word ruột ‘intestine’ in the

[3] The capital letters indicate concepts not words or linguistic expressions (see Lakoff 1987).

first source. The definitions of the lexical units in those dictionaries do not provide sufficient information to explain why Vietnamese speakers think of, understand, and express ruột 'intestine' in the ways they do, therefore this study collected contexts in which the emotion occurs from six Vietnamese e-news websites including: *vnexpress.net, ngoisao.net, vietnamnet.vn, tuoitre.com.vn, thanhnien.com.vn, tintuconline.com.vn.* (alexa.com 2008) in order to provide a scene of the intestine which shows how Vietnamese speakers use the word ruột 'intestine'. Within the framework of Cognitive Linguistics, this paper analyzes the Vietnamese expressions of ruột 'intestine' in order to establish the references of ruột 'intestine' accurately.

2.2. Methodology

This study aims to establish conceptual metaphors of *ruột* 'intestine' in understanding and shaping several abstract concepts. According to Lakoff (1993), conceptual metaphor is a set of conceptual mappings between a source and a target domain. The source domain is associated with tangible and physical experiences therefore it is typically concrete; the target domain is associated with abstract experiences such as emotions, thoughts, life, arguments, etc. therefore, it is more abstract than the source domain. The conceptual relations between the source domain and the target domain can be captured in the formula: A IS B (A is understood in terms of B) (Lakoff 1993).

In order to establish metaphors of *ruột* 'intestine', this study uses the metaphor identification procedure which was developed in Tran (2018) to identify conceptual metaphors. This procedure is based on the principles of the MIP (Pragglejaz Group 2007) which was designed to improve the identification of conceptual metaphors, especially those in discourse contexts.

The metaphorical identification procedure for this study consists of the following steps:

- Step 1: Read the whole context to establish its general meaning
- Step 2: Based on the meaning of the context, determine whether the lexical units *ruột* 'intestine' and its compounds are metaphorical. If the basic meaning of the lexical unit contrasts with its contextual meaning but can be understood in comparison with it, then it is metaphorical. Vietnamese dictionaries are used to decide on the basic meaning of a lexical unit in this procedure.
- Step 3: Identify the source and target domain
  Source domain: The source domain is RUỘT (intestine) which is used to understand more abstract concepts in this study.
  Target domain: Based on the general meaning of a group of lexical units evoked by the context in step 2 to identify the target domain, for example, the meaning of the lexical units obtained in step 2 refers to anger, the target domain is ANGER.
- Step 4: Formulate conceptual metaphors by the formula: Target domain is Source domain – A is B.

3. Ruột 'intestine' in Vietnamese culture

In Vietnamese culture, *ruột* 'intestine' is believed to consist of pieces. The number of the pieces is not clearly specified or fixed. It can be nine pieces as in *khi vò chín khúc, khi chau đôi mày* (Nguyen Du, 1820/2015, verse 488). This verse describes a man who feels sadness due to the sad music he was listening to. The sad music made his intestines crumple *khi vò chín khúc* (when -crumple - nine-piece) and his eyebrows squeeze (*khi chau đôi mày* when - squeeze – eyebrows) (Việt Nam Tự điển 1970) . The number of

the intestinal pieces can be 100 as in a folk poem: “*Gặp nhau bỏ rối cho nhau, Một trăm khúc ruột nó đau như dần*” (lit. Meeting- each other- make - the other – entangle-. One-hundred- pieces- of- intestine- which-pain-like-being beaten) (Vũ Dung, et al 2000:232). This stanza describes a meeting between a man and a woman which made both of them be in love with each other. They feel disturbed, restless and abnormal emotions which they never felt before, therefore one hundred pieces of their intestine hurt like they were beaten.

In Vietnamese folk belief, *ruột* ‘intestine’ is considered to be where human life begins. A mother gives birth to her children by breaking her intestines into pieces which is expressed in the daily language. Consider the following examples:

(1) *Mẹ dứt ruột sinh con ra vào một ngày mùa đông.*
Mother break intestine give birth you out on one day winter
‘I gave birth to you (lit. I broke my intestine to give birth to you) on a winter day.’

(2) *Con là khúc ruột của mẹ nên mẹ luôn lolắng cho con.*
You be piece intestine of mother so mother always care about child
‘You are my child (lit. a piece of my intestine) so I always care about you.’

(3) *Chị em là khúc ruột trên khúc ruột dưới phải yêu thương nhau.*
Siblings be piece intestine up piece intestine down must care each.other
‘Siblings are different pieces of the same intestines so they must care about each other.’

Examples (1-2) present the Vietnamese speakers’ understanding of the intestine as the place where human life begins. Children come from their mother’s intestine. Each of the mother’s intestinal pieces represents each of her children. The order of every intestinal piece manifests the birth order of children within a family as shown in example (3).

In Vietnamese culture, the intestine is also believed to be the place for thoughts and other mental activities[4]. This belief is recorded in Vietnamese folktales which reflect “thought patterns” of the Vietnamese (Hy Tuệ, cited in Nguyễn Đổng Chi 2000:1478). The folk story, “The medicinal resurrected plant, or the story of Cuội on the Moon” tells about an intellectual change in human beings when their intestines were replaced. In the folktale, a man’s wife was killed by robbers. They threw her intestine into a river. The husband wanted to revive his wife by replacing her intestine with a dog intestine. Because of the dog’s intestine in her body, her mind changed. She always did the opposite of what she was told to do or she even completely forgot about the instructions (Nguyễn Đổng Chi 2000:777). This story highlights the influence of the Vietnamese cultural model of *ruột* ‘intestines’ in shaping the understanding of the intestine as the seat of thoughts.

Besides, the intestine is also viewed as the container of emotions and valuable objects. This understanding is described in a Vietnamese folk tale in Buddhism “The Buddhist flag banner in pagodas”. The story is about a bad man who robbed and killed

[4] This belief is reflected in a custom of not offering rice noodle to the deceased because the rice noodle will make their intestine entangled then they will not be able to find way home (Hoàng, 2020)

people for living. One day he caught a monk who was passing by. The monk tried to teach him about Buddhism and about good and bad. The man came to be awakened and to repent about what he had done. He wanted to offer something in worship to the Buddha to show his repentance, but he had nothing. He cut open his belly and took out his intestine to the monk for him to bring to the Buddha. The monk received the man's intestine but when he reached a stream, he threw it into the water. A crow saw everything. He flew to the Buddha and squealed loudly. The Buddha understood the whole story. He rewarded the crow, punished the monk and sent the man to Nirvana to become a Buddha. Since then, pagodas began to use flag banners to remind everyone about this story. On the banner is a picture of a crow holding a silk sheet with his beak. The silk represents the intestine of the man who slit his belly to offer his intestine in worship to Buddha (Nguyễn Đổng Chi 2000:147). The story shows that the bad man's awakening and repentance are represented with his intestine, and his intestine is a valuable object which was used to offer in worship to the Buddha. As such, his thoughts and emotions are understood in terms of the intestine which is considered as a valuable object.

This section has shown that the intestine is viewed as the seat of human life and the container of both emotions and thoughts. This cultural understanding of the intestine does not correspond with the Western medicine which is practiced in health clinics and health centers in Vietnam. However, the manifestation of this belief still remains in the language and this points out the role of the Vietnamese cultural model in shaping the way Vietnamese speakers understand emotional and mental activities.

4. Ruột 'intestine' in the Vietnamese language

This section presents linguistic expressions in which *ruột* 'intestines' is conceptualized as the seat of emotional and mental activities, and as representations of other various abstract concepts including human character traits and behaviors, family ties, family behaviors, importance, and valuable possessions. This section shows that such conceptualizations of *ruột* 'intestines' are rooted in the Vietnamese folk belief that *ruột* 'intestines' is regarded as the seat of human life. These conceptualizations are compared to those of the English 'heart' and 'head' to reveal the conceptual similarities and differences of those organs and the body part in the two languages.

4.1. Intestine is a container of emotions

In Western cultures, emotions are associated with the heart (Niemeier 2000, 2008). My data shows that in Vietnamese, *ruột* 'intestine' is regarded as a container of several emotions, including happiness, unreciprocated love, anger, sadness, worry, impatience and regret.

Happiness, contained in the intestine, is a positive emotion which brings comfort to the intestine. This conceptualization corresponds to the use of *ruột* 'intestine' in the following examples:

(4)

Nghe	*hàng xóm*	*khen*	*con*	*bố mẹ*	*mát*	*hết cả*	*ruột.*
hear	neighbor	praise	children	parents	cool	it.all	intestine

'The neighbor gave compliments to their children that made the parents happy (lit. their entire intestine is cool).'

(5) *Bà mẹ* *nở* *từng* *khúc* *ruột* *vì* *con* *được* *khen*
Mother expand each piece intestine because child PASS praise
xinh đẹp.
pretty
‘The mother feels happy (lit. every piece of her intestines expands) because her daughter was praised for her beauty.’

Happiness of the parents (4) and of the mother (5) is described in terms of ‘the intestine is cool’ and ‘every single piece of the intestine is expanded’ due to the compliments for their children. As presented here, it is the intestine that feels happy, i.e., the speakers are happy. This conceptualization of happiness in terms of *ruột* ‘intestine’ is in line with the metaphor: HAPPINESS IS A PLEASURABLE PHYSICAL SENSATION found in Western cultures (Kövecses 2010:99).

The emotions: unreciprocated love, anger, sadness, worry, impatience and regret are negative emotions because they cause the speakers to experience unpleasant and disruptive reactions. Consider the following examples.

(6) *Thương* *em* *đứt* *ruột* *nhưng* *giả vờ* *ngó lơ.*
Love you break intestine but pretend ignore
‘I love you much (lit. I love you much so my intestine was broken) but I pretended to ignore you.’

(7) *Thằng bé* *đá* *con* *chó* *làm* *bà* *lộn* *ruột.*
Boy kick CLF dog make 3SG.FEM upside.down intestine
Seeing the boy kick the dog made her upset (lit. her intestine was upside down).

(8) *Anh Gò* *buồn* *héo* *ruột* *vì* *gà* *nhà* *anh*
Mr. Gò sad wither intestine because chicken house 3SG.MASC
chết *nhiều.*
die a.lot.
Mr. Gò was so sad (lit. his intestine was withered) because his chicken died a lot.

(9) *Bà* *rối* *ruột* *vì* *không* *biết* *con* *đang* *ở* *đâu.*
3SG.FEM entangle intestine because NEG know child PROG stay where
‘She was worried (lit. her intestine entangled) because she did not know where her child was.’

(10) *Ông* *sốt* *ruột* *vì* *không* *thể* *chờ* *tàu* *đến* *muộn.*
3SG.MASC fever intestine because NEG wait can train come late
‘(He) was impatient (lit. his intestine is in fever) because he could not wait for the train which comes late.’

(11) *Mai* *xót* *ruột* *vì* *lỡ* *cơ hội* *mua* *nhà*
Mai sharp.pain intestine because miss chance buy house
giá *rẻ.*
price cheap
Mai regretted a lot (lit. her intestine was in pain like she is having a wound in the intestine which is rubbed with salt) because she missed the chance to buy a house with a cheap price.

The emotions: unreciprocated love (6), anger (7), sadness (8), worry (9), impatience (10) and regret (11) show great impacts on the intestine when these emotions occur. The speaker's intestine could be 'broken' (6), 'upside down' (7), 'wither' (8), 'entangle' (9), 'in fever' (10), and 'in pain with a wound in the intestine which is rubbed with salt' (11). Those adjectives describe the intensity of the emotions and point to the imaginative effects of emotion-evoking experiences. These examples are manifestations of the more general metaphors EMOTIONS ARE FORCES (Kövecses 2010: 289), EMOTIONS ARE HEAT (Lakoff et al 1991) and PASSIVE EXPERIENCES ARE THE PHYSICAL EFFECTS OF FORCES (Kövecses 2004: 42) established in Western cultures.

There is a considerable similarity between the conceptualization of the Vietnamese 'intestine' and that of the Western 'heart' where the heart is perceived of as "the centre of emotions" (Niemeier 2008:352) which is expressed as, for example: broken heart, aching heart, my heart is bleeding, heart burning, heart-rending (Niemeier 2008:353). Different from the Western 'heart' as the center of various emotions, the Vietnamese 'intestine' is regarded as the seat of several emotions. Although the imaginative effects of the heart and the intestine may be similar (broken, burning, rending, etc.) as shown by the linguistic expressions, their reference to the emotions are different. For example, 'the broken heart' in Western cultures refers to a great sadness because a love affair has ended unhappily (Collins Dictionary online) while 'the broken intestine' refers to unreciprocated love and regret. The uses of similar effects on the Western heart and the Vietnamese intestine show that the speakers of the two cultures selected and assigned the similar effects of the emotions to the 'heart' and the 'intestine' to understand the emotions. However, the effects on the Western 'heart' are both physiological and cultural experiences (see more in Niemeier 2008), while those on the Vietnamese 'intestine' are cultural experiences which indicate that the conceptualization of the Vietnamese 'intestines' as the container of emotions is culture-based.

4.2. Intestine is a container of thoughts and knowledge

As well as seen as the seat of emotions above, the intestine is found to be associated with thoughts and other mental activities.

(12) *Ta* *đọc* *sách* *đến* *đâu,* *chữ* *cứ như* *chôn* *vào* *ruột.*
1SG read book to where, letter like bury in intestine
'As many books as I read, I remembered all of them (lit. the words are like to be buried in my intestine.'

(13) *Chúng tôi* *nhớ* *như* *chôn* *vào* *ruột*
We remember like bury in intestine
lời *mẹ* *dặn* *trước khi* *mất.*
word mother remind before die
'We remembered every single word (lit. like to bury words in our intestine) our mother reminded us before she passed away.'

(14) *Mộc* *lú* *ruột* *rồi,* *đã* *mua* *hành* *mà*
Mộc absent-minded intestine already PST buy onion, but
tưởng *là* *chưa* *mua.*
think be not yet buy
Mộc was absent-minded (lit. his intestine was forgetful). He had bought the onions but he thought he did not.

Examples (12-13) describe knowledge and other mental activities associated with the intestine. The intestine is viewed as a storage for the received knowledge. The person (12) remembers all the books he reads like he stores them in his intestine. The persons in (13) buried what their mother said, before dying, in their intestine for them to remember. Example (14) is about absent-mindedness which is understood in terms of the forgetful intestine. Mộc forgot that he had bought the onions. His forgetfulness is caused by his absent-minded intestine. This conceptualization of the ‘intestine’ is comparable with that of the Western ‘head’ where the head is viewed as “the centre of rational judgment” (Niemeier 2008:365) as shown in: *empty-headed, Put these weird thoughts out of your head. Who’s put such bizarre ideas into your head?* (p.363, 364).

However, unlike the Western ‘head’ as the center of thoughts which is manifested in a large number of expressions, the conceptualization of the Vietnamese intestine is associated with two concepts including knowledge and absent-mindedness and exemplified with rather limited expressions. However, the similar conceptualizations between the Vietnamese ‘intestine’ and the Western ‘heart’ and ‘head’ is evident. The two languages use the internal organs and the body part: the heart, the head, and the intestine to describe emotions and thoughts. The uses of the Vietnamese ‘intestine’ and the Western ‘heart’ and ‘head’ suggest that the speakers of the two languages select the internal organs and the body part along with either their actual physiological responses or imaginative responses to the emotions and thoughts which are meaningful for them to understand and talk about such experiences. The conceptualizations of the Western ‘heart’ and ‘head’ as shown in Niemeier (2008)’s study are both physiological and culture based while that of the Vietnamese ‘intestine’ is structured based on imaginative reactions of the intestine to the emotions and thoughts which indicate the culture-based of the Vietnamese ‘intestine’.

4.3. Intestine is an indicator of human characteristics / human behaviors

This section discusses how the ‘intestine’ is conceptualized as an indicator of ‘human characteristics’ and ‘human behaviors’. In this conceptualization, the ‘intestine’ is understood as an ‘object’ which can be placed on the skin, and as a ‘person’ who is in fever.

(15) *Cô ấy* *ruột* *để* *ngoài* *da,* *nghĩ* *sao* *nói* *vậy.*
3PS.FEM intestine place out skin think what speak same
‘She wears her heart on sleeve (lit. she places her intestine on her skin). What she says is exactly the same thing what she thinks’

(16)	*Sau*	*khi cưới*	*2 tháng,*	*anh ấy*	*sốt*	*ruột*	*muốn*	*có*
	After	marry	2 month,	3SG.MASC	fever	intestine	want	have
	con	*ngay.*						
	child	right						

'After two months of marriage, he was impatient (lit. his intestine was in fever), he wanted to have a baby right away.'

Examples (15-16) prove the association of the 'intestine' with 'human characteristics' and 'human behaviors'. The person who places her intestine on her skin is honest but rude. She always speaks without any consideration even what she speaks can be the truth but can displease others (Vũ Dung, et al. 2000). The explanation for this characteristic is the displacement of the intestine *ruột để ngoài da* (one's intestine is placed on his skin). The intestine is supposed to be within the human body, so the displacement of the intestine refers to something unusual, unacceptable and here is the person's characteristics: rude honesty. Due to these characteristics, the way the person often acts or talks is improper and therefore it is not socially acceptable.

Example (16) shows the illness of the intestine *sốt ruột* (the intestine is in fever) that refers to impatience and carelessness that go along with wanting to get things done quickly. The illness of the intestine is associated with impatience and carelessness which are also not socially acceptable in Vietnamese culture.

The displacement and the illness of the intestine referring to rude honesty, impatience and carelessness suggest a cultural value which is added to this aspect of the intestine – the concept of 'socially acceptable' in Vietnamese.

The conceptualization of *ruột* 'intestines' is similar to that of the English 'heart'. In English, the heart is perceived to have a default location in the chest but it can be displaced. This displacement of the heart refers to particular feelings or character traits (Foolen 2009:381). The English expression: *wear your heart on your sleeve* meaning to show your true feelings openly (Foolen 2009:382) can be seen as equivalent to the Vietnamese expression *ruột để ngoài da* (the intestine is placed on the skin). The displacement of the 'heart' in English and the 'intestine' in Vietnamese indicates a common understanding at times: "it is necessary to stifle the expression of the felt emotion and to use another expression as a mask" because 'it is not advantageous to always 'wear one's heart on one's sleeve'." (Riggio 2106:241).

4.4. Intestine is an indicator for family members and close relatives.

As mentioned in the previous sections, the intestine is regarded as the seat of life; it is the intestine where the children come from. For this reason, family members and relatives are seen to come from the same intestine. This is how the intestine's meaning is extended to represent family ties. Therefore, the 'intestine' is viewed as an indicator of identifying family members and relatives. For example:

(17)	*Bố mẹ*	*ruột*	*của*	*cô ấy*	*đang*	*ở*	*Việt Nam,*
	Parent	intestine	of	she	PROG	stay	Vietnam
	còn	*bố mẹ*	*chồng*	*đang*	*ở*	*Mỹ.*	
	and	parent	husband	PROG	stay	America.	

Her parents (lit. her intestine parents) are in Vietnam and her parent-in-laws are in America

(18) *Chị Hoa là con ruột, không phải là con nuôi của bà Mai.*
Ms. Hoa BE child intestine NEG must be child adopt of Mrs. Mai
‘Hoa is Mrs. Mai’s birth daughter, not her adopted daughter.

(19) *Họ là anh chị em ruột.*
3PL be brother sister younger intestine
‘They are siblings (lit. they are intestine siblings).’

(20) *Cô, chú, bác, ruột của anh ấy sống ở cùng một phố.*
Aunt uncles intestine of 3SG live in same one street
‘His aunts and uncles (lit. his intestine aunts and uncles) live in the same street.’

(21) *Cháu ruột của bác Hoa là sinh viên đại học.*
Niece intestine of Mrs. Hoa be student university
‘Mrs. Hoa’s niece (lit. Mrs. Hoa intestine niece) is a college student.’

(23) *Họ là bà con ruột rà với nhau có chung tổ tiên.*
3PL be relative CLF intestine RDP together have share ancestor
‘They are relative to each other (lit. they are intestine relatives). They have the same ancestor.’

Examples (17-22) show that birth parents are called “intestine parents’’, birth children are “intestine children”, siblings are ‘intestine siblings’, one’s uncles and aunts are “intestine uncles and aunts”, one’s nieces and nephews are “intestine nieces and nephews”, relatives are “intestine relatives”. These examples show that family members are seen as linked to the same intestine. This indicates the role of the intestine as an indicator in identifying family members and relatives and reveals a Vietnamese cultural value: Family members and close family ties are defined by the same intestine.

This cultural value provides guidance for treating one’s family members: Family members must live in harmony and care about others. For example:

(23) *Quan hệ giữa tôi và anh trai là*
Relationship between 1SG and brother be
chảy ruột mềm”. Chúng tôi luôn giúp đỡ
shed intestine soft” 1PL always help
“máu *nhau ruột*
blood each.other intestine
‘My brother and I are very close. We care about each other (lit. if blood is shed, the intestine gets soft). We always help each other.’

(24) *Mẹ tôi dạy “tay đứt ruột xót”,*
Mother 1SG teach “hand cut intestine sting”
nên chị em phải sống thuận hòa với nhau.
so sibling must live harmony with each.other
My mother taught us “if there are cuts on hands, the intestine will sting”, so siblings must live in harmony with each other.

In examples (23-24), *máu* ‘blood’, *tay* ‘hand’, and *ruột* ‘intestine’ are considered parts of the body and they symbolize family members and relatives. These examples show

that what happens to one part of the body (the family members) must impact the other part of the body (the others): when the blood is shed, the intestine softens and when the hand gets a cut, the intestine stings. That is, one must feel affection for his/her siblings, and other family members or relatives. Therefore, they always care about each other, and they must live in harmony.

The possible damage to the intestine suggests that this understanding of the 'intestine' is, to some degree, similar to that of the Western 'heart'. The Western 'heart' is found to feel 'soft', 'tender', 'warm', 'swelling', etc. as in: *soft heart, tender heart, heart-swelling, to have a warm heart*. For instance, people would feel pity, compassion; affection would be said that they have a special kind of heart which is perceived to be made out of soft material. Therefore, their heart easily gets indented or affected. This is when the person feels empathy, or concern about others (Niemeier 1997:88-89).

This section shows that there is a similarity between the conceptualization of the Vietnamese 'intestine' and the Western 'heart' in the sense that the two organs are perceived to feel touched and to care about others.

4.5. The uses of ruột 'intestines' to denote valuable possessions and important items

This section shows that the 'intestine' is regarded as a place where people keep their valuable possessions or as a representation of important items. This understanding of the intestine is influenced by the Vietnamese folk belief about the intestine, as the seat of life and as the valuable materials, as mentioned in the two folktales in section 4. This understanding of the intestine still remains in Vietnamese today and is manifested in the following examples:

(25)	*"Đồng tiền liền*	*khúc*	*ruột",*	*con*	*phải*	*giữ gìn*
	'Money	piece	intestine',	child	must	safeguard
	tiền	*kiếm được.*				
	money	make				

'Money must be attached to the intestine". You must safeguard the money you earned.'

(26)	*"Của*	*là*	*cuống*	*ruột"*	*nên*	*bà*
	property	be	stem	intestine	so	she
	luôn	*chú ý*	*giữ gìn*	*của cải*	*của*	*gia đình.*
	always	pay.attention	safeguard	riches	of	family

'Riches are important (lit. are the stem of the intestine) so she always pays attention to safeguard her family's riches.'

(27)	*Đàn bà*	*lo*	*cho*	*con*	*vì*	*con cái*	*là*	*cuống*	*ruột*
	women	care	for	child	because	child RDP	be	stem	intestine
	của	*họ.*							
	of	3PL							

'Women care about their children because their children are the apple of their eyes (lit. the stem of their intestine).'

Examples (25-27) show that money, valuable materials and children are considered as important as the intestine - the seat of life - in Vietnamese culture. They must be

attached to the intestine for the purpose of safety or to be protected as the stem of one’s intestine. The association between the intestine and money, valuable possessions and children suggests the concepts ‘importance’ and ‘valuable’ which are other aspects of the ‘intestine’: Being valuable is being attached to the intestine and being important is being attached to the intestine. This aspect of the ‘intestine’ reveals the cultural model of the ‘intestine’ in identifying what is important and valuable and how to safeguard them. This makes the intestine the storage of important and valuable objects.

This section has presented that the conceptualizations of the Vietnamese ‘intestine’ are rooted in the cultural model of the ‘intestine’ which actually influences the language use and motivates a number of abstract concepts such as ‘emotions’, ‘thoughts’, ‘characteristics’, ‘behaviors’, ‘important and valuable objects’, ‘family ties’ and ‘cultural values’.

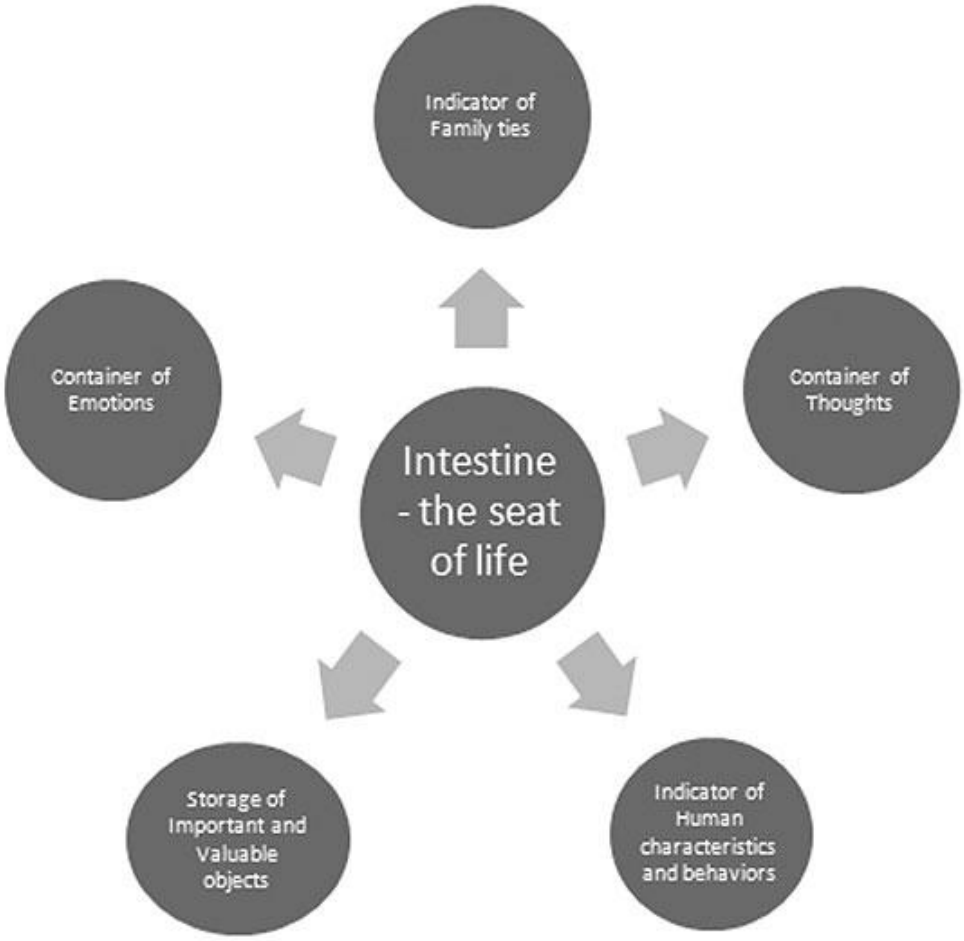


Figure 1: The cultural model of the Vietnamese ‘intestine’

The intestine metaphors in this study presents the correspondences between the source domain (the intestine) and the target domain (abstract concepts) as culture-specific, as presented in the previous sections. That is, the correspondences require cultural explanations that are not apparent to the unconscious thoughts / unconscious metaphors which come from the nature of our body (Lakoff and Johnson 1999). This evidence argues for the cultural embodiment of the ‘intestine’ in Vietnamese, that is, the intestine metaphors are culture-based then the cultural model of the ‘intestine’ is culturally embodied.

The figure 1 sums up the cultural model of the 'intestine' which motivates conceptual metaphors of the 'intestine'. It should be noted that the cultural model of the Vietnamese intestine is structured by only cultural conceptual metaphors of the intestine. This fact suggests that the Vietnamese cultural model only views the cultural experiences associated with the intestine to be meaningful to describe the concepts discussed in this study.

5 Conclusions

This study has presented an account for the culturally constructed conceptualizations of the Vietnamese *ruột* 'intestine' and indicated the role of *ruột* 'intestine' in the Vietnamese speakers' conceptualizations of the internal experiences. The conceptualizations of the 'intestine' as the container of emotional, mental activities, characteristics, behaviors, etc. are not arbitrary, but reflect and motivated by the Vietnamese cultural model of the 'intestine'. Its origin can be found in the Vietnamese folk beliefs of the 'intestine'.

The analysis of linguistic expressions concerning the Vietnamese intestine presented in this study reveals that *ruột* 'intestine' is a culturally significant concept that Vietnamese speakers recruit in thinking and talking about such abstract concepts. The comparison of the conceptualization of the Vietnamese 'intestine' and the Western 'heart' and 'head' indicates the conceptual differences and similarities in the uses of these organs and the body part to construct those abstract concepts in the two cultures. The comparison of the metaphors in the two cultures reveals the significant roles of cultures in motivating such conceptualizations. Cultural embodiment is based in culturally salient concepts which form the cultural basis for the intestine metaphors in Vietnamese.

References

Ansah, Gladys N. 2011. *The Cultural Basis of Conceptual Metaphors: The Case of Emotions in Akan and English*. Papers from the Lancaster University Postgraduate Conference in Linguistics & Language Teaching Volume 5 Papers from LAEL PG 2010.

Collins Dictionary online. https://www.collinsdictionary.com/

Foolen, Ad. 2008. The heart as a source of semiosis: The case of Dutch. In *Culture, Body, and Language Conceptualizations of Internal Body Organs across Cultures and Languages*, 373-394. Edited by Farzad Sharifian, Rene Dirven, Ning Yu and Susanne Niemeier. Mouton De Gruyter: Berlin·New York.

Gibbs, Raymond. W. Jr.. 1999. Taking metaphor out of our heads and putting it into the cultural world. In *Metaphor in cognitive linguistics*, 146–166. Edited by Raymond W. Gibbs Jr., and Gerard J. Steen. Amsterdam: John Benjamins.

Hoàng, Hằng V. 2020. *Một cách tiêu pha thời gian*. https://reatimes.vn/mot-cach-tieu-pha-thoi-gian-202200406185102224.htm. Accessed 13-05-2024.

Holland, Dorothy and Quinn, Naomi (1987). *Cultural models in language and thought*. Cambridge: Cambridge University Press.

Hy Tuệ. 1997. A treasure chest of Vietnamese Folklore from a researcher's perspective. In Nguyễn Đổng Chi (1958/2000), *Kho tàng truyện cổ tích Việt Nam*. 1475-1482. Nxb Giáo Dục. Hà Nội.

Johnson, Mark. 1987. The Body in the mind: The bodily basis of meaning, imagination, and reason. Chicago: University of Chicago Press.

Kövecses, Zoltán. 1990. *Emotion Concepts*. New York: Springer Verlag.

Kövecses, Zoltán. 2004. *Metaphor and Emotion. Language, Culture, and Body in Human Feeling*. 2nd edition. Cambridge University Press.

Kövecses, Zoltán. 2005. *Metaphor in Culture: Universality and Variation*. Cambridge: Cambridge University Press.

Kövecses, Zoltán. 2010. *Metaphor: A practical introduction* (2nd ed.). Oxford, New York: Oxford University Press.

Kövecses, Zoltán. 2017. *The interplay between metaphor and culture*. https://www.researchgate.net/publication/321314977_The_interplay_between_metaphor_and_culture. DOI:10.3726/b10542. Accessed July 3, 2023.

Lakoff, George. 1987. *Women, Fire, and dangerous things: What categories reveal about the mind*. Chicago and London: The University of Chicago Press.

Lakoff, George and Johnson, Mark. 1980. *Metaphors We Live By*. Chicago: The University of Chicago Press.

Lakoff, George and Johnson, Mark. 1999. Philosophy in the flesh: The embodied mind and its challenge to Western thought. New York: Basic Books.

Lakoff, George and Kövecses, Zoltán. 1987. The cognitive model of anger inherent in American English. In *Cultural models in language and thought*, 195-221. Edited by Dorothy Holland and Naomi Quinn. Cambridge: Cambridge University Press.

Lakoff, George. 1993. The contemporary theory of metaphor. In *Metaphor and Thought*, 82-132. Edited by Andrew Ortony. Cambridge and New York: Cambridge University Press.

Lakoff, George, Jane Espenson, and Adele Goldberg. 1991. *Master metaphor list*. http://araw.mede.uic.edu/~alansz/metaphor/METAPHORLIST.pdf Accessed July 3, 2023.

Lê, Văn Đức and Lê Ngọc Trụ. 1970.*Việt Nam Tự Điển*. NXB Khai Trí.

Maalej, Zouhair. 2004. Figurative Language in Anger Expression in Tunis-Arabic: An Extended View of Embodiment. *Metaphor and Symbol*, 19(1), 51-75.

Maalej, Zouhair. 2008. The heart and cultural embodiment in Tunisian Arabic. In *Culture, Body, and Language Conceptualizations of Internal Body Organs across Cultures and Languages*, 395-424. Edited by Farzad Sharifian, René Dirven, Ning Yu and Susanne Niemeier. Mouton De Gruyter: Berlin·New York.

Nguyễn, Đổng Chi. 1958/2000. *Kho tàng truyện cổ tích Việt Nam*. Nxb Giáo Dục. Hà Nội.

Nguyễn, Du. 1820/2022. *Truyện Kiều*. (Tái bản). NVB Văn học. Hà Nội.

Niemeier, Susanne. 1997. To have one’s heart in the right place – Metonymic and metaphorical evidence for the folk model of the heart as the site of emotions in English. In *Human Contact through Language and Linguistics*, 87–106. Edited by Birgit Smieja and Meike Tasch. Frankfurt: Peter Lang.

Niermeier, Susanne. 2008. To be in control: kind-hearted and cool-headed. The head-heart dichotomy in English. In *Culture, Body, and Language Conceptualizations of Internal Body Organs across Cultures and* Languages, 349-372. Edited by Farzad Sharifian, René Dirven, Ning Yu and Susanne Niemeier. Mouton De Gruyter: Berlin·New York.

Pragglejaz Group. 2007. MIP: A method for identifying metaphorically used words in discourse. *Metaphor and Symbol*, 22(1), 1-39.

Radden, Günter, Kövecses, Zoltán. 1999. Towards a theory of metonymy. In *Metonymy in Language and Thought*, 17-59. Edited by Klaus-Uwe Panther and Günter Radden. Amsterdam / Philadelphia: John Benjamins Publishing Company.

Riggio, R. E. 2016. Charisma. In *Encyclopedia of Mental Health* (2nd ed.), 239-244. Edited by Howard S. Friedman. NY: Elsevier.

Sharifian, Farzad; René Dirven, Ning Yu, and Susanne Niemeier. 2008. Culture and language: Looking for the "mind" inside the body. In *Culture, Body, and Language Conceptualizations of Internal Body Organs across Cultures and Languages*, 3-23. Edited by Farzad Sharifian, René Dirven, Ning Yu and Susanne Niemeier. Mouton De Gruyter: Berlin·New York

Siahaan, Poppy. 2008. Did he break your heart or your liver? A contrastive study on metaphorical concepts from the source domain ORGAN in English and in Indonesian. In *Culture, Body, and Language Conceptualizations of Internal Body Organs across Cultures and Languages*, 45-74. Edited by Farzad Sharifian, René Dirven, Ning Yu and Susanne Niemeier. Mouton De Gruyter: Berlin·New York.

The free dictionary online https://www.thefreedictionary.com/

Tran, HienT. 2018. *Conceptual Structures of Vietnamese Emotions*. Unpublished dissertation. University of New Mexico, USA.

Viện Ngôn Ngữ học. 2000. *Từ điển Tiếng Việt*. Nhà xuất bản Đà nẵng.

Vũ Dung, Vũ Thúy Anh ,Vũ Quang Hào. 2000. *Từ điển thành ngữ tục ngữ Việt Nam*. Nxb Giáo dục. Hà Nội.

Yu, Ning. 1995. Metaphorical expression of anger and happiness in English and Chinese. *Metaphor and Symbolic Activity*, 10, 223–245.

Yu, Ning. 2008. The Chinese heart as the central faculty of cognition. In *Culture, Body, and Language Conceptualizations of Internal Body Organs across Cultures and Languages*, 123-168. Edited by Farzad Sharifian, René Dirven, Ning Yu and Susanne Niemeier. Mouton De Gruyter: Berlin·New York.

Yu, Ning. 2014. Embodiment, culture, and language. In *The Routledge handbook of language and culture.* 227–239. Edited by Farzad Sharifian. London: Routledge.